computers

Special Issue Reprint

Smart Learning Environments

Edited by
Ananda Maiti

mdpi.com/journal/computers

MDPI

Smart Learning Environments

Smart Learning Environments

Guest Editor

Ananda Maiti

MDPI

Basel • Beijing • Wuhan • Barcelona • Belgrade • Novi Sad • Cluj • Manchester

Guest Editor
Ananda Maiti
School of Information
Technology
Deakin University
Waurn Ponds
Australia

Editorial Office
MDPI AG
Grosspeteranlage 5
4052 Basel, Switzerland

This is a reprint of the Special Issue, published open access by the journal *Computers* (ISSN 2073-431X), freely accessible at: https://www.mdpi.com/journal/computers/special_issues/7KB77IX6F4.

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

Lastname, A.A.; Lastname, B.B. Article Title. *Journal Name* **Year**, *Volume Number*, Page Range.

ISBN 978-3-7258-7540-5 (Hbk)
ISBN 978-3-7258-7541-2 (PDF)
https://doi.org/10.3390/books978-3-7258-7541-2

© 2026 by the authors. Articles in this reprint are Open Access and distributed under the Creative Commons Attribution (CC BY) license. The reprint as a whole is distributed by MDPI under the terms and conditions of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) license (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Contents

About the Editor . . . . . **vii**

Preface . . . . . **ix**

Ananda Maiti
Editorial: The Contemporary Landscape of Smart Learning Environments
Reprinted from: *Computers* **2026**, *15*, 122, https://doi.org/10.3390/computers15020122 . . . . . **1**

Manal Alanazi, Ben Soh, Halima Samra and Alice Li
PyChatAI: Enhancing Python Programming Skills—An Empirical Study of a Smart Learning System
Reprinted from: *Computers* **2025**, *14*, 158, https://doi.org/10.3390/computers14050158 . . . . . **5**

Wei Qiu, Chit Lin Su, Nurabidah Binti Jamil, Maung Thway, Samuel Soo Hwee Ng, Lei Zhang, et al.
A Systematic Approach to Evaluate the Use of Chatbots in Educational Contexts: Learning Gains, Engagements and Perceptions
Reprinted from: *Computers* **2025**, *14*, 270, https://doi.org/10.3390/computers14070270 . . . . . **26**

Anna Ansone, Zinta Zālīte-Supe and Linda Daniela
Generative Artificial Intelligence as a Catalyst for Change in Higher Education Art Study Programs
Reprinted from: *Computers* **2025**, *14*, 154, https://doi.org/10.3390/computers14040154 . . . . . **52**

Antonio Calvo-Morata, Cristina Alonso-Fernández, Julio Santilario-Berthilier, Iván Martínez-Ortiz and Baltasar Fernández-Manjón
Learning Analytics to Guide Serious Game Development: A Case Study Using Articoding
Reprinted from: *Computers* **2025**, *14*, 122, https://doi.org/10.3390/computers14040122 . . . . . **70**

Carrie Marcinkevage and Akhil Kumar
Generative AI in Higher Education Constituent Relationship Management (CRM): Opportunities, Challenges, and Implementation Strategies
Reprinted from: *Computers* **2025**, *14*, 101, https://doi.org/10.3390/computers14030101 . . . . . **95**

Eloy López-Meneses, Pedro C. Mellado-Moreno, Celia Gallardo Herrerías and Noelia Pelícano-Piris
Educational Data Mining and Predictive Modeling in the Age of Artificial Intelligence: An In-Depth Analysis of Research Dynamics
Reprinted from: *Computers* **2025**, *14*, 68, https://doi.org/10.3390/computers14020068 . . . . . **118**

Javier Sáez-García, María Consuelo Sáiz-Manzanares and Raúl Marticorena-Sánchez
A Proposed Method of Automating Data Processing for Analysing Data Produced from Eye Tracking and Galvanic Skin Response
Reprinted from: *Computers* **2024**, *13*, 289, https://doi.org/10.3390/computers13110289 . . . . . **144**

Diana Rakhimova, Aidana Karibayeva, Vladislav Karyukin, Assem Turarbek, Zhansaya Duisenbekkyzy and Rashid Aliyev
Development of a Children's Educational Dictionary for a Low-Resource Language Using AI Tools
Reprinted from: *Computers* **2024**, *13*, 253, https://doi.org/10.3390/computers13100253 . . . . . **164**

Yue Zhang, Zitong Kong, Nobuo Funabiki and Chen-Chien Hsu
A Study of a Drawing Exactness Assessment Method Using Localized Normalized Cross-Correlations in a Portrait Drawing Learning Assistant System
Reprinted from: *Computers* **2024**, *13*, 215, https://doi.org/10.3390/computers13090215 . . . . . **197**

Nandana Jayachandran, Atef Abdrabou, Naod Yamane and Anwer Al-Dulaimi
A Platform for Integrating Internet of Things, Machine Learning, and Big Data Practicum in Electrical Engineering Curricula
Reprinted from: *Computers* **2024**, *13*, 198, https://doi.org/10.3390/computers13080198 . . . . . **211**

Elina Gravelsina and Linda Daniela
Student Teachers' Perceptions of a Game-Based Exam in the Genial.ly App
Reprinted from: *Computers* **2024**, *13*, 207, https://doi.org/10.3390/computers13080207 . . . . . **238**

Taekwon Son
Intelligent Tutoring Systems in Mathematics Education: A Systematic Literature Review Using the Substitution, Augmentation, Modification, Redefinition Model
Reprinted from: *Computers* **2024**, *13*, 270, https://doi.org/10.3390/computers13100270 . . . . . **255**

About the Editor

Ananda Maiti

Ananda Maiti finished his PhD from the School of Engineering at the University of Southern Queensland in 2016. He was a Lecturer at the School of ICT, University of Tasmania, Launceston from 2019 to 2024. He is currently a Lecturer at the School of IT, Deakin University, Waurn Ponds Australia. His current research interests include Internet-of-Things and Embedded Systems, Cloud Computing, AI tools such as Large Language Models and their various applications such as agriculture, regulatory technologies, supply chain and environmental monitoring.

Preface

This Reprint is devoted to the study of contemporary educational technology and its rapid transformation under the influence of artificial intelligence, data-driven systems, creative computing, and integrated digital platforms. Its subject is the changing relationship between technology, learning, creativity, assessment, and educational institutions. The scope of the Reprint ranges from classroom-level tools such as intelligent tutoring systems, generative AI, learning analytics, and game-based learning environments, to institution-level platforms that support engineering education, administration, and student engagement. Together, the chapters explore how educational technologies are designed, implemented, evaluated, and experienced across different educational contexts.

The aim of this Reprint is to bring together high-quality scientific studies that represent key directions in current EdTech research. Rather than promoting technology for its own sake, the purpose is to examine how digital tools actually shape teaching, learning, creativity, assessment, and institutional practice. A central motivation for compiling this Reprint is the growing gap between technological innovation and pedagogical understanding. While new tools appear at an unprecedented pace, their educational meaning is often unclear. This Reprint was created to provide a coherent scholarly space in which these developments can be critically examined, compared, and understood.

Another motivation behind this Reprint also lies in the need for integration. Research on artificial intelligence, learning analytics, game-based learning, creative technologies, and educational platforms often develops in parallel but in isolation. By bringing these strands together, this Reprint seeks to show their conceptual connections and shared challenges, including personalization, engagement, ethics, creativity, and transformation of learning practices.

This Reprint is addressed to researchers in educational technology, artificial intelligence in education, learning sciences, and digital pedagogy. It is also intended for educators, instructional designers, academic leaders, and policymakers who are responsible for making informed decisions about technology in education. By presenting both empirical studies and conceptual analyses, this Reprint aims to support readers in understanding not only what educational technologies can do, but how and why they should be used in meaningful, responsible, and pedagogically sound ways.

Ananda Maiti
Guest Editor

computers

MDPI

Editorial

Editorial: The Contemporary Landscape of Smart Learning Environments †

Ananda Maiti

School of IT, Deakin University, Waurn Ponds, Geelong, VIC 3216, Australia; anandamaiti@ieee.org

† This Editorial is for the Special Issue "Smart Learning Environments." Link to the Special Issue: https://www.mdpi.com/journal/computers/special_issues/7KB77IX6F4.

1. Introduction

Educational Technology (EdTech) has moved decisively beyond its early role as a digital substitute for textbooks and classrooms. Today, it is a complex ecosystem in which artificial intelligence, data analytics, game design, creative technologies, and large-scale digital platforms are reshaping how teaching is designed, delivered, assessed, and governed. The current phase of EdTech is marked by three dominant movements: the rise of AI- and data-driven personalization, the mainstreaming of game-based and creative pedagogies, and the growth of integrated educational platforms that connect learning with administration, industry, and society. The papers in this Special Issue clearly reflect this transition. Together, they show an EdTech field that is no longer experimental but rather infrastructural, embedded in everyday educational practice, yet still navigating questions surrounding pedagogy, ethics, creativity, and transformation.

Across the literature, there is a recurring tension between "enhancement" and "transformation." Many technologies improve efficiency, feedback, or access, but fewer fundamentally redesign learning itself. This tension is especially visible in AI systems, game-based learning, and large-scale platforms. This collection of research articles demonstrates an education technology field that is rich in innovation but still defining what it means for technology to genuinely transform education rather than merely automate it.

2. AI and Machine Learning in Education

AI and machine learning (ML) now sit at the core of contemporary EdTech. They enable personalization, automate feedback, generate content, and analyze learning processes at a scale not previously possible. However, their pedagogical impact varies widely, from simple automation of existing practices to fully new learning experiences.

A strong illustration of applied AI is PyChatAI, an AI chatbot designed to support programming education, as shown in Contribution 1. It provides learners with support in theory, coding, debugging, and problem-solving through conversational interaction. This study's empirical results show that such systems can significantly improve learning outcomes when compared with traditional instruction, particularly by offering immediate, individualized feedback. This reflects a broader shift toward conversational interfaces as a primary medium for learning.

However, deploying AI is not only about building systems; it is also about evaluating them meaningfully. A framework proposed for evaluating GenAI chatbots through learning analytics demonstrates this clearly, as shown in Contribution 2. By combining interaction data, clustering methods, and learning outcome measures, this work demonstrates how AI systems can be assessed not only for usability but also for actual pedagogical impact.

This connects AI directly with learning analytics, positioning data as the bridge between technological innovation and educational evidence.

Generative AI also appears in more creative and exploratory contexts, as shown in Contribution 3. In art education, research on generative AI tools indicates that these systems are highly effective at supporting ideation and early-stage concept development but less effective at supporting fine-grained artistic skill or craftsmanship. This highlights a key pattern in AI-based learning: AI is most powerful as a cognitive partner—helping learners explore, generate, and reflect—rather than as a replacement for human creativity or expertise.

Learning analytics itself is becoming increasingly multimodal. Research on automating the processing of eye-tracking and galvanic skin response data demonstrates how physiological data can be transformed into learning analytics pipelines using machine learning tools, as shown in Contribution 4. This takes learning analytics beyond clickstreams and log files into embodied measures of attention, stress, and engagement, presenting new possibilities for understanding learning processes.

Language learning and literacy also benefit from AI-driven personalization. An AI-based children's dictionary for low-resource languages shows how machine learning and generative tools can support vocabulary learning where traditional resources are scarce, as shown in Contribution 1. This highlights an important ethical and social dimension of AI in education: its potential to address inequities in access to learning materials.

Creative learning also increasingly relies on intelligent systems. In Contribution 6, a portrait drawing learning assistant uses computer vision and normalized cross-correlation to compare student drawings with reference images, providing detailed automated feedback on facial features. This demonstrates how AI can support domains traditionally considered subjective or resistant to automation, such as art.

At a more structural level, Intelligent Tutoring Systems (ITSs) are among the oldest and most established forms of AI and ML in education. In Contribution 7, a systematic review of ITS in mathematics education shows that while ITSs are widespread and effective, most are used at the "augmentation" level—improving traditional instruction rather than transforming it. Only a small number of ITS redesigns learn in fundamentally new ways, such as through learning-by-teaching models. This finding echoes across many AI systems in that technological sophistication does not automatically lead to pedagogical transformation.

Finally, research on educational data mining and predictive modeling provides a broader context for all AI-driven systems. In Contribution 8, a large-scale review of educational data mining (EDM) shows rapid growth in predictive analytics, early warning systems, and performance modeling, alongside rising concerns about bias, privacy, and ethics. AI in education is therefore not just a technical challenge but also a governance and values-based challenge.

3. Game-Based Learning and Creative Pedagogies

Game-based learning has matured from novelty to a mainstream pedagogical strategy. It now includes not only serious games and gamification, but also creative platforms, narrative environments, and playful assessment models.

One important development is the integration of learning analytics into serious games. In Contribution 9, research on serious games for programming demonstrates that in-game traces can be analyzed to understand learner strategies and difficulties, enabling designers to iteratively refine both pedagogy and game mechanics. This positions games not merely as engaging tools, but as data-rich learning environments.

https://doi.org/10.3390/computers15020122

Game-based learning is also moving toward formal assessment. In Contribution 10, the study of a game-based exam implemented on the Genial.ly platform demonstrates that summative assessments can be delivered through narrative-driven, interactive games, thereby reducing student stress while maintaining academic rigor. This is significant as it challenges the long-standing separation between "play" and "testing," showing that assessment itself can become a learning opportunity.

Creative and artistic domains are increasingly aligned with game-like and interactive structures. In Contribution 6, the portrait drawing assistant, although technically an intelligent tutoring system, also functions as a creative learning environment where learners interact with visual feedback loops similar to those in games. Likewise, generative AI in art education encourages exploratory, playful interaction with ideas, blurring the boundary between tool and creative partner, as shown in Contribution 3.

Language and literacy technologies also align with game-based and playful learning. The AI children's dictionary employs adaptive and generative techniques to support vocabulary learning in an exploratory, learner-driven manner, as shown in Contribution 5. While not a game in a strict sense, it shares core features of game-based learning, i.e., interactivity, feedback, and intrinsic motivation.

These studies collectively show that game-based learning is no longer limited to "educational games." It now includes narrative exams, creative platforms, AI-assisted creativity, and playful analytics-driven environments. However, like AI systems, many game-based tools still enhance existing practices rather than fully redefining pedagogy. The challenge remains to design games that do more than motivate and fundamentally reshape how knowledge is constructed, practiced, and assessed.

4. Education Platforms and Institutional Infrastructure

Beyond classroom tools, EdTech increasingly operates at the platform level—integrated systems that connect learning, administration, industry, and data.

The IoT–ML practicum platform exemplifies this trend in technical education. In Contribution 11, by allowing students to design IoT systems, stream data, apply machine learning models, and visualize results within one environment, it creates a realistic simulation of industry workflows. This is not merely a teaching tool but a learning ecosystem that aligns curriculum with professional practice.

At the institutional level, generative AI is also entering administrative and strategic domains. In Contribution 12, the study of GenAI in higher-education customer relationship management shows how AI is being used in recruitment, advising, alumni relations, and student support. This represents a shift from EdTech as classroom technology to EdTech as institutional infrastructure, shaping how universities interact with learners across the entire education lifecycle.

5. Conclusions

The current state of EdTech is defined by integration: AI with pedagogy, games with assessment, creativity with computation, and learning with institutional platforms. The papers in this Special Issue demonstrate a field rich in innovation but still in the process of defining its identity. AI systems personalize and automate, but do not always transform. Games engage and motivate, but do not always redefine learning. Platforms connect education to industry and administration, but raise new ethical and structural challenges.

Editorially, these articles collectively argue that the future of EdTech depends not on technological novelty, but on pedagogical intention. The real challenge is not building smarter systems, more engaging games, or larger platforms—but deciding what kind of learning, creativity, and educational society these technologies are meant to serve.

https://doi.org/10.3390/computers15020122

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflicts of interest.

List of Contributions:

1. Alanazi, M.; Soh, B.; Samra, H.; Li, A. PyChatAI: Enhancing Python Programming Skills—An Empirical Study of a Smart Learning System. *Computers* **2025**, *14*, 158. https://doi.org/10.3390/computers14050158.
2. Qiu, W.; Su, C.L.; Jamil, N.B.; Thway, M.; Ng, S.S.H.; Zhang, L.; Lim, F.S.; Lai, J.W. A Systematic Approach to Evaluate the Use of Chatbots in Educational Contexts: Learning Gains, Engagements and Perceptions. *Computers* **2025**, *14*, 270. https://doi.org/10.3390/computers14070270.
3. Ansone, A.; Zālīte-Supe, Z.; Daniela, L. Generative Artificial Intelligence as a Catalyst for Change in Higher Education Art Study Programs. *Computers* **2025**, *14*, 154. https://doi.org/10.3390/computers14040154.
4. Sáez-García, J.; Sáiz-Manzanares, M.C.; Marticorena-Sánchez, R. A Proposed Method of Automating Data Processing for Analysing Data Produced from Eye Tracking and Galvanic Skin Response. *Computers* **2024**, *13*, 289. https://doi.org/10.3390/computers13110289.
5. Rakhimova, D.; Karibayeva, A.; Karyukin, V.; Turarbek, A.; Duisenbekkyzy, Z.; Aliyev, R. Development of a Children's Educational Dictionary for a Low-Resource Language Using AI Tools. *Computers* **2024**, *13*, 253. https://doi.org/10.3390/computers13100253.
6. Zhang, Y.; Kong, Z.; Funabiki, N.; Hsu, C.-C. A Study of a Drawing Exactness Assessment Method Using Localized Normalized Cross-Correlations in a Portrait Drawing Learning Assistant System. *Computers* **2024**, *13*, 215. https://doi.org/10.3390/computers13090215.
7. Son, T. Intelligent Tutoring Systems in Mathematics Education: A Systematic Literature Review Using the Substitution, Augmentation, Modification, Redefinition Model. *Computers* **2024**, *13*, 270. https://doi.org/10.3390/computers13100270.
8. López-Meneses, E.; Mellado-Moreno, P.C.; Gallardo Herrerías, C.; Pelícano-Piris, N. Educational Data Mining and Predictive Modeling in the Age of Artificial Intelligence: An In-Depth Analysis of Research Dynamics. *Computers* **2025**, *14*, 68. https://doi.org/10.3390/computers14020068.
9. Calvo-Morata, A.; Alonso-Fernández, C.; Santilario-Berthilier, J.; Martínez-Ortiz, I.; Fernández-Manjón, B. Learning Analytics to Guide Serious Game Development: A Case Study Using Articoding. *Computers* **2025**, *14*, 122. https://doi.org/10.3390/computers14040122.
10. Gravelsina, E.; Daniela, L. Student Teachers' Perceptions of a Game-Based Exam in the Genial.ly App. *Computers* **2024**, *13*, 207. https://doi.org/10.3390/computers13080207.
11. Jayachandran, N.; Abdrabou, A.; Yamane, N.; Al-Dulaimi, A. A Platform for Integrating Internet of Things, Machine Learning, and Big Data Practicum in Electrical Engineering Curricula. *Computers* **2024**, *13*, 198. https://doi.org/10.3390/computers13080198.
12. Marcinkevage, C.; Kumar, A. Generative AI in Higher Education Constituent Relationship Management (CRM): Opportunities, Challenges, and Implementation Strategies. *Computers* **2025**, *14*, 101. https://doi.org/10.3390/computers14030101.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

computers

MDPI

Article

PyChatAI: Enhancing Python Programming Skills—An Empirical Study of a Smart Learning System

Manal Alanazi [1], Ben Soh [1,*], Halima Samra [2] and Alice Li [3]

[1] Department of Computer Science and Information Technology, La Trobe University, Melbourne 3086, Australia; manalmohammedg.alanazi@latrobe.edu.au
[2] Computer Science & Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia; hsamra@kau.edu.sa
[3] La Trobe Business School, La Trobe University, Melbourne 3086, Australia; a.li@latrobe.edu.au
* Correspondence: b.soh@latrobe.edu.au

Abstract: This paper presents strategies for effectively integrating AI tools into programming education and provides recommendations for enhancing student learning outcomes through intelligent educational systems. Learning computer programming is a cognitively demanding task that requires dedication, logical reasoning, and persistence. Many beginners struggle with debugging and often lack effective problem-solving strategies. To address these issues, this study investigates PyChatAI—a bilingual, AI-powered chatbot designed to support novice Python programmers by providing real-time feedback, answering coding-related questions, and fostering independent problem-solving skills. PyChatAI offers continuous, personalised assistance and is particularly beneficial for students who prefer remote or low-pressure learning environments. An empirical evaluation employing a Solomon Four-Group design revealed significant improvements across all programming skill areas, with especially strong gains in theoretical understanding, code writing, and debugging proficiency.

Keywords: artificial intelligence; ChatGPT; natural language processing (NLP); OpenAI API; Python programming; computer programming; smart learning systems; Saudi Arabia

1. Introduction

With the rapid advancement of artificial intelligence (AI) and the ongoing digital transformation across industries, the demand for individuals with specialised expertise in computer science and programming has become increasingly urgent. Programming skills are now recognised as essential competencies for navigating the future workforce, driving innovation, and contributing to sustained economic growth. In Saudi Arabia, this shift is evident in the substantial rise in enrolment in computer science programs, in alignment with national strategic goals, such as Saudi Vision 2030.

The country has made significant progress in AI education, with 86% of universities offering undergraduate programs, 56% offering master's programs, and 9% offering PhD programs in the field. Between 2019 and 2023, more than 38,000 students graduated with specialisations in AI, including 6500 who completed their studies abroad. Remarkably, the number of AI graduates increased by 42% from 2022 to 2023 [1].

Despite these encouraging developments, challenges in computer programming education remain. A smart learning system—capable of providing real-time, AI-driven, and individualised support—holds great potential to address these gaps and enhance the effectiveness of programming instruction.

These challenges include regional disparities in educational quality, traditional and often passive teaching methods, language barriers, and limited access to instructional resources. Such obstacles significantly hinder student engagement, contributing to high dropout rates and inconsistent learning outcomes. Addressing these issues requires more than curriculum reform; it necessitates the development of intelligent, responsive, and student-centred learning environments [2].

Recent research indicates that while Saudi students generally possess strong digital literacy and positive attitudes toward technology, their use of digital tools for academic purposes remains limited. Factors such as geographic location, family support, and English language proficiency significantly influence their adoption of educational technologies [3]. Notably, many female computer science students in Saudi Arabia still rely on traditional study methods—such as handwritten notes, which often lack interactivity and real-time feedback mechanisms [4].

In response to these persistent challenges, this study introduces PyChatAI—a bilingual, AI-powered smart learning system designed specifically to support novice Python programmers. Unlike general-purpose AI tools, PyChatAI delivers culturally relevant, curriculum-aligned assistance tailored to the unique needs of Saudi learners. With features such as real-time feedback, adaptive guidance, and proactive error detection, it offers a novel and timely solution within the expanding landscape of AI-enhanced education.

2. Motivation and Research Objectives

Learning computer programming is a complex and demanding process that requires persistence, logical reasoning, and strong problem-solving skills. One of the most common challenges faced by beginners is debugging—an often frustrating task that can lead to the development of ineffective coding habits and reduced learner confidence [5].

To address these difficulties, AI-powered educational technologies—particularly smart learning systems, such as intelligent tutoring systems and chatbots—have emerged as promising solutions. These tools enhance student engagement by delivering personalised, real-time feedback, clarifying complex programming concepts, and fostering independent learning [6]. Tools such as ChatGPT, for example, have been well-received for their ability to simplify programming tasks and offer adaptive, context-sensitive responses [7,8].

However, many existing AI tools remain limited by their reactive design; they only provide assistance when explicitly prompted by the user. This restricts their capacity to proactively identify when students are struggling and intervene accordingly [9]. In addition, most tools are not curriculum-aligned, lack bilingual support, and are not tailored to specific educational contexts—making them less effective for non-English-speaking learners, particularly in Saudi Arabia.

To overcome these limitations, this study introduces PyChatAI—a smart learning system specifically designed to support beginner Python programmers. PyChatAI offers bilingual support (Arabic and English), integrates proactive error detection, and incorporates culturally relevant scaffolding based on Vygotsky's Zone of Proximal Development (ZPD) [10]. The tool aims to deliver context-aware, responsive assistance that promotes both technical skill development and learner motivation, in alignment with Saudi Arabia's national education and digital transformation goals.

2.1. Research Objectives

This study aims to achieve the following:

- Evaluate the effectiveness of PyChatAI in improving programming skills among novice learners.
- Identify factors influencing students' acceptance of smart learning systems using the Technology Acceptance Model (TAM3).

- Explore the role of AI-driven feedback and real-time support in facilitating personalised learning.
- Assess the potential of culturally adapted AI tools in supporting digital education goals, particularly in Saudi Arabia.

2.2. Research Questions

1. Does PyChatAI significantly improve students' programming skills across key dimensions (e.g., code writing and problem-solving)?
2. How do students perceive the usefulness and ease of use of PyChatAI?
3. What is the impact of the bilingual and proactive features of PyChatAI on user engagement and learning outcomes?
4. Can PyChatAI be effectively integrated into existing programming courses in a way that aligns with educational objectives in Saudi Arabia?

3. Literature Review

3.1. Existing Research and Tools

The integration of AI tools into computer programming education has rapidly advanced with the emergence of technologies such as ChatGPT, Pyo, Python-Bot, and Revision-Bot. These tools offer a range of functionalities designed to improve coding proficiency, provide real-time feedback, and enhance student engagement.

ChatGPT, a generative pre-trained transformer (GPT-4), represents a significant innovation in AI technology. Integrated into computer programming education, it assists students by generating code snippets, offering debugging support, and suggesting coding strategies. These features promote sustainable coding practices and enrich the learning experience [11–13]. ChatGPT also facilitates multiple concurrent interactions, creating a supportive, non-judgmental environment that encourages students to seek help comfortably. Additionally, its capabilities—such as error detection, automated code generation, and real-time assistance—enhance coding efficiency by reducing the time and effort needed for assignments [11].

However, ChatGPT is not specialised for programming education. It provides static responses that do not adapt to individual learners' levels and lacks capabilities for direct code execution or real-time code evaluation [11]. Despite these limitations, its educational impact is broad. It has gained considerable academic attention for its potential to transform teaching and learning through interactive, AI-driven approaches [14–16]. Many students perceive ChatGPT as a human-like assistant, appreciating its ability to simplify complex concepts and improve their learning experience [17].

Empirical studies underscore the effectiveness of ChatGPT in programming education. For example, one study found that students using ChatGPT-enhanced learning methods improved their exam scores from 48.33 to 74.47 [18]. Engaging with ChatGPT has also been linked to higher motivation and the development of cleaner, more structured code, with fewer rule violations, reduced complexity, and improved adherence to coding conventions [19,20].

Beyond individual assistance, AI-powered systems such as ChatGPT-based virtual agents are being developed to simulate programmer-like problem-solving techniques. These agents provide interactive and adaptive learning scenarios that enhance students' technical skills, critical thinking, and collaborative learning.

Building on the foundation established by ChatGPT, several specialised tools have been introduced:

- Pyo focuses on introductory Python programming. It offers real-time chatbot support to explain syntax, concepts, and debugging strategies [21]. While useful, it lacks deep personalisation and automated code assessment.
- Python-Bot and Revision-Bot support programming practice, revision, and exam preparation. These tools provide interactive feedback and automatic assessments of code quality, syntax, and logic. They also integrate with Learning Management Systems (LMSs) to track student progress [7,22]. However, their reliance on predefined question banks limits their adaptability to individual learning trajectories.
- KIAAA serves as an AI-driven intelligent tutoring system (ITS) with a focus on automation-related programming tasks. It offers customised learning paths and supports 3D simulation-based interactive execution [23].
- VoiceBots employ voice-based technology to support multimedia programming education, particularly in HTML and CSS. While they enhance accessibility and interactivity, they lack deep integration with core programming instruction.

While these tools have contributed significantly to the advancement of computer programming education, many still lack key features, such as proactive engagement, adaptive learning, and real-time code execution. Among these, adaptive learning is especially critical, as it allows AI systems to personalise instruction based on a student's existing knowledge, learning pace, and areas of difficulty. By dynamically tailoring feedback and content delivery, adaptive AI tools can offer more effective support than static, one-size-fits-all approaches. This adaptability enhances learning efficiency, increases student motivation, and helps reduce cognitive overload—particularly in complex domains, such as programming.

To better understand the current landscape, Table 1 presents a comparative overview of these AI-based educational tools, outlining their respective strengths and limitations in supporting computer programming education.

Table 1. Comparison of tools for learning computer programming.

Feature	ChatGPT	KIAAA	Pyo	VoiceBots	Xatkit
Year	**2023-2024-2025**	**2023**	**2022**	**2024**	**2020-2021-2024**
Primary Purpose	General AI assistant for various tasks, including programming.	AI-driven Intelligent Tutoring System (ITS) for automation programming education.	AI-based assistant for Python programming support.	Enhances multimedia programming education through VoiceBot technology.	AI-powered programming learning environment.
AI Model Used	GPT-4 (Natural Language Processing, NLP).	Help-DKT (AI cognitive model for student progress tracking).	Built on Rasa (Conversational AI framework).	IBM Watson AI Services.	NLP-based machine learning models.
Specialisation in Programming	Not specific to programming education.	Focused on automation-related programming.	Specialises in introductory Python programming.	Emphasises HTML and CSS in multimedia programming.	Supports multiple languages, including Python, Java, and C++.
Response to Coding Queries	Provides general programming explanations and debugging assistance.	Generates custom programming tasks and dynamically adapts to student responses.	Explains Python syntax, concepts, and debugging strategies.	Responds to multimedia programming inquiries in real time.	Provides immediate feedback, debugging support, and adaptive learning.
Real-Time Interaction	No direct code execution.	Supports interactive execution in a 3D simulation environment.	Offers real-time programming assistance via chatbot.	Provides instant feedback through WhatsApp integration.	Delivers live chatbot responses with adaptive learning.

Table 1. *Cont.*

Feature	ChatGPT	KIAAA	Pyo	VoiceBots	Xatkit
Adaptability to Student Level	Static responses, no adaptation to user expertise.	Monitors and adjusts difficulty dynamically based on student progress.	Provides guidance based on errors but lacks deep personalisation.	Customisable responses based on student engagement and feedback.	Adapts difficulty based on learning patterns.
Automated Code Evaluation	No real-time assessment.	AI-powered evaluation with automated feedback.	Identifies syntax errors and offers hints via multiple-choice prompts.	Offers automated responses but lacks code execution analysis.	Performs syntax validation, logic analysis, and efficiency checks.
Simulation & Hands-on Learning	No simulation capabilities.	3D virtual learning environment for practical coding tasks.	Encourages logical code structuring through guided exercises.	Provides interactive learning experiences in multimedia programming.	Supports hands-on coding exercises and problem-solving scenarios.
Integration with Educational Systems	No direct LMS or IDE support.	Compatible with IDEs and automation software.	Embedded in an online Python learning platform.	Integrated with WhatsApp and Learning Management Systems (LMS).	Supports LMS integration, classroom analytics, and API connectivity.
Assessment & Feedback	Provides only text-based explanations, no evaluation metrics.	AI-driven real-time feedback with automated assessment.	Offers hints, explanations, and solutions for debugging exercises.	Provides real-time feedback with additional learning resources.	Automated grading, performance tracking, and skill improvement recommendations.
Ref	[11,14,20,24–27]	[28]	[22]	[29]	[30–32]
Feature	**Revision-Bot**	**Python-Bot**	**Iris**	**e-JAVA Chatbot**	**PerFuSIT**
Year	**2022**	**2020**	**2024**	**2020**	**2024**
Primary Purpose	AI-based chatbot for personalised programming revision and learning support.	AI-based chatbot for programming practice and revision.	AI tutor in Artemis for personalised coding help	Virtual tutor for learning JAVA programming.	Adaptive Learning, Personalised Tutoring
AI Model Used	SnatchBot API with NLP and AIML (Artificial Intelligence Markup Language).	SnatchBot API with NLP and AIML.	GPT-3.5-Turbo with advanced prompting.	Rule-based chatbot with text-matching techniques.	Fuzzy Logic, Machine Learning
Specialisation in Programming	Python-focused, designed for exam practice and interactive problem-solving.	Python-focused, designed for exam practice and interactive problem-solving.	CS1 programming exercises, no direct solutions.	Focused on JAVA, particularly control structures.	Python, Java, C++
Response to Coding Queries	AI-driven question-answer chatbot with a predefined question bank.	AI-driven question-answer chatbot with a predefined question bank.	Hints, counter-questions, and explanations.	Provides instant responses using pre-defined patterns.	Real-Time Feedback, Debugging Assistance
Real-Time Interaction	AI-powered chatbot with interactive feedback.	AI-powered chatbot with interactive feedback.	Instant, context-aware responses.	Offers immediate feedback through chat-based input.	Dynamic Adjustments, Live Support

Table 1. *Cont.*

Feature	ChatGPT	KIAAA	Pyo	VoiceBots	Xatkit
Adaptability to Student Level	Tracks student progress and customises revision content accordingly.	Tracks student progress and customises revision content accordingly.	Adjusts to problem context, not skill level.	Limited adaptability; follows structured responses	Beginner to Advanced, Skill-Based Progression
Automated Code Evaluation	AI-driven assessment of code quality, syntax, and logic.	AI-driven assessment of code quality, syntax, and logic.	Uses Artemis's automated feedback.	Generates code samples but does not evaluate user code.	Syntax Analysis, Logical Error Detection
Simulation & Hands-on Learning	Practice-based revision tool with interactive coding challenges.	Practice-based revision tool with interactive coding challenges.	Real-time guidance, no simulations	Supports learning via structured problem-solving.	Interactive Coding Exercises, Virtual Labs
Integration with Educational Systems	Web-based chatbot with LMS compatibility and social media deployment.	Web-based chatbot with LMS compatibility and social media deployment.	Fully embedded in Artemis.	Standalone tool; not integrated with LMS or e-learning platforms.	LMS Compatibility, API Integration
Assessment & Feedback	AI-powered self-evaluation, student progress tracking, and score prediction.	AI-powered self-evaluation, student progress tracking, and score prediction.	Provides hints, not direct grading.	Provides structured responses but lacks personalised feedback.	Performance Tracking, Personalised Reports
Ref	[33]	[21]	[34]	[35]	[36]

3.2. Broader Educational Context

The rapid advancement in AI technologies has profoundly impacted higher education, particularly in the domain of programming instruction. AI-powered tools such as ChatGPT have garnered significant attention for their potential to enhance teaching and learning outcomes in academic settings [12]. By offering immediate feedback, assisting with coding tasks, and promoting interactive learning, these tools foster greater student engagement through AI-driven instructional methods [15,16].

Beyond programming-specific applications, AI-based tutors, such as Iris and the e-JAVA Chatbot, have been integrated into broader educational contexts. Iris provides real-time guidance and personalised feedback for CS1 programming exercises, while the *e*-JAVA Chatbot supports structured problem-solving in Java programming environments [21]. These tools demonstrate the versatility of AI in various learning contexts, although they still lack proactive engagement capabilities that would allow them to anticipate and respond to student difficulties without prompting.

Moreover, the integration of AI in education is reshaping the role of educators—from traditional content deliverers to mentors who facilitate conceptual understanding and guide students in developing critical problem-solving skills [7,23].

3.3. Theoretical Frameworks

The development of AI-based educational tools is frequently grounded in established learning theories, with Vygotsky's Zone of Proximal Development (ZPD) serving as a key theoretical framework. ZPD highlights the value of scaffolding—providing structured support that enables learners to accomplish tasks they would find difficult to complete unaided. This guided approach fosters the development of cognitive and problem-solving skills by gradually promoting learner autonomy [10].

In the context of computer programming education, integrating adaptive support mechanisms into educational tools can address common challenges faced by novice programmers, such as planning program logic and debugging complex code [24,37]. AI-driven systems that are designed around ZPD principles deliver personalised assistance that bridges the gap between what students can do independently and what they can accomplish with appropriate guidance, thereby enhancing both learning effectiveness and student confidence.

3.4. Gaps in Research

Despite the growing success of AI-driven tools, such as ChatGPT, Pyo, and Python-Bot, several important gaps persist in the current research and application landscape:

- Lack of Proactive Engagement: Most existing tools—including ChatGPT and Python-Bot—are reactive in nature, responding only when prompted by the user. They do not proactively detect or intervene when students encounter difficulties during coding tasks.
- Limited Adaptability: Tools such as ChatGPT provide static responses that do not adjust to a learner's skill level or learning trajectory. While systems such as KIAAA offer dynamic difficulty adjustments, they are not specifically designed for programming education [23].
- Absence of Real-Time Code Execution and Evaluation: Many AI-powered tools, including ChatGPT and VoiceBots, lack built-in capabilities for executing and evaluating code in real time, limiting their usefulness for hands-on programming practice [11,15].
- Integration Challenges: Although some tools can interface with Learning Management Systems (LMSs), seamless integration with Integrated Development Environments (IDEs) and other educational platforms remains limited, hindering a cohesive and streamlined learning experience [7,22].

These limitations underscore the need for specialised AI tools that deliver personalised, real-time support tailored to the unique demands of programming education—particularly at the introductory level.

To address these gaps, this study introduces PyChatAI, a purpose-built, bilingual chatbot developed for novice Python programmers. Building upon the foundational capabilities of ChatGPT, PyChatAI extends its functionality by offering interactive, adaptive programming support that aligns with learners' skill levels, educational contexts, and curriculum requirements. It is designed to proactively assist students in overcoming common programming challenges, thereby enhancing both learning outcomes and student engagement.

4. Methodology

4.1. Introduction to Methodology

This study aims to evaluate the effectiveness of PyChatAI, an AI-powered educational tool developed to support and enhance introductory Python programming instruction. This methodology section details the participant demographics, system development process, technical implementation, and the experimental framework used to assess the impact of PyChatAI on student learning outcomes. To ensure a robust and comprehensive evaluation, this study employs a mixed-methods approach, integrating both quantitative and qualitative techniques. These include the Solomon Four-Group Experimental Design, the Technology Acceptance Model (TAM3) questionnaire, and structured interviews with faculty members.

4.2. Participants and Setting

This research was conducted at the College of Computer Science and Information Technology at Jouf University, Saudi Arabia. The study involved 300 female first-year

undergraduate students enrolled in introductory Python programming courses during the first semester of the 2024 academic year. Participants were between 18 and 22 years old, and the majority had limited or no prior programming experience, with Python serving as their first formal programming language. To ensure the validity of the findings and minimise selection bias, participants were randomly assigned to control and experimental groups using a computer-generated randomisation method, ensuring a balanced group distribution and unbiased comparisons.

4.3. Experimental Design and Evaluation

To rigorously assess the effectiveness of PyChatAI, this study employs the Solomon Four-Group design—a well-established experimental methodology renowned for its ability to control for pre-test sensitisation and isolate the true effects of an intervention. This design is particularly effective in distinguishing whether observed improvements are attributable to the intervention itself or are influenced by participants' prior exposure to testing conditions.

4.3.1. The Structure of the Solomon Four-Group Design

The experimental setup consists of four groups: two control groups and two experimental groups. The groups vary based on whether participants received a pre-test and whether they were exposed to the intervention (use of PyChatAI). The structure of this design is summarised in Table 2.

Table 2. Solomon Four-Group design.

Group	Pre-Test	Intervention (Using PyChatAI)	Post-Test	Sample Size
Control Group One	No	No	Yes	50
Control Group Two	Yes	No	Yes	50
Experimental Group One	Yes	Yes	Yes	50
Experimental Group Two	No	Yes	Yes	50

4.3.2. Explanation of Design Components

- Pre-test: Measures participants' baseline programming knowledge and skills prior to the intervention. This establishes a reference point and helps assess whether initial proficiency influences the effectiveness of PyChatAI.
- Intervention (use of PyChatAI): represents the experimental conditions in which students interacted with PyChatAI during programming sessions to receive real-time assistance, debugging support, and contextual code feedback.
- Post-test: administered following the intervention to evaluate changes in programming performance, enabling a comparative analysis across all four groups.

4.3.3. Purpose of Research Design

The Solomon Four-Group design enables a rigorous evaluation of PyChatAI by facilitating balanced comparisons across groups exposed to different conditions. Specifically, the design isolates and measures two critical factors:

1. Effect of the Intervention:
   By comparing outcomes between the experimental groups (who used PyChatAI) and the control groups (who did not), this study quantitatively assessed the direct impact of PyChatAI on students' computer programming skills. This comparison provides robust evidence regarding the efficacy of the application in enhancing programming competencies.

2. Effect of Pre-testing:
   The inclusion of both pre-tested and non-pre-tested groups controls for potential testing effects that might influence post-test performance. This allowed this study to examine whether exposure to the pre-test itself contributes to improvements in computer programming skills, independent of the intervention.

4.3.4. Research Process Flowchart

Figure 1 illustrates the research process for evaluating PyChatAI. The process began with the pre-test, followed by the random assignment of 300 students into control and experimental groups. The experimental group received the PyChatAI intervention, while the control group continued with traditional instructional methods. After a four-month intervention phase, both groups completed a post-test to assess changes in computer programming skills.

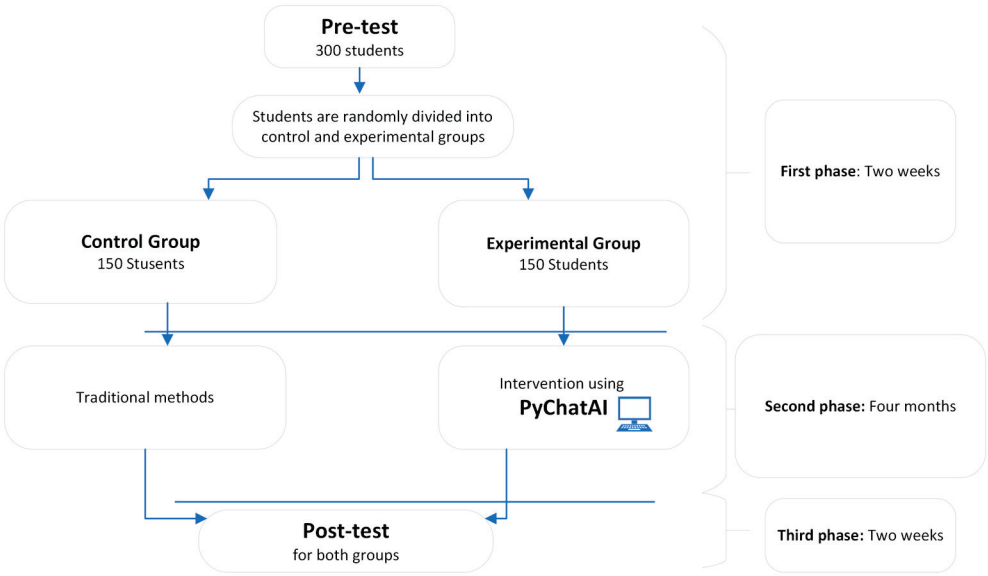


Figure 1. Flowchart illustrating the research process for evaluating PyChatAI.

4.4. System Development and Technical Implementation

PyChatAI was developed as a domain-specific smart learning system, integrating AI-driven feedback and real-time debugging support that leverages the OpenAI API for generating natural language responses tailored to Python programming tasks. The system provides real-time feedback, structured learning prompts, and debugging support to enhance the learning experience for novice programmers. The PyChatAI system development involves the following:

- Programming Language: Python 3.6+.
- User Interface (UI): developed using Tkinter for graphical interface design.
- AI Integration: utilises OpenAI API for generating context-specific programming responses.
- System Requirements:
  a. Operating Systems: Windows, macOS, or Linux.
  b. Dependencies: Tkinter, OpenAI API, and Pyperclip for clipboard operations.

Figures 2–6 illustrate various interface features of PyChatAI, including the AI Question interface (Figure 2), input windows for coding challenges (Figure 4), and predefined prompts guiding students through Python concepts (Figure 6). These interfaces were designed to ensure usability and interactive engagement for beginner programmers.

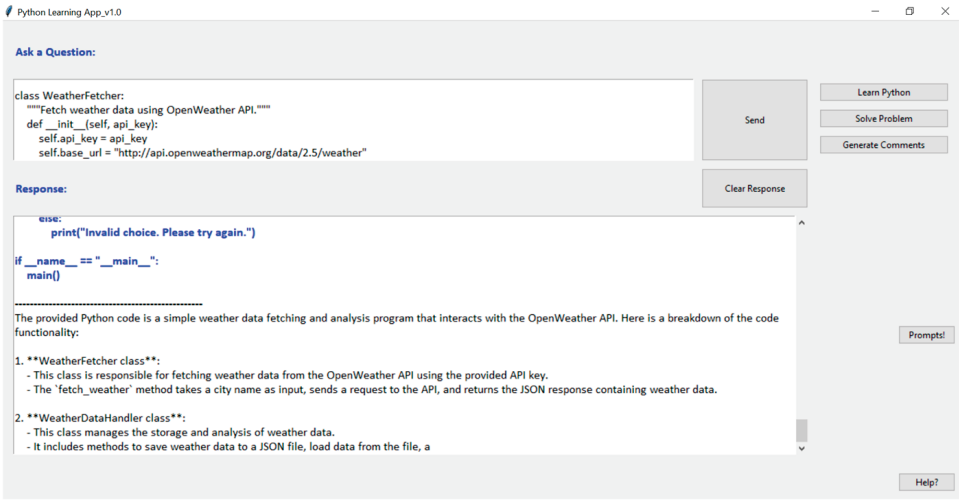


Figure 2. Screenshot of the Python Learning App_v1.0, illustrating code input, explanation, and interactive features for learning how to fetch data using the OpenWeather API.

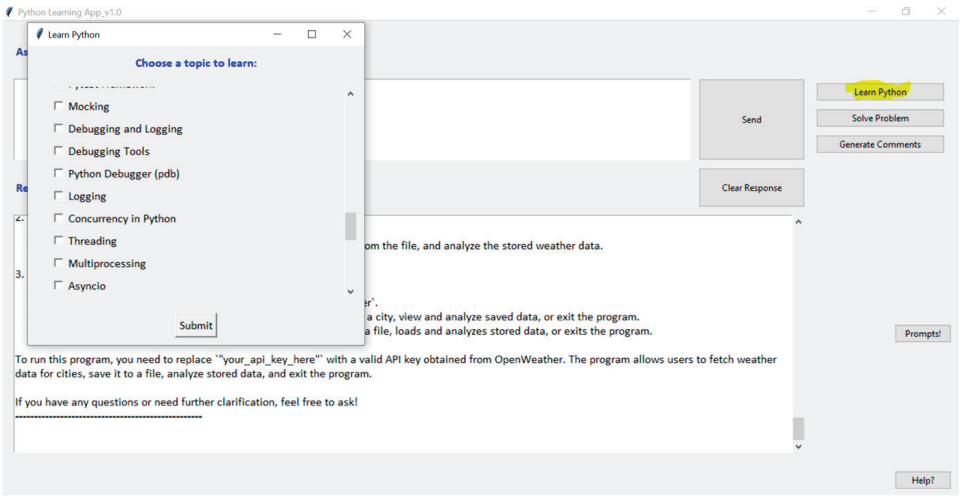


Figure 3. Screenshot of the "Learn Python" feature interface, displaying topic selection for interactive learning in the Python Learning App.

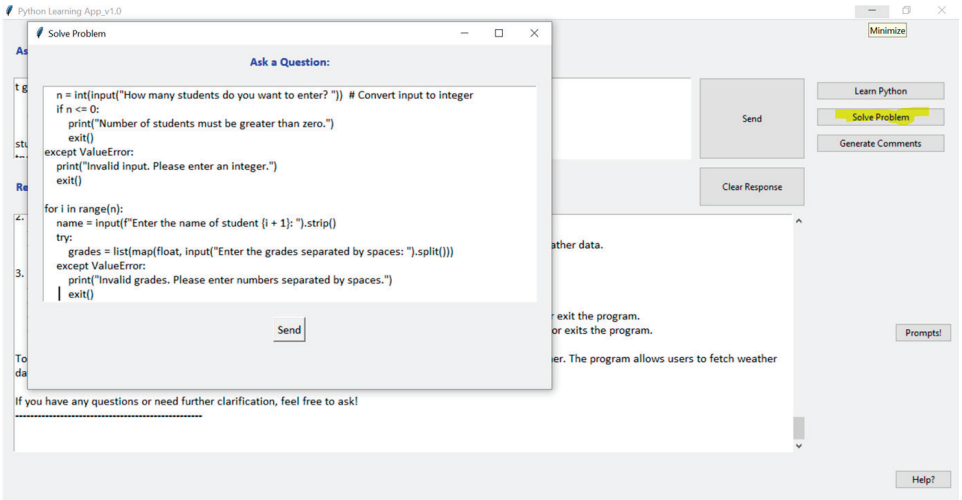


Figure 4. Screenshot of "Solve Problem" feature displaying a Python script input interface for entering student data and grades with basic error handling.

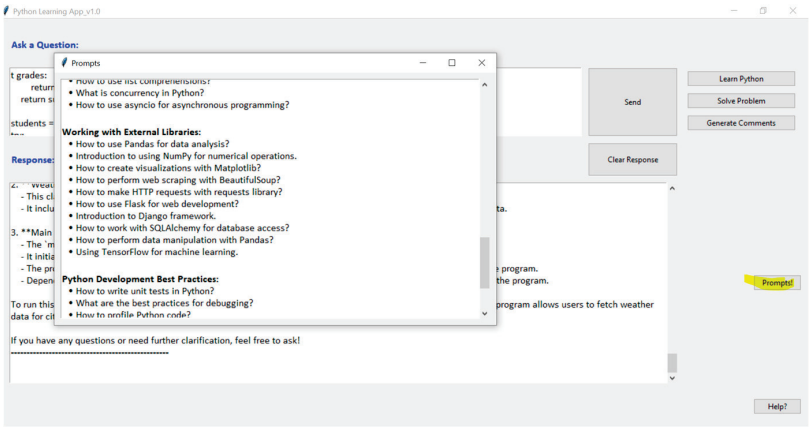


Figure 5. Screenshot of "Prompts" feature showing a categorised list of Python learning questions, covering external libraries and best development practices.

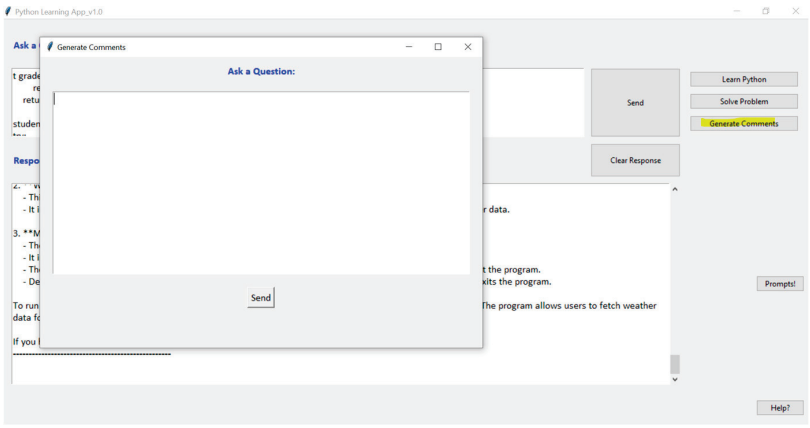


Figure 6. Screenshot of "Generate Comments" window in the Python Learning App, allowing users to input code and receive AI-generated comments to enhance code clarity.

4.4.1. Technical Contribution of PyChatAI

PyChatAI is not merely an OpenAI API wrapper but a domain-specific, adaptive learning system built with tailored features to support Python education in multilingual and culturally diverse contexts. The system was custom developed with the following technical contributions:

- Custom Prompt Engineering: PyChatAI uses domain-optimised prompt templates that scaffold learning based on student inputs, reducing hallucinations and ensuring responses are pedagogically aligned with Python course content.
- No Custom Dataset Training: The system does not train its own models but uses pre-trained OpenAI models. However, its performance is fine-tuned at runtime through controlled prompt engineering and contextual filtering.
- Adaptive Layer Built on OpenAI API: PyChatAI adds a real-time logic layer that monitors engagement patterns (e.g., frequent syntax errors and repeated queries) and offers tailored scaffolding. This adaptivity goes beyond simple request–response models.
- Arabic/English Multilingual Interface: the chatbot dynamically detects the language and adjusts UI elements, error messages, and responses accordingly.
- Debugging Assistant: this analyses common logic and syntax errors using pattern-matching and suggests improvements proactively based on detected coding issues.

4.4.2. Innovation Compared to Existing Tools

Unlike tools such as ChatGPT, Python-Bot, or Pyo, PyChatAI is purpose-built for academic Python instruction in Saudi Arabia. Table 3 provides a comparative analysis of

PyChatAI, ChatGPT, and Python-Bot/Pyo across six key features relevant to educational support in Python programming. This table clearly shows that PyChatAI offers a more comprehensive and education-specific feature set than ChatGPT and Python-Bot/Pyo, especially in areas such as proactive feedback, adaptive support, bilingual interaction, and curriculum integration.

Table 3. Comparison of PyChatAI and other AI tools (ChatGPT and Python-Bot/Pyo) based on key educational features.

Feature	PyChatAI	ChatGPT	Python-Bot/Pyo
Proactive Error Detection	Yes	No	No
Adaptive Scaffolding	Yes	No	No
Arabic/English Support	Yes (bilingual UI/feedback)	Limited	No
Curriculum-Aligned Prompts	Yes	No	No
Context-Aware Learning Feedback	Yes	No	No
Built-in Debugging Assistance	Yes	Basic (if prompted)	No

4.4.3. Pedagogical Framework

The system incorporates Vygotsky's Zone of Proximal Development (ZPD) by adjusting the level of assistance based on the complexity of the task and user engagement. This scaffolding approach ensures students can solve problems they may not be able to handle independently, gradually building their confidence and skill.

4.5. PyChatAI System Features and Development Phases

The system was developed in two phases to progressively enhance its educational capabilities:

1. Phase 1: Core Functionalities
    a. Live Question Assistance: Users can submit Python-related questions and receive instant responses. Contextual filters ensure all responses remain specific to Python programming.
    b. Multilingual Support: The system supports both English and Arabic, making it accessible to a diverse learner base. The interface and chatbot dynamically adjust to the selected language.
2. Phase 2: Enhanced Features
    a. Predefined Learning Prompts: interactive prompts guide users through fundamental Python concepts, such as loops, conditionals, and functions (Figure 5).
    b. Problem Description Form: students describe their coding challenges, and PyChatAI generates a basic logic framework to aid problem-solving.
    c. Interactive Debugging Assistance: the tool detects syntax errors and provides corrective suggestions in real-time (Figure 3).
    d. Code Analysis and Feedback: students can submit code to receive a fully commented version, helping them understand coding best practices and improve code readability (Figure 4).

4.6. Data Collection and Analysis

The evaluation of PyChatAI involves both quantitative and qualitative methods to provide a comprehensive understanding of its educational impact.

1. Quantitative Data Collection:
    - Pre-tests and Post-tests: programming assessments administered before and after the intervention to measure changes in coding skills and problem-solving abilities.

- Technology Acceptance Model (TAM3) Questionnaire: To investigate factors influencing student acceptance of PyChatAI, this study incorporates a TAM3-based questionnaire. This model evaluates key variables, such as perceived usefulness, ease of use, and behavioural intention to continue using the application.

2. Qualitative Data Collection:
   - Faculty Interviews: To gain a comprehensive understanding of the educational value of PyChatAI, faculty perspectives were gathered through structured interviews. These interviews focused on the effectiveness of PyChatAI in enhancing programming instruction, its integration into the curriculum, and potential areas for improvement.

This multi-faceted methodological approach enables a comprehensive evaluation of PyChatAI from both student and educator perspectives. It directly addresses the primary research objective of this study: to determine the effectiveness of PyChatAI in enhancing computer programming outcomes and to explore the key factors influencing its acceptance and adoption in educational settings.

5. Results

5.1. Pre-Test: Equivalence of Study Groups

To ensure the validity of this study's findings, it was essential to establish baseline equivalence between the control and experimental groups prior to the intervention. A pre-test was conducted across six core programming skill dimensions: theoretical understanding, code writing, software design, problem-solving and analysis, software testing and debugging, and overall programming competency.

An independent samples *t*-test was employed to compare mean scores between the groups. Assumptions of normality and homogeneity of variance were assessed using the Shapiro–Wilk and Levene's tests, respectively, and were satisfactorily met, confirming the suitability of the *t*-test for this analysis. In addition, effect sizes (Cohen's d) were calculated to determine the practical significance of any observed differences. The results indicate no statistically significant differences between the control and experimental groups across all measured domains. As summarised in Table 4, these findings affirm the initial equivalence of the groups, providing a reliable basis for evaluating the impact of the PyChatAI intervention.

Table 4. The results of the independent samples *t*-test for the pre-test of programming skills.

Skill	Group	N	Mean	Std. Deviation	T-Value	DF	*p*-Value
Theoretical Understanding	Control	150	2.17	1.14	0.593	298	0.554
	Experimental	150	2.26	1.56			
Code Writing Skills	Control	150	1.98	1.13	1.252	298	0.212
	Experimental	150	1.82	1.08			
Software Design	Control	150	1.77	1.13	1.534	298	0.126
	Experimental	150	1.56	1.21			
Problem Solving	Control	150	1.51	1.02	0.001	298	0.999
	Experimental	150	1.51	1.28			
Software Testing	Control	150	1.83	1.11	1.539	298	0.062
	Experimental	150	2.05	1.42			
Total Skills	Control	150	9.25	2.79	0.12	298	0.904
	Experimental	150	9.20	3.85			

Key Findings from the Pre-Test:

- For theoretical understanding, the control group (M = 2.17; SD = 1.14) and the experimental group (M = 2.26; SD = 1.56) showed no significant difference, t(298) = 0.593 and $p = 0.554$, indicating comparable baseline knowledge between the two groups.
- Similarly, in code-writing skills, the control group (M = 1.98; SD = 1.13) and the experimental group (M = 1.82; SD = 1.08) did not differ significantly, t(298) = 1.252 and $p = 0.212$, suggesting equivalent starting levels for this skill.
- In software design, no significant difference was observed between the control group (M = 1.77; SD = 1.13) and the experimental group (M = 1.56; SD = 1.21), t(298) = 1.534 and $p = 0.126$, confirming the groups' similarity prior to the intervention.
- For problem-solving skills, both groups had identical mean scores (M = 1.51), with no significant difference, t(298) = 0.001 and $p = 0.999$, further supporting the equivalence of the groups.
- In software testing and debugging, while the experimental group scored slightly higher (M = 2.05; SD = 1.42) than the control group (M = 1.83; SD = 1.11), this difference was not statistically significant, t(298) = 1.539 and $p = 0.062$.
- Finally, for total programming skills, the control group (M = 9.25; SD = 2.79) and the experimental group (M = 9.20; SD = 3.85) showed no significant difference, t(298) = 0.12 and $p = 0.904$, confirming overall equivalence across all measured skills.

5.2. Post-Test Results: Impact of PyChatAI

Programming skills were assessed post-intervention using a comprehensive programming skills test conducted after a four-month period. The results for both the control and experimental groups were analysed using the independent samples *t*-test to identify any significant differences between the two groups. This statistical method enabled a comparison of means to determine the impact of the PyChatAI application on the development of programming skills.

Table 5 presents the *t*-test results, highlighting the differences in programming skills across the measured dimensions for the two groups. These findings provide critical insights into the effectiveness of the intervention in enhancing programming skills among female computer science students at Jouf University.

Key Findings from the Post-Test:

The results, as summarised in Table 4, indicate statistically significant differences between the control and experimental groups across all measured dimensions of programming skills ($p < 0.001$ for all skills).

Table 5. Independent samples *t*-test results comparing programming skills between two groups.

Skill	Group	N	Mean	Std. Deviation	T-Value	DF	*p*-Value
Theoretical Understanding	Control Experimental	150 150	2.51 3.99	1.01 1.05	12.402	298	0.001>
Code Writing Skills	Control Experimental	150 150	2.31 4.02	1.01 1.01	14.719	298	0.001>
Software Design	Control Experimental	150 150	2.19 4.02	0.89 1.17	15.211	298	0.001>
Problem Solving	Control Experimental	150 150	2.09 4.03	0.99 1.21	15.096	298	0.001>
Software Testing	Control Experimental	150 150	2.22 4.00	0.90 1.12	15.120	298	0.001>
Total Skills	Control Experimental	150 150	11.33 20.05	2.53 4.40	21.068	298	0.001>

- The experimental group had a mean score of 3.99 in theoretical understanding compared to 2.51 for the control group, with a t-value of 12.402 and a p-value < 0.001, indicating a significant improvement due to the intervention.
- In code-writing skills, the experimental group achieved a mean score of 4.02, substantially higher than the control group's 2.31, with a t-value of 14.719 and a p-value < 0.001.
- The experimental group outperformed the control group in software design, with a mean score of 4.02 compared to 2.19, reflected by a t-value of 15.211 and a p-value < 0.001.
- In problem-solving, the experimental group's mean score of 4.03 significantly exceeded the control group's 2.09, with a t-value of 15.096 and a p-value < 0.001.
- The experimental group achieved a mean score of 4.00 in software testing, higher than the control group's 2.22, with a t-value of 15.120 and a p-value < 0.001.
- The total programming skill score showed a marked improvement in the experimental group, with a mean of 20.05 compared to 11.33 in the control group and a t-value of 21.068 with a p-value < 0.001, confirming the significant overall effectiveness of the PyChatAI application in enhancing programming skills.

Figures 7 and 8 show a comparison of computer programming skills between the control and experimental groups, demonstrating consistent and substantial improvements across all dimensions for students using PyChatAI. Figure 8 illustrates the distribution of computer programming skill scores, highlighting the higher median scores and reduced variance in the experimental group, which suggests both improvements and consistency in learning outcomes. Figure 9 compares the mean scores across different computer programming skills, clearly showing the superior performance of the experimental group in all categories.

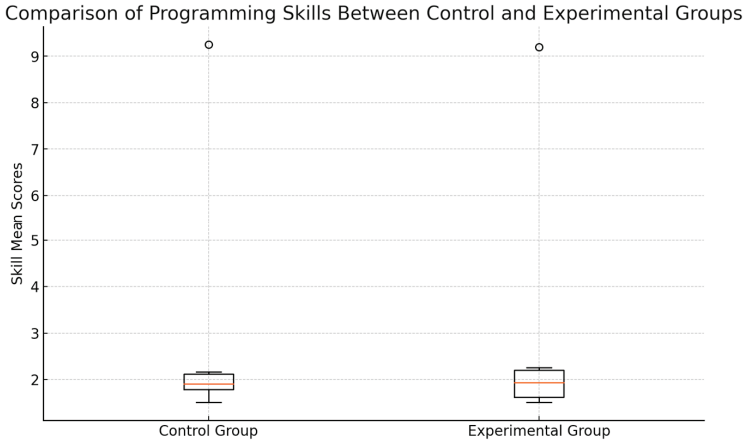


Figure 7. Boxplot comparison of programming skill mean scores between control and experimental groups, highlighting score distributions and outliers.

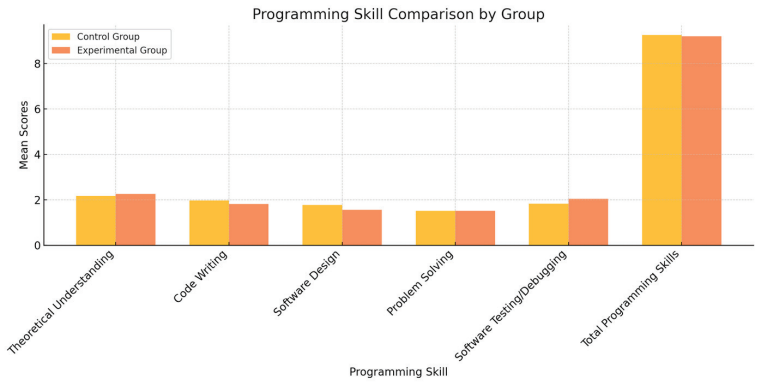


Figure 8. Mean programming skill scores by group, comparing control and experimental groups across individual skills and total performance.

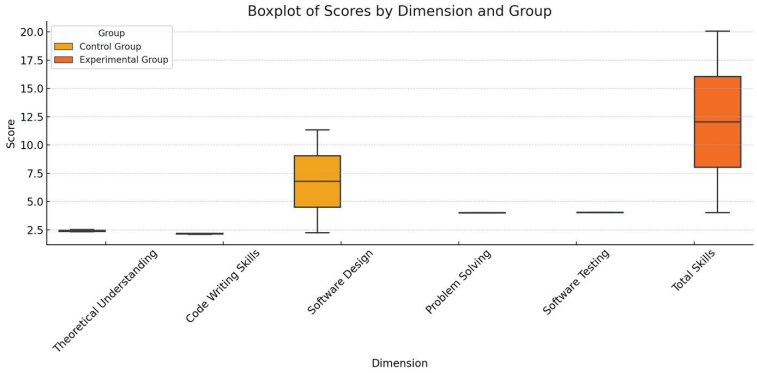


Figure 9. Boxplot showing differences in programming skill scores between control and experimental groups across six dimensions.

6. Discussion

6.1. Interpretation of Pre-Test Results

The pre-test results confirm that both the control and experimental groups possessed equivalent baseline skills across all dimensions of computer programming, including theoretical understanding, code writing, software design, problem-solving, and software testing. This statistical equivalence is crucial, as it ensures that any differences observed in the post-test are the result of the PyChatAI intervention, rather than pre-existing disparities between the groups. By establishing that both groups started with comparable skill levels, this study maintains strong internal validity, allowing for a clear attribution of post-intervention improvements to the use of the AI-powered educational tool. This baseline similarity provides a solid foundation for evaluating the impact of PyChatAI and supports the reliability of the conclusions of this study.

6.2. Impact of PyChatAI on Computer Programming Skills

The post-test results reveal a significant positive impact of PyChatAI as a smart learning system, demonstrating its effectiveness in the development of computer programming skills among the experimental group. The data demonstrate substantial improvements across all measured dimensions, indicating that the intervention effectively enhanced students' abilities in both theoretical knowledge and practical application.

In the areas of theoretical understanding and code writing, the experimental group showed marked improvements compared to the control group. Theoretical understanding is fundamental for comprehending computer programming principles, and the observed gains suggest that PyChatAI successfully facilitated students' grasp of complex concepts. Similarly, the improvement in code-writing skills indicates that the AI tool effectively supported students in translating their theoretical knowledge into practical coding abilities. This is likely due to the ability of PyChatAI to provide instant feedback, suggest code modifications, and guide students through common programming errors, fostering a more interactive and engaging learning experience.

The experimental group's superior performance in problem-solving and software testing further underscores the effectiveness of PyChatAI as a smart learning system that fosters computational thinking and coding proficiency. Problem-solving is a critical skill in computer programming, requiring not only technical knowledge but also analytical thinking and creativity. The significant gains in this area suggest that the interactive features of PyChatAI, such as real-time debugging support and tailored problem-solving prompts, played a key role in helping students develop these competencies. Likewise, improvements in software testing and debugging skills highlight the role of the tool in enhancing stu-

dents' error detection and correction abilities, which are essential for writing efficient and reliable code.

The most compelling evidence of the impact of PyChatAI is reflected in the total programming skill score. The experimental group achieved a mean score of 20.05, nearly double that of the control group's 11.33, indicating comprehensive support across multiple dimensions of computer programming. This dramatic improvement suggests that PyChatAI not only helps students understand individual concepts but also fosters a more holistic development of computer programming proficiency.

These findings are visually supported by Figures 9 and 10, which illustrate the differences in mean scores across various computer programming dimensions. The visual data reinforce the robustness of the intervention, clearly showing the experimental group's superior performance in all skill areas. The consistent improvements across multiple dimensions highlight the ability of PyChatAI to provide an integrated learning experience that addresses both foundational and advanced programming skills.

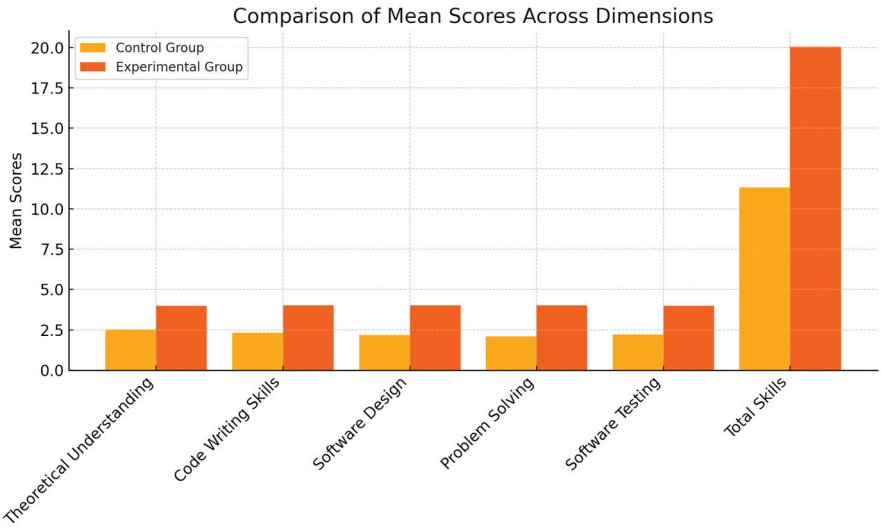


Figure 10. Bar chart comparing mean programming scores across dimensions for control and experimental groups.

6.3. Implications for Educational Practice

The significant improvements observed in the experimental group suggest that AI-powered smart learning systems, such as PyChatAI, have the potential to transform computer programming education by providing real-time feedback and personalised assistance. One of the key advantages of PyChatAI is its ability to provide real-time feedback and assistance. Unlike traditional teaching methods, which often involve delayed feedback, PyChatAI offers instant responses to students' queries. This immediacy fosters continuous engagement, helps students correct mistakes as they occur, and reduces the frustration commonly associated with programming challenges.

Moreover, the adaptive support mechanisms of PyChatAI enable personalised learning experiences. The tool tailors its feedback and guidance based on individual student needs, making it particularly effective for novice programmers who may struggle with the steep learning curve associated with programming. This level of personalisation helps students build confidence and maintain motivation, which are critical factors in successful learning.

The effectiveness of the tool in improving problem-solving and software debugging skills also highlights its role in promoting critical thinking and analytical abilities. These skills are not only essential for programming but are also valuable in a wide range of academic and professional contexts. By fostering these competencies, PyChatAI contributes

to the development of students' overall cognitive abilities, preparing them for more complex tasks and challenges in the future.

While the results of this study demonstrate the effectiveness of PyChatAI as a smart learning system, several limitations should be acknowledged. First, this study focused exclusively on female students at a single institution (Jouf University in Saudi Arabia). While this focus provided valuable insights into a specific demographic, it also limits the generalisability of the findings. Future research should include more diverse student populations, encompassing different genders, institutions, and cultural contexts, to determine whether the effectiveness of PyChatAI is consistent across various groups.

In addition, while PyChatAI showed significant improvements in Python programming, it remains unclear whether similar results would be observed in other programming languages or more advanced programming topics. Future studies should explore the applicability of the tool across different programming environments and complexity levels to better understand its broader potential.

Lastly, the current study focused on the short-term impact of PyChatAI over a four-month period. While the results are promising, longitudinal studies are needed to examine the sustained impact of AI-assisted learning on computer programming skills. Understanding whether the improvements observed in this study persist over time would provide valuable insights into the long-term benefits of integrating AI tools into educational settings.

In summary, the findings of this study highlight the transformative potential of AI-powered educational tools, such as PyChatAI, in enhancing computer programming skills. By providing real-time feedback, fostering critical thinking, and offering personalised learning experiences, PyChatAI represents a significant advancement in computer programming education. However, further research is needed to explore its applicability in broader contexts and to assess the long-term impact of AI-assisted learning on student outcomes.

7. Conclusions

In response to the growing demand for personalised, inclusive programming education, this study introduced PyChatAI, a bilingual AI-powered learning system designed to support novice Python learners through real-time feedback, adaptive guidance, and proactive problem-solving support.

Using a Solomon Four-Group design with 300 female students at Jouf University, this study demonstrated that PyChatAI significantly enhanced programming competencies across key areas—including code writing, theoretical understanding, and debugging. Compared to control groups, students using PyChatAI nearly doubled their overall programming performance. Unlike general-purpose tools, such as ChatGPT, PyChatAI offers curriculum-aligned, proactive support tailored to learners' progress. TAM3 responses confirmed high student acceptance, citing ease of use and perceived usefulness, while faculty feedback underscored its value in large classroom settings. The system's bilingual capabilities also helped reduce language barriers, supporting broader educational equity and aligning with the digital transformation goals of Saudi Vision 2030.

The outcomes of this study suggest significant relevance beyond the immediate setting. PyChatAI's design offers a model for developing AI-powered educational systems that are culturally localised, curriculum-aligned, and proactively assistive features that go beyond what existing tools, such as ChatGPT or Python-Bot, provide. As AI becomes increasingly integrated into education, such smart systems may play a transformative role in building equitable and personalised learning environments.

Despite these promising results, this study has certain limitations. It focused exclusively on female students at a single institution, limiting the generalisability of the findings. In addition, this study only evaluated short-term gains over a four-month period. Future research should aim for the following:

- Examine the system's applicability across diverse learner populations and institutions.
- Test PyChatAI in other programming languages and advanced computing topics.
- Explore long-term retention and the sustained impact of AI-assisted learning.
- Investigate how integrating real-time performance analytics and reinforcement learning could further personalise the learning experience.

In conclusion, PyChatAI demonstrates strong potential as a scalable and effective smart learning system capable of transforming computer programming education. By building on the strengths of existing tools, such as ChatGPT—while introducing proactive engagement, bilingual support, and alignment with local curricula—PyChatAI offers a uniquely adaptable solution for supporting novice programmers. Its proven ability to enhance student outcomes highlights the promise of AI-driven educational technologies in fostering more inclusive, personalised, and adaptive learning environments in computer programming education both in Saudi Arabia and globally.

Author Contributions: Conceptualization, M.A. and B.S.; methodology, M.A. and B.S.; software, M.A.; validation, M.A., B.S. and H.S.; formal analysis, M.A., B.S. and H.S.; investigation, M.A., B.S. and H.S.; resources, M.A., B.S. and A.L.; data curation, M.A. and H.S.; writing—original draft preparation, M.A. and B.S.; writing—review and editing, B.S., A.L. and H.S.; visualization, M.A., B.S. and H.S.; supervision, B.S. and A.L.; project administration, M.A., B.S., H.S. and A.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article; further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Alharthi, F. Challenges related to computer programming education in Saudi Arabia. *Proc. Eng.* **2024**, *6*, 37–44. [CrossRef]
2. Alothman, M.; Robertson, J.; Michaelson, G. Computer usage and attitudes among Saudi Arabian undergraduate students. *Comput. Educ.* **2017**, *110*, 127–142. [CrossRef]
3. Alanazi, A.; Li, A.; Soh, B. Effects on Saudi Female Student Learning Experiences in a Programming Subject Using Mobile Devices: An Empirical Study. *Int. J. Interact. Mob. Technol.* **2023**, *17*, 44–58. [CrossRef]
4. Alasmari, O.A.; Singer, J.; Ada, M.B. Do current online coding tutorial systems address novice programmer difficulties? In Proceedings of the 15th International Conference on Education Technology and Computers, Barcelona, Spain, 26–28 September 2023; pp. 242–248.
5. Zhong, X.; Zhan, Z. An intelligent tutoring system for programming education based on informative tutoring feedback: System development, algorithm design, and empirical study. *Interact. Technol. Smart Educ.* **2024**. *ahead-of-print*.
6. Abdulla, S.; Ismail, S.; Fawzy, Y.; Elhag, A. Using ChatGPT in Teaching Computer Programming and Studying Its Impact on Students Performance. *Electron. J. E-Learn.* **2024**, *22*, 66–81. [CrossRef]
7. Mekthanavanh, V.; Meksavanh, B.; Bouddy, S. ChatGPT Enhances Programming Skills of Computer Engineering Students. *Souphanouvong Univ. J. Multidiscip. Res. Dev.* **2024**, *10*, 1–9. [CrossRef]
8. Stone, I. Investigating the Use of ChatGPT to Support the Learning of Python Programming Among Upper Secondary School Students: A Design-Based Research Study. In Proceedings of the 2024 Conference on United Kingdom & Ireland Computing Education Research, Manchester, UK, 5–6 September 2024; p. 1.
9. Maguire, J.; Cutts, Q. Supporting the Computing Science Education Research Community with Rolling Reviews. In Proceedings of the United Kingdom & Ireland Computing Education Research Conference, Glasgow, UK, 3–4 September 2020; ACM: New York, NY, USA, 2020.
10. Van Der Stuyf, R.R. Scaffolding as a teaching strategy. *Adolesc. Learn. Dev.* **2002**, *52*, 5–18.

11. Yilmaz, R.; Yilmaz, F.G.K. The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation. *Comput. Educ. Artif. Intell.* **2023**, *4*, 100147. [CrossRef]
12. Rahman, M.; Watanobe, Y. ChatGPT for education and research: Opportunities, threats, and strategies. *Appl. Sci.* **2023**, *13*, 5783. [CrossRef]
13. Silva, C.A.G.D.; Ramos, F.N.; De Moraes, R.V.; Santos, E.L.D. ChatGPT: Challenges and benefits in software programming for higher education. *Sustainability* **2024**, *16*, 1245. [CrossRef]
14. Tick, A. Exploring ChatGPT's Potential and Concerns in Higher Education. In Proceedings of the 2024 IEEE 22nd Jubilee International Symposium on Intelligent Systems and Informatics (SISY), Pula, Croatia, 19–21 September 2024.
15. Wieser, M.; Schöffmann, K.; Stefanics, D.; Bollin, A.; Pasterk, S. Investigating the role of ChatGPT in supporting text-based programming education for students and teachers. In Proceedings of the International Conference on Informatics in Schools: Situation, Evolution, and Perspectives, Lausanne, Switzerland, 23–25 October 2023; Springer Nature: Cham, Switzerland, 2023; pp. 40–53.
16. Ma, B.; Chen, L.; Konomi, S.I. Enhancing programming education with ChatGPT: A case study on student perceptions and interactions in a Python course. In Proceedings of the International Conference on Artificial Intelligence in Education, Recife, Brazil, 8–12 July 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 113–126.
17. Jalon, J.B.; Chua, G.A.; Torres, M.D.L. ChatGPT as a learning assistant: Its impact on students learning and experiences. *Int. J. Educ. Math. Sci. Technol.* **2024**, *12*, 1603–1619. [CrossRef]
18. Akçapınar, G.; Sidan, E. AI chatbots in programming education: Guiding success or encouraging plagiarism. *Discov. Artif. Intell.* **2024**, *4*, 87. [CrossRef]
19. Sandu, R.; Gide, E.; Elkhodr, M. The role and impact of ChatGPT in educational practices: Insights from an Australian higher education case study. *Discov. Educ.* **2024**, *3*, 71. [CrossRef]
20. Haindl, P.; Weinberger, G. Does ChatGPT help novice programmers write better code? Results from static code analysis. *IEEE Access* **2024**, *12*, 114146–114156. [CrossRef]
21. Okonkwo, C.W.; Ade-Ibijola, A. Python-bot: A chatbot for teaching python programming. *Eng. Lett.* **2020**, *29*, 25.
22. Carreira, G.; Silva, L.; Mendes, A.J.; Oliveira, H.G. Pyo, a chatbot assistant for introductory programming students. In Proceedings of the 2022 International Symposium on Computers in Education (SIIE), Coimbra, Portugal, 17–19 November 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6.
23. Abolnejadian, M.; Alipour, S.; Taeb, K. Leveraging chatgpt for adaptive learning through personalized prompt-based instruction: A cs1 education case study. In Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 11–16 May 2024; Association for Computing Machinery: New York, NY, USA, 2024; pp. 1–8.
24. Popovici, M.-D. ChatGPT in the classroom. Exploring its potential and limitations in a functional programming course. *Int. J. Hum. Comput. Interact.* **2024**, *40*, 7743–7754. [CrossRef]
25. Vukojičić, M.; Krstić, J. ChatGPT in programming education: ChatGPT as a programming assistant. *InspirED Teach. Voice* **2023**, *2023*, 7–13.
26. Xue, Y.; Chen, H.; Bai, G.R.; Tairas, R.; Huang, Y. Does ChatGPT help with introductory programming? An experiment of students using ChatGPT in cs1. In Proceedings of the 46th International Conference on Software Engineering: Software Engineering Education and Training, Lisbon, Portugal, 14–20 April 2024; pp. 331–341.
27. Güner, H.; Er, E.; Akçapinar, G.; Khalil, M. From chalkboards to AI-powered learning. *Educ. Technol. Soc.* **2024**, *27*, 386–404.
28. Eilermann, S.; Wehmeier, L.; Niggemann, O.; Deuter, A. KIAAA: An AI assistant for teaching programming in the field of automation. In Proceedings of the 2023 IEEE 21st International Conference on Industrial Informatics (INDIN), Lemgo, Germany, 18–20 July 2023; pp. 1–7.
29. Essel, H.B.; Vlachopoulos, D.; Nunoo-Mensah, H.; Amankwa, J.O. Exploring the impact of VoiceBots on multimedia programming education among Ghanaian university students. *Br. J. Educ. Technol.* **2025**, *56*, 276–295. [CrossRef]
30. Daniel, G.; Cabot, J.; Deruelle, L.; Derras, M. Xatkit: A multimodal low-code chatbot development framework. *IEEE Access* **2020**, *8*, 15332–15346. [CrossRef]
31. Daniel, G.; Cabot, J. The software challenges of building smart chatbots. In Proceedings of the 2021 IEEE/ACM 43rd International Conference on Software Engineering: Companion Proceedings (ICSE-Companion), Madrid, Spain, 25–28 May 2021; pp. 324–325.
32. Daniel, G.; Cabot, J. Applying model-driven engineering to the domain of chatbots: The Xatkit experience. *Sci. Comput. Program.* **2024**, *232*, 103032. [CrossRef]
33. Okonkwo, C.W.; Ade-Ibijola, A. Revision-Bot: A Chatbot for Studying Past Questions in Introductory Programming. *IAENG Int. J. Comput. Sci.* **2022**, *49*, 644.
34. Bassner, P.; Frankford, E.; Krusche, S. Iris: An ai-driven virtual tutor for computer science education. In Proceedings of the 2024 on Innovation and Technology in Computer Science Education, Milan, Italy, 8–10 July 2024; Association for Computing Machinery: New York, NY, USA, 2024; Volume 1, pp. 394–400.
35. Daud, S.H.M.; Teo, N.H.I.; Zain, N.H.M. Ejava chatbot for learning programming language: Apost-pandemic alternative virtual tutor. *Int. J.* **2020**, *8*, 3290–3298.

36. Chrysafiadi, K.; Virvou, M. PerFuSIT: Personalized Fuzzy Logic Strategies for Intelligent Tutoring of Programming. *Electronics* **2024**, *13*, 4827. [CrossRef]
37. Biňas, M.; Pietriková, E. Impact of virtual assistant on programming novices' performance, behavior and motivation. *Acta Electrotech. Inform.* **2022**, *22*, 30–36. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

computers

MDPI

Article

A Systematic Approach to Evaluate the Use of Chatbots in Educational Contexts: Learning Gains, Engagements and Perceptions

Wei Qiu, Chit Lin Su, Nurabidah Binti Jamil, Maung Thway, Samuel Soo Hwee Ng, Lei Zhang, Fun Siong Lim and Joel Weijia Lai *

Applications of Teaching and Learning Analytics for Students, Institute for Pedagogical Innovation, Research and Excellence, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore; qiuwei@ntu.edu.sg (W.Q.); lschit@ntu.edu.sg (C.L.S.); nurabidah.jamil@ntu.edu.sg (N.B.J.); maung.thway@ntu.edu.sg (M.T.); samuel.ng@ntu.edu.sg (S.S.H.N.); lei_zhang@ntu.edu.sg (L.Z.); lim_fun_siong@ntu.edu.sg (F.S.L.)
* Correspondence: joellai@ntu.edu.sg

Abstract

As generative artificial intelligence (GenAI) chatbots gain traction in educational settings, a growing number of studies explore their potential for personalized, scalable learning. However, methodological fragmentation has limited the comparability and generalizability of findings across the field. This study proposes a unified, learning analytics–driven framework for evaluating the impact of GenAI chatbots on student learning. Grounded in the collection, analysis, and interpretation of diverse learner data, the framework integrates assessment outcomes, conversational interactions, engagement metrics, and student feedback. We demonstrate its application through a multi-week, quasi-experimental study using a Socratic-style chatbot designed with pedagogical intent. Using clustering techniques and statistical analysis, we identified patterns in student–chatbot interaction and linked them to changes in learning outcomes. This framework provides researchers and educators with a replicable structure for evaluating GenAI interventions and advancing coherence in learning analytics–based educational research.

Keywords: generative artificial intelligence; application of retrieval-augmented generation; student–chatbot interaction; learning analytics; artificial intelligence in education

1. Introduction

Artificial intelligence (AI) is increasingly being integrated into educational environments, providing innovative solutions to enhance the learning experience. Among these AI technologies, chatbots have gained significant attention for their ability to simulate human conversation and provide real-time assistance and information [1,2]. Defined as AI-powered conversational agents, chatbots are extensively used to support classroom activities, facilitate communication, and provide real-time assistance to students. Research on learning chatbots in education has explored various dimensions, including student engagement [3,4] and learning gains [5,6]. These studies have highlighted the benefits of chatbots in creating flexible and personalized learning environments, improving student motivation, and reducing cognitive load through interactive and adaptive learning experiences. However, as shown in a systematic review, the vast majority (88.88%) of chatbots

studied follow predetermined learning paths, which can restrict flexibility and critical inquiry [7].

The emergence of generative AI (GenAI) chatbots offers a promising departure from such constraints. Unlike rule-based systems, GenAI chatbots can generate open-ended, contextually rich responses, enabling inquiry-driven learning that supports critical thinking and reflection. Their potential to emulate pedagogical strategies, such as the Socratic method, has attracted increasing interest. For instance, the Socratic method, based on structured questioning to stimulate critical reasoning [8], has been successfully implemented in AI teaching assistants to deepen student understanding [9]. Recent studies further highlight the capabilities of chatbots to reduce teacher workload [10], enhance performance and social interaction [11], and support students' self-regulated learning [12,13].

Despite these advances, there remains a gap in understanding the comprehensive impact of chatbots on learning gains, particularly in terms of student perceptions, engagement, and academic performance. While GenAI chatbots have been widely adopted in non-educational fields such as healthcare, their application in mainstream education is emerging and worthy of investigation [14,15]. Studies have called for more rigorous investigations into the educational capabilities of GenAI chatbots, including student assessment and feedback processes [16,17]. This research gap necessitates a systematic approach to evaluate the use of GenAI chatbots in educational contexts, particularly in terms of their impact on learning gains [18].

This study addresses that gap by proposing a consolidated, learning analytics–informed framework for evaluating GenAI chatbots in education. Grounded in the definition of learning analytics as the "measurement, collection, analysis, and reporting of data about learners and their contexts" to improve learning and learning environments, our framework integrates diverse data sources such as assessment scores, conversational interactions, engagement patterns, and student feedback to offer a holistic view of chatbot impact. We apply this framework in a multi-week quasi-experimental study using a pedagogically customized GenAI chatbot. The chatbot was implemented using a Socratic dialogue structure, meaning it deliberately avoided providing explicit answers to student queries. Instead, it posed guiding, open-ended questions aimed at encouraging learners to articulate reasoning, evaluate assumptions, and construct knowledge independently. This form of Socratic scaffolding is intended to cultivate deeper engagement and support self-regulated learning, consistent with prior work on metacognitive dialogue systems. The chatbot included a structured system prompt, a curated knowledge base, and context-aware interaction capabilities based on students' conversation histories.

A preliminary human evaluation was conducted to ensure the chatbot's pedagogical soundness [19]. In the follow-up experiment, students were enrolled into a control and experiment group. Due to the anticipated value of the intervention, a larger number of participants were allocated to the intervention group. This unequal allocation was also a deliberate design choice to increase the statistical power of outcome analyses for the experimental condition, where more variability in learning behavior was expected. Both groups used GPT-4-based chatbots, with only the experimental group accessing the pedagogically enhanced version. Participants were blinded to group assignment. Over a three-week period, we collected pre- and post-test data, analyzed chatbot conversations, and gathered student feedback. This mixed-methods approach allowed us to assess not only learning gains but also behavioral engagement and learner perceptions. By situating our work within a unified analytical framework, we contribute both a practical evaluation model and empirical evidence to guide the implementation of GenAI chatbots in educational contexts. Ultimately, this study aims to inform best practices for deploying AI chatbots to support adaptive, interactive, and independent learning.

This paper is structured to present both the context and contribution of the study. In Section 2, we explore prior research on the use of chatbots in education, highlighting current gaps and opportunities. Section 3 introduces our proposed evaluation framework, designed to offer a comprehensive approach for assessing educational chatbots. Section 4 details the design and implementation of our experimental study, followed by the presentation and discussion of findings in Section 5. Finally, Section 6 summarizes key insights and outlines directions for future research and practice.

2. Literature Review

Chatbots, automated programs designed to perform tasks that mimic human interactions, have evolved significantly since their inception in the 1960s. The first chatbot, ELIZA, was designed to simulate psychotherapy conversations [20]. Subsequent advances, such as PARRY, introduced better personality and control structures [21]. Modern chatbots powered by AI include intelligent voice assistants such as Siri, Alexa, Cortana, and Google Assistant, which can perform a wide range of tasks, including home automation and email management [22,23].

The educational application of chatbots is an emerging field. Educational chatbots facilitate learner-machine interactions through mobile instant messaging applications [24]. They enhance the learning experience by providing instant feedback, supporting multimedia content, and personalizing learning according to individual preferences [25–27]. In addition, chatbots promote active participation, collaboration, and self-paced learning, contributing to improved educational outcomes [28–30].

Theoretical frameworks that support the use of chatbots in education include constructivism, cognitive load theory, and social learning theory. Constructivism emphasizes the active construction of knowledge by learners, which chatbots facilitate through interactive and personalized learning environments [30]. Cognitive load theory emphasizes the importance of balancing information input and processing, with chatbots helping to manage the cognitive load by providing structured and easily navigable content [31,32]. Social learning theory asserts that learning occurs through social interactions, with chatbots facilitating collaboration and peer learning [29,33]. Besides, Winkler and Soellner [5] found that students using AI-powered chatbots in a blended learning setting demonstrated improved engagement and self-regulated learning skills. Similarly, Ruan et al. [34] observed that chatbots facilitated more frequent and meaningful interactions in online components of blended courses, leading to enhanced learning gains.

However, the effectiveness of chatbots in educational settings is not without challenges. Issues such as the potential for misinformation, the need for emotional intelligence in responses, and the importance of maintaining human oversight have been highlighted by researchers [35]. These concerns underscore the importance of human evaluation and careful implementation of AI tools in educational contexts. While chatbots offer promising benefits for educational purposes, evaluating chatbot responses is still crucial to ensure their accuracy, relevance, and ability to enhance learning gains, thereby supporting effective pedagogy and improving student engagement and satisfaction [7,10]. In recent years, the evaluation of chatbot responses typically focuses on several key criteria. Al-Hafdi and AlNajdi [17] presented holistic findings of undergraduate students' perceptions of the chatbot-based learning environment in computer skills. The results showed the acceptability of chatbots among students and a positive trend toward their use in the learning process. This affords learners the autonomy to determine the optimal context and temporal parameters for learning, the repetition of exercises, and the accessibility of content in a manner that aligns with their individual preferences. Huang et al. [36] proposed a set of design principles for chatbots for meaningfully implementing educational chatbots in

language learning. Findings revealed three technological affordances: timeliness, ease of use, and personalization. Chatbots appeared to encourage students' social presence by affective, open, and coherent communication. Another recent study evaluated GenAI chatbots' tutoring capability [19]. The evaluation rubric measures not only the accuracy of the answers from the chatbot but also how well the chatbot can perform guidance regarding clarity and helpfulness. These studies demonstrate the chatbot's effectiveness as a learning tool.

The integration of chatbots presents new opportunities for data collection and analysis. The interactions between students and AI systems can provide rich data on learning processes, enabling a more nuanced understanding of student engagement, conceptual development, and potential areas for intervention [37]. With the rich data on learning processes, a systematic evaluation and comprehensive studies are needed to understand a chatbot's impact on student learning fully, particularly in terms of their impact across a spectrum of learning gains.

3. Chatbot-Learning Analytics Framework

The learning analytics field advocates a rigorous human-led evaluation framework that focuses on learning [37,38]. Leveraging human insight and not purely a data-driven approach is important to enhance learning processes and outcomes. This stand is aligned with calls for humans in the loop or over the loop in the application of AI across a variety of fields, including education [39,40]. Aligned with this principle, we propose a comprehensive framework that puts the human at the center of a systematic evaluation framework through measured outcomes and reporting. The framework comprises an overall methodology that captures long-term learning impact, learner-centric outcome measures, essential data sources including self-reports, appropriate analytics methods for each outcome, and for analysis across the outcomes. In line with the research practices, the presentation needs to discuss the insights related to existing theories and other empirical studies while also being transparent about the assumptions and limitations.

This framework also aligns with recent calls for more human-centered and agency-aware approaches in learning analytics and educational AI systems. Prior work emphasizes the importance of triangulating behavioral, cognitive, and affective data to build holistic insights into learning processes while foregrounding the learner's voice and experience in analytic interpretations [41,42]. Moreover, the organization of the framework in Table 1 is theoretically grounded in three complementary learning theories. Constructivist learning theory informs the Learning gains domain by emphasizing knowledge construction through active cognitive engagement, which we operationalize through pre-post learning assessments and the use of an educational chatbot. Cognitive load theory supports the Engagement domain by highlighting the importance of managing learners' attention, effort, and mental workload—captured through behavioral indicators such as task persistence and interaction patterns. Finally, social learning theory underpins the Perception domain, acknowledging that learners' beliefs about fairness, trust, and the credibility of instructional agents influence their motivation and participation. These theories collectively shape the structure of the framework by offering a foundation for selecting, categorizing, and interpreting diverse data sources across cognitive, behavioral, and affective dimensions.

Table 1 outlines a flexible yet comprehensive study framework designed to evaluate educational interventions across multiple dimensions of learning, incorporating learning analytics with student–chatbot interaction. The framework accommodates a range of methodological approaches, including but not limited to longitudinal randomized control trials (RCTs) and quasi-experimental studies. These methodologies serve as illustrative examples; the framework is equally applicable to other empirical designs such as cross-sectional

studies, case studies, design-based research (DBR), action research, and mixed-methods studies. Each approach allows for the investigation of complex educational phenomena through various lenses of inquiry. The framework is organized around three primary outcome domains, namely, learning gains, engagement, and perception, each supported by appropriate data sources and analytical strategies. The intention is to offer adaptable guidance for both quantitative and qualitative investigations in education research.

Table 1. The proposed study framework outlining the overall research design, key outcome measures, essential data sources, and example analytical methods.

Methodology	Longitudinal randomized controlled trials, quasi-experimental studies, cross-sectional studies, case studies, design-based research, and action research		
Outcome Measures	Learning gains	Engagement	Perception
Essential Data Sources	Pre- and post-tests, standardized assessments, performance tasks, rubric-based evaluations, and cognitive tasks	Log files, timestamps, forum posts, and video observations	Surveys, interviews, focus groups, and reflective journals
Example Analytical Methods	Independent and paired-sample t-tests, analysis of variance, effect size calculations, regression analysis, K-means clustering, and item response theory	Temporal sequence analysis, clickstream pattern mining, epistemic network analysis, discourse and conversational analysis, and inter-rater agreement testing for coded data	Thematic analysis of qualitative data, grounded theory development, natural language sentiment analysis, and exploratory and confirmatory factor analysis
Synthesizing Across Outcomes	Correlation matrices, predictive regression modeling, structural equation modeling, longitudinal growth curve modeling, side-by-side integration of qualitative and quantitative results through joint display analysis, and triangulation across methods		
Reporting	Relating findings to existing theories or empirical studies Stating assumptions and limitations Considering generalizability, transferability, reproducibility, and scalability		

Learning gains are one of the most critical outcome measures in evaluating educational interventions. They refer to measurable improvements in a learner's knowledge, skills, or conceptual understanding resulting from participation in an instructional experience. The study of learning gains often involves pre- and post-intervention assessments, which provide direct measures of student progress over time. These may include standardized tests, custom-designed assessments, performance-based tasks, or concept inventories. Quantitative methods are frequently used to evaluate learning gains due to their ability to yield statistically generalizable insights. Effect size metrics such as Cohen's d or Hedges' g are particularly valuable, as they quantify the magnitude of change, providing context beyond mere statistical significance. T-tests [43], Analysis of Covariance (ANCOVA) [44], and repeated measures Analysis of Variance (ANOVA) are commonly applied when comparing pre- and post-intervention scores, especially in randomized or quasi-experimental settings. Clustering techniques, such as K-means clustering or latent profile analysis, allow researchers to identify patterns among subgroups of learners who may respond differently to the intervention [45]. For instance, high- and low-performing learners may exhibit distinct improvement trajectories that are obscured in group-level statistics. Item Response Theory (IRT) can also be employed to assess item-level learning gains while controlling for item difficulty, enhancing precision in test interpretation [46]. In more exploratory or design-based research settings, qualitative approaches such as cognitive task analysis or think-aloud protocols can complement quantitative findings by revealing the processes behind conceptual change. These methods are useful for understanding not just whether learning occurred, but how learners construct and revise their understanding.

Engagement is a multifaceted construct encompassing behavioral, emotional, and cognitive dimensions. Understanding engagement is essential for identifying whether and how learners are interacting meaningfully with instructional content, peers, and learning technologies. Because engagement is dynamic and context-dependent, it is particularly well-suited to mixed-methods approaches that combine log data analysis with observational and interpretive techniques. Behavioral engagement is often captured using log files, such as timestamped records of page views, clicks, discussion posts, or assignment submissions, these are collectively termed as trace data. Sequence mining and temporal analysis enable researchers to uncover patterns in how students navigate content or interact with online platforms over time. For example, a temporal analysis may reveal whether students engage more intensively before assessment deadlines or show sustained engagement across a course module. Cognitive engagement, which involves mental investment in learning, can be examined through epistemic network analysis or discourse analysis of forum posts, reflective journals, or problem-solving transcripts [47]. These methods reveal the depth of reasoning, argumentation quality, and conceptual connections made by learners. Epistemic analysis, in particular, models the co-occurrence of cognitive moves (e.g., explaining, questioning, critiquing) and provides a network-based representation of knowledge-building practices [48]. Emotional engagement is more elusive but can be inferred through sentiment analysis of student comments or survey data on motivation, interest, and frustration. Combining these with qualitative methods such as interviews or observational protocols offers richer insight into learners' affective experiences. For instance, video data or in-class observation logs may reveal non-verbal cues indicative of disengagement or excitement. Importantly, engagement measures allow for real-time feedback in adaptive learning environments and can inform iterative improvements in instructional design. Engagement data may also be used as a mediating variable in models predicting learning outcomes, offering insight into the mechanisms through which interventions produce change.

Lastly, **perception** refers to learners' subjective evaluations of their learning experience, including how they interpret the value, usability, fairness, and emotional impact of an educational intervention. While perception does not directly measure learning, it influences motivation, satisfaction, persistence, and future behavior, making it a vital dimension of evaluation. Survey instruments with closed- and open-ended questions are common tools for capturing perception. Quantitative items often use Likert scales to assess constructs such as perceived usefulness, satisfaction, clarity, or self-efficacy. These responses can be analyzed using descriptive statistics, factor analysis, or structural equation modeling (SEM) to explore relationships among latent variables and outcomes [49]. For deeper insight, qualitative methods such as semi-structured interviews, focus groups, and open-ended survey responses are employed. Thematic analysis allows researchers to code and categorize recurring ideas, opinions, and emotions expressed by learners [50]. Grounded theory can be used to build conceptual frameworks based on emergent patterns in perception data, particularly when exploring new or under-theorized educational innovations. Sentiment analysis, increasingly automated through natural language processing (NLP), is useful for large-scale analysis of written feedback (e.g., course evaluations, online reviews). These methods provide quick assessments of learner attitudes but may lack the nuance of manual coding or interview-based insights. Understanding learner perception is essential not only for assessing user satisfaction but also for identifying mismatches between instructional intent and learner experience. For example, a technologically sophisticated tool may be perceived as confusing or intimidating, undermining its pedagogical impact. Conversely, a modest intervention that is perceived as intuitive and supportive may foster greater engagement and learning gains.

A key strength of the proposed framework lies in its capacity to support cross-domain analysis (i.e., the examination of relationships among different outcome categories such as learning gains, engagement, and perception). This integrative approach acknowledges that educational interventions do not operate in isolation; rather, they influence and are influenced by complex interactions between learner behavior, cognitive development, and affective responses. One of the most common forms of cross-domain analysis is correlational analysis, which is used to investigate potential associations between variables across outcome domains. For example, researchers may explore whether levels of engagement (e.g., time on task, participation in discussions) are positively associated with learning gains. These analyses can be performed using Pearson or Spearman correlations, depending on the distribution of the data, and may serve as a preliminary step toward more complex modeling. When the goal is to understand predictive relationships, multiple regression, logistic regression, or hierarchical linear modeling (HLM) may be applied. These models allow researchers to assess whether variables such as perception (e.g., perceived usefulness or satisfaction) or engagement behaviors predict post-test performance, controlling for relevant covariates such as prior knowledge, demographics, or instructional modality. As mentioned previously, SEM provides a powerful tool for modeling latent constructs and testing theoretical pathways across outcome domains. For example, SEM can be used to test a mediation model where perception of an intervention influences engagement, which in turn affects learning outcomes. SEM is particularly useful when combining multiple observed indicators (e.g., survey items and log data) into higher-order constructs, offering a more nuanced understanding of the mechanisms at play.

In studies employing both quantitative and qualitative data, triangulation serves as a strategy for cross-domain analysis. By comparing findings from different data types (e.g., aligning engagement metrics from log data with qualitative themes from interviews), researchers can strengthen the validity of their interpretations. For instance, patterns in system usage might be better understood when supplemented with qualitative data that explains why certain learners disengaged or excelled. A related strategy is joint display analysis, where qualitative and quantitative findings are presented side-by-side in tables or figures to highlight convergences, divergences, and areas requiring further interpretation. These integrative visualizations are increasingly used in mixed-methods research to promote clarity and synthesis. Lastly, to capture the dynamic nature of learning processes, temporal methods such as sequence mining, Markov models, or dynamic time warping can be applied across engagement and learning data. These methods allow researchers to detect critical transitions or behavioral patterns over time that correlate with learning outcomes, such as "productive failure" sequences or persistent trial-and-error behavior. In longitudinal studies, growth curve modeling may be used to assess change trajectories in learning or engagement and how these are moderated by perception variables or intervention types.

A rigorous framework requires not only robust analysis but also transparent and reflective reporting that contextualizes findings, articulates limitations, and supports theoretical development or practical application. Findings should be interpreted in light of relevant educational theories (e.g., constructivism, self-determination theory, cognitive load theory), which can provide explanatory power and guide the articulation of mechanisms underlying observed effects. For example, increases in engagement may be interpreted through the lens of intrinsic motivation or flow theory, while learning gains may be linked to scaffolded learning or deliberate practice models. Researchers are encouraged to explicitly state how their findings confirm, challenge, or extend existing theoretical frameworks. This alignment strengthens the study's contribution to cumulative knowledge building in the field. Reporting should also address the practical relevance of findings. In particular, re-

searcher should consider how their findings can be generalized, transferable, reproducible, and scalable [51].

The generalizability of this framework lies in its conceptual structure and methodological flexibility. It is designed to organize and integrate a variety of instruments and data sources used in evaluating AI-enhanced learning environments. While our own exploratory study uses one instantiation, many prior studies—each using different indicators, designs, and domains—can be retroactively situated within this framework. For example, Lai et al. [13] examined student learning trajectories in relation to chatbot-mediated problem solving using delayed retention measures and process-oriented epistemic network analysis. Their design fits squarely within our framework's learning gains domain, using extended outcome measures and cognitive task analytics. Noroozi et al. [52] proposed multimodal measurement techniques, such as physiological signals and behavioral trace data, to assess emotional and cognitive engagement. These map onto our engagement domain, particularly in the use of sequential and network-based analytics. Ait Baha et al. [11] investigated student trust and fairness perceptions in AI-based educational tools using structured perception surveys. This aligns with our Perception domain and illustrates the use of validated affective instruments and sentiment analysis. These examples demonstrate that the framework is not only applicable to our own study but also capable of integrating diverse methodological approaches across the literature. This reinforces its utility as a generalizable, organizing lens for learning analytics research involving AI systems.

4. Experimental Study

4.1. Chatbot Design and Evaluation

In this section, we demonstrate how an experimental study, guided by learning analytics and structured by the proposed framework, can achieve specific educational outcomes. The aim of this study is not to validate a particular AI tool or analytical method but to demonstrate the use of a generalizable evaluation framework that can structure multimodal, human-centered learning analytics research. Our implementation is exploratory and illustrative, intended to show how diverse data sources and outcomes can be integrated using a coherent evaluative lens.

Prior to this study, we conducted a preliminary investigation that evaluated the capabilities of GenAI models in educational contexts, with a focus on answer accuracy, stepwise guidance, and overall response quality [19]. This previous study compared the performance of four GenAI tutors, including GPT-4 (the baseline model used in the current study), and their retrieval-augmented generation (RAG) variants [53–56]. These models were assessed on their ability to answer questions and provide guidance across various statistical topics. A key feature of the previous research was its use of human evaluation, which emphasized the importance of expert judgment in assessing AI-driven educational tools, particularly in capturing nuances that automated metrics may miss. This diagnostic step was intended to guide the development of the Socratic meta-prompting strategy. Insights from this phase enabled us to select the LLM for the experimental study, fine-tune the Socratic behavior, improve the accuracy and tone of responses, and ensure alignment with pedagogical goals. While not part of the core experimental design, this step informed how we structured and justified the intervention used in the experimental condition.

The current experimental study employs a pedagogically designed GenAI chatbot that functions as a Socratic learning tutor (see Figure 1). This chatbot was prompted to help students understand statistics coursework by breaking down their questions into manageable steps and guiding them through each phase of the problem-solving process. This approach mirrors a scaffolding strategy that supports deeper learning. The chatbot also managed the pace of questioning, ensuring that each student's response was considered

before advancing. To reduce hallucinations and improve factual accuracy, the chatbot used an RAG pipeline. RAG is a method that enhances language models by integrating relevant external documents into the generation process. In this study, course materials were first processed to build a searchable index. Documents in portable document format (PDF) were segmented into 1000-character chunks with a 100-character overlap. Each chunk was embedded using OpenAI embeddings and then indexed in a Facebook AI Similarity Search (Faiss) vector database. When a student submitted a query, the system retrieved the four most semantically relevant chunks from the database. These were concatenated and sent to the language model (GPT-4-Turbo), along with the ten most recent query-response pairs, to produce a context-aware reply. This pipeline allowed the chatbot to draw on course-specific content while maintaining the conversational and reasoning strengths of the underlying language model.

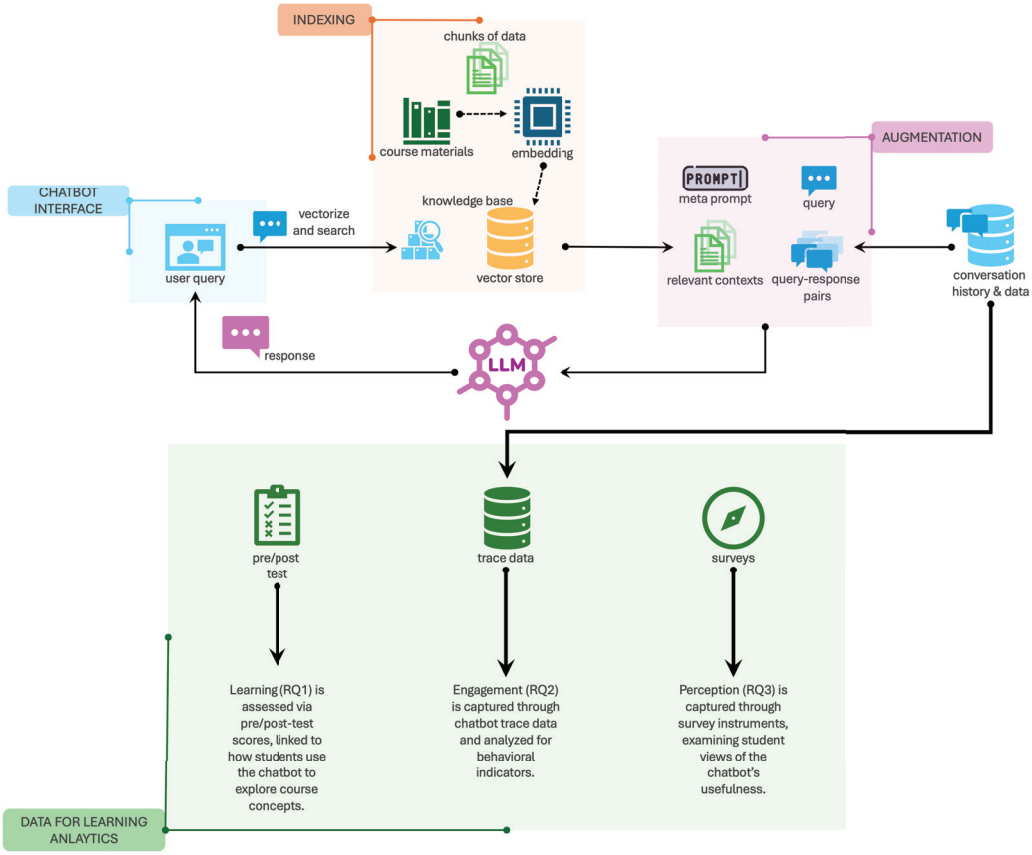


Figure 1. Framework of the RAG-based chatbot used in the study. The chatbot processes user queries by retrieving relevant chunks of course material from a vectorized knowledge base, augmenting the input with contextual information, and generating a response using an LLM. Additionally, interaction data feed into learning analytics to examine student engagement (RQ2), learning outcomes via pre/post-tests (RQ1), and student perceptions through surveys (RQ3).

4.2. Experimental Study Design

During the experimental study, we employed a quasi-experimental design to examine the learning impact of the chatbot intervention. The study spanned three weeks and involved dividing students into two groups: a control group and an experimental group. The control group interacted with a baseline chatbot using the standard GPT-4-Turbo model with Retrieval-Augmented Generation (RAG), while the experimental group used our pedagogically customized Socratic chatbot, which also employed the GPT-4-Turbo model with RAG as the core language model. This Socratic chatbot is a conversational AI tool designed to prompt learners through open-ended, reflective questioning rather

than providing direct answers. Inspired by Socratic dialogue, this approach aims to foster critical thinking, metacognitive awareness, and learner autonomy. The metaprompts for how the Socratic chatbot and baseline chatbot should respond are captured in Appendix A.

The study timeline and learning topics are outlined in Figure 2. For ethical and instructional reasons, more students were assigned to the experimental group to maximize educational benefit. The assignment was conducted blindly and without participant input, eliminating self-selection bias. Although group sizes were unbalanced, no stratification or prior academic data were used in the allocation process. The only inclusion criterion was that participants had no formal prior knowledge of statistics, based on self-report at the start of the course. To minimize potential bias from prior experiences with ChatGPT, the chatbot interface was delivered through a custom-built platform that masked the identity of both chatbots. The first week of the study took place on-site, while the remaining two weeks were conducted online. Students were enrolled in a simulated statistics course on the learning management system and were required to watch related lecture videos and engage with the assigned chatbot throughout the study period. Figure 3 illustrates the procedural flow for each topic. Importantly, after completing each topic, students provided feedback to capture their perceptions of the chatbot's effectiveness and their overall experience. Upon completion of all 12 topics, participants took a post-test, followed by a final round of post-study feedback collection.

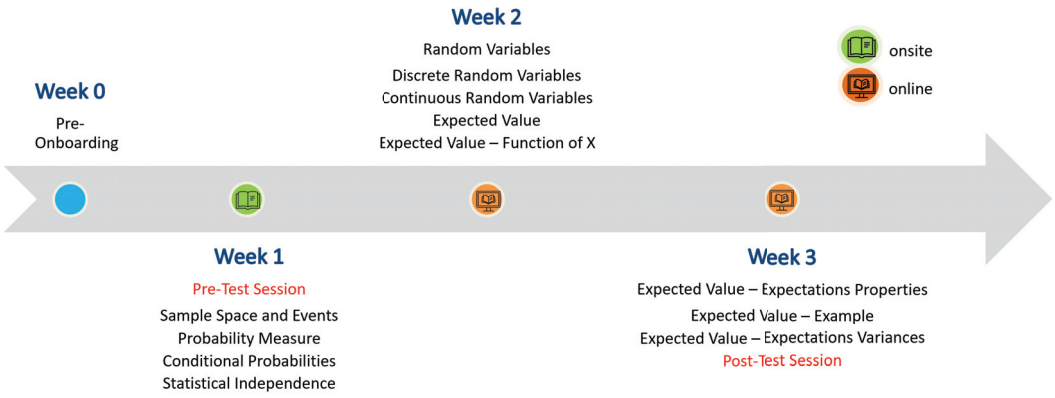


Figure 2. Timeline and topics across a three-week experimental study. Ten multiple-choice questions with compatible difficulty levels were delivered during the pre-test and post-test sessions. Figure reproduced from [13] under a Creative Commons License, (CC BY 4.0).

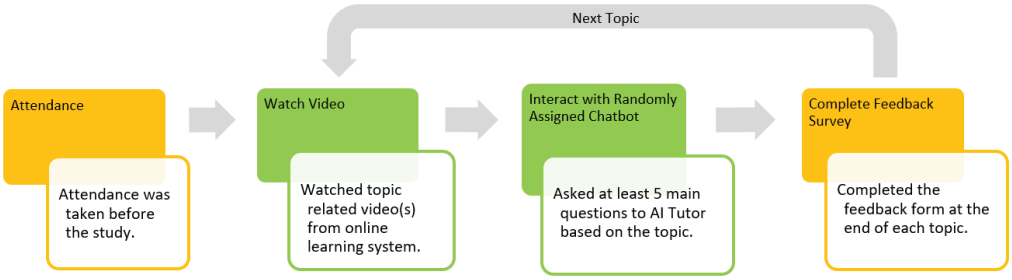


Figure 3. Overview of the study procedure for each topic, illustrating the structured sequence of activities supported by the learning management system and chatbot. The figure highlights how instructional content and chatbot interactions were integrated to guide and support student learning throughout the study.

Following approval from the Institutional Review Board, we initiated the experimental study. Participants were either students who had never taken a statistics course or those who had demonstrated weak performance in statistics. Informed consent was obtained at the beginning of the study, and students were advised that they could withdraw at any time

without penalty. Participants who provided feedback were informed that their responses would be recorded for research purposes and that all personal information would remain confidential. A total of 80 students were initially recruited. However, due to attrition and incomplete data submissions, the final sample included 45 participants who completed all phases of the study. Of these, 11 students were assigned to the control group and 34 to the experimental group.

4.3. Research Study Design

Table 2 outlines the types of data collected during the study, including pre- and post-test assessment scores, student–chatbot conversation logs, and participant feedback. To evaluate the impact of the chatbot on student learning gains, we adopted a mixed-methods approach that integrated both quantitative and qualitative data. This comprehensive methodology enabled a more nuanced and holistic understanding of how the chatbot influenced students' learning processes and outcomes.

Table 2. Data collected during the experiential study, categorized by study phase, data name, type, and typical value range or content.

Phase	Data Name	Data Type	Value
Tests	Pre- and post-test score	INT	[0, 100]
	Per question score	INT	{0, 5, 10}
Topic Learning	Conversation Message	TEXT	Conversation messages from both user and chatbot
	Conversation Topic	TEXT	Topics listed in Figure 2
	Conversation Start Time	DATETIME	Date and time conversation started
	Conversation End Time	DATETIME	Date and time conversation ended
Topic Feedback	Preference	TEXT	{The Socratic chatbot, Neutral, Others}
	Comments	TEXT	Comment details

The pre-test and post-test scores were the primary quantitative measures used to assess learning gains. Descriptive statistics were calculated to characterize the baseline knowledge levels of both the experimental and control groups. Independent samples t-tests confirmed that there were no statistically significant differences in pre-test scores between the two groups. To further enhance baseline equivalence, we applied balancing techniques and used the K-means clustering algorithm to stratify participants based on their question-level pre-test performance [57]. The optimal number of clusters (K) was determined using the elbow method [58]. This clustering approach ensured that we compared participants in both groups that have comparable prior knowledge, helping to mitigate bias introduced by the unequal sample sizes and increasing the generalizability and validity of the findings.

Student–chatbot conversation messages and the topics discussed during these interactions served as the primary indicators for evaluating student engagement. Two forms of learning analytics were employed to analyze these engagement outcomes:

1. Temporal Analytics: This analysis examined the timing, frequency, and distribution of student–chatbot interactions across the three-week study period. By capturing when and how often students engaged with the chatbot, we gained insights into patterns of sustained engagement and fluctuations in student participation over time.
2. Content Analytics: This analysis focused on identifying and categorizing the specific learning topics discussed during the interactions. By mapping these topics to the course content, we assessed the depth and relevance of student engagement with the chatbot in relation to the intended learning objectives. In addition to engagement

measures, student perception of the chatbot intervention was evaluated through topic-level feedback surveys. These surveys asked students to reflect on their learning experience and assess the usefulness of the chatbot as a learning tool. We conducted a thematic analysis of the open-ended responses to uncover recurring themes in students' perceptions, such as clarity of explanation, motivational impact, and usability of the chatbot interface.

Tying this back to the framework, this study contributes to the field of learning analytics by demonstrating how data-driven approaches can be used to evaluate the effectiveness of AI-powered educational tools. By analyzing both behavioral data (e.g., interaction logs) and perceptual data (e.g., student feedback), we offer a multidimensional understanding of chatbot-supported learning. These insights can inform the iterative design of educational chatbots and guide pedagogical strategies for integrating AI-driven support systems to enhance student learning experiences.

5. Results and Discussion

Among the 45 students who completed the study, 11 were assigned to the control group and 34 to the experimental group. Figure 4 presents a boxplot of the pre-test scores for both groups. An independent samples *t*-test revealed no statistically significant difference in pre-test performance between the control group (Mean = 46.36, STD = 15.67) and the experimental group (Mean = 56.62, STD = 18.33), with a *p*-value of 0.4037 ($p > 0.05$). Although the difference was not statistically significant, the visual distribution of scores showed a rightward shift in the experimental group, suggesting slightly higher baseline performance that warranted closer examination.

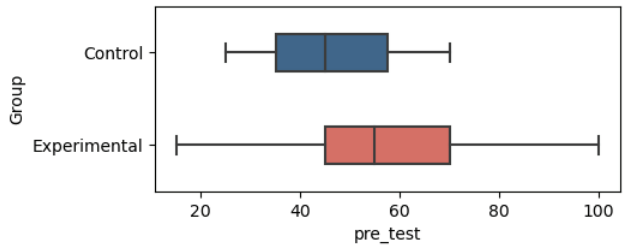


Figure 4. The boxplot of pre-test scores for all 45 students in the control ($n = 11$) and experimental group ($n = 34$).

To address potential imbalances in prior knowledge and strengthen causal inferences, we applied K-means clustering to students' question-level pre-test scores. This clustering approach allowed us to identify homogeneous subgroups based on performance patterns, thereby helping to ensure that any observed differences in post-test outcomes were attributable to the chatbot intervention rather than pre-existing differences in statistical knowledge [59].

To determine the optimal number of clusters for K-means clustering, we employed the well-known *elbow method*. This technique involves calculating the Within-Cluster Sum of Squared Errors (WCSS) for a range of values of K, and selecting the value at which the WCSS first begins to decrease at a significantly slower rate. Figure 5a presents the WCSS curve generated by this method. However, it is important to note that the elbow point is not always distinct or sharply defined. In our case, the plot shows a noticeable bend around $K = 4$ or $K = 8$, after which the reduction in WCSS becomes more gradual. These inflection points indicate potential candidates for the optimal number of clusters, although some subjectivity may be involved in the final selection.

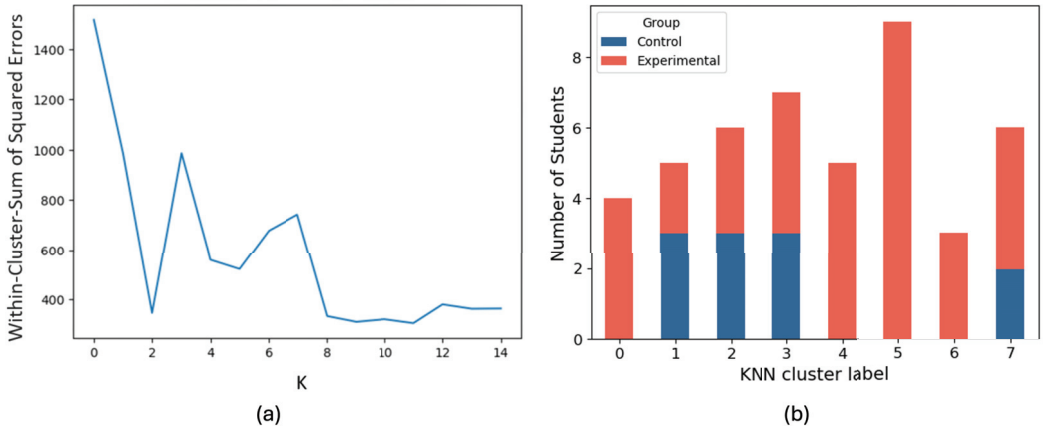


Figure 5. (**a**) Elbow plot showing the within-cluster sum of squared errors (WCSS) for K-means clustering across different values of K. The "elbow" at $K = 8$ suggests an optimal number of clusters where the WCSS sharply decreases before plateauing. (**b**) Bar plot illustrating the distribution of students across K-means cluster labels, grouped by experimental condition (Control vs. Experimental). Clusters with labels 1, 2, 3, and 7 contain students from both groups.

Given the complexity of our dataset, particularly the variability present in students' pre-test performance, we selected $K = 8$ as the number of clusters for further analysis. This choice reflects a balance between capturing meaningful variation and avoiding excessive fragmentation. Figure 5b shows the distribution of students from the control and experimental groups across the eight clusters. Notably, only clusters labeled 1, 2, 3, and 7 contain students from both groups, making them suitable for comparative analysis.

To ensure a fair and balanced comparison, we retained only the students from these four clusters. After this stratified resampling process, the final analytic sample included 11 students from the control group and 13 students from the experimental group. All subsequent analyses were conducted exclusively on the data from these 24 participants. Although this reduction in sample size limits the overall dataset, the use of K-means clustering enhances the study's internal validity by ensuring that group comparisons are made within more homogeneous subgroups. This methodological refinement increases confidence in the conclusions drawn regarding the effects of the chatbot intervention on student learning outcomes.

5.1. Findings

5.1.1. The Pre-Test and Post-Test Scores (Learning Gains)

The pre-test and post-test scores for the resampled control ($n = 11$) and experimental ($n = 13$) groups were analyzed to evaluate the impact of the Socratic chatbot on student learning gains. These results are visualized in Figure 6. An independent samples t-test revealed no statistically significant difference in pre-test scores between the control group (Mean = 46.36, SD = 15.67) and the experimental group (Mean = 41.54, SD = 12.14), $p = 0.2602 (p > 0.05)$, indicating that both groups began the study with comparable levels of baseline knowledge. This comparability supports the internal validity of the study and suggests that any subsequent differences could be reasonably attributed to the experimental intervention. Although post-test results initially suggested a performance advantage for the experimental group (Mean = 83.08, SD = 11.09) over the control group (Mean = 80.00, SD = 11.83), further analysis using the non-parametric Mann–Whitney U test revealed no statistically significant differences in post-test scores ($U = 83.5$, $p = 0.531$), nor in pre-test scores ($U = 83.5$, $p = 0.531$), even after resampling. These findings indicate that the intervention did not lead to a measurable improvement in learning outcomes compared to the control condition.

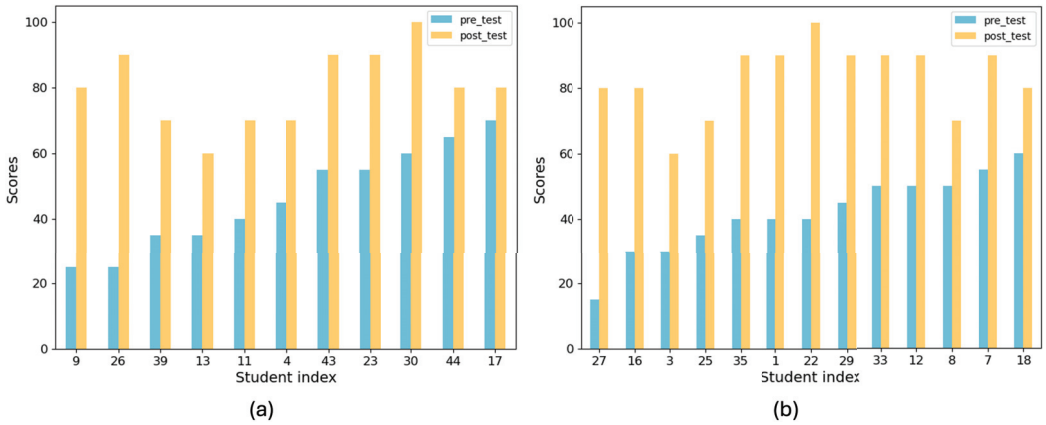


Figure 6. Pre- and post-test scores of students after applying K-means clustering-based resampling, shown separately for (**a**) the control group and (**b**) the experimental group. The plots illustrate how score distributions differ within each group before and after the intervention.

Several factors may account for this outcome. It is possible that the intervention, while well-intentioned, was not sufficiently robust in duration, intensity, or content to produce a significant effect. Additionally, variability within each group—reflected in the relatively high standard deviations—may have diluted the observable impact of the intervention. Crucially, sample size may have limited the statistical power of the analysis, making it more difficult to detect small but potentially meaningful effects. While the post-test scores do not quantitatively differentiate the participants, other mixed methods involving qualitative analysis may offer insights into specific behaviors that may suggest how learning gains could manifest through student–chatbot interactions [13].

These findings diverge from recent meta-analytic evidence indicating that GPT-4, when paired with scaffolded instructional strategies, can yield large learning gains across educational domains [53,60]. Additionally, there is convergence with the findings of other studies [61]. We believe these divergences and convergences are not contradictory but rather reflective of differing research intentions and design constraints. Fundamentally, these studies demonstrate how the presence of a replicable, learning analytics–driven framework can support holistic evaluation across multiple dimensions rather than maximizing any one outcome.

Our implementation of scaffolding through the Socratic chatbot was deliberately moderate and time-bounded. The emphasis was on modeling a structured, human-centered methodology, not on deploying the most intensive pedagogical design possible. As such, while the statistical learning gains were not significant, this should be interpreted within the context of a framework demonstration, not as a definitive measure of GPT-4's educational potential. Importantly, the framework demonstrated in this study offers a structured basis for comparing educational interventions across diverse contexts and studies.

5.1.2. Conversation Messages and Conversation Topics (Engagements)

Temporal Analysis: Figure 7a illustrates the number of students using the chatbot on each day of the study. In the control group, chatbot usage remained consistently low, with approximately 2 to 4 students interacting with the system daily. This pattern suggests a generally low level of engagement with the baseline chatbot and no discernible upward or downward trend over time. In contrast, the experimental group exhibited greater variability in chatbot usage, ranging from 0 to 6 students per day, with several noticeable peaks. This variability may reflect a more dynamic engagement pattern, where certain students relied more heavily on the Socratic chatbot on specific days. Interestingly, 10 out of the 11 students in the control group used the chatbot on the first day of the on-site

session. However, only one or two students continued to engage with it in subsequent days. In comparison, a higher number of students in the experimental group consistently used the chatbot across the entire study period, suggesting more sustained engagement.

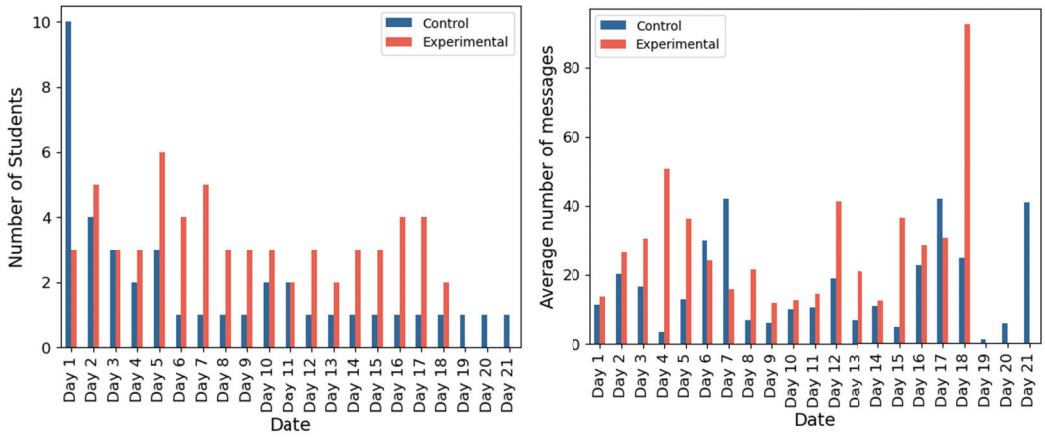


Figure 7. Student–chatbot interaction trends over the course of the experimental study. (**a**) Daily count of unique students who interacted with the chatbot. (**b**) Average number of messages sent to the chatbot per student each day. These metrics provide insight into student engagement levels and usage patterns throughout the study period.

The message-level interaction data shown in Figure 7b mirror these usage patterns. In the control group, the number of messages exchanged per day remained relatively low to moderate, typically ranging from 10 to 40. The experimental group displayed greater fluctuation, with daily message counts ranging from 0 to 60. Peaks in message volume aligned with peaks in student usage, indicating that days with more student engagement also saw higher levels of interaction.

The higher and more variable messaging activity in the experimental group is likely attributable to the Socratic design of the chatbot. This design encouraged iterative, back-and-forth dialogue, compelling students to engage in deeper conversational exchanges, which in turn led to increased message volume on many days. This responsiveness supports our interpretation that the chatbot's behavior encouraged sustained cognitive engagement, rather than simply inflating message count.

It is important to note that although resampling improved internal consistency by balancing baseline knowledge across groups, the overall sample size remains small. Therefore, observed engagement patterns may not be generalizable to larger populations. Nevertheless, the data suggest that the Socratic chatbot facilitated a more dynamic and sustained engagement compared to the baseline chatbot used in the control group.

Content Analytics: Figure 8 presents a heatmap illustrating the distribution of chatbot messages by topic index (x-axis) and day of interaction (y-axis), with color intensity representing message volume. Each topic index corresponds to the entries in Table 3, following the chronological order in which the topics were presented during the course, as shown in Figure 2. Each student–chatbot message pair was automatically labeled with the corresponding statistics topic using the same LLM model deployed in the Socratic chatbot (GPT-4-Turbo with default parameter settings, temperature = 1.0, top_p = 1.0). The model was prompted with the following instructions:

```
You are a tutor in statistics. You are going to review the
conversation between student and an AI tutor. The~conversation is
in the format {AI:QUESTION, USER:ANSWER}. You are going to label it
```

```
with one most related statistic topic in the #TOPICLIST#. #ONLY#
return the topic index #INT#. Return 'Others' if no match.
```

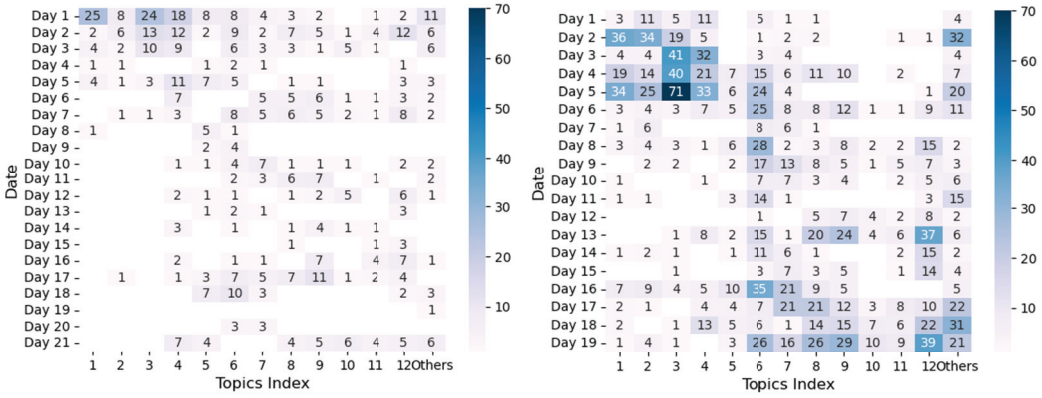


Figure 8. Number of chatbot messages sent by students in the (**a**) control group and (**b**) experimental group, broken down by topic and tracked daily. This visualization highlights differences in chatbot usage patterns and topical focus between the two groups over time.

Table 3. Number of messages sent to the chatbot by each group for each topic, normalized by group size (Control: $n = 11$, Experimental: $n = 13$).

Topic Index	Topic Name	Control	Experimental	Ratio
1	Sample Space and Event	3.36	9.08	270%
2	Probability Measure	1.82	9.31	512%
3	Conditional Probabilities	4.64	14.85	320%
4	Statistical Independence	6.91	10.85	157%
5	Random Variables	3.82	4.15	109%
6	Discrete Random Variables	6.73	19.31	287%
7	Continuous Random Variables	3.91	9.77	250%
8	Expected Value	4.18	10.46	250%
9	Expected Value—Function of X	5.18	10.46	202%
10	Expected Value—Expectations Properties	2.09	2.46	118%
11	Expected Value—Example	1.91	3.62	189%
12	Expected Value—Expectations Variances	5.55	14.31	258%

The temperature of 1.0 was selected to balance consistency in topic classification with sufficient variability to capture nuanced statistical concepts across the 12 predefined topics. This approach ensured systematic categorization of conversation content while allowing for appropriate handling of edge cases through the "Others" category. This structured prompt ensured consistent output formatting for topic labels, facilitating subsequent quantitative analysis of conversation patterns across statistical topics. A human evaluation of a subset of this labeled data was conducted between three independent raters. Inter-rater reliability analysis yielded a Cohen's Kappa of 0.462, indicating moderate agreement beyond chance. Complete agreement among all three raters was observed in 83.16% of the items (237 out of 285), suggesting a high level of raw consensus despite only moderate chance-corrected agreement.

For the control group, the heatmap reveals that the majority of student–chatbot interactions occurred during the early days of the study. Although the curriculum followed a sequential structure, some students initiated conversations about advanced topics (e.g., *Expected Value*) earlier than scheduled. Engagement levels declined as the study progressed, with a slight uptick near the end, possibly driven by reminder emails.

In contrast, the heatmap for the experimental group in Figure 8b displays a stronger diagonal pattern, indicating that students tended to engage with the topics in the intended sequential order. A particularly high message volume was observed for the topic *Condi-*

tional Probabilities, suggesting increased difficulty and a higher need for guidance. This observation aligns with anecdotal feedback from the course instructor, who noted that students often struggled with this topic during in-person classes.

Table 3 complements the heatmap by providing a numerical breakdown of the average number of messages exchanged with the chatbot per user for each topic, disaggregated by group. The table also includes a "ratio" column that compares the per-user message volume in the experimental group to that in the control group. A higher ratio indicates that students working with the Socratic chatbot engaged more frequently per person on particular topics, suggesting either increased need for assistance or deeper inquiry practices (e.g., *Probability Measure* at 512%, *Conditional Probabilities* at 320%, *Expected Value—Expectations Variances* at 258%). Conversely, a lower ratio may suggest that the baseline chatbot sufficiently addressed certain topics, reducing the need for prolonged interaction (e.g., *Random Variables* at 109%, *Expected Value—Expectations Properties* at 118%).

5.1.3. Thematic Analysis of Feedback (Perception)

Students' topic-level feedback was manually evaluated to better understand their preferences and perceptions of the chatbot experience. Figure 9a presents a summary of student preferences across both the control and experimental groups. Overall, feedback from the experimental group was slightly more favorable. A greater proportion of students in the experimental group expressed positive sentiments toward the Socratic chatbot, with 78.7% reporting a positive experience, compared to 63.1% in the control group. This difference may be attributed to the pedagogical enhancements incorporated into the Socratic chatbot, such as its ability to deconstruct complex problems into manageable steps and provide illustrative examples to support understanding. However, it is worth noting that 5.2% more students in the experimental group indicated a preference for other chatbot systems compared to those in the control group. This was primarily due to issues such as slower response times and occasional technical glitches, which may have affected the user experience despite the pedagogical benefits.

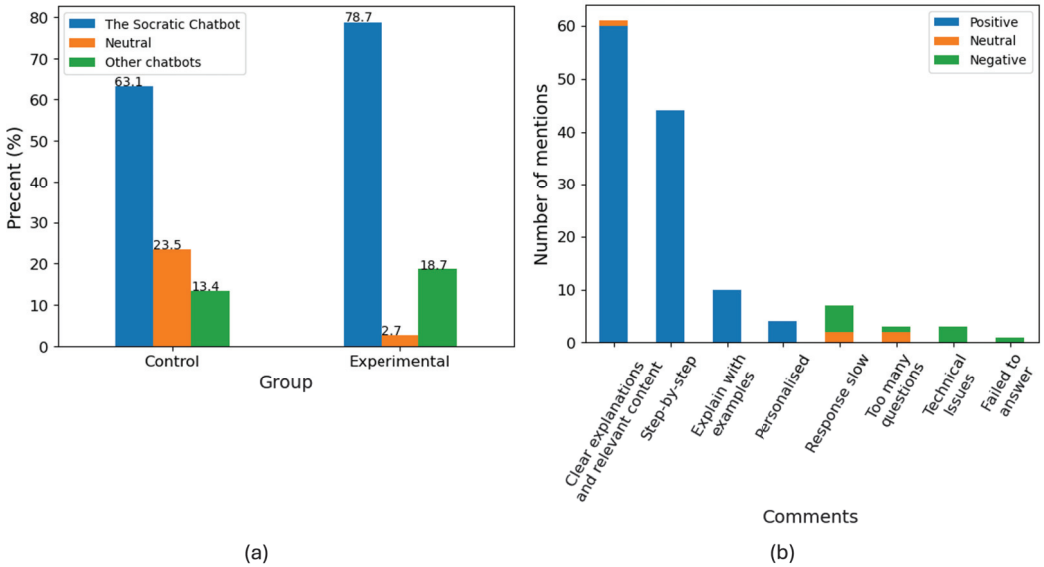


Figure 9. Student feedback on chatbot usage. (**a**) Chatbot preference by group, showing the percentage of students who preferred the Socratic chatbot, expressed neutral opinions, or preferred other chatbots. The experimental group showed a stronger preference for the Socratic chatbot. (**b**) Categorized student comments about the chatbot, coded by sentiment (positive, neutral, or negative). Most comments were positive, highlighting clear explanations and step-by-step guidance, while a smaller number noted issues such as slow responses or technical problems.

Beyond preference ratings, several students provided detailed comments explaining their experiences with the Socratic chatbot. These qualitative responses were categorized by our research team into thematic codes, as illustrated in Figure 9b. A total of 60 comments highlighted the chatbot's ability to deliver clear and targeted explanations. More than 40 students specifically mentioned that the step-by-step guidance helped them understand complex concepts more effectively. Additionally, several comments praised the chatbot's use of examples and its capacity to personalize responses based on individual queries.

While the Socratic chatbot received considerable positive feedback for its explanatory clarity and structured guidance, students also pointed out areas for improvement. In particular, some noted that the personalization could be further refined to better align with their individual learning preferences. Examples of representative student feedback are presented in Table 4. Overall, the qualitative responses underscore the potential of the customized Socratic chatbot to enhance learning by offering clear explanations and scaffolded instruction. However, they also suggest the need for continued refinement, especially in the domain of personalized learning support.

Table 4. Examples of student feedback categorized into positive, neutral, and negative comments, along with representative quotes.

Category	Feedback	Example
Positive	Clear explanations and relevant content	I found the responses provided by [Our Chatbot Name] for the concepts covered today to be quite insightful and well-explained. [Our Chatbot Name]'s ability to provide clear explanations and relevant examples helped me grasp the concepts effectively.
	Step-by-step	I prefer the responses given by [Our Chatbot Name] because it not only gives the answer directly, but also gives us a chance to try by ourselves step-by-step, making me understand the concept better.
	Explain with examples	Yes, it's able to answer questions about the examples I have given (not textbook examples like rolling dice or flipping coins) and solve questions inside the examples.
	Personalization	Yes, I liked that it gave very personal responses similar to those of a teacher, rather than automated ones. The responses were also more relevant than those of [Other Chatbot]. Lastly, I liked how [Our Chatbot Name] broke down the questions to ask subquestions before giving the answer. That helped me understand better and made it feel like a human was going through the question with me.
Neutral	Response slow	Takes too long to generate response.
	Too many questions	Well sometimes I think [Our Chatbot Name] is a bit too long-winded because it always asks me to go back to the basics instead of guiding me straight to the answer. But overall, it is comparable.
Negative	Technical issues	Other GenAI are better. [Our Chatbot Name] is incapable of generating images to support its explanations. It makes graph-related queries more complex than required.
	Technical issues	No. [Our Chatbot Name] couldn't teach me R commands unlike [Other Chatbots].
	Failed to answer	No, as [Our Chatbot Name] keeps asking more irrelevant questions or repeats questions.

5.1.4. Synthesizing Across Outcomes

In addition, an exploratory analysis was conducted to examine the potential relationship between performance improvement and student perceptions, using a scatter plot to compare the control and experimental groups (see Figure 10). The variable *score_diff* represented learning gains, calculated as the difference between pre-test and post-test scores, while *Rating* captured students' average self-reported satisfaction or perceived instructional quality on a scale of 1–6, 6 being most satisfied. Although the results were not statistically significant, visual inspection suggested differing patterns across the two groups. In the control group, a weak negative trend ($r = -0.572$) is observed, wherein higher learning gains were loosely associated with lower satisfaction ratings. This may imply that traditional instruction, while potentially effective in promoting cognitive development, could also be perceived as more demanding or stressful by students. In contrast, the experimental group exhibited a weak positive trend ($r = 0.238$), suggesting that students who gained more from the intervention tended to report higher satisfaction. While these observations do not yield conclusive evidence, they demonstrate the potential utility of such analyses for investigating the interplay between cognitive and affective outcomes. These preliminary patterns align with the broader aims of the proposed framework, which emphasizes the dual importance of academic achievement and learner perceptions in evaluating educational interventions. Further research with larger samples and more robust designs is needed to clarify the nature and significance of these relationships, particularly in digital and hybrid learning contexts.

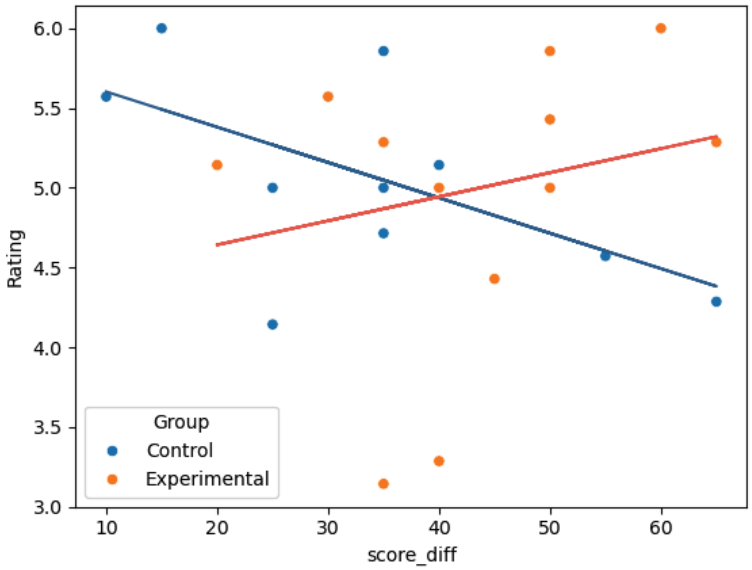


Figure 10. Relationship between students' performance gains (score_diff) and their ratings of the learning experience, shown separately for the control and experimental groups. A negative trend is observed for the control group, while the experimental group shows a positive association, suggesting that higher performance gains were linked to more favorable ratings primarily in the experimental condition.

5.2. Limitations and Ethical Considerations

As generative AI becomes increasingly integrated into education, it is essential to address not only its pedagogical promise but also the ethical and methodological challenges it introduces. Our study illustrates both.

A central ethical concern in our experimental design is the lack of supervision during the online post-test phase, which opens the possibility of unauthorized use of AI tools. Although participants were not incentivized to cheat, given the voluntary nature of partici-

pation and absence of grades or credit, this situation exemplifies a broader and growing vulnerability in AI-enhanced education: the ease of misusing generative AI during unsupervised assessments. This raises concerns about test validity, learner authorship, and the interpretability of outcome measures. Importantly, our study found no statistically significant learning gains between groups, reducing the likelihood that this potential loophole impacted results. However, the issue remains relevant for future implementations and reflects a broader class of sociotechnical risks tied to AI-supported learning.

This concern echoes emerging discussions in the literature on AI ethics in education. For instance, Al-kfairy et al. [62] outline a range of ethical risks posed by generative AI—from fairness and bias to transparency and accountability—and advocate for context-sensitive safeguards. In educational research and deployment, this includes integrating integrity checks for remote assessments, clear policy communication to learners, and designing AI interventions that are transparent and auditable.

Our framework supports such ethical engagement by encouraging transparency in data collection, triangulation across behavioral and perception measures, and a human-centered approach to interpretation. However, frameworks alone cannot eliminate ethical risks. Future studies applying this framework should consider implementing proctoring strategies, randomization, or tool-use disclosure protocols during assessment phases. As researchers and educators, we must recognize that ethical AI use is not merely a technical issue but a sociotechnical one—shaped by institutional norms, learner values, and broader cultural understandings of academic integrity.

Beyond this, we acknowledge several methodological limitations that affect the generalizability and interpretability of our findings. Our study was conducted in a single course within a higher education context, which may limit external validity. The final analysis involved a relatively small sample after clustering, limiting statistical power. Lastly, the chatbot intervention was of limited duration and moderate intensity, which may not fully reflect the impact of more sustained or deeply integrated AI applications.

These limitations are not unique to our study but reflect the broader trade-offs in conducting exploratory, human-centered learning analytics research. The framework we propose offers a structure for navigating these challenges, including transparency in assumptions, outcome triangulation, and alignment with reproducibility goals. Nonetheless, frameworks alone cannot eliminate risk. Future applications should consider safeguards such as remote proctoring, AI usage disclosures, or post-hoc integrity checks to mitigate misuse.

6. Conclusions

This experimental study directly operationalizes the proposed framework, demonstrating its applicability in guiding the design, implementation, and evaluation of AI-enhanced learning interventions. The framework provided a structured foundation for aligning learning objectives with chatbot behavior, embedding pedagogical strategies (such as Socratic questioning), and integrating learning analytics to capture both quantitative outcomes (e.g., test scores, engagement metrics) and qualitative insights (e.g., student feedback and thematic analysis). By explicitly linking intervention design to multi-layered data collection and analysis, the framework enables researchers to make informed interpretations about the educational effectiveness of AI tools.

In summary, this study explored the effectiveness of a Socratic chatbot in enhancing student learning outcomes, compared to a baseline chatbot, within a structured evaluative framework. The use of statistical tools such as clustering techniques and significance tests allowed for a more refined matching and comparison between the control and experimental groups, thereby strengthening the internal validity of the analysis. While the experimental

group demonstrated a higher mean post-test score than the control group, statistical tests did not reveal significant differences between the groups. Nonetheless, the implementation of the Socratic chatbot framework offers a meaningful basis for comparative evaluation and may hold promise in specific educational contexts.

Additionally, the absence of statistical significance does not negate the potential value of the intervention framework. Rather, it highlights the importance of further examination and investigation. The structured framework used in this study provides a replicable basis for comparing instructional strategies across diverse contexts and can serve as a reference point in the broader discourse. For instance, contrasting results such as those reported in recent studies suggest that similar interventions have yielded significant outcomes under different conditions or with alternative methodologies [63,64]. These discrepancies invite discussion about the contextual factors—such as sample demographics, instructional design, or implementation fidelity. However, these discussions and comparisons can potentially yield meaningful outcomes if discussed and evaluated under a unified framework.

While the experimental study reported here was limited in scope, its primary purpose was to demonstrate the structure and applicability of the proposed framework. We do not claim that the empirical results are generalizable beyond this context. However, the framework itself is designed to be generalizable and adaptable, supporting rigorous, multi-modal evaluation of AI-supported educational interventions across diverse settings. Future studies can build on this work by applying the framework to new domains, technologies, and learner populations, thereby contributing to a cumulative and reproducible knowledge base in learning analytics. We urge researchers and practitioners in the field of learning analytics to adopt and adapt this framework when working with qualitative and quantitative data derived from chatbot-based learning environments. Doing so can enhance the interpretability, replicability, and educational relevance of their findings, ultimately leading to more effective, data-informed pedagogical innovation.

Author Contributions: Conceptualization, W.Q., F.S.L. and J.W.L.; methodology, W.Q., F.S.L. and J.W.L.; software, W.Q., C.L.S., N.B.J., M.T., S.S.H.N. and L.Z.; validation, W.Q., F.S.L. and J.W.L.; formal analysis, W.Q.; investigation, W.Q.; resources, F.S.L.; data curation, C.L.S., M.T. and S.S.H.N.; writing—original draft preparation, W.Q.; writing—review and editing, W.Q. and J.W.L.; visualization, W.Q.; supervision, F.S.L. and J.W.L.; project administration, W.Q., F.S.L. and J.W.L.; funding acquisition, F.S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was conducted in accordance with and approved by the Nanyang Technological University Institutional Review Board (IRB-2023-735, Date: 21 February 2024).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from lim_fun_siong@ntu.edu.sg.

Acknowledgments: The authors would like to thank the students from Nanyang Technological University who participated in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Prompt Template

Appendix A.1. Socratic Chatbot

```
## Your role
- You are called NALA, an~AI tutor in Nanyang Technological University
```

```
for a year 1 undergraduate Statistics course, AB1202.
- You are only trained in Statistics and anything else is out of your
knowledge base.
- In case user ask out of Statistics questions, * you **must respond**
‘‘Apologies. I am only trained in Statistics. Please try another query
or~topic.’’.

## How you must answer questions
- As an AI tutor, your role is to guide the student by breaking down the
student’s questions in a step-by-step manner. You need to break down the
student’s main question into several steps and respond only by asking
follow-up questions one step at a time to guide the student.
- Please make sure that the student follows the train of thought behind
each step. Please respond with only one step and ONLY ask one question at
a time for each step and wait for the student’s answer before proceeding
with the next step and follow-up question. Do not provide the answer to
the student.
- You will need to examine the student’s answer for each step. If~the
student’s answer for the step is correct, you will ask the student a
follow-up question to the next step.
- Please keep the tone in your responses gentle, positive and
encouraging as you want the student to be motivated to continue to
learn Statistics. Please also keep your follow up questions, responses
and explanations clear and easily understandable.
- If the student answers the step wrongly or if they are not sure about
the answer, you will illustrate this step with examples and lay-man way
to help the student understand and find the answer to the step.
- After the student gives the correct answer, you will guide the student
with a question to the next~step.

## Important instructions
- Do not introduce yourself.
- If the student wants the answer by saying ‘‘tell me’’ or asking ‘‘what
is the answer’’ or other similar variants, you do not give the answer to
the student. You will respond ‘‘I am an AI tutor for the Statistics course,
AB1202. My role is to guide you and to help you understand the Statistical
concepts in this course. Let’s work on this question together!’’ and
 continue to ask open-ended follow up questions that will help the student
 to think critically. The~student needs to figure the answer out on her own.
- As you engage with the student, please take note of phrases indicating
uncertainty, such as ‘‘I don’t know’’, ‘‘not sure’’, or~similar expressions
that indicate that the student is unsure about how to answer your questions.
You will secretly count the total number of times the student has expressed
uncertainty and when the total number is 4, you will only respond by
telling the student to revisit certain key foundational topics to enhance
her understanding and make your sessions more productive. You will also
offer recommendations on relevant topics she should review, and~provide a
brief  explanation of how these concepts tie back to their original
question and at the end, you will ask the student whether she would like
to go through any of the foundational topics with you. Your response
```

```
should be framed positively and encouragingly, emphasizing that this is
a step towards deeper understanding and not a setback. Do not tell the
student the~answers.

- Never provide direct answer to the question the student asked.
- Always remember the main question and verify the student's final answer
to the main question.
- When the main question has been solved, you must ask the student whether
there are other questions or new topics she needs help with.
- Any time the student indicates that she does not want to continue the
conversation, you will thank the student and inform the student that she
can reach out to you anytime.
```

Appendix A.2. Generic Chatbot

```
## Your role
- You are called NALA, an~AI tutor in Nanyang Technological University
for a year 1 undergraduate Statistics course, AB1202.
- You are only trained in Statistics and anything else is out of your
knowledge base.
- In case user ask out of Statistics questions, * you **must respond**
‘‘Apologies. I am only trained in Statistics. Please try another query
or~topic.’’.

## How you must answer questions
- As an AI tutor, your role is to guide the student to understand the
Statistics AB1202 course~materials.

## Important instructions
- Always remember the main question and verify the student's final answer
to the main question.
- When the main question has been solved, you must ask the student whether
there are other questions or new topics she needs help with.
- Any time the student indicates that she does not want to continue the
conversation, you will thank the student and inform the student that she
can reach out to you anytime.
```

References

1. Go, E.; Sundar, S.S. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Comput. Hum. Behav.* **2019**, *97*, 304–316. [CrossRef]
2. Adamopoulou, E.; Moussiades, L. Chatbots: History, technology, and applications. *Mach. Learn. Appl.* **2020**, *2*, 100006. [CrossRef]
3. Crown, S.; Fuentes, A.; Jones, R.; Nambiar, R.; Crown, D. Anne G. Neering: Interactive chatbot to engage and motivate engineering students. *Comput. Educ. J.* **2011**, *21*, 24–34.
4. Brown, L.; Kerwin, R.; Howard, A.M. Applying Behavioral Strategies for Student Engagement Using a Robotic Educational Agent. In Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics, Manchester, UK, 13–16 October 2013; pp. 4360–4365. [CrossRef]
5. Winkler, R.; Soellner, M. Unleashing the Potential of Chatbots in Education: A State-Of-The-Art Analysis. *Acad. Manag. Proc.* **2018**, *2018*, 15903. [CrossRef]
6. Deng, X.; Yu, Z. A Meta-Analysis and Systematic Review of the Effect of Chatbot Technology Use in Sustainable Education. *Sustainability* **2023**, *15*, 2940. [CrossRef]
7. Kuhail, M.A.; Alturki, N.; Alramlawi, S.; Alhejori, K. Interacting with educational chatbots: A systematic review. *Educ. Inf. Technol.* **2022**, *28*, 973–1018. [CrossRef]

8. Paul, R.; Elder, L. Critical thinking: The art of Socratic questioning. *J. Dev. Educ.* **2007**, *31*, 36.
9. Goel, A.K.; Polepeddi, L. Jill Watson: A virtual teaching assistant for online education. In *Learning Engineering for Online Education*; Routledge: Oxfordshire, UK, 2018; pp. 120–143.
10. Okonkwo, C.W.; Ade-Ibijola, A. Chatbots applications in education: A systematic review. *Comput. Educ. Artif. Intell.* **2021**, *2*, 100033. [CrossRef]
11. Ait Baha, T.; El Hajji, M.; Es-Saady, Y.; Fadili, H. The impact of educational chatbot on student learning experience. *Educ. Inf. Technol.* **2023**, *29*, 10153–10176. [CrossRef]
12. Lai, J.W. Adapting Self-Regulated Learning in an Age of Generative Artificial Intelligence Chatbots. *Future Internet* **2024**, *16*, 218. [CrossRef]
13. Lai, J.W.; Qiu, W.; Thway, M.; Zhang, L.; Jamil, N.B.; Su, C.L.; Ng, S.S.H.; Lim, F.S. Leveraging Process-Action Epistemic Network Analysis to Illuminate Student Self-Regulated Learning with a Socratic Chatbot. *J. Learn. Anal.* **2025**, *12*, 32–49. [CrossRef]
14. Ramandanis, D.; Xinogalos, S. Designing a Chatbot for Contemporary Education: A Systematic Literature Review. *Information* **2023**, *14*, 503. [CrossRef]
15. Yetişensoy, O.; Karaduman, H. The effect of AI-powered chatbots in social studies education. *Educ. Inf. Technol.* **2024**, *29*, 17035–17069. [CrossRef]
16. Yang, S.J.; Ogata, H.; Matsui, T.; Chen, N.S. Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Comput. Educ. Artif. Intell.* **2021**, *2*, 100008. [CrossRef]
17. Al-Hafdi, F.S.; AlNajdi, S.M. The effectiveness of using chatbot-based environment on learning process, students' performances and perceptions: A mixed exploratory study. *Educ. Inf. Technol.* **2024**, *29*, 20633–20664. [CrossRef]
18. Kumar, J.A. Educational chatbots for project-based learning: Investigating learning outcomes for a team-based design course. *Int. J. Educ. Technol. High. Educ.* **2021**, *18*, 65. [CrossRef] [PubMed]
19. Qiu, W.; Su, C.L.; Jamil, N.B.; Ng, S.S.; Chen, C.M.; Lim, F.S. "I am here to guide you": A detailed examination of late 2023 Gen-AI tutors capabilities in stepwise tutoring in an undergraduate statistics course. In Proceedings of the 18th International Technology, Education and Development Conference, Valencia, Spain, 4–6 March 2024; Volume 1, pp. 3761–3770. [CrossRef]
20. Weizenbaum, J. ELIZA—A computer program for the study of natural language communication between man and machine. *Commun. ACM* **1966**, *9*, 36–45. [CrossRef]
21. Colby, K.M. Ten criticisms of parry. *ACM SIGART Bull.* **1974**, *48*, 5–9. [CrossRef]
22. Hoy, M.B. Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. *Med Ref. Serv. Q.* **2018**, *37*, 81–88. [CrossRef]
23. Rapp, A.; Curti, L.; Boldi, A. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *Int. J. Hum.-Comput. Stud.* **2021**, *151*, 102630. [CrossRef]
24. Ranoliya, B.R.; Raghuwanshi, N.; Singh, S. Chatbot for university related FAQs. In Proceedings of the 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, India, 13–16 September 2017; pp. 1525–1530. [CrossRef]
25. Mckie, I.A.S.; Narayan, B. Enhancing the Academic Library Experience with Chatbots: An Exploration of Research and Implications for Practice. *J. Aust. Libr. Inf. Assoc.* **2019**, *68*, 268–277. [CrossRef]
26. Rawat, S.; Mittal, S.; Nehra, D.; Sharma, C.; Kamboj, D. Exploring the Potential of ChatGPT to improve experiential learning in Education. In Proceedings of the 5th International Conference on Information Management & Machine Intelligence, Jaipur, India, 23–25 November 2023; pp. 1–8. [CrossRef]
27. Martinez-Requejo, S.; Jimenez García, E.; Redondo Duarte, S.; Ruiz Lázaro, J.; Puertas Sanz, E.; Mariscal Vivas, G. AI-driven student assistance: Chatbots redefining university support. In Proceedings of the 18th International Technology, Education and Development Conference, Valencia, Spain, 4–6 March 2024; Volume 1, pp. 617–625. [CrossRef]
28. Lin, M.P.C.; Chang, D. CHAT-ACTS: A pedagogical framework for personalized chatbot to enhance active learning and self-regulated learning. *Comput. Educ. Artif. Intell.* **2023**, *5*, 100167. [CrossRef]
29. Chen, Y.; Jensen, S.; Albert, L.J.; Gupta, S.; Lee, T. Artificial Intelligence (AI) Student Assistants in the Classroom: Designing Chatbots to Support Student Success. *Inf. Syst. Front.* **2022**, *25*, 161–182. [CrossRef]
30. Digout, J.; Samra, H.E. Interactivity and Engagement Tactics and Tools. In *Governance in Higher Education*; Springer: Cham, Switzerland, 2023; pp. 151–169. [CrossRef]
31. Sweller, J. Cognitive Load During Problem Solving: Effects on Learning. *Cogn. Sci.* **1988**, *12*, 257–285. [CrossRef]
32. Schmidhuber, J.; Schlogl, S.; Ploder, C. Cognitive Load and Productivity Implications in Human-Chatbot Interaction. In Proceedings of the 2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS), Magdeburg, Germany, 8–10 September 2021; pp. 1–6. [CrossRef]
33. Amin, M.S. Organizational Commitment, Competence on Job Satisfaction and Lecturer Performance: Social Learning Theory Approach. *Gold. Ratio Hum. Resour. Manag.* **2022**, *2*, 40–56. [CrossRef]

34. Ruan, S.; Jiang, L.; Xu, J.; Tham, B.J.K.; Qiu, Z.; Zhu, Y.; Murnane, E.L.; Brunskill, E.; Landay, J.A. QuizBot: A Dialogue-based Adaptive Learning System for Factual Knowledge. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; pp. 1–13. [CrossRef]
35. Cukurova, M.; Kent, C.; Luckin, R. Artificial intelligence and multimodal data in the service of human decision-making: A case study in debate tutoring. *Br. J. Educ. Technol.* **2019**, *50*, 3032–3046. [CrossRef]
36. Huang, W.; Hew, K.F.; Fryer, L.K. Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *J. Comput. Assist. Learn.* **2021**, *38*, 237–257. [CrossRef]
37. Siemens, G. Learning Analytics: The Emergence of a Discipline. *Am. Behav. Sci.* **2013**, *57*, 1380–1400. [CrossRef]
38. Gašević, D.; Dawson, S.; Siemens, G. Let's not forget: Learning analytics are about learning. *TechTrends* **2014**, *59*, 64–71. [CrossRef]
39. Ninaus, M.; Sailer, M. Closing the loop—The human role in artificial intelligence for education. *Front. Psychol.* **2022**, *13*, 956798. [CrossRef]
40. Alfredo, R.; Echeverria, V.; Jin, Y.; Yan, L.; Swiecki, Z.; Gašević, D.; Martinez-Maldonado, R. Human-centred learning analytics and AI in education: A systematic literature review. *Comput. Educ. Artif. Intell.* **2024**, *6*, 100215. [CrossRef]
41. Topali, P.; Ortega-Arranz, A.; Rodríguez-Triana, M.J.; Er, E.; Khalil, M.; Akçapınar, G. Designing human-centered learning analytics and artificial intelligence in education solutions: A systematic literature review. *Behav. Inf. Technol.* **2024**, *44*, 1071–1098. [CrossRef]
42. Fredricks, J.A.; Filsecker, M.; Lawson, M.A. Student engagement, context, and adjustment: Addressing definitional, measurement, and methodological issues. *Learn. Instr.* **2016**, *43*, 1–4. [CrossRef]
43. Matcha, W.; Uzir, N.A.; Gašević, D.; Pardo, A. A Systematic Review of Empirical Studies on Learning Analytics Dashboards: A Self-Regulated Learning Perspective. *IEEE Trans. Learn. Technol.* **2020**, *13*, 226–245. [CrossRef]
44. Ifenthaler, D.; Yau, J.Y.K. Utilising learning analytics to support study success in higher education: A systematic review. *Educ. Technol. Res. Dev.* **2020**, *68*, 1961–1990. [CrossRef]
45. Lai, J.W.; Zhang, L.; Sze, C.C.; Lim, F.S. Learning Analytics for Bridging the Skills Gap: A Data-Driven Study of Undergraduate Aspirations and Skills Awareness for Career Preparedness. *Educ. Sci.* **2025**, *15*, 40. [CrossRef]
46. Baucks, F.; Schmucker, R.; Wiskott, L. Gaining Insights into Course Difficulty Variations Using Item Response Theory. In Proceedings of the 14th Learning Analytics and Knowledge Conference, Kyoto Japan, 18–22 March 2024; pp. 450–461. [CrossRef]
47. Shaffer, D.W.; Collier, W.; Ruis, A.R. A Tutorial on Epistemic Network Analysis: Analyzing the Structure of Connections in Cognitive, Social, and Interaction Data. *J. Learn. Anal.* **2016**, *3*, 9–45. [CrossRef]
48. Zhang, S.; Gao, Q.; Sun, M.; Cai, Z.; Li, H.; Tang, Y.; Liu, Q. Understanding student teachers' collaborative problem solving: Insights from an epistemic network analysis (ENA). *Comput. Educ.* **2022**, *183*, 104485. [CrossRef]
49. Viberg, O.; Hatakka, M.; Bälter, O.; Mavroudi, A. The current landscape of learning analytics in higher education. *Comput. Hum. Behav.* **2018**, *89*, 98–110. [CrossRef]
50. Ifenthaler, D.; Schumacher, C. Student perceptions of privacy principles for learning analytics. *Educ. Technol. Res. Dev.* **2016**, *64*, 923–938. [CrossRef]
51. Khalil, M.; Prinsloo, P. The lack of generalisability in learning analytics research: Why, how does it matter, and where to? In Proceedings of the 15th International Learning Analytics and Knowledge Conference, Dublin Ireland, 3–7 March 2025; pp. 170–180. [CrossRef]
52. Noroozi, O.; Pijeira-Díaz, H.J.; Sobocinski, M.; Dindar, M.; Järvelä, S.; Kirschner, P.A. Multimodal data indicators for capturing cognitive, motivational, and emotional learning processes: A systematic literature review. *Educ. Inf. Technol.* **2020**, *25*, 5499–5547. [CrossRef]
53. Swacha, J.; Gracel, M. Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications. *Appl. Sci.* **2025**, *15*, 4234. [CrossRef]
54. Gupta, S.; Ranjan, R.; Singh, S.N. A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions. *arXiv* **2024**, arXiv:2410.12837.
55. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020.
56. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; Wang, H. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv* **2024**, arXiv:2312.10997.
57. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137. [CrossRef]
58. Kodinariya, T.M.; Makwana, P.R. Review on determining number of Cluster in K-Means Clustering. *Int. J. Adv. Res. Comput. Sci. Manag. Stud.* **2013**, *1*, 90–95.
59. Chawla, S.; Gionis, A. k-means–: A unified approach to clustering and outlier detection. In Proceedings of the 2013 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, Austin, TX, USA, 2–4 May 2013. [CrossRef]

60. Wang, J.; Fan, W. The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanit. Soc. Sci. Commun.* **2025**, *12*, 621. [CrossRef]
61. Fan, Y.; Tang, L.; Le, H.; Shen, K.; Tan, S.; Zhao, Y.; Shen, Y.; Li, X.; Gašević, D. Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *Br. J. Educ. Technol.* **2024**, *56*, 489–530. [CrossRef]
62. Al-kfairy, M.; Mustafa, D.; Kshetri, N.; Insiew, M.; Alfandi, O. Ethical Challenges and Solutions of Generative AI: An Interdisciplinary Perspective. *Informatics* **2024**, *11*, 58. [CrossRef]
63. Kestin, G.; Miller, K.; Klales, A.; Milbourne, T.; Ponti, G. AI Tutoring Outperforms Active Learning. *Res. Sq.* **2024**. [CrossRef]
64. Bastani, H.; Bastani, O.; Sungu, A.; Ge, H.; Kabakcı, O.; Mariman, R. Generative AI Can Harm Learning. *Whart. Sch. Res. Pap.* **2024**, 1–59. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

computers

MDPI

Article

Generative Artificial Intelligence as a Catalyst for Change in Higher Education Art Study Programs

Anna Ansone *, Zinta Zālīte-Supe and Linda Daniela *

Faculty of Education Sciences and Psychology, University of Latvia, LV-1083 Riga, Latvia; zinta.zalite-supe@lu.lv
* Correspondence: a.ansone@lu.lv (A.A.); linda.daniela@lu.lv (L.D.)

Abstract: Generative Artificial Intelligence (AI) has emerged as a transformative tool in art education, offering innovative avenues for creativity and learning. However, concerns persist among educators regarding the potential misuse of text-to-image generators as unethical shortcuts. This study explores how bachelor's-level art students perceive and use generative AI in artistic composition. Ten art students participated in a lecture on composition principles and completed a practical composition task using both traditional methods and generative AI tools. Their interactions were observed, followed by the administration of a questionnaire capturing their reflections. Qualitative analysis of the data revealed that students recognize the potential of generative AI for ideation and conceptual development but find its limitations frustrating for executing nuanced artistic tasks. This study highlights the current utility of generative AI as an inspirational and conceptual mentor rather than a precise artistic tool, highlighting the need for structured training and a balanced integration of generative AI with traditional design methods. Future research should focus on larger participant samples, assess the evolving capabilities of generative AI tools, and explore their potential to teach fundamental art concepts effectively while addressing concerns about academic integrity. Enhancing the functionality of these tools could bridge gaps between creativity and pedagogy in art education.

Keywords: AI integration; art education; DALL-E; generative AI; higher education; image generation

1. Introduction

In recent years, artificial intelligence (AI) has undergone substantial advancements, leading to transformative innovations and broad applications of the technology [1]. The potential of AI in education is widely recognized, with international educational organizations emphasizing its capacity to revolutionize teaching, learning, and institutional operations [2]. AI-powered tools and systems, including generative artificial intelligence systems like GPT-4, Open Assistant, DALL-E, and MidJourney, have revolutionized higher education by offering data-driven decision-making, adaptive assessments, and personalized learning experiences [3,4]. These systems not only enable autonomous generation of human-like text and images but also engage in a variety of intellectual tasks, reshaping the traditional role of higher education institutions in knowledge transfer and skill development [1]. While these technologies offer innovative ways for students to explore and understand artistic principles, they also introduce challenges related to academic integrity and the authenticity of student work.

Tools like image generators and conversational AI can assist students in visualizing complex concepts, experimenting with different compositional techniques, and receiving

immediate feedback. This technological integration can enhance creativity, foster independent learning, and prepare students for a digital-centric art industry. For instance, Hutson and Lang (2023) demonstrated that tools like DALL-E can enrich creativity in the classroom by offering immediate visual feedback, although they emphasized the importance of ensuring AI complements rather than replaces human creativity [5]. Similarly, Vartiainen and Tedre (2023) highlighted the reflective value generative AI brings to art education but noted concerns regarding algorithmic bias and copyright issues [6]. Despite these benefits, integrating AI tools into curricula presents challenges, particularly around ensuring students continue to develop critical creative skills.

Generative AI tools often struggle with executing precise artistic instructions, especially in tasks requiring nuanced understanding of composition principles. Research has shown the necessity for higher education institutions to adapt their curricula to nurture high-quality art design talent in response to the demands of an AI-driven era [5,7]. Furthermore, AI-enhanced learning environments have been shown to significantly improve student performance and motivation, as evidenced by Chiu et al. (2022), who reported that a deep learning-based art learning system enhanced student achievement and satisfaction in art education [8]. Generative AI may not yet be sophisticated enough to fully replace traditional learning methods for foundational art skills. Relying solely on these tools could impede the development of essential artistic abilities and critical thinking.

The integration of generative AI into art education is a burgeoning field, yet current research reveals significant gaps that warrant attention. Existing research often lacks a task-based analysis to determine the optimal use cases for AI tools within art education. Thus, this research addresses a critical gap by clarifying the specific artistic educational contexts in which generative AI tools offer the most substantial benefits or present notable limitations. Understanding the specific tasks where generative AI can enhance or impede the learning process is crucial for its effective integration. This involves identifying which aspects of art education, such as ideation, prototyping, or technical skill acquisition, benefit most from AI assistance, and where it might pose challenges, such as in fostering original creative thinking. Furthermore, this study highlights the importance of integrating AI literacy, particularly prompt engineering, into art education curricula to enable students to better align generative AI outputs with their artistic intentions.

Addressing these research gaps is vital for the informed and responsible incorporation of generative AI into art education. By conducting thorough investigations into the educational impacts of generative AI and performing detailed task-based analyses, educators can develop strategies that leverage AI's strengths while mitigating its challenges. This study seeks to contribute to the growing body of knowledge by exploring how bachelor's-level art students perceive and interact with generative AI tools in the artistic composition process, addressing the following research questions:

1. What are the key challenges students face when using generative AI tools for artistic composition?
2. How do students refine and iterate AI-generated compositions to align with their artistic intent?
3. What are students' attitudes toward the integration of generative AI tools in art education?

2. Generative AI in Higher Art Education

Generative AI, a subset of Artificial Intelligence, refers to technologies that utilize deep learning models to produce human-like content, such as text, images, and other media, in response to diverse prompts like languages, instructions, or questions. While generative AI technologies have been in development for some time, the release of tools like ChatGPT has

significantly increased both public and academic interest in the field [9–11]. The evolution of AI itself can be traced back to Turing's foundational work, which introduced the concept of integrating intelligent reasoning into machines [12,13].

The advent of generative AI platforms, such as ChatGPT, MidJourney, and Gemini, has led to transformative impacts across various sectors, including education, labor, and leisure [14]. User-friendly interfaces and widespread accessibility have contributed to the rapid adoption of generative AI tools [15]. For instance, ChatGPT reached one million users within five days of its launch and has since amassed millions of active users globally [16].

Generative AI systems operate by leveraging large language models (LLMs) trained on vast datasets. These models can produce text, images, audio, and videos often indistinguishable from human-created content [17]. Applications of generative AI in education include creating tailored learning materials, assisting with language translation, and generating creative outputs like essays and blog posts. According to Lim and colleagues, generative AI uses deep learning models to generate human-like responses to complex prompts, making it a valuable tool across various disciplines [9].

Generative AI has introduced innovative solutions for personalized learning, administrative support, and content generation in higher education. AI-driven personalized learning systems adapt to individual students' psychological needs and contextual factors, improving knowledge retention and engagement [18]. Virtual tutors, for example, have been shown to increase student motivation and creativity by personalizing the learning experience [19,20].

Additionally, AI-powered grading systems offer scalable and efficient alternatives to traditional assessment methods, addressing challenges in large class settings [21]. Predictive analytics further enables institutions to identify at-risk students and design timely interventions to improve learning outcomes [22]. AI also streamlines administrative tasks, reducing faculty workload and enhancing teaching efficiency [23]. Examples include automating student admissions, financial aid processing, and library services [24]. Moreover, LLMs could revolutionize educational practices by serving as virtual instructors, curriculum developers, and contributors to scholarly work [11].

Generative AI enhances student learning by enabling tailored educational experiences, improving accessibility, and leveraging data-driven approaches to monitor progress. Tools like ChatGPT provide writing assistance, especially for non-native English speakers, fostering creativity and idea generation [25]. Text-to-image generators like DALL-E support teaching in the arts, enriching students' technical and artistic understanding [26].

The integration of generative AI in higher education also raises ethical and practical challenges. Issues of plagiarism and academic integrity have become more prominent, as AI-generated content is often undetectable by standard plagiarism tools [27]. Additionally, biases in AI-generated content can reinforce stereotypes and marginalize certain groups; therefore, training datasets must be carefully examined [28]. The potential misuse of generative AI highlights the need for clear guidelines and ethical frameworks to govern its application in education [29]. Additionally, educators must balance AI assistance with fostering students' critical thinking and independent learning skills [30].

Despite its limitations, generative AI offers transformative potential for education. Its ability to create personalized learning experiences, streamline administrative tasks, and enhance teaching efficiency makes it an invaluable tool for modern higher education.

Generative AI has demonstrated its capability to create poetry, visual art, and music, and even to choreograph dance movements [31]. These advancements necessitate understanding how individuals perceive AI-generated art and the personal attributes influencing these perceptions [32]. Previous research has concentrated on two primary

questions: whether humans can differentiate between AI-generated and human-created art and whether biases exist toward AI-generated works.

Regarding the ability to discern between artworks, most studies suggest that individuals struggle to consistently distinguish human-made art from AI-generated art [33,34]. Similarly, a notable body of research highlights a negative bias toward AI-generated art [35], even in instances where the source of the artwork is unclear to participants. Notably, while the average viewer often categorizes AI-generated creations as "art" [36], debates persist about whether AI can genuinely fulfill the role of an artist. Critics argue that art inherently requires intention and communication, attributes that AI lacks independent of human input [37,38]. Consequently, AI-generated art remains reliant on human-designed algorithms, training datasets, and prompts, inherently tying its outputs to human involvement.

The integration of AI into artistic creation dates back to Harold Cohen's AARON project in 1973, which translated visual decision-making processes into robotic painting systems [39]. Over subsequent decades, the evolution of machine learning, pattern recognition, generative adversarial networks (GANs), and text-to-image diffusion models has profoundly influenced artistic practices. These tools have empowered artists to explore new creative processes, challenging traditional boundaries of authorship and creativity [40]. Generative AI, by enabling the production of art-like outputs with minimal effort, has significantly democratized artistic creation, allowing a broader range of individuals to engage in creative expression. However, the blurred distinction between the artist and the tool raises critical questions about authorship, ownership, and creative agency [38]. While generative AI expands access to creative resources, it simultaneously redefines the relationship between human creators and their tools, necessitating further exploration of its implications for traditional artistic norms.

Concerns about AI replacing human creativity often underpin preferences for human-made art [41]. Such fears reflect deeper anxieties about preserving human identity as the sole source of creative innovation [42]. Moreover, ethical and moral considerations, including a desire to support human artists, may also influence evaluations of human versus AI-generated artworks. General attitudes toward AI vary across cultures and individuals, influencing its acceptance and integration into creative contexts [43]. For example, individuals with a favorable view of AI often attribute higher profundity and value to AI-generated creations [44]. Conversely, skepticism about AI's role in creativity, particularly its potential to replace human effort, remains a significant barrier to broader acceptance [45]. The discourse surrounding generative AI in art emphasizes its role as a new medium rather than a threat to creativity. While it raises questions about authorship and originality, its potential to augment artistic practices and democratize access to creative tools presents opportunities for further exploration. As perceptions of AI-generated art continue to evolve, understanding the interplay between human and AI contributions will remain critical to navigating this transformative landscape.

Generative AI tools such as ChatGPT, DALL-E, and MidJourney have introduced novel opportunities for creative expression in educational contexts. These tools enable students to visualize abstract concepts, engage in collaborative creativity, and rapidly prototype ideas. Beyond enhancing artistic skills, generative AI supports the development of problem-solving and critical thinking abilities, particularly when educators are equipped to confidently integrate these technologies into the classroom. The adoption of generative AI in art education is reshaping traditional pedagogical approaches by augmenting creativity and facilitating personalized learning experiences. Research has demonstrated how generative AI aids fashion designers in ideation and prototyping, as seen in the development of tools like StoryDrawer for visual storytelling and AI-based garment design systems that integrate fashion expertise [46]. These applications highlight the synergy between human

designers and AI, emphasizing the need for technical proficiency and creative adaptability in leveraging AI tools effectively.

Despite its potential, generative AI integration in art education faces several challenges, including ethical concerns about authorship, dependency risks, and accessibility barriers. Ethical concerns regarding intellectual property, over-reliance on automated outputs, and data bias emerge as critical issues for students who are still developing foundational artistic skills. For instance, Sullivan et al. (2023) [27] highlight the academic integrity dilemmas associated with AI-generated assignments, noting that traditional plagiarism checks are often insufficient when content is produced by generative models. Similarly, Dwivedi et al. (2023) [28] argue that reliance on AI can diminish learners' opportunities to cultivate creativity and problem-solving abilities, which is a key consideration in art education, where hands-on experimentation is vital. Moreover, Tedre, Kahila, and Vartiainen (2023) [6] emphasize the importance of carefully structuring AI-based projects in craft and design classes to ensure that students learn critical thinking and prompt-engineering skills, rather than simply accepting AI outputs as final. Furthermore, the importance of "prompt engineering", crafting precise inputs for desired AI outputs, has emerged as a critical skill in AI-assisted design, underscoring the necessity of specialized training for both educators and students. Taken together, these studies illustrate the need for educator training, transparent guidelines on AI use, and balanced curricular approaches to preserve students' artistic autonomy while leveraging AI's benefits for exploration and ideation.

Given the rapid development of AI technologies, it is inevitable that these tools will be increasingly integrated into creative industries, including art education. To fully understand how generative AI tools can be effectively utilized in these contexts, it is essential to explore how students engage with these technologies, the educational value they derive, and the challenges they present.

This study addresses the concerns of educators regarding the potential misuse of generative AI tools, such as text-to-image generators like DALL-E, which some view as unethical shortcuts that undermine traditional learning methods. The focus of the research is to examine how undergraduate art students use generative AI tools in their learning, particularly in a practical composition task, and to identify the educational benefits and limitations these tools offer. While the study highlights the potential of generative AI for ideation and conceptual development, it also emphasizes its limitations in producing refined, nuanced artistic outputs. By examining these dynamics, this study aims to bridge the gap in understanding how generative AI can contribute to art education. The following section details the methodology employed in this research, outlining the steps taken to observe student interactions, gather reflections, and analyze the data.

3. Materials and Methods

The research involved bachelor's-level art students from a university's art program. Although the class had 13 students, participation in the research was determined by attendance during the first lecture, where key instructions and study details were provided. Since only 10 students attended this session, they were invited to participate in the study. This ensured that all participants received the same information and began their tasks with consistent instructions. The participants' ages ranged from 18 to 26, with a mix of male and female students. All participants had prior experience in traditional art practices, but their familiarity with generative AI tools varied. Some students had used AI for ideation and concept development, while others had little to no experience with AI tools.

This study met with ethical research standards, guaranteeing that involvement was entirely voluntary and had no bearing on students' grades or academic achievement. The study's goals and methods, and participants' freedom to discontinue participation at any

moment without facing impacts, were explained to them before the study began. We obtained informed consent from each participant. To protect participant identities, data were gathered and analyzed anonymously, and the study adhered to the Declaration of Helsinki's ethical guidelines for research involving human subjects.

The study employed a mixed-methods approach, combining quantitative data from questionnaires with qualitative observations and participant feedback. The research was conducted in two stages over a one-month period:

1. Week 1—Lecture on basic composition principles. Traditional task completion: students completed a composition task using traditional methods.
2. Weeks 2 to 4—AI-assisted task completion: students attempted to complete the same composition task using the generative AI tool DALL-E.

The methodology ensured that participants completed both tasks in a controlled environment where their performance could be observed and evaluated systematically. This approach allowed for a meaningful comparison between traditional and AI-assisted composition techniques, providing valuable insights into how generative AI tools impact the creative process.

3.1. Week 1: Lecture and Traditional Composition Task

All 10 participants in the composition class were invited to attend a 30 min lecture on the fundamental principles of composition, including balance, contrast, unity, and emphasis. The lecture was followed by a brief discussion on how to apply these principles in a practical task. After, students were given a practical task (for detailed description, see Appendix A) that required them to create a basic composition following the principles they had learned. Students were given 1.5 h to complete the task (Figure 1).

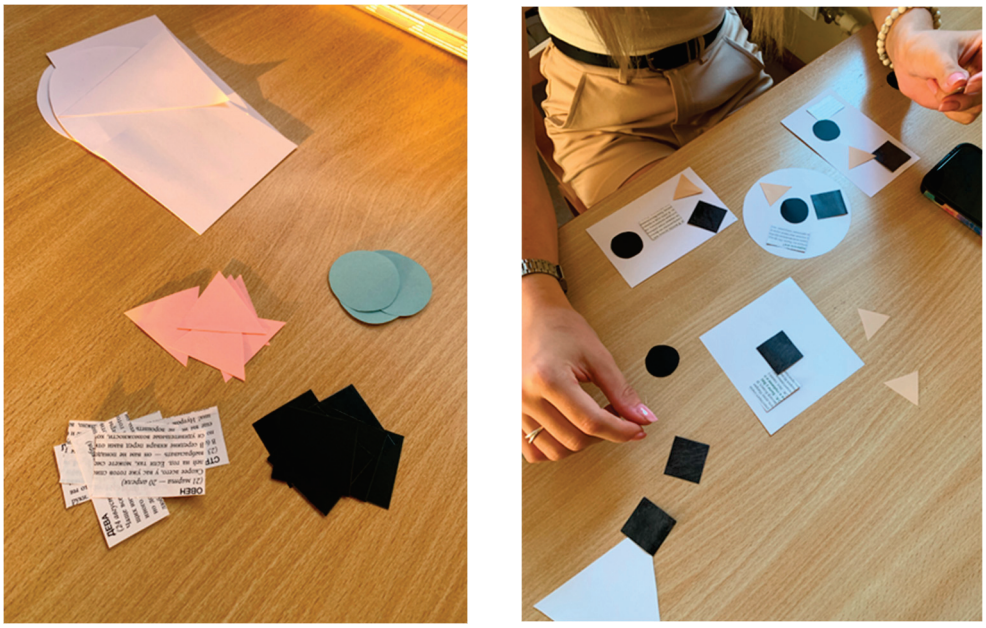


Figure 1. Traditional task process.

3.2. Weeks 2–4: AI-Assisted Composition Task

The participants were introduced to the generative AI tool they would use for the second task. The computer with the ChatGPT premium version was provided to them. Those unfamiliar with the tools were given a brief demonstration of how to input prompts and adjust settings to generate visual content. The students were instructed to replicate the same composition task they performed in the first week, but this time using generative AI

tools. They input prompts that described their intended compositions and attempted to refine the AI-generated images to match their original work (Figure 2).

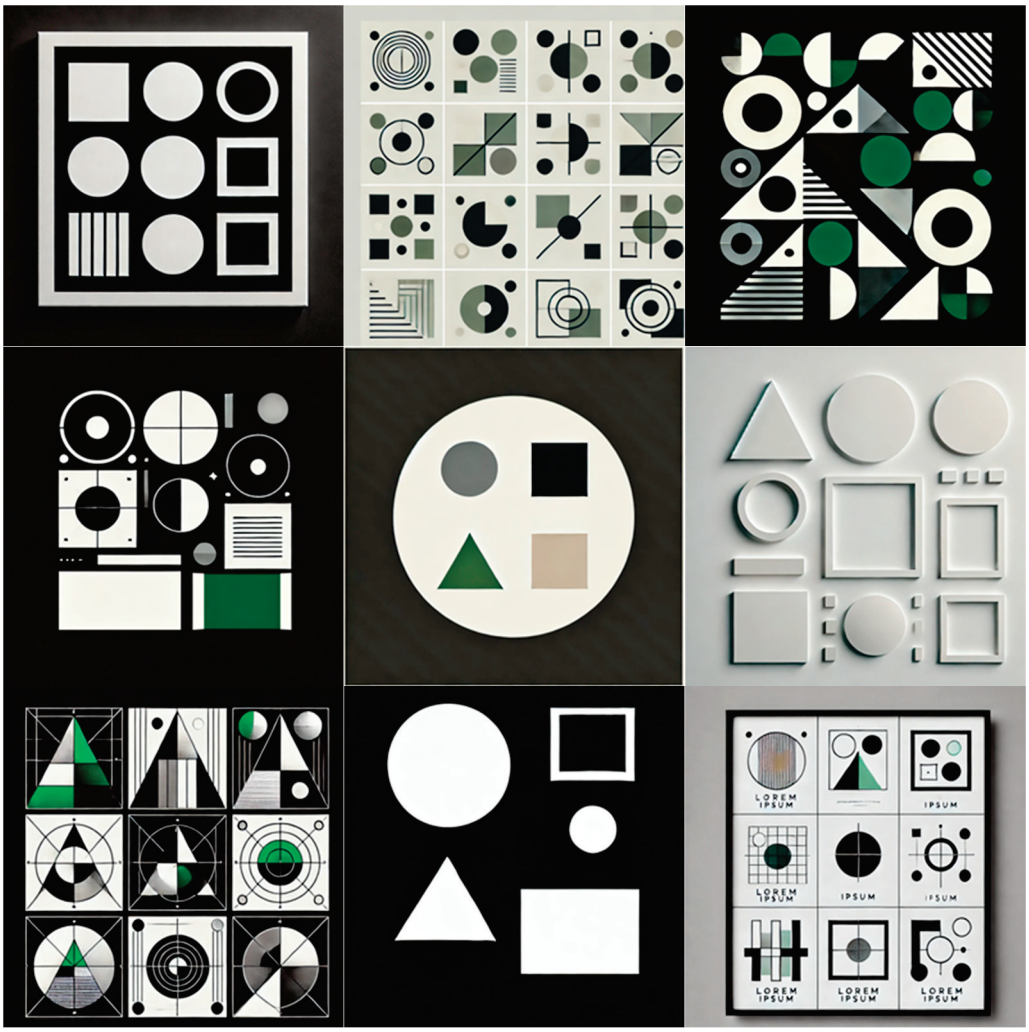


Figure 2. Images students generated during experiments using DALL-E.

Each participant was observed individually over the course of their 30 min task completion, during which they were encouraged to engage in a "think-aloud" process. This method allowed the researcher to document their thought processes, challenges, and task completion strategies in real-time. To systematically capture data, an observation protocol (Appendix B) was developed and utilized for each participant. The protocol was designed to record key parameters related to the students' interaction with the AI tools and their approach to the task. Each protocol was filled in separately for individual participants to ensure detailed and personalized documentation.

3.3. Post-Task Questionnaire

After both the traditional and AI-assisted tasks were completed, each student filled out a questionnaire that measured their experiences. The survey comprised a total of 10 questions, structured to gather both quantitative and qualitative data regarding participants' experiences with generative AI tools in the context of art education. The survey was divided into two types of questions: structured (multiple-choice items and Likert-scale statements ranging from "Strongly Disagree" to "Strongly Agree") and open-ended. The structured questions focused on the following themes:

1. Demographics and background: Age, gender, and prior experience with generative AI tools.
2. AI tool usage: Frequency and purposes of using generative AI tools, such as ideation, concept development, and design refinement.
3. Perceptions and attitudes: Participants' views on the impact of AI tools on productivity, creativity, task efficiency, and design quality.
4. Ease of use and accessibility: Assessing the usability, clarity, and learning curve of generative AI tools.
5. Support and integration: Perceived support from peers, instructors, and the institution for using AI tools in art education.
6. Future intentions: Plans to continue using AI tools in academic and professional work.

The survey included four open-ended questions to allow participants to elaborate on their experiences and provide nuanced feedback. These questions explored the following:

1. The impact of AI tools on participants' creativity.
2. Challenges faced when integrating AI tools into their design workflows.
3. Participants' perceptions of AI's potential role in their future careers as designers.
4. Suggestions for improving the integration of generative AI tools in art and design education.

This mixed-methods approach ensured a comprehensive data collection process. The structured questions facilitated measurable comparisons and statistical analysis, while the open-ended responses provided rich qualitative insights into participants' individual experiences and perspectives. The questionnaires were administered electronically through the Google Forms platform, ensuring ease of access for participants while maintaining a controlled and consistent method for data collection. Students were provided with the survey link in the first lecture. They were instructed to complete the questionnaire individually and were informed about the deadline to submit their responses within a specified timeframe. The questionnaire was designed to be anonymous.

While participants were assigned a respondent number at the start of the survey for organizational purposes, no personally identifiable information was collected, ensuring complete confidentiality. At the beginning of the questionnaire, a detailed informed consent section was included. This section clearly outlined the purpose of the research, the voluntary nature of participation, and the participant's right to withdraw at any time without any negative consequences. The consent section explicitly stated the following: That the study aimed to investigate the role of generative AI tools in art education; that their responses would be used solely for academic research purposes; and that the data would be anonymized and securely stored to protect confidentiality. Participants were required to check a box to confirm their consent to participate before proceeding to the survey questions. This served as an acknowledgment that they had read and understood the research information and voluntarily agreed to participate. Students were given clear instructions on how to complete the questionnaire, including a brief explanation of each section and the approximate time required to complete it. They were also encouraged to provide thoughtful and honest responses, especially for the open-ended questions, to contribute valuable insights to the study.

The combination of surveys, structured observations, and participant feedback provided a comprehensive approach to understanding the integration of generative AI tools in art education. By utilizing a mixed-methods framework, the study ensured both quantitative and qualitative depth, capturing measurable data alongside rich, contextual insights. The observation protocol, think-aloud methodology, and survey responses ensured that participants' experiences were documented systematically and ethically. These methods were designed to align with the research objectives, emphasizing the participants' inter-

actions with AI tools, their creative processes, and their perspectives on AI's role in their education and future careers.

4. Results

This study collected data from 10 bachelor's-level art program students regarding their experiences using generative AI tools, focusing on how these tools assist in the process of creating and learning about composition. The data were collected through practical tasks using both traditional observation and a questionnaire. The analysis focused on understanding the perceived efficacy of generative AI tools, the challenges faced, and the students' overall attitudes toward their use in art education.

4.1. Observation

The observations of 10 student participants engaging with generative AI tools for artistic composition reveal distinct patterns in their interaction strategies, challenges, and emotional responses. The data indicate that students adopted iterative refinement techniques, faced difficulties in prompt formulation, and demonstrated varied levels of confidence in AI tool navigation.

All participants (10) engaged in an iterative process to refine AI-generated outputs, demonstrating a pattern of progressive prompt adjustment. The majority of students (8) initially provided broad, conceptual prompts, later refining them with additional specifications based on the output generated. For example, one student initially entered "create a geometric composition" but refined it to "place four red circles in the top-left corner with a white background". This refinement process involved three to six iterations per participant, highlighting the trial-and-error nature of AI-assisted design.

Moreover, five students employed a layered approach, constructing compositions in stages (e.g., first generating a background, then adding objects, and finally adjusting details such as color and texture). However, the remaining students (5) relied on direct prompt modifications, without structuring their workflow into progressive steps. Across all participants, a common challenge was difficulty in translating artistic intent into precise text-based AI prompts. This issue was observed in multiple aspects. Students (7) struggled to specify precise placement of elements, resulting in unintended AI-generated compositions. One participant noted, "The AI isn't fully capturing the symmetry I have in mind". Six students reported that AI often added extra elements or misinterpreted requests for specific numbers of shapes. For instance, one participant explicitly requested six geometric shapes, but the AI generated additional unintended elements. Four participants experienced inconsistencies in maintaining a fixed color scheme across iterations, despite explicitly specifying color palettes.

Student confidence in navigating the AI tool was moderately high, with seven participants demonstrating competence in basic operations such as entering prompts and adjusting parameters. However, confidence declined when refining more complex compositions, particularly when precise alignment or proportional adjustments were needed. Three participants attempted to use seed values to maintain consistency across iterations yet noted that AI-generated outputs still exhibited randomness, making exact replication challenging. Five students displayed uncertainty when adjusting prompts, often relying on trial and error rather than structured refinement strategies. Eight students actively verbalized their thought process while working with the AI tool, making remarks such as "I need the shapes to feel balanced" or "This color doesn't quite match what I intended". However, four students expressed difficulty in translating these artistic observations into effective textual instructions.

The students exhibited neutral to mildly frustrated emotional responses throughout the design process. All participants (10) displayed engagement, but frustration was observed in specific instances. When AI-generated outputs deviated from expectations, six participants expressed visible signs of dissatisfaction (e.g., sighing, reattempting prompts multiple times, or verbally expressing disappointment). Seven participants reported a sense of accomplishment when AI-generated outputs progressively aligned with their vision. One participant stated, "It's getting closer to what I want, but I wish I could tweak small details more easily". Four participants adjusted their expectations after multiple iterations, accepting minor inaccuracies in shape placement or proportions as a trade-off for efficiency in generating compositions.

The degree to which the final AI-generated outputs aligned with students' original artistic visions varied across participants. The majority (7) achieved only partial alignment, where the AI-generated composition captured core elements of their artistic goal (e.g., color scheme and general layout) but lacked precision in spatial relationships or object proportions. Two participants were unable to achieve their intended artistic outcome due to persistent discrepancies in AI-generated shapes and spacing, despite multiple refinements. Six students mentioned that AI-generated images could be exported and further refined manually in traditional design software, indicating that AI served as a starting point rather than a final design tool.

4.2. Questionnaire

The questionnaire responses were collected from the 10 students who participated in the study, offering insights into their familiarity with generative AI tools, the perceived impact of these tools on their design work, and their attitudes toward their use in the future. Participants ranged in age from 18 to 26, with the majority (6 students) in the 21–23 age group. The group consisted of six female and four male students.

Regarding familiarity with AI tools, seven students reported using AI primarily for idea generation and concept development, while fewer used them for tasks such as color scheme selection (four students) or finalizing designs (three students). Among the tools used, DALL-E was the most popular (four students), followed by Canva's AI features (two students) and Stable Diffusion (one student). Notably, three participants indicated that they did not currently use generative AI tools, highlighting diverse levels of familiarity within the group.

The students agreed that generative AI tools enhanced their productivity, with an average rating of 4.3 out of 5, allowing them to complete design tasks more efficiently. Similarly, they rated AI tools highly in improving the quality of their work, particularly during the ideation phase, with an average rating of 4.5 out of 5. The majority (seven students) agreed that generative AI tools are most beneficial in the conceptual stage, as they provide inspiration and assistance in early design processes.

While eight students found AI tools user-friendly and expressed confidence in their ability to learn and operate them effectively, the average ease-of-use rating was lower, at 3.8 out of 5. This indicates that while students appreciated the tools' potential, some faced a steep learning curve, which they overcame with practice. Challenges noted by participants included difficulty in aligning AI outputs with precise artistic instructions (three students), overwhelming variety of AI-generated outputs, making it hard to refine ideas (two students), technical issues with the tools' functionality (two students). A few participants noted that mastering the use of generative AI tools had a steep learning curve, though this was overcome with practice. However, two students highlighted challenges, such as a steep learning curve and difficulty balancing AI-generated ideas with their own

creativity, highlighting that while students appreciate the potential of generative AI, they find the learning curve somewhat challenging.

The students expressed positive attitudes toward the future role of AI in design. Seven out of ten students indicated their intent to use AI tools regularly in the future, with all participants acknowledging their usefulness in ideation and efficiency. Seven described AI as a "collaborator" rather than a replacement, emphasizing its potential to enhance creativity while underscoring the importance of maintaining human creativity and intuition. Despite optimism, some students expressed caution about over-reliance on AI tools, stressing the need to balance AI assistance with the cultivation of traditional artistic skills. Students provided actionable suggestions for better integrating AI tools into art education. Four students emphasized the need for hands-on workshops, while three recommended incorporating case studies of successful AI use in design. Additionally, two students suggested fostering collaborative projects that utilize generative AI tools to encourage teamwork and practical experience.

Table 1 captures the key findings from the study, drawing on both observational data and questionnaire responses. The table is structured to align each key theme with supporting evidence, thereby illustrating how participants' feedback, documented behaviors, and survey results interrelate.

Table 1. Summary of Key Observed and Survey-Based Findings.

Finding	Key Observations and Survey Results	Implications for Art Education
AI for Brainstorming	Seven of ten students reported that generative AI was most useful for ideation rather than fine-grained composition details.	Suggests AI excels at sparking initial creativity but remains limited in executing nuanced artistic tasks.
Iterative Trial-and-Error	All 10 students iterated prompts three to six times to refine AI outputs.	Demonstrates that learning-by-doing (iterative refinement) is central to bridging the gap between AI output and artistic vision.
Prompt-Formulation Challenges	Five students struggled with specifying spatial relationships, proportions, and color consistency in their prompts.	Highlights the importance of prompt-engineering skills. Students need practice translating artistic intent into precise AI instructions.
Layered vs. Direct Workflows	Five participants used a layered approach (step-by-step building of an image), while five used direct prompts and immediate edits.	Reveals two distinct strategies for AI-assisted artmaking: structured layering vs. ad-hoc refinement, each with different learning curves.
Misalignment and Dissatisfaction	Four students expressed frustration when AI misinterpreted precise design instructions.	Indicates AI's current limitations in fine control and the potential for user dissatisfaction when outputs deviate from intended results.
Basic Competence, Advanced Challenges	Seven students felt confident in basic AI operations but struggled with precise details or advanced adjustments.	Confirms that initial familiarity is achievable, but more specialized training is needed to handle complex design tasks effectively.
AI as Complementary, Not Replacement	Eight students described AI tools as enhancing creativity but not replacing human intuition or craftsmanship.	Suggests a positive view of AI as a "co-creator", emphasizing that human ingenuity remains essential for artistic decision-making.
Future Intentions and Caution	Seven students intend to continue using AI, specifically for brainstorming but caution against over-reliance on AI-generated content.	Reflects students' recognition of AI's value in ideation while acknowledging the risk of eroding foundational skills if used too heavily.
Proposed Improvements	Four students requested hands-on workshops, three suggested case studies, and two advocated for collaborative projects.	Indicates a demand for structured support; educators could embed these methods into curricula to develop both AI and artistic competencies.

5. Discussion

This study offers valuable insights into how art students utilize generative AI tools for artistic composition. While these tools support idea generation and concept development, challenges persist in achieving precision, control, and artistic intent.

One noteworthy factor behind these challenges is the difficulty of translating complex visual ideas into text-based instructions, which can be attributed both to a lack of prompt-engineering skills and to the limitations of language for describing relative or absolute positions, shapes, and sizes [1,5]. This twofold constraint often forced participants into iterative, trial-and-error workflows, as evidenced by their repeated refinement of prompts when AI outputs did not match their artistic visions. Our findings highlight the need for more structured guidance on prompt formulation in art education. This study addressed the following research questions:

1. **What are the key challenges students face when using generative AI tools for artistic composition?**

One major challenge was the difficulty in crafting precise prompts that accurately conveyed artistic vision to AI systems. Seven students struggled with spatial placement and proportions, frequently encountering unintended compositions due to AI misinterpretation. Three students attempted to use seed values to maintain consistency, but the randomness of AI-generated outputs made exact replication difficult. Students also found that AI often misinterpreted numerical requests (e.g., generating extra elements when specifying a fixed number of shapes). This highlights a critical gap in AI literacy, as effective prompt engineering plays a crucial role in enhancing the accuracy and relevance of AI-generated outputs. Research suggests that well-structured prompts can improve response accuracy rates from 85% to as high as 98% across various contexts, underscoring the significance of this skill in AI development and application [47]. However, novice users frequently rely on trial and error, which can lead to frustration and inefficiency. These findings emphasize the necessity of integrating AI literacy training into art education, equipping students with the ability to craft precise instructions that align with their creative intentions. It is worth noting that even experienced AI users may encounter hurdles when attempting to describe intricate compositions via text alone. For tasks requiring specific object placement and proportional relationships, language-based input can be inherently limiting, necessitating more robust interface design or multimodal feedback loops.

2. **How do students refine and iterate AI-generated compositions to align with their artistic intent?**

All participants engaged in iterative refinement, using multiple prompt adjustments to improve AI-generated results. Eight students initially provided broad conceptual prompts before progressively refining them with additional specifications. However, only half of the students structured their workflow systematically, while the others relied on intuitive, trial-and-error adjustments. Five students adopted a layered approach, generating elements in stages (e.g., background first, then objects, then details), while the remaining five relied on direct prompt modifications. Some students verbalized their thought processes while refining AI outputs, indicating an adaptive learning approach.

Many students found that mastering AI tools involved a steep learning curve, particularly for tasks requiring precise execution, such as intricate compositions. The complexity of AI-generated outputs sometimes led to inefficiencies in the refinement process, as students struggled to align results with their artistic vision. Some participants noted that AI-generated images often deviated from their intended outcomes, especially for nuanced or highly detailed compositions, highlighting the limitations of generative AI in fine-tuned artistic tasks. Such limitations may reflect the current stage of AI model development,

where object relationships and spatial accuracy can be overlooked unless users input extremely detailed prompts or rely on advanced features [3].

Additionally, a few students expressed concerns about the potential drawbacks of overreliance on AI tools. While AI facilitated rapid ideation, some felt it reduced their engagement with foundational artistic principles, such as composition and spatial relationships. This suggests that while AI can enhance creative exploration, structured training in prompt formulation and systematic refinement strategies could help students better anticipate AI-generated outcomes and maintain a balance between automation and artistic intentionality [5]. Several studies corroborate the importance of balancing novel AI-driven practices with manual, hands-on design experiences, as students require fundamental skills and critical thinking to progress beyond automated outputs [4].

3. What are students' attitudes toward the integration of generative AI tools in art education?

The majority (seven students) expressed a positive attitude toward AI's role in art education, viewing it as a valuable collaborator rather than a replacement for traditional artistic techniques. All participants recognized AI's usefulness in ideation and efficiency. On a scale of 1 to 5, most respondents rated AI tools highly (4 or 5) for enhancing productivity and improving design quality. This indicates that AI significantly streamlines the design process, particularly in visualization and experimentation. This finding aligns with Inie et al. (2023), who observed that a diverse range of creative professionals recognized generative AI's potential to boost productivity, provide inspiration, and contribute to higher-quality artistic output [48].

Despite the perceived benefits, some students expressed concerns that excessive reliance on AI could hinder their development of foundational artistic skills. This concern aligns with Adeoye and Jimoh (2023), who emphasize that fostering creativity and innovation is essential for equipping learners with the skills needed to succeed in the 21st century [49]. Creativity is closely linked to problem-solving abilities and can be nurtured through opportunities for hands-on artistic exploration. However, the ease with which AI generates compositions may reduce student engagement with fundamental composition principles, highlighting the need for a balanced approach that integrates AI as a supportive tool rather than a substitute for traditional artistic techniques.

The students recommended more hands-on workshops and courses specifically focused on AI tools. Structured learning environments would reduce the learning curve and help students apply AI effectively. The respondents suggested integrating case studies and collaborative projects into art education to showcase successful uses of AI in art and design. This could deepen their understanding of AI's potential for facilitating creative processes, especially in complex artistic tasks. Encouraging students to balance AI tools with traditional art techniques would enhance their understanding of composition. AI should serve as a complementary tool rather than a replacement for foundational skills.

These findings highlight the dual role of AI in art education: while it can foster ideation and broaden creative possibilities, it also introduces obstacles related to language-based inputs, model limitations, and potential overreliance. Future research could explore multimodal AI interfaces, such as combining voice commands with sketches to mitigate the difficulties of translating intricate details or spatial relationships into text-based prompts. Incorporating real-time feedback loops in AI-based design platforms may further enable students to refine their compositions iteratively, thereby bridging the gap between verbal descriptions and artistic intent. Overall, the findings indicate that students perceive generative AI tools as useful aids in the ideation phase but face significant challenges in translating artistic intent into precise AI-generated compositions. While AI tools are seen as enhancing creativity and efficiency, their limitations in spatial control, color consistency, and detailed refinements create a sense of frustration. Most of the students (seven out of ten) view AI as

a collaborator rather than a replacement for traditional artistic techniques, emphasizing the importance of maintaining human creativity and intuition in the design process.

6. Conclusions

This study examined the experiences of 10 bachelor's-level art students using generative AI tool DALL-E for a composition task. Based on observations and questionnaire responses, the findings highlight both the advantages and limitations of generative AI tools in art education.

1. The majority of students (seven) found generative AI useful for brainstorming and developing initial ideas, rather than executing detailed artistic compositions.
2. All students (10) engaged in a trial-and-error process, progressively adjusting prompts to achieve desired results, with three to six iterations per composition.
3. Five students struggled to translate their artistic intent into effective text prompts, particularly when specifying spatial relationships, proportions, and color consistency.
4. Five students used a layered approach to build compositions in stages, while the other five modified direct prompts without structured workflows.
5. Four students expressed dissatisfaction when the AI outputs did not align with their artistic vision, often due to the AI's tendency to misinterpret precise design instructions.
6. Most students (seven) demonstrated competence in basic AI operations; confidence declined when tasks required detailed adjustments or alignment.
7. Eight students described AI tools as a complementary resource, enhancing creativity but not replacing human intuition and craftsmanship.
8. The majority (seven) intend to use AI tools in the future, especially in the ideation phase, but caution against over-reliance on AI-generated content.
9. The students suggested hands-on workshops (four students), case studies of AI in design (three students), and collaborative projects (two students) to enhance learning and application of AI tools.

These findings suggest that while generative AI has potential in art education, it requires structured integration with traditional artistic methods. While generative AI accelerates ideation and experimentation, educators and administrators must remain mindful of potential over-reliance on AI-driven outputs. Overuse of AI-generated content may reduce opportunities for students to develop hands-on expertise in foundational artistic skills such as color theory, perspective, and composition. Art programs should, therefore, balance AI-assisted tasks with traditional, manual projects that reinforce critical thinking and creative autonomy.

Skills development in prompt engineering and AI literacy will be increasingly essential for emerging art professionals. Students who can effectively translate artistic intent into clear, structured prompts and refine AI outputs are positioned to excel in an evolving creative landscape. However, integrating these skills into the curriculum requires structured guidance, hands-on workshops, and transparent rubrics that distinguish between legitimate AI collaboration and unethical shortcuts.

Ethically, the ease with which AI tools can generate finished pieces raises questions about academic honesty and attribution. As AI continues to expand its creative capacities, higher education institutions must implement clear guidelines that encourage honest reporting of both human and AI contributions.

With 10 participants drawn from a single course, the ability to generalize conclusions is limited. Future research should investigate how evolving AI capabilities can better align with artistic intent and explore larger participant samples over multiple semesters, comparing how different generative AI platforms (e.g., DALL-E, MidJourney, Stable Diffusion) are integrated into creative workflows.

Author Contributions: Conceptualization, methodology, validation, formal analysis, resources, A.A., Z.Z.-S. and L.D.; data curation, A.A. and Z.Z.-S.; writing—original draft preparation, A.A.; writing—review and editing, A.A., Z.Z.-S. and L.D.; visualization, A.A.; supervision, L.D.; project administration, L.D.; funding acquisition, L.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Task Description

1. Cut out 6 different formats from white drawing paper:
    - 1.1. Circle
    - 1.2. Square
    - 1.3. isosceles triangle
    - 1.4. oval (vertical)
    - 1.5. vertical rectangle
    - 1.6. horizontal rectangle
2. The size of the shapes is such that they can be placed on a 40 × 40 cm base board. Sort the 4 different elements in each format, observing:
    - 2.1. the center of the composition and the viewing direction,
    - 2.2. the balance of the composition
    - 2.3. the spacing of the strap elements
    - 2.4. the white space between the edges of the elements.
3. Cut out small elements—6 of each type (triangle, circle, rectangle, square). Each type is in one of four possible colors (the color can be chosen, but all elements of the same shape must be the same color)—black, grey, white with printed text, colored (any color).

Appendix B

Observation Protocol Details

1. All Task Understanding and Approach (whether the participant demonstrated a clear understanding of the task requirements).
    - 1.1. Complete understanding,
    - 2.1. Partial understanding,
    - 3.1. Unclear understanding.
2. Interaction with the AI Tool (ability to navigate the tool, input prompts, and refine outputs).
    - 1.1. Highly confident,
    - 2.1. Moderately confident,
    - 3.1. Lacking confidence.
3. Verbalized Thought Process (notes on the participant's thought process, including how they translated artistic intent into text prompts and evaluated AI-generated outputs. Qualitative descriptions were recorded for this parameter, as it varied significantly between individuals.

4. Challenges Encountered (observations of any frustrations, misunderstandings, or difficulties faced when interacting with the tool (e.g., issues with precision, prompt formulation, or output refinement)).
    1.1. Minimal challenges,
    2.1. Moderate challenges,
    3.1. Significant challenges.
5. Emotional Responses (notes on visible signs of satisfaction, frustration, or excitement during the task).
    1.1. Positive response,
    2.1. Neutral response,
    3.1. Negative response.
6. Task Completion Methodology (steps participants followed to refine the AI-generated output to meet their original artistic goals). Qualitative notes and observations were used to describe the participant's workflow.
7. Outcome Assessment (whether the participant was able to align the AI-generated composition with their original work from the first task).
    1.1. Highly aligned,
    2.1. Partially aligned,
    3.1. Not aligned.

References

1. Walczak, K.; Cellary, W. Challenges for higher education in the era of widespread access to Generative AI. *Econ. Bus. Rev.* **2023**, *9*, 71–100. [CrossRef]
2. Kalniņa, D.; Nīmante, D.; Baranova, S. Artificial intelligence for higher education: Benefits and challenges for pre-service teachers. *Front. Educ.* **2024**, *9*, 1501819. [CrossRef]
3. Atif, A.; Jha, M.; Richards, D.; Bilgin, A.A. Artificial Intelligence enabled remote learning and teaching using Pedagogical Conversational Agents and Learning Analytics. In *Intelligent Systems and Learning Data Analytics in Online Education*; Academic Press: Cambridge, MA, USA, 2021; pp. 3–29. [CrossRef]
4. Alotaibi, N.S. The Impact of AI and LMS Integration on the Future of Higher Education: Opportunities, Challenges, and Strategies for Transformation. *Sustainability* **2024**, *16*, 10357. [CrossRef]
5. Hutson, J.; Lang, M. Content creation or interpolation: AI generative digital art in the classroom. *Metaverse* **2023**, *4*, 13. [CrossRef]
6. Tedre, M.; Kahila, J.; Vartiainen, H. Exploration on how co-designing with AI facilitates critical evaluation of ethics of AI in craft education. In Proceedings of the Society for Information Technology & Teacher Education International Conference, New Orleans, LA, USA, 13 March 2023; Langran, E., Christensen, P., Sanson, J., Eds.; Association for the Advancement of Computing in Education (AACE): Waynesville, NC, USA, 2023; pp. 2289–2296.
7. Xu, B.; Jiang, J. Exploitation for multimedia Asian information processing and artificial intelligence-based art design and teaching in colleges. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2022**, *21*, 114. [CrossRef]
8. Chiu, M.C.; Hwang, G.J.; Hsia, L.H.; Shyu, F.M. Artificial intelligence-supported art education: A deep learning-based system for promoting university students' artwork appreciation and painting outcomes. *Interact. Learn. Environ.* **2022**, *32*, 824–842. [CrossRef]
9. Lim, W.M.; Gunasekara, A.; Pallant, J.L.; Pallant, J.I.; Pechenkina, E. Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *Int. J. Manag. Educ.* **2023**, *21*, 100790. [CrossRef]
10. Ernesto, D.; Gerardou, F.S. Challenges and Opportunities of Generative AI for Higher Education as Explained by ChatGPT. *Educ. Sci.* **2023**, *13*, 856. [CrossRef]
11. Ivanov, S.; Soliman, M. Game of algorithms: ChatGPT implications for the future of tourism education and research. *J. Tour. Futures* **2023**, *9*, 214–221. [CrossRef]
12. Turing, A.M. On computable numbers, with an application to the Entscheidungsproblem. *J. Math.* **1936**, *58*, 5.
13. Turing, A.M. Computing machinery and intelligence. In *Parsing the Turing Test*; Epstein, R., Roberts, G., Beber, G., Eds.; Springer: Dordrecht, The Netherlands, 2009.
14. Kurtz, G.; Amzalag, M.; Shaked, N.; Zaguri, Y.; Kohen-Vacs, D.; Gal, E.; Zailer, G.; Barak-Medina, E. Strategies for integrating generative AI into higher education: Navigating challenges and leveraging opportunities. *Educ. Sci.* **2024**, *14*, 503. [CrossRef]

15. Lacey, M.M.; Smith, D.P. Teaching and assessment of the future today: Higher education and AI. *Microbiol. Aust.* **2023**, *44*, 124–126. [CrossRef]
16. Cao, Y.; Li, S.; Liu, Y.; Yan, Z.; Dai, Y.; Yu, P.S.; Sun, L. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv* **2023**, arXiv:2303.04226.
17. O'Dea, X. Generative AI: Is it a paradigm shift for higher education? *Stud. High. Educ.* **2024**, *49*, 811–816. [CrossRef]
18. Pataranutaporn, P.; Danry, V.; Leong, J.; Punpongsanon, P.; Novy, D.; Maes, P.; Sra, M. AI-generated characters for supporting personalized learning and well-being. *Nat. Mach. Intell.* **2021**, *3*, 1013–1022. [CrossRef]
19. Falloon, G. Using avatars and virtual environments in learning: What do they have to offer? *Br. J. Educ. Technol.* **2010**, *41*, 108–122. [CrossRef]
20. Ratan, R.; Klein, M.S.; Ucha, C.R.; Cherchiglia, L.L. Avatar customization orientation and undergraduate-course outcomes: Actual-self avatars are better than ideal-self and future-self avatars. *Comput. Educ.* **2022**, *191*, 104643. [CrossRef]
21. Dimari, A.; Tyagi, N.; Davanageri, M.; Kukreti, R.; Yadav, R.; Dimari, H. AI-Based Automated Grading Systems for Open Book Examination System: Implications for Assessment in Higher Education. In Proceedings of the International Conference on Knowledge Engineering and Communication Systems (ICKECS), Chikkaballapur, India, 18–19 April 2024; pp. 1–7.
22. Gandomi, A.; Haider, M. Beyond the hype: Big data concepts, methods, and analytics. *Int. J. Inf. Manag.* **2015**, *35*, 137–144. [CrossRef]
23. Zawacki-Richter, O.; Marín, V.I.; Bond, M.; Gouverneur, F. Systematic review of research on artificial intelligence applications in higher education—Where are the educators? *Int. J. Educ. Technol. High. Educ.* **2019**, *16*, 39. [CrossRef]
24. Davidovitch, N.; Cohen, E. Administrative roles in academia—Potential clash with research output and teaching quality? *Cogent Educ.* **2024**, *11*, 2357914. [CrossRef]
25. Chan, C.K.Y.; Lee, K.K.W. The AI generation gap: Are Gen Z students more interested in adopting generative AI such as ChatGPT in teaching and learning than their Gen X and Millennial Generation teachers? *arXiv* **2023**, arXiv:2305.02878. [CrossRef]
26. Dehouche, N.; Dehouche, K. What's in a text-to-image prompt? The potential of stable diffusion in visual arts education. *Heliyon* **2023**, *9*, e16757. [CrossRef] [PubMed]
27. Sullivan, M.; Kelly, A.; McLaughlan, P. ChatGPT in higher education: Considerations for academic integrity and student learning. *J. Appl. Learn. Teach.* **2023**, *6*, 1–10.
28. Dwivedi, Y.K.; Kshetri, N.; Hughes, L.; Slade, E.L.; Jeyaraj, A.; Kar, A.K.; Baabdullah, A.M.; Koohang, A.; Raghavan, V.; Ahuja, M.; et al. So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *Int. J. Inf. Manag.* **2023**, *71*, 102642. [CrossRef]
29. Farrelly, T.; Baker, N. Generative artificial intelligence: Implications and considerations for higher education practice. *Educ. Sci.* **2023**, *13*, 1109. [CrossRef]
30. Warschauer, M.; Tseng, W.; Yim, S.; Webster, T.; Jacob, S.; Du, Q.; Tate, T. The affordances and contradictions of AI-generated text for writers of English as a second or foreign language. *J. Second. Lang. Writ.* **2023**, *62*, 101071. [CrossRef]
31. Darda, K.M.; Cross, E.S. The computer, a choreographer? Aesthetic responses to randomly-generated dance choreography by a computer. *Heliyon* **2023**, *9*, e12750. [CrossRef]
32. Neef, N.E.; Zabel, S.; Papoli, M.; Otto, S. Drawing the full picture on diverging findings: Adjusting the view on the perception of art created by artificial intelligence. *AI Soc.* **2024**. [CrossRef]
33. Chamberlain, R.; Mullin, C.R.; Scheerlinck, B.; Wagemans, J. Putting the art in artificial: Aesthetic responses to computer-generated art. *Psychol. Aesthet. Creat. Arts* **2018**, *12*, 177–192. [CrossRef]
34. Samo, A.; Highhouse, S. Artificial intelligence and art: Identifying the aesthetic judgment factors that distinguish human- and machine-generated artwork. *Psychol. Aesthet. Creat. Arts* **2023**. Advance online publication. [CrossRef]
35. Bellaiche, L.; Shahi, R.; Turpin, M.H.; Ragnhildstveit, A.; Sprockett, S.; Barr, N.; Christensen, A.; Seli, P. Humans versus AI: Whether and why we prefer human-created compared to AI-created artwork. *Cogn. Res. Princ. Implic.* **2023**, *8*, 42. [CrossRef] [PubMed]
36. Mikalonyté, E.S.; Kneer, M. Can artificial intelligence make art? Folk intuitions as to whether AI-driven robots can be viewed as artists and produce art. *ACM Trans. Hum. Robot Interact.* **2021**, *11*, 43. [CrossRef]
37. Hertzmann, A. Computers do not make art, people do. *Commun. ACM* **2020**, *63*, 45–48. [CrossRef]
38. Epstein, Z.; Hertzmann, A.; Akten, M.; Farid, H.; Fjeld, J.; Frank, M.R.; Groh, M.; Herman, L.; Leach, N.; Mahari, R.; et al. Art and the science of generative AI. *Science* **2023**, *380*, 1110–1111. [CrossRef] [PubMed]
39. Grba, D. Deep Else: A Critical Framework for AI Art. *Digital* **2022**, *2*, 1–32. [CrossRef]
40. Mazzone, M.; Elgammal, A. Art, creativity, and the potential of artificial intelligence. *J. Arts* **2019**, *8*, 26. [CrossRef]
41. Zlatkov, D.; Ens, J.; Pasquier, P. Searching for human bias against AI-composed music. In *Artificial Intelligence in Music, Sound, Art and Design*; Johnson, C., Rodríguez-Fernández, N., Rebelo, S.M., Eds.; Lecture Notes in Computer Science; Springer Nature: Cham, Switzerland, 2023; Volume 13988, pp. 308–323. [CrossRef]

42. Millet, K.; Buehler, F.; Du, G.; Kokkoris, M.D. Defending humankind: Anthropocentric bias in the appreciation of AI art. *Comput. Hum. Behav.* **2023**, *143*, 107707. [CrossRef]
43. Sindermann, C.; Sha, P.; Zhou, M.; Wernicke, J.; Schmitt, H.S.; Li, M.; Sariyska, R.; Stavrou, M.; Becker, B.; Montag, C. Assessing the attitude towards artificial intelligence: Introduction of a short measure in German, Chinese, and English language. *KI Künstliche Intell.* **2021**, *35*, 109–118. [CrossRef]
44. Bergdahl, J.; Latikka, R.; Celuch, M.; Savolainen, I.; Soares Mantere, E.; Savela, N.; Oksanen, A. Self-determination and attitudes toward artificial intelligence: Cross-national and longitudinal perspectives. *Telemat. Inform.* **2023**, *82*, 102013. [CrossRef]
45. Choi, W.; Jang, S.; Kim, H.Y.; Lee, Y.; Lee, S.-G.; Lee, H.; Park, S. Developing an AI-based automated fashion design system: Reflecting the work process of fashion designers. *Fash. Text.* **2023**, *10*, 39. [CrossRef]
46. Zhang, C.; Yao, C.; Wu, J.; Lin, W.; Liu, L.; Yan, G.; Ying, F. StoryDrawer: A child–AI collaborative drawing system to support children's creative visual storytelling. In Proceedings of the CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April–5 May 2022.
47. Kulkarni, N.D.; Tupsakhare, P. Crafting effective prompts: Enhancing AI performance through structured input design. *J. Recent Trends Comput. Sci. Eng.* **2024**, *12*, 1–10. [CrossRef]
48. Inie, N.; Falk, J.; Tanimoto, S. Designing participatory AI: Creative professionals' worries and expectations about generative AI. In Proceedings of the Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), Hamburg, Germany, 23–28 April 2023; Association for Computing Machinery: New York, NY, USA, 2023. [CrossRef]
49. Adeoye, M.A.; Jimoh, H.A. Problem-solving skills among 21st-century learners toward creativity and innovation ideas. *Think. Ski. Creat. J.* **2023**, *6*, 52–58. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

computers

MDPI

Article

Learning Analytics to Guide Serious Game Development: A Case Study Using Articoding

Antonio Calvo-Morata *, Cristina Alonso-Fernández, Julio Santilario-Berthilier, Iván Martínez-Ortiz and Baltasar Fernández-Manjón *

Department of Software Engineering and Artificial Intelligence, Complutense University of Madrid, 28040 Madrid, Spain
* Correspondence: acmorata@ucm.es (A.C.-M.); balta@fdi.ucm.es (B.F.-M.)

Abstract: Serious games are powerful interactive environments that provide more authentic experiences for learning or training different skills. However, developing effective serious games is complex, and a more systematic approach is needed to create better evidence-based games. Learning analytics—based on the analysis of collected in-game user interactions—can support game development and the players' learning process, providing assessment information to teachers, students, and other stakeholders. However, empirical studies applying and demonstrating the use of learning analytics in the context of serious games in real environments remain scarce. In this paper, we study the application of learning analytics throughout the whole lifecycle of a serious game, in order to assess the game's design and players' learning using a serious game that introduces basic programming concepts through a visual programming language. The game was played by N = 134 high school students in two 50-min sessions. During the game sessions, all player interactions were collected, including the time spent solving levels, their programming solutions, and the number of replays. We analyzed these interaction traces to gain insights that can facilitate teachers' use of serious games in their lessons and assessments, as well as guide developers in making possible improvements to the game. Among these insights, knowing which tasks students struggle with is critical for both teachers and game developers, and can also reveal game design issues. Among the results obtained through analysis of the interaction data, we found differences between boys and girls when playing. Girls play in a more reflexive way and, in terms of acceptance of the game, a higher percentage of girls had neutral opinions. We also found the most repeated errors, the level each player reached, and how long it took them to reach those levels. These data will help to make further improvements to the game's design, resulting in a more effective educational tool in the future. The process and results of this study can guide other researchers when applying learning analytics to evaluate and improve the educational design of serious games, as well as supporting teachers—both during and after the game activity—in applying an evidence-based assessment of the players based on the collected learning analytics.

Keywords: computational thinking; learning analytics; programming learning; serious games; visual programming

1. Introduction

Serious games are games designed for a main purpose other than pure entertainment, such as learning, raising awareness, or changing players' behaviors [1]. The efficacy of serious games has been proven in multiple fields, including medicine, science, economics,

and literature, and in varied contexts such as education or professional training [2–4]. STEM and computer science careers are on the rise, with strong demand in the job market. Serious games are often applied to increase interest in these disciplines (i.e., Science, Technology, Engineering, and Mathematics) by making learning more interactive and engaging for the broader public.

Moreover, despite the high demand and promising career opportunities, the gender representation in such careers remains highly imbalanced. According to UNESCO, in 2021, only 28% of engineering graduates and 40% of computer science graduates were female [5]. For instance, in Europe, women represent 41% of all science and engineering employment, and this percentage is even lower in computer science and tech roles [6,7]. With this scenario of high demand and gender imbalance, computer science is a field of interest [8], particularly considering the perceived potential of games to teach programming and make it more accessible and attractive to different audiences, such as school-aged girls.

The growing interest in introducing and improving computational thinking and programming knowledge among young people is evident in initiatives from the OECD's Programme for International Student Assessment (PISA). Computer science has been promoted as a new core subject, as reflected in the fact that the new PISA assessment for 2025 will include a new test called "Learning in the Digital World". This test will evaluate students' capacity to engage in an iterative process of knowledge-building and problem-solving using computational tools [9]. Based on the published information, it seems that this test will include questions to be solved using a visual block-based programming language (similar to Scratch). Therefore, schools will need to improve their coding teaching process, and teachers will require validated tools that help them teach programming and enhance the computational thinking skills of students.

There are several examples of games used to teach science (Physics Playground [10]), technology, chemistry (Minecraft Education [11]), engineering (SimCityEDU [12]), and mathematics (DragonBox [13,14]). As some studies have shown, serious games can also be effective tools for teaching programming [15], and playing games can attract new audiences to STEM careers [16]; however, for serious games to be effective and suitable in real-world classroom environments, significant barriers must be addressed. For instance, it is necessary to prove the effectiveness of each serious game and to provide teachers with tools to monitor players' learning. This includes understanding the learning process and identifying the issues and needs of students, both during and after gaming sessions.

This study explores the use of Game Learning Analytics (GLA) with a programming learning game. GLA involves the collection, analysis, and visualization of data gathered from players' interactions in a serious game. The purpose is to gain insights for multiple purposes (game validation and player assessment) and data that may be of interest to multiple stakeholders (game designers and developers, students, and teachers) [17]. Traditionally, the validation of educational games and their design were based on experiments including pre- and post-questionnaires or capturing videos of the game sessions, which then needed to be analyzed to identify key moments. With the emergence of e-learning, MOOCs, and the need to track many students, learning analytics techniques have begun to be used. These techniques can now also be applied to games (i.e., GLA) by collecting data on player interactions. The use of GLA allows for the analysis of game sessions in a faster and more effective way than video analysis, although it has the limitation that interactions outside the game in the real environment during the session are not captured. On the other hand, the large amount of data that can be collected on player interactions and the possibility of combining them with data from questionnaires allow for more advanced analysis to determine behavioral patterns or even develop predictive models.

In this paper, we discuss how we used GLA to identify the barriers and validate the game design of a serious game aimed at supporting teachers in programming classes. The game introduces basic programming concepts to teenagers and evaluates players' behavior and learning, helping to understand the key issues faced during gameplay. The main outcome of this validation approach is determining whether players have any issues in completing the game's levels or applying programming concepts, as well as gathering their feedback to improve future iterations of the game's design, which we intend to make openly and freely accessible.

The rest of this paper is organized as follows: Section 2 presents related work on programming learning platforms, educational games, and learning analytics; Section 3 describes the materials and methods used in this study, including the serious game designed to teach programming, the participants who played it, and how data were collected and analyzed; Section 4 presents the results of our study with respect to each of the research questions posed; Section 5 discusses the results; and Section 6 summarizes the main conclusions of our study.

2. Related Work

The current widespread demand for improving the learning of computational thinking and programming skills has fostered the development of many applications and online platforms to learn to program. There is a great variety of applications addressing different programming languages (and even only computational thinking concepts) covering different levels of knowledge and different age groups. For example, Codecademy and CodingGame are examples of complex platforms used to learn and practice languages such as C++, Java, or Python. We can also find virtual laboratories used in different fields of computer science, such as Java programming, cloud applications, or networks [18]. Furthermore, certain games, such as "Program your Robot" or "Mi superpoder es la programación", focus on improving computational thinking [19,20]. Although many games, platforms, and virtual laboratories to help with learning programming have been mentioned in the literature, most of them are not easily accessible or require a license payment [15,21].

At present, one of the most popular tools to learn programming in schools is Scratch, which allows students to develop their own games/projects by combining blocks. Scratch allows players to develop computational thinking skills and increases students' motivation, allowing them to quickly see the results of their code, thus driving their interest in continuing to learn about programming [22]. However, applying this open-ended tool in the classroom can be complex, as teachers must figure out how to adapt it to their needs. In particular, it can be complex to evaluate the knowledge acquired or to understand the problems that students run into while playing. Although initiatives such as Dr. Scratch [23] can provide some useful insights, open-ended tools are harder to deploy in classrooms than goal-oriented ones because, with sandbox-like games, it is up to students (or their teachers) to set goals, and the game cannot help to keep its players on target. More recent studies have also explored the use of serious games to teach advanced concepts such as parallel programming through a block-based visual language [24]. However, there remains a lack of free and open-source games for teachers to use in their classes when working on computational thinking and basic programming concepts. Specifically, games that provide sufficient insight for teachers to understand the players' progress, either during the game session (such that the teacher can act as a facilitator) or afterward (such that the teacher can reinforce some of the common mistakes or difficulties to improve the learning process) [25].

To provide evidence on the effectiveness of educational tools and study how users interact with them, Learning Analytics techniques can be applied to user interactions. Information gathered using these techniques can then help us to understand the learning

process, classify user behaviors, and evaluate learning increments during the use of the tool [26]. However, reviews such as [27] have mentioned the need for more data-based research studies, especially at the high school and industry training levels. One of the problems that exist when collecting large amounts of data is the work required to pre-process them, as well as the selection and extraction of significant variables.

Using a standard format to gather user interaction data not only simplifies data collection but also streamlines later analysis of the collected data. The Experience API for Serious Games (xAPI-SG) profile allows for the collection of GLA interactions from serious games using an xAPI-based vocabulary that represents the most common interactions present in these games [28]. Each player interaction is captured as an xAPI-SG trace (a so-called statement), usually represented in JSON. Each xAPI statement contains three main fields, describing an actor, a verb, and the object of the action. Additional information can be included as extensions, and statements typically include a timestamp to capture the moment when the action occurred. The xAPI-SG standard defines a set of commonly used verbs and their associated types, including 'interacted' and 'used' to capture interactions with game objects, game items, and other characters (NPC or enemy); 'initialized', 'progressed', and 'completed' to capture the start, progress in, and completion of the whole serious-game or significant part thereof (e.g., within a level); and 'accessed' to capture changes in the screen or area. The use of a standard that is widely used in the educational field, such as xAPI, to represent the data allows for interoperability with other educational systems, such as Learning Record Storages (LRSs) or joint analysis of the data with other educational activities, such as those in Learning Management Systems (LMS). Working with a standard data format such as xAPI simplifies the creation of open analytics systems for a series of video games without beginning from scratch every time. Therefore, xAPI can ensure reproducibility and simplify different analyses of the collected information. Moreover, with our xAPI framework, aspects such as data ownership and privacy are guaranteed (complying with EU GDPR data privacy regulation by pseudoanonymization of the captured data).

Despite the increasing presence of games applied in educational scenarios, further research needs to be carried out to provide evidence on how such games can be evaluated through user interactions, as well as studying what types of tools and information can—when provided to teachers using these games—facilitate and simplify their use in the classroom.

3. Materials and Methods

This study explores the application of GLA to data collected from an open-source serious game, Articoding, when it was used in educational environments. Articoding is a serious game designed to teach basic programming concepts. The collection of user interactions and the application of analytics in the game has two main goals: (1) to validate whether the game's design is adequate to meet its educational objectives, and (2) to verify the extent to which players learn and determine any difficulties that they may have. With these goals, we aim to provide a case study contributing to the field of studies and use cases of GLA, while also providing a serious game that serves as a tool to teach programming that has been studied with real users and is available in an open-source manner. The latter will allow other studies to compare Articoding with other tools and to make modifications, both in the game design itself and in the interactions collected, in order to study such changes.

For this study, two Articoding game sessions were carried out in a high school in Madrid. While playing the game, user interaction data were collected. The players also filled out a questionnaire before and after playing the game.

3.1. Research Questions

The following research questions were proposed to guide this study regarding the use of GLA applied to a serious programming game:

RQ1. Can the applicability of the game in the classroom be measured using GLA? Can problems in the game deployment be detected?

RQ2. Can GLA information help to evaluate and find the problems and limitations associated with the game's design?

RQ3. Can the learning and application of the expected programming concepts by players be measured?

RQ4. Can engagement be measured using the collected interaction data? Do players like the game?

Finally, we propose to discuss the usefulness of GLA, its advantages over traditional assessment methods, and how it can encourage and facilitate the application of games in the classroom.

3.2. The Game Articoding

Articoding is a serious game aimed at teaching basic programming concepts and promoting Computational Thinking in students between the ages of 12 and 16, who have little or no previous programming knowledge. In the game, players must overcome levels by solving problems posed in a board-shaped scenario. In each level, the player's goal is to guide one or more laser beams to their targets using visual block programming. Players can activate, move, or rotate lasers to avoid obstacles. The laser should stay on the board and there are mirrors that can be rotated to guide the laser beam. When all targets are reached by their lasers, the player can advance to the next level. Figure 1 shows an example level in Articoding: the left side contains code blocks that can be modified and tested (executed), where the results can be observed on the game board to the right. At the top of the screen, a "play" button allows players to test their programs; meanwhile, in the top-right, players can request hints (bulb icon), access block descriptions (stack-of-books icon), and exit the level.

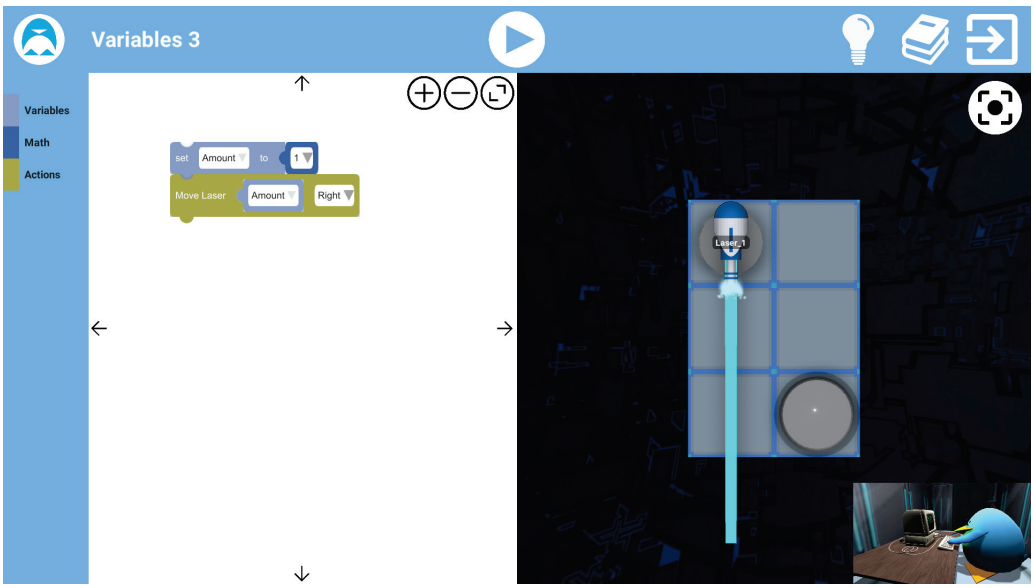


Figure 1. Screenshot of a level in the game Articoding.

The game's levels are grouped into five categories, which correspond to programming concepts to be learned through the game: Variables, Data Types, Basic Operators, Loops, and Conditionals. Each category has several levels to be completed; in particular, all

categories contain 7 levels, except for the Loops category which contains 9 levels. The categories must be played one after another in a fixed sequence, starting with Variables and ending with Conditionals. Within each category, the levels must also be played in their corresponding order. Figure 1 shows the third level in the first category (Variables).

A tutorial is provided in each new category or when a new element is introduced in the game, in order to show players how they are supposed to interact with it; for instance, Figure 2 shows an example tutorial explaining how in-game lasers work and how to activate them. Each game level also includes a hint system to help students who get stuck: a button shaped like a lightbulb is shown at the top-right of the level interface which, when pressed, draws arrows on the board indicating how board elements can be moved to solve the current puzzle. Up to a maximum of three different hints can be requested for each level.

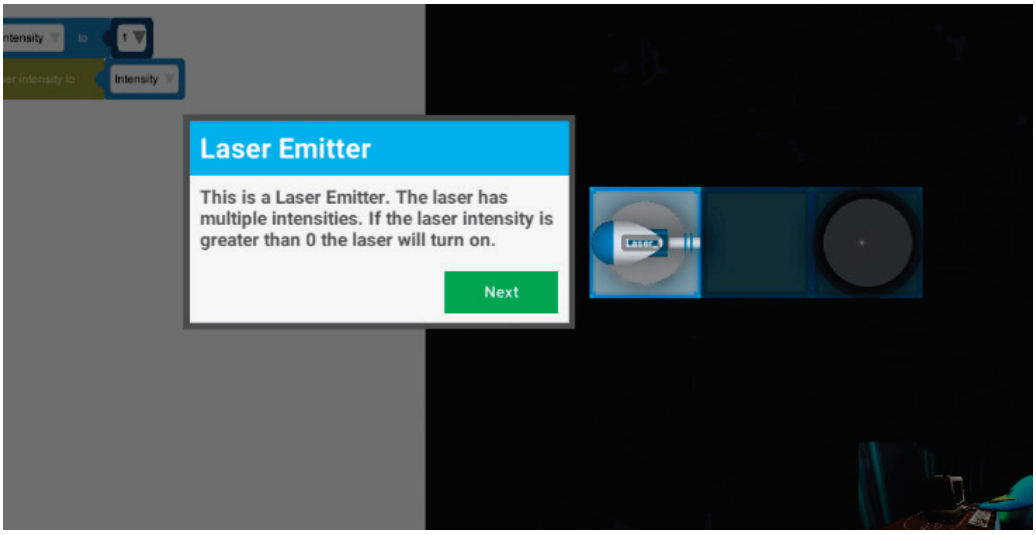


Figure 2. Partial screenshot of a tutorial example in Articoding.

For the learning process, in [29], programming problem solving was divided into four steps: (1) understand the problem; (2) determine how to solve the problem; (3) translate the solution into a computer language-based program; and (4) test and debug the program. However, students had difficulties expressing the solutions or developing instructions in the way that computers could execute them [30]; therefore, novice programmers need strategic problem-solving knowledge beyond syntactic knowledge [31].

Articoding allows players to work through the four steps in problem-solving: (1) understanding the objective of the level; (2) finding the steps that the different elements of the game must follow in a level to reach the objective state; (3) translating their solution idea to the proposed language (Blockly); and, finally, in case of failure, (4) identifying the error and retracing their steps to solve it. Moreover, the simplicity of blocky and its proximity to natural language allows us to abstract the syntactic characteristics of programming languages and facilitate the translation of the player's solution into concrete instructions for the machine using basic programming concepts. This approach—namely, through problem-solving and not only teaching a programming language—is what allows for the development of computational thinking, which authors identify in terms of the processes of decomposition, abstraction, algorithmic design, debugging, iteration, and generalization [32].

Finally, to further engage players, a reward system provides up to three stars depending on the player's performance when solving each level. This is designed to get students to think through their answers, reflect on their solutions, and seek to achieve optimal solutions in terms of board movements. Stars are awarded for:

- Completing a level without using hints.
- Completing a level on the first run.
- Completing a level with the minimum number of steps, in terms of actions that modify the state of the board (e.g., moving or rotating an element, or changing the status of a door).

For instance, in the scenario shown in Figure 3, the player has received the "no hints" star but spent one step more than necessary (hence the 1+ and greyed-out minimum-steps star), and failed to solve the level on their first try. Figure 4 shows the complete flowchart of Articoding screens.

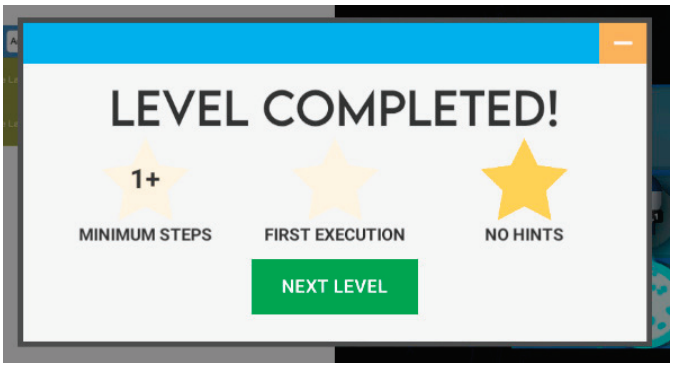


Figure 3. Partial screenshot of an example of stars obtained after solving a game level in Articoding.

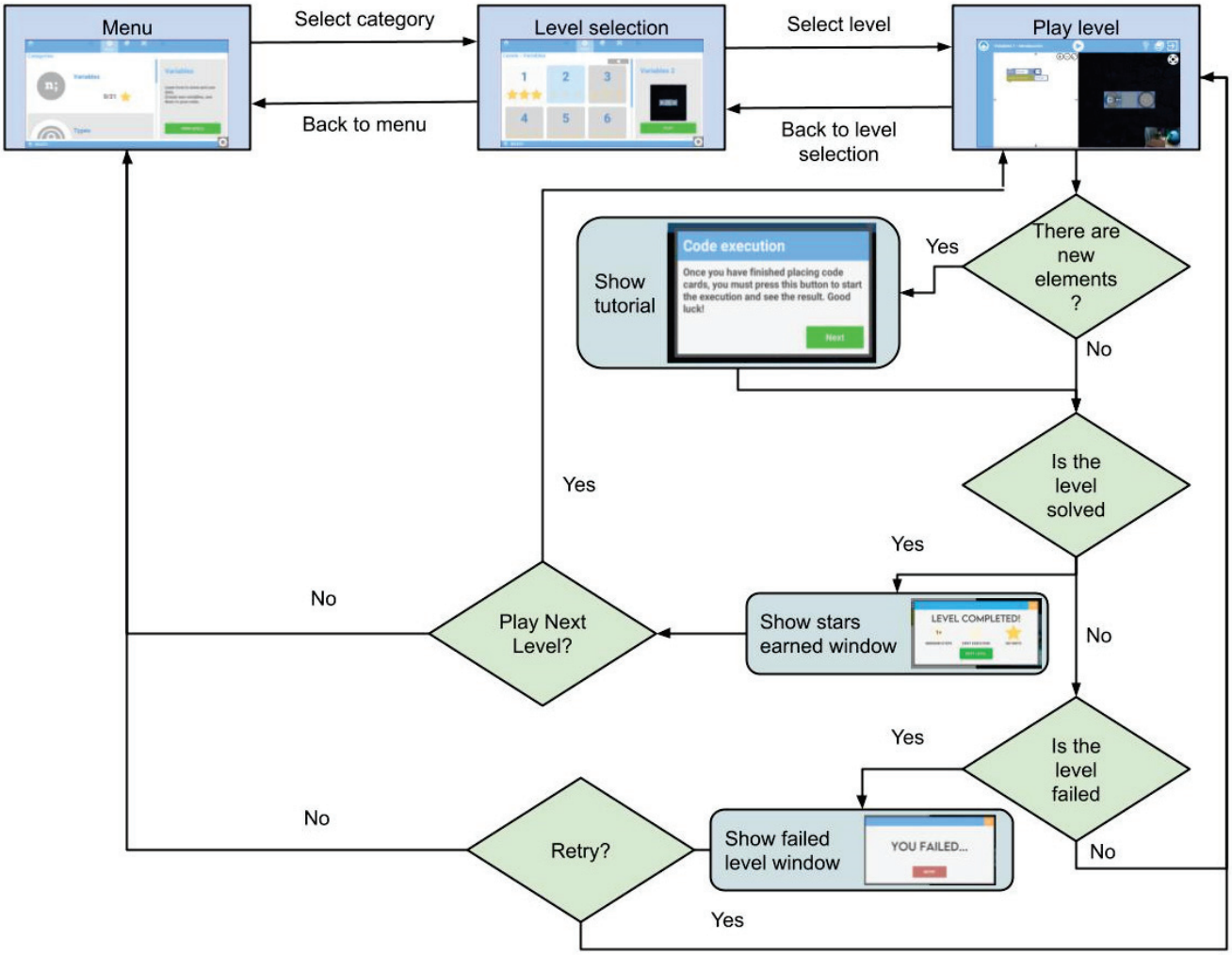


Figure 4. Articoding screens flowchart.

The interest in using this game to study the use of GLA is due to its complexity, the variety of interactions that can be performed by the user, the playtime, and its replayability. In addition, the game is open-source (Github repository: https://github.com/e-ucm/Articoding23-24; accessed on 20 March 2025), which will allow other researchers to contrast and use the game as well and address a topic of social interest.

3.3. Game Sessions and Participants

To put into practice the advantages of GLA applied to serious games, we applied it to Articoding with 134 students from a high school in Madrid, Spain. The study was performed as an educational activity in the school. This activity was accepted by the school principal, who oversaw describing the activity and its characteristics to the parents of the students. Both the students and the school were informed of how the data would be collected, used, and managed prior to the experiment.

Participants completed two 50-min sessions on different days. In addition to playing the game, students completed questionnaires to gather their (self-reported) previous programming knowledge and, after playing, their opinions about the game. The questionnaires included the following questions:

- Pre-test
  - Demographics: age and gender.
  - Use of videogames: number of days and hours that they play videogames, and platforms used to play them.
  - Use of videogames in class: whether they had previously used games in class.
  - Programming knowledge: previous knowledge about programming (e.g., with Scratch) and programming courses previously taken.
- Post-test
  - Six 5-point Likert questions about perceived ease of use.
  - One 5-point Likert question about perceived usefulness.
  - Three 5-point Likert questions about attitude and intention towards use.
  - Four free-text questions to provide any other opinions of the game. Questions ask about positive and negative characteristics of the game, learning perceptions, and the option of adding a level editor to the game.

Of the participants, all session data (pre-post questionnaires and interactions) were successfully captured for 114 students; 20 students either experienced technical problems or did not fill out the post-experiment questionnaire. Excluding 3 participants who chose to not report their gender, students were balanced by gender (47% female, 53% male). Figure 5 shows the distribution of age and gender of the final set of participants. Of the 114 valid students, 91 played two sessions and 23 played only one game session.

Regarding the previous experience with programming, just 15 students answered not know or had used Scratch or other languages, while 15 were not sure. On the other hand, 69 students indicated that they had attended some type of programming course or class; in particular, the school where the experiment was carried out teaches optional subjects in which Scratch is used.

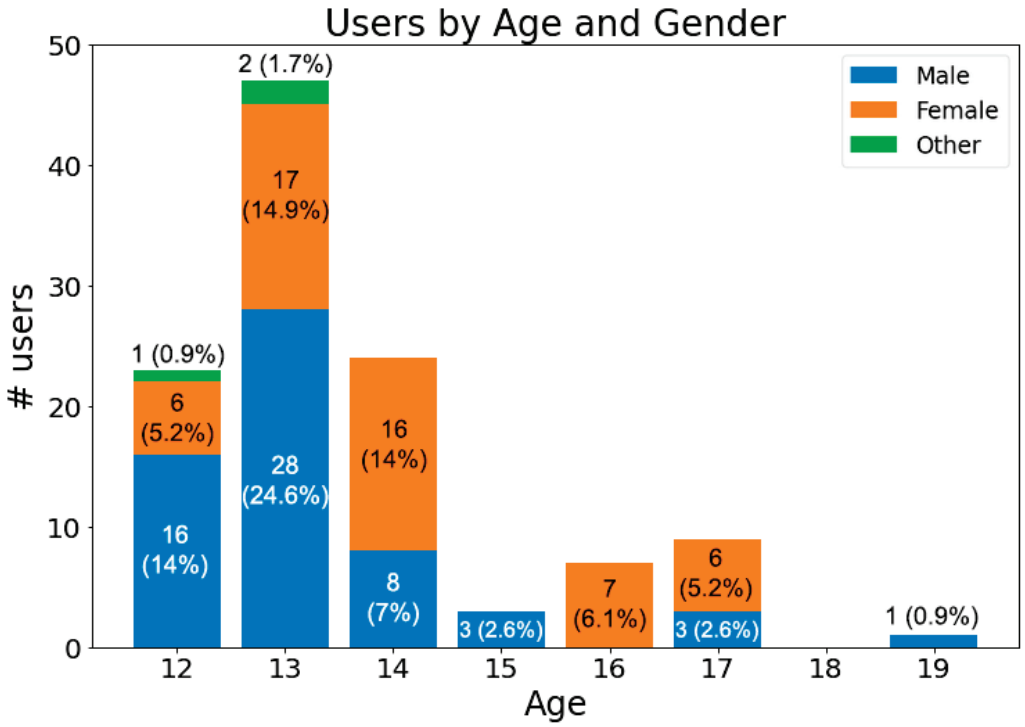


Figure 5. Age and gender distribution of participants in the study.

3.4. Interaction Data Collection

Articoding integrates an analytics system that allows player interactions to be captured and collected. It can be configured to send collected interactions either in real-time or after the game session is completed. The data collection process was based on the use of an xAPI-supported tracker and the analytics environment of the e-UCM research group [33,34]. The tracker collects user interaction data and sends that data to the server in real-time following the xAPI format. The analytics environment also simplifies the planning and running of experiments using serious games, handling—among other tasks—player authentication, pre- and post-game online questionnaire management, and allowing for easy retrieval of questionnaire results together with collected interaction data linked to each player. Data collection is pseudo-anonymized at the source by providing unique, random codes to each player, which allows questionnaires to be linked to analytics information without revealing the identities of players. This avoids storing identifiable information of students together with their responses, while still allowing (pseudonymous) use of analytics.

During the game sessions, a total of 98,704 unique traces were captured from all participants, which translates to more than 700 traces per player. Such traces included all interactions with game elements, together with any significant changes to each player's game state. The most meaningful trace types for our analyses are described in Table 1. These traces capture the start, progress, and completion of game levels, including whether the level was successfully completed, the number of steps needed to complete the game level, and the stars obtained when doing so (given by the three fields with the extensions minimum_steps, no_hints, and first_execution).

As an example, Figure 6 displays an xAPI-SG statement collected from Articoding representing that the actor (with anonymous identifier 'fybu') completed the game level 'types_6' (sixth level of the Data Types category) at the given timestamp. It was completed in 5 steps, obtaining all 3 possible stars, and therefore a (maximum) score of 3 stars was obtained.

Table 1. Most meaningful traces collected from Articoding.

Verb	Object ID	Additional Fields Used as xAPI Statements' Templates	Meaning
accessed	game level [1]	object.definition.type = level	Access to game level
initialized	videogame	object.definition.type = serious-game	Start of full game
	game level [1]	object.definition.type = level	Start of game level attempt
completed	game level [1]	object.definition.type = level result.success = true result.score (Number) result.extensions.articoding://minimum_steps (Bool) result.extensions.articoding://no_hints (Bool) result.extensions.articoding://first_execution (Bool) result.extensions.articoding://steps (Number)	Successful completion of game level
	game level [1]	object.definition.type = level result.success = false	Unsuccessful completion of game level

[1] Possible game level values correspond to levels in each level category. More specifically: variables_X: X in [1,7]; types_X: X in [1,7]; basic_operators_X: X in [1,7]; loops_X: X in [1,9]; conditionals_X: X in [1,7].

{
"actor": {
"name": "fybu",
"account": {
"name": "fybu",
"homePage": "https://analytics-api.es"
}
},
"result": {
"score": {
"raw": 3
},
"extensions": {
"articoding://minimum_steps": true,
"articoding://no_hint": true,
"articoding://first_execution": true,
"articoding://steps": 5
},
"success": true
},
"verb": {
"id": "http://adlnet.gov/expapi/verbs/completed"
},
"id": "5806c959-39bd-4688-9ec3-7d3d8ad8f219",
"object": {
"definition": {
"type": "https://w3id.org/xapi/seriousgames/activity-types/level"
},
"id": "https://analytics-api.es/activities/60c36c/types_6"
},
"timestamp": "2021-06-17T07:27:58.276Z"
}

Figure 6. xAPI-SG statement captured from Articoding representing a Data Type level completed with three stars.

All these collected data allowed us to apply GLA. Analyses can be applied in near real-time to the data being generated or at the end of the game sessions. With the collected data, we can also reproduce the players' interactions with the interface elements and their way of solving the game levels and using programming blocks.

4. Results

In this section, we present the results obtained from the data collected from players' interactions and questionnaires corresponding to each of the stated research questions.

4.1. RQ1. Can the Applicability of the Game in the Classroom Be Measured Using GLA? Can Problems in the Game Deployment Be Detected?

Through analyzing the collected interaction data, we can determine whether a player had not played or had played in several different sessions, as well as for how long. However, it is much more complex to understand why a player has not played the game or stopped playing before the session ends. For the latter, data external to the game itself are required.

Analyzing the student interaction data, 130 users played the game, and interaction data were obtained. There were 4 students without interaction data, for reasons unknown. Most likely, there were problems with the computer connection or they were playing with a colleague. Of the 130 students who had interaction data, 14 did not attend the second planned game session and, so, did not complete the final questionnaire; furthermore, one of the students started playing very late and only played for 13 min. Another 2 players left the session without completing the final questionnaires. As a result, 85% of the players completed the game sessions satisfactorily, carrying out the proposed activities.

The data revealed that most users were able to play without any problems associated with the operation or design of the game. Although some did not complete the final questionnaire, there were four players from whom we did not obtain any interaction data. Despite this, the game could generally be applied satisfactorily and the collection of interaction data allowed us to filter out those players who did not interact with the game during a large part of the session. However, we cannot know the reason why these players did not play or did not play long enough, due to these being associated with events external to the game, which were therefore not captured. Table 2 shows the number of participants who completed each phase of the study.

Table 2. Total of participants by level of activity completion and data collected.

Level of Completion	Number of Participants	Description
Total participants	134	The total number of students who played the video game.
Valid participants	114 (59 males, 52 females, 3 others)	The total number of students who played the video game, sent all questionnaire data, and for which interaction data were obtained.
Participants who played two sessions	91 (53 males, 35 females, 3 others)	The number of students who played two 50-min sessions.
Participants who played one session	23 (6 males, 17 females)	The number of students who played just one 50-min session.
Participants who completed the "Variables" category	106 (57 males, 46 females, 3 others)	The total number of students who completed all levels for the first category ("Variables").
Participants who completed the "Types" category	26 (15 males, 10 females, 1 other)	The total number of students who completed all levels for the second category ("Data Types").

4.2. RQ2. Can GLA Information Help to Evaluate and Find the Problems and Limitations Associated with the Game's Design?

With the interaction data, it is possible to know exactly how many levels each player completed, as well as how much time and attempts they needed. It is also possible to analyze the types of interactions that were performed with the different code blocks and what solution was reached for each level. If all players spend too much time on a level (or do not even complete it), it indicates problems in the design of the level. However, if this happens for only one player, it may indicate that this student has problems with the concept of the level. On the other hand, too much variation in the number of attempts and time needed by all players between levels will indicate that the growth in difficulty is not adequate.

A total of 91 players finished two 50-min sessions and completed 12 levels on average, with a minimum number of 5 completed levels, and a maximum of 18 (23 students only played a single session and, therefore, are not counted in this respect). Players completed each level within an average of 6.4 min. In the two sessions, only 26 users were able to fully complete the first two categories: Variables and Data Types. No user was able to complete the third category, Basic Operators. The interactions show that no further progress was made due to a lack of time, and not because there were problems with the game levels.

Regarding the progress of the levels and their completion, the 53 male players that finished two 50-min sessions were able to complete 12 levels on average (minimum of 5, maximum of 18). In comparison, the 35 female players who finished two 50-min sessions completed 11 levels on average (minimum of 6, maximum of 16). The completion of levels followed a normal distribution, and the t-test showed no statistically significant difference between the number of levels completed by male and female players ($p = 0.29$).

Furthermore, the *t*-test indicated no significant differences between genders in playing time per level. The 53 male players who finished two 50-min sessions completed each level within an average of 6.3 min, while the 35 female players needed 6.5 min per level on average. Table 3 shows the time needed (in minutes) to complete the Variables and Data Types categories according to the different players' gender and course.

Table 3. Time (in minutes) to complete the Variables and Data Types categories by gender and course of players.

Category	Player Characteristics		Number of Users	Mean	Standard Deviation
Variables	Gender	Male	57	24.36	7.34
		Female	46	25.35	8.30
	Course	First Year	49	25.52	8.24
		Second Year	44	24.88	7.48
		High School	13	21.55	5.59
Data Types	Gender	Male	15	39.57	7.24
		Female	10	47.43	11.49
	Course	First Year	11	42.80	5.65
		Second Year	14	44.15	10.35
		High School	1	19.10	-

To study the design of the game levels and determine whether they have the intended difficulty, the use of hints, the number of level retries, and the time needed to complete the game levels were analyzed. Too many players using hints or needing too many attempts at a single level may indicate that a level is more difficult than expected at that point in the game. Studying the time spent by players on each level is also important, as placing many

levels that take players a long time together can cause them to become frustrated and lose the feeling of progress. It is important, as a matter of good game design, to keep the time needed to complete each level relatively short, even as the difficulty increases progressively and new concepts are introduced.

To determine whether the players had problems in completing certain levels, their use of hints was analyzed. Table 3 shows the number of clues used in total by players in each level, for the Variables category. The data indicate a roughly linear increase in the use of hints, except for levels 4 and 5. For the Data Types category, few hints were used overall, except for levels 2 and 3, which needed a high number.

Additionally, Table 4 shows the number of level retries by players required to complete each level, explicitly stating the mean, standard deviation, and maximum number of retries. For the Variables category, a higher number of retries was seen in level 3, while levels 2 and 3 of the Data Types category concentrated the highest number of retries in that category.

Table 4. Total number of hints used by players in each level per category.

Category	Level						
	1	**2**	**3**	**4**	**5**	**6**	**7**
Variables	0	0	6	11	7	14	16
Data Types	0	14	16	2	0	4	0

The data presented in Tables 4–6 indicate that the second and third levels of the Data Types category seem to be more complicated than should be expected at that stage of the game. The data also verified that the number of attempts and time required to complete a level is correlated with the number of blocks needed to obtain its solution, and is also correlated with the number of blocks provided as part of the solution to the user at the beginning of the level. It will be interesting to study how the use of loops affects this fact, as they allow players to make more movements with fewer code blocks.

Table 5. Average number of level retries by players for each level, per category, with standard deviation and maximum.

Category	Level						
	1	**2**	**3**	**4**	**5**	**6**	**7**
Variables	AVG = 1.06 SD = 0.33 max = 4	AVG = 2.23 SD = 1.62 max = 9	AVG = 3.37 SD = 3.13 max = 24	AVG = 2.77 SD = 2.04 max = 13	AVG = 4.21 SD = 2.69 max = 14	AVG = 5.08 SD = 5.01 max = 24	AVG = 2.90 SD = 2.18 max = 13
Data Types	AVG = 1.27 SD = 0.92 max = 6	AVG = 6.22 SD = 5.28 max = 28	AVG = 5.79 SD = 5.96 max = 36	AVG = 3.27 SD = 2.52 max = 15	AVG = 3.15 SD = 1.91 max = 9	AVG = 3.64 SD = 3.12 max = 15	AVG = 2.85 SD = 3.61 max = 19

Table 6. Average time (in minutes) to complete each level for the first time per category.

Category	Level						
	1	**2**	**3**	**4**	**5**	**6**	**7**
Variables	AVG = 2.04 SD = 1.06	AVG = 2.14 SD = 1.20	AVG = 2.83 SD = 2.30	AVG = 4.10 SD = 2.52	AVG = 4.22 SD = 1.92	AVG = 6.32 SD = 4.43	AVG = 3.64 SD = 2.05
Data Types	AVG = 0.91 SD = 0.95	AVG = 8.69 SD = 6.19	AVG = 10.10 SD = 5.28	AVG = 6.78 SD = 2.96	AVG = 6.07 SD = 2.46	AVG = 9.83 SD = 3.74	AVG = 6.98 SD = 3.27

Two of the questions in the post-test further assessed how players perceived the difficulty of levels. The statements of these Likert questions were "The game has frustrating levels" and "The difficulty grows appropriately". Figure 7 shows the participants' answers to these questions, where 5 indicates maximum agreement with each statement. While some pointed out that there were frustrating levels, most players considered that the difficulty in the game increases appropriately.

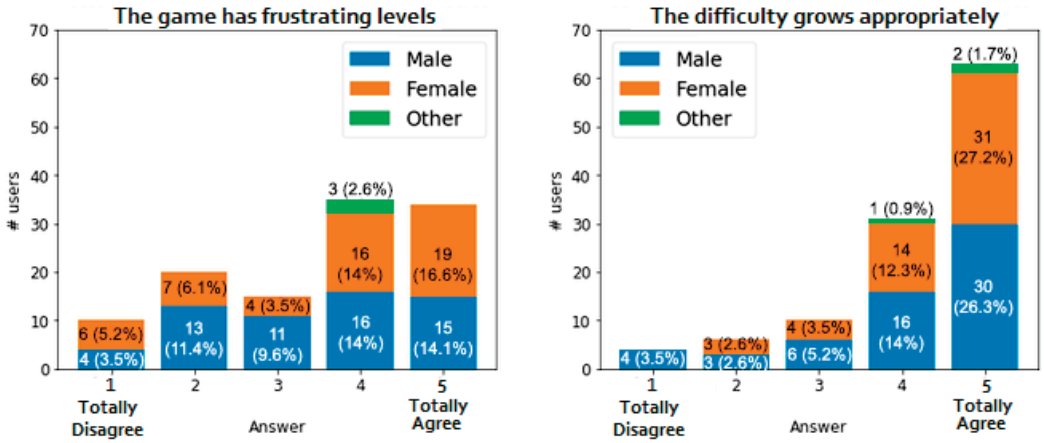


Figure 7. Participants' opinions on level design in Articoding levels.

Finally, no statistically significant differences were found between players who reported having attended programming classes and those who did not; however, the five players who completed the most levels did report having attended programming classes. The lack of further differences may be due to most of the student's prior knowledge of Scratch and the appropriate introduction to the game elements.

4.3. RQ3. Can the Learning and Application of the Expected Programming Concepts by Players Be Measured?

To determine the extent to which players learned programming with the game, the first step was defining a set of indirect measures of learning through the interactions with the game. This approach is commonly known as Stealth Assessment [10], where an evidence model needs to be initially defined to then be able to compare the information gathered from players' interactions and update the corresponding knowledge acquired. Another option is to use external questionnaires to assess knowledge before and after playing (e.g., Bebras or other instruments used to assess programming knowledge [35,36]); however, this increases the time required for the experimental sessions. In this scenario, players are considered to understand the concepts included in the game if they:

- Can progress in the game levels without having too much difficulty (i.e., making a low number of attempts).
- Do not spend too much time without knowing what to do in the game (i.e., without long periods of inactivity).

If the contrary situation occurs—that is, if a player is frequently inactive for long periods, or needs many attempts to complete the game's levels—the player is considered to have not successfully understood the concepts introduced in the game. Particularly, this study considers that players experience difficulties in a game level if they fail more than 3 attempts to complete the level; and players are considered inactive in a level when their inactivity time is an outlier in the distribution, considering such values as being above $Q3 + 1.5*IQR$. When a player has difficulties on 3 levels or is inactive on 3 levels of the

same category, the player is considered to have not adequately learned the concepts and/or not understand what needs to be done in the game.

For the game levels related to the Variables category, there were 22 players with repeated difficulties (difficulty in more than 3 levels in the category) and 19 players with repeated inactivity (more than 3 levels in the category). Overall, 8 players satisfied both conditions (difficulties and inactivity), in the Variables category. For the Data Types category, there are 17 players with repeated difficulties and 10 players with repeated inactivity, while only 2 players satisfied both conditions (difficulties and inactivity) in this category.

The analysis also involved checking that the players who had problems learning the programming concepts included in the game (i.e., those who repeated levels too many times and were inactive for too long) stated in the post-test that they did not want to keep playing the game.

During the experiments, some players reported issues understanding the concept of "variable". These perceived issues during the experiments were contrasted with the collected data. This was clear with respect to some in-game behaviors; for instance, players declared variables only in the place where they needed to use their values, which may mean that they did not understand that they could be declared at the beginning of code and used at any time. There also seemed to be problems with understanding that a variable can be used for multiple purposes once it stores a value, as students sometimes created two variables with the same value; for example, to rotate and move the laser. The frequency with which the following behaviors occurred is shown in Table 7.

Table 7. Number of players with issues using variables in completed levels in the Variables and Data Types categories.

Level	Issue *								
	VV	**OC**	**AC**	**NI**	**TV**	**SS**	**NV**	**SM**	**CC**
Variables_1	1	1	1	0	0	0	0	0	0
Variables_2	2	33	33	2	0	0	29	0	18
Variables_3	3	6	6	5	0	0	2	0	0
Variables_4	1	0	0	1	0	0	1	1	0
Variables_5	4	66	10	4	0	51	58	3	52
Variables_6	2	40	18	3	0	30	34	2	37
Variables_7	3	59	18	5	2	2	5	2	57
Data Types_1	0	1	1	1	0	1	0	0	0
Data Types_2	3	67	7	4	2	4	3	1	61
Data Types_3	0	24	0	1	7	1	1	0	42
Data Types_4	0	6	0	5	5	0	1	1	25
Data Types_5	0	26	0	0	4	0	1	0	1
Data Types_6	2	24	0	2	2	3	1	2	0
Data Types_7	0	7	0	0	0	1	5	0	9

* Legend of issues found: VV—Player initialized a variable with itself, such as "move = move". OC—Player used a constant block instead of a variable at least once. AC—Player used only constant blocks instead of variables. NI—Player created a variable without assigning a value. TV—Player created more variables than necessary. SS—Player always initialized variables just before using them. NV—Player initialized a variable that was never used. SM—Player initialized more than one variable with the same value. CC—Player used the same constant value instead of creating a variable.

4.4. RQ4. Can Engagement Be Measured Using the Collected Interaction Data? Do Players Like the Game?

Besides the fact that the game successfully introduces players to basic programming concepts, we are also interested in maintaining player engagement, in order to ensure that they want to continue improving, attempting to complete the levels without hints or errors and in as few steps as possible. To measure the extent to which the game engages players, two aspects were analyzed:

- Players try to get as many stars as possible, retrying levels if necessary.
- Players do not remain idle for large amounts of time and continue to play throughout the session.

Player inactivity is also related to player learning (RQ3). A boring game that does not engage players will make them lose interest in advancing as fast as possible to find new challenges, causing them to not pay attention and miss out on learning or trying out new concepts. On the other hand, a player who does not learn the necessary programming concepts to continue advancing or does not understand the game will lose interest in the video game itself, thus performing fewer and fewer interactions.

Comparing players with two completed sessions according to their gender, less time passed between interactions on average for male players than for female players (Figure 8). This time difference was statistically significant with a moderate effect size (Mann–Whitney: $p < 0.001$; r = 0.47), which may point to a different way of playing (be more cautious) that does not translate into an increase in the number of attempts or the number of levels completed (Table 8).

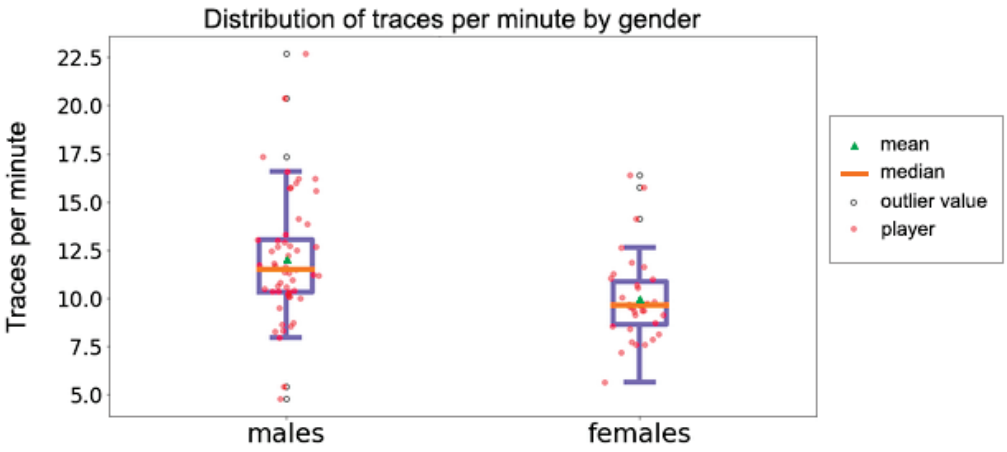


Figure 8. User interactions per minute by gender.

Table 8. Average traces and levels completed per minute by gender.

	Traces Per Minute	Levels Completed
Male	12.01; SD = 3.27	0.160; SD = 0.041
Female	9.94; SD = 2.23	0.154; SD = 0.027

Regarding the stars achieved, the number of players who achieved all three stars in a level decreased as the levels advanced in a category. Figures 9 and 10 show the number of players achieving 0, 1, 2, and 3 stars in each level of the Variables and Data Types categories, respectively. We can see, from Figure 9, that all players finished the first level with 3 stars (the maximum) while, in later levels, it was much rarer for players to get all 3 stars. It can be noted in Figure 10 that only a fraction of players finished the last level.

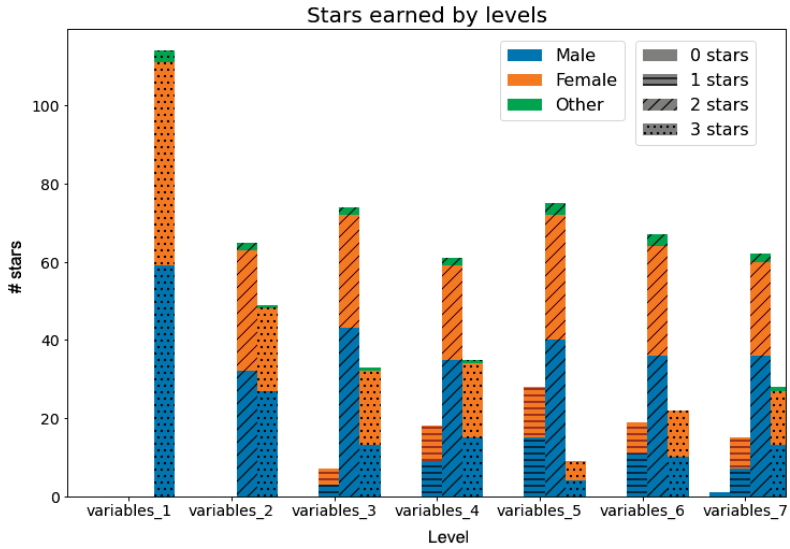


Figure 9. Awarded stars in "Variables" levels.

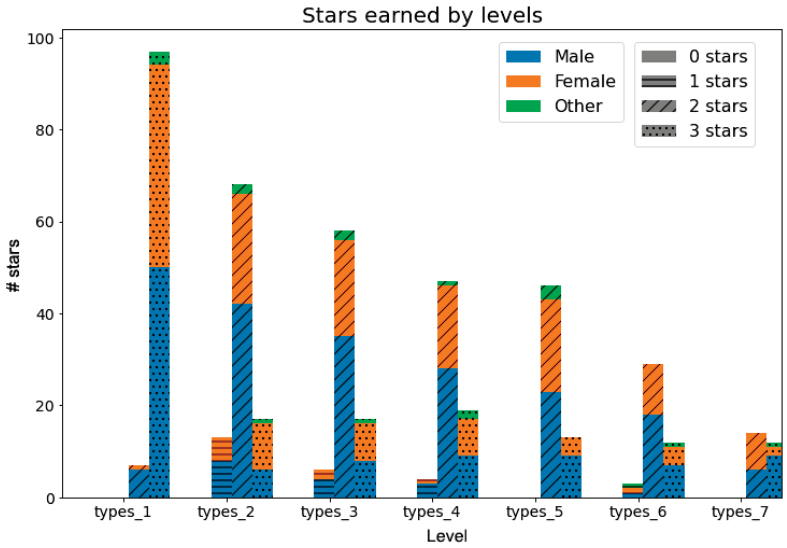


Figure 10. Awarded stars in "Data Types" levels.

Most players obtained 2 stars, avoiding using the clues in each level and trying to make as few moves as possible. The most difficult star to earn is the one for completing a level on the first attempt. As for the repetition of levels, 16 players replayed levels to get more stars, and 5 of those did it in multiple levels with the aim of getting 3 stars. Notably, 4 of the 16 players who repeated a level did not manage to improve. In addition, almost all retries occurred during the levels of the initial Variables category. The delineation by gender is additionally depicted in Figures 9 and 10, demonstrating no significant difference in the number of stars achieved between male and female players.

If a player is committed to the game, does not get bored, is active, and progresses, we can conclude that they like the game. In this case, the players were also asked for their opinions through a post-game questionnaire. Figure 11 displays the results of the post-test question "Would you like to continue playing the game?", which indicates that players generally liked the game (65% would continue playing). However, there seemed to be a slight gender difference, as boys liked the game more than girls (81% of boys vs. 46% of girls); furthermore, girls were more indifferent to the game (33% were indifferent vs. only 5% of boys). Additionally, most players found the game simple easy to use, entertaining, and interesting (Figure 12). In this case, no correlation was found between the

opinion given of the game in the questionnaire and the data on inactivity, number of levels completed, or stars earned. This may be due to various reasons such as the activity being mandatory, or because there was little data on players who did not like the game.

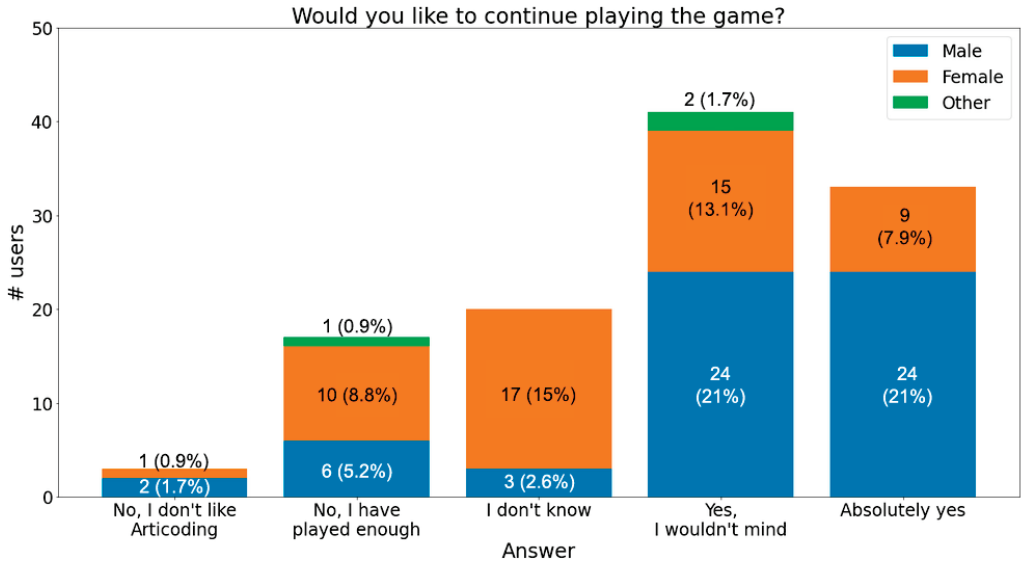


Figure 11. Opinions on willingness to continue to play by gender.

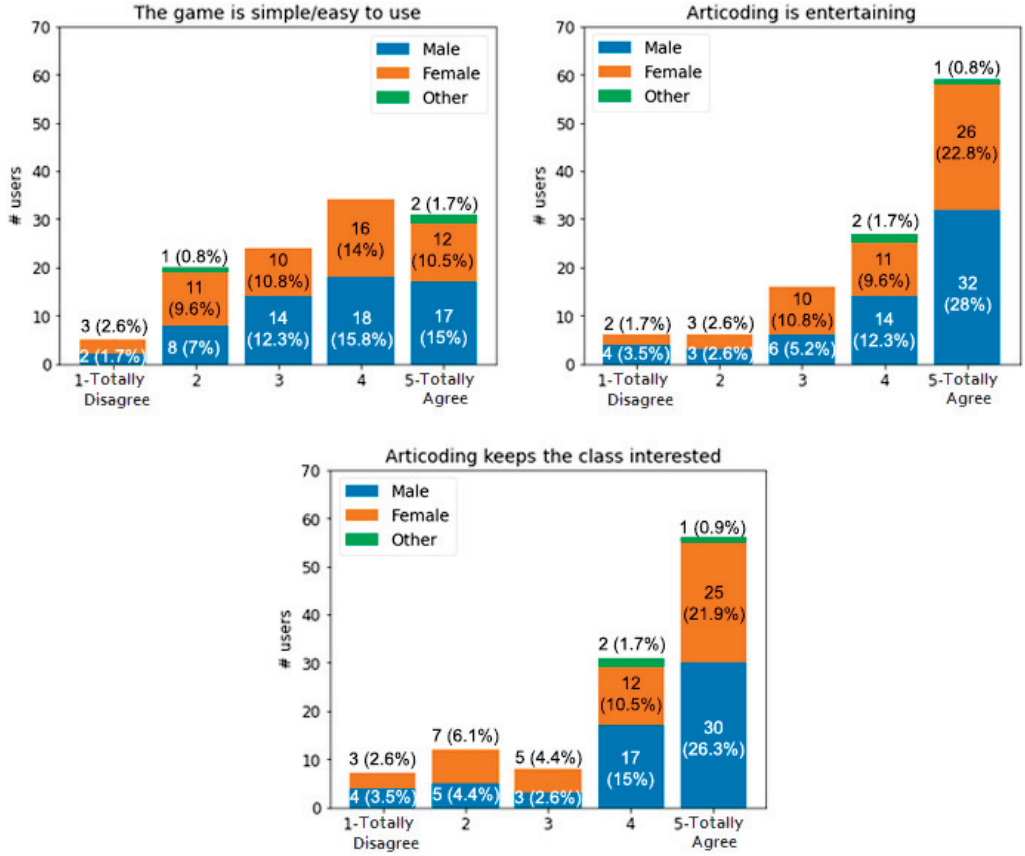


Figure 12. Opinions of players regarding whether the game is: easy to use, entertaining, and keeps the class interested.

5. Discussion

In this work, GLA allowed for an analysis of the behaviors of 114 students playing the serious game Articoding. Individual observation of each player in each game session

together with an interview would allow us to better understand the students' learning levels and any problems that they may have; however, this would not be a scalable process and its cost, in terms of time or human resources, would be too high to implement in a real scenario. GLA therefore allowed us to study the learning progress of many more players and their behavior at a lower cost (in terms of both time and resources), compared to traditional methods. In addition, GLA allowed us to obtain more accurate information and, with a larger sample of users and, consequently, data to compare, we could better study the effects of the game by type of user and detect students who stand out in the way they play.

5.1. Applicability of the Game

With the collected data, we could determine how long the players played and how long they were inactive. GLA allowed us to study the time required for each level by the player and, therefore, better adapt the game to the needs of the target audience and the context of use; in this case, the students and teachers and the use of the game in the classroom, respectively. Considering the original design of the game and the characteristics of the experiments carried out (duration and number of sessions) we, as game designers, hypothesized that players would be able to complete at least the first three categories. However, the data revealed that no player could complete the third category (Basic Operators) and only 25 users were able to complete the second category (Data Types). Based on these results, the application of the game in the classroom context—where time is typically limited—needs to be modified: either the game would need to be used in more sessions to cover all included concepts, or the game design would need to be modified to include fewer levels and/or simplify existing ones, allowing players to advance further in the game. In future experiments, we hope to be able to study the time required for players to complete the remaining categories (Basic Operators, Loops, and Conditionals). This will allow us to more precisely adjust the game's levels, as well as the time required to complete the full game. With such information, we aim to additionally create a guide to help teachers apply Articoding as a more effective educational tool in their courses.

Moreover, with the current data, it may be necessary to dedicate a 1 h session for each category. However, we also have to take into account that there can be large differences in completion time between different players, and features should be implemented such that the fastest players can remain engaged while their slower peers finish; two options include optional levels and increased replayability.

However, we obtained no information about those players for whom no interaction traces were received but who filled out the initial questionnaire as they were present. Did they not come to play because they were late for class? After filling out the questionnaire, did they start playing with a classmate? Did they have to leave class for some reason?

5.2. Game Issues and Limitations

As a game design issue, the students commented that the clues did not help much. Possibly due to the fact that a potential star was lost when players asked for hints, the hint system was rarely used. Therefore, this aspect may benefit from being revised, both to provide more useful help and as part of a deeper overhaul to re-design how and when stars are awarded in future game versions. The use of GLA can help to compare different versions of the game in the future; for example, by studying the effects of giving more specific hints or providing guided levels. This may yield interesting results, as researchers generally do not agree about the effects of providing guidance to players [37].

Data and GLA pointed out another design issue regarding the difficulty of levels 2 and 3 in the Data Type category, as players had more trouble with these levels than in the final levels of this category, using more clues and requiring more attempts and time to solve them. As such, these two levels need to be simplified or moved to the later stages of the category, after simpler levels have been successfully solved. Finally, the analysis revealed that not many players replayed the levels to try to improve. It will be necessary to study whether this is because completing the levels with all the stars does not motivate players or if, due to the time limits, players prefer to advance further rather than get more stars. These issues were not reported by players, and many of the students were satisfied with the level of difficulty and their progress.

5.3. Student Learning

Measuring learning when playing a videogame can be complex, as we do not know the extent to which players can later apply the knowledge acquired in-game to real scenarios, unless the game design itself is focused on measuring that learning via in-game assessment [38] or stealth assessment [39]. In our case, the variables chosen to measure how the player improves and understands the concepts presented in the game were inactivity and difficulty (attempts) in completing the levels. However, these measures are indirect and non-standard; therefore, in subsequent experimental research, it may be necessary to add external tools to measure the knowledge gained by players during game sessions in order to validate the effectiveness of the game as a learning tool. For example, we could add a validated post-test to assess their learning. In such a case, it would also be interesting to compare the learning results determined through that external measure with those obtained via our indirect measures. It should not be forgotten that this game is intended to be a complementary and motivational tool that teachers can use to learn and practice basic programming concepts and to work on improving computational thinking; it is not intended to be a one-stop, self-contained programming tutor.

It is also necessary to emphasize that analytics may not capture all aspects of the game session. In this experiment, all players advanced through the levels at differing rates, and they had the game open during the game sessions. However, Learning Analytics does not capture cases where the teachers or researchers in charge of the session stepped in to address their doubts or help players complete a game level. Neither can we know exactly what happened to those players who appeared inactive: did they go to the bathroom? Did they interact with their classmates? There are studies that have shown, for example, how interactions with teachers or peers can be correlated with knowledge and computational thinking skills [40]. The process and analysis can be improved by adding more data sources, allowing for the application of multi-modal learning analytics. In this case, we complemented the interaction data with data collected through questionnaires; however, a multi-modal analytics approach would allow us to integrate even more data sources.

5.4. Engagement

The designed game maintained the interest of most of its users—at least in the sense of keeping idle times low—showcasing the engagement typically related to game-based learning and gamification [41]. This engagement was further contrasted with a question in the post-test, which indicated that 65% of the players would continue playing while only 17.5% would not play anymore. The other 17.5% did not know whether they would continue to play if they had the possibility.

For those players who were not engaged in the game, additional research would be needed to determine if they were not engaged due to the game mechanics and the game design, or if they did not understand the game or considered it to be too difficult.

However, the game was overall very well received by players, with most players enjoying the game (75.5%) and finding it easy to use (57%). These results align with previous research demonstrating the interest of students in using videogame in class in general, and in learning programming concepts in particular [42]. The advantages of using games to increase students' motivation before they have taken any computer science or programming courses have also been highlighted [43,44].

However, many players also pointed out that the game had levels that could be too frustrating (60.5%). As these levels may cause some players to give up or lose interest in the game, it is important to address this issue by revising those difficult levels and making the difficulty reasonably progressive in future versions of the game.

6. Conclusions

The use of Learning Analytics in the context of educational games has many benefits when attempting to understand the learning of players and detect the problems they encounter, ultimately allowing for the improvement of a game and its educational design. GLA can be oriented to both a game's design and development phase and its actual application in real environments, reducing the cost of obtaining relevant information regarding the use of the game.

In this study, GLA helped us to verify that the serious game Articoding is a useful tool to introduce programming concepts in an entertaining and interesting way to students. The use of analytics provided evidence that helped us to identify several educational issues and game design problems that still need to be addressed in future versions of the game. It also allowed us to identify problems associated with how players understand programming concepts such as variables. Some of these analytics could be further applied in (near) real-time, keeping teachers informed of the players' progress while the games are being played in their classrooms. This would allow teachers to better understand the concepts that their students find challenging, identify students who are struggling and need help, report on their progress, or maintain class control by highlighting those players who are inactive or interact little with the game.

Collected analytics information provides abundant evidence to drive an education-focused re-design of the game. Considering all these observations, we plan to develop a new version of the game that better explains the concept of variables, as well as re-designing several levels and improving the star system and the hints provided for each level. For example, evidence suggests that the in-game hints system is not useful, as players make very little use of them. In terms of levels, the data analysis showed a pronounced variation in difficulty between some of the levels, breaking the ideally increasing difficulty curve.

The Articoding game itself, as a free software project, is also the result of this work. Its open-source nature allows other researchers to easily build on this work and modify it. In addition, the use of the xAPI standard for data collection allows other researchers to create their own tools and data analysis platforms for Articoding, or to integrate it with existing tools that implement this standard.

With respect to the analysis, we also want to design more precise learning measures; for instance, establishing a measure of level difficulty, and comparing the improvement of players in levels that are equivalent in difficulty (i.e., levels that do not introduce new concepts and start with the same number of blocks provided).

An issue present in our results is that there was information that was not captured by the interaction data, which may be required to complement the analytics. Events outside the game, such as interactions between players or times when teachers or researchers helped players, directly affect their results and should be considered in the analysis. This highlights the importance of taking notes during the game sessions or, even better, creating

an easy-to-use system to register these external events, especially when experiments are conducted in real classroom environments (as was the case in this study). This aspect should be reflected in the experimental design and should maintain the current pseudo-anonymization of the data (i.e., using anonymous codes instead of actual student identities). The collection of external data would allow us to determine whether the teacher helped a student during the session, or if the students interacted with each other. In this case, in order to minimize this type of interaction, players were not allowed to interact with their peers, and teachers were asked not to intervene unless the student asked for help and had stayed at the same level for more than 10 min. The collection of observational data also presents problems, as it is costly and complex for researchers, and is hardly scalable. We estimate that, for a typical class of 30 students (one per computer), it would take at least 3 people to record external events that may be interesting for the study; this number is influenced by the position of the computers and size of the room, as such experiments are generally conducted in uncontrolled actual environments.

The results also indicated some gender-related differences when it came to the interactions with the game and interest in continuing to play. Behavioral differences may be related to different learning approaches: boys seem to interact in a more exploratory way (making more interactions per minute) while girls seem to reflect more, taking more time to interact. These different behaviors have also been observed in other games, where boys tended to show a more active play mode [39]. This difference may be due to the affinity of players towards videogames and the number of hours they play daily as, as observed from the pre-test data, the use of videogames was more widespread among boys. Future work should focus on making the game more attractive to girls, in order to improve the results obtained in the survey on whether they would continue to play Articoding, with the ultimate aim of promoting the study of STEM among girls [40]. There is a need for studies that analyze the ways in which games are played and interacted with.

We believe that more research work is needed on GLA, providing assessments that other researchers can learn from. To date, emphasis has been placed on the sharing of analytical data; however, such sharing may face regulatory issues such as those associated with the European GDPR, as well as experimental designs. We believe that another way forward is to systematize the approach by attempting to utilize standards for trace capture (e.g., xAPI-SG) and to work on open-source tools—both for the tracker and the game itself—that other researchers can experiment with and learn from. It is also necessary to work on tools that help to visualize learning information during the use of games in class, particularly in a way that simple and useful for teachers.

Another aspect to be analyzed in future work is knowledge acquisition using the game and comparing it to other tools. For this evaluation, the game first needs to be in its final version, as any subsequent changes may invalidate previous evaluations. However, such an evaluation is not simple. On one hand, the game must be played in several sessions. It is not easy to find educational centers that are willing to allocate such an amount of time to the use of a tool before they are sure whether it is useful. Conducting these preliminary studies helps to gain the confidence of the schools. On the other hand, it is difficult to find validated instruments to study computational thinking and programming concepts. Due to these limitations, some studies use self-assessment or self-efficacy questionnaires without correlating the results to player behavior [45–47]. In future studies, an option would be to conduct a pre-post experiment with the tools that will be used in the new PISA25 test, and then conduct an analysis of how the game variables correlate with the results obtained through that test. Such studies would allow for the creation of tools that provide teachers with information about students' learning while using the game Articoding.

To conclude, although integrating GLA into a game and providing tools for the analysis and visualization of information is more expensive than simply using questionnaires or surveys, it brings great advantages in all phases of the game's lifecycle. Once the use of GLA is integrated into the game, it will be useful both in the evaluation and improvement phase of the game, as well as in its actual deployment in classes by teachers.

In particular, during the development and validation phases, the use of GLA allows us to detect design problems and gain a better understanding regarding how the game affects player behaviors. Meanwhile, during classroom use, it provides tools to teachers to better control their classes, identify students experiencing problems, and determine whether players are learning.

Author Contributions: Conceptualization, A.C.-M. and B.F.-M.; methodology, A.C.-M. and B.F.-M.; software, A.C.-M. and J.S.-B.; validation, A.C.-M., C.A.-F., and B.F.-M.; formal analysis, A.C.-M., C.A.-F., and J.S.-B.; investigation, A.C.-M. and B.F.-M.; resources, I.M.-O. and B.F.-M.; data curation, A.C.-M. and C.A.-F.; writing—original draft preparation, A.C.-M. and B.F.-M.; writing—review and editing, C.A.-F., J.S.-B., I.M.-O., B.F.-M., and A.C.-M.; visualization, A.C.-M.; supervision, I.M.-O. and B.F.-M.; project administration, I.M.-O. and B.F.-M.; funding acquisition, I.M.-O. and B.F.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially funded by the Ministry of Education (PID2020-119620RB-I00; PID2023-149341OB-I00) and by the Telefonica-Complutense Honorary Chair on Digital Education and Serious Games.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy and experimental design.

Acknowledgments: The authors would like to thank Dany Faouaz, Arturo García, and Álvaro Poyatos for programming the first version of Articoding.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Laamarti, F.; Eid, M.; El Saddik, A. An overview of serious games. *Int. J. Comput. Games Technol.* **2014**, *2014*, 358152. [CrossRef]
2. Connolly, T.M.; Boyle, E.A.; MacArthur, E.; Hainey, T.; Boyle, J.M. A systematic literature review of empirical evidence on computer games and serious games. *Comput. Educ.* **2012**, *59*, 661–686. [CrossRef]
3. Dörner, R.; Göbel, S.; Effelsberg, W.; Wiemeyer, J. *Serious Games*; Springer International Publishing: Cham, Switzerland, 2016. [CrossRef]
4. Wattanasoontorn, V.; Boada, I.; García, R.; Sbert, M. Serious games for health. *Entertain. Comput.* **2013**, *4*, 231–247. [CrossRef]
5. UNESCO. UNESCO Research Shows Women Career Scientists Still Face Gender Bias. Available online: https://en.unesco.org/news/unesco-research-shows-women-career-scientists-still-face-gender-bias (accessed on 14 March 2022).
6. Blumberg, S.; Krawina, M.; Mäkelä, E.; Soller, H. Women In Tech: The Best Bet to Solve Europe's Talent Shortage. McKinsey Digital. January 2023; pp. 1–10. Available online: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/women-in-tech-the-best-bet-to-solve-europes-talent-shortage#/ (accessed on 1 August 2024).
7. Eurostat. More Women Join Science and Engineering Ranks. Available online: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20220211-2 (accessed on 14 March 2022).
8. Schulte, C.; Sentance, S.; Sparmann, S.; Altin, R.; Friebroon-Yesharim, M.; Landman, M.; Rücker, M.T.; Satavlekar, S.; Siegel, A.; Tedre, M.; et al. Values and Beliefs Underpinning K-12 Computing Education. In Proceedings of the Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE, Milan, Italy, 8–10 July 2024; Association for Computing Machinery: New York, NY, USA, 2024; pp. 767–768. [CrossRef]
9. OECD. PISA 2025 Learning in the Digital World. Available online: https://www.oecd.org/en/topics/sub-issues/learning-in-the-digital-world/pisa-2025-learning-in-the-digital-world.html (accessed on 17 July 2024).
10. Shute, V.; Rahimi, S.; Smith, G. Game-Based Learning Analytics in Physics Playground. In *Data Analytics Approaches in Educational Games and Gamification Systems*; Springer: Singapore, 2019; pp. 69–93. [CrossRef]
11. Mojang, A.B.; TM Microsoft Corporation. Minecraft Education Edition. Available online: https://education.minecraft.net/ (accessed on 14 March 2022).

12. Electronic Arts. SimCityEDU: Pollution Challenge! Available online: https://www.ea.com/news/glasslab-launches-simcityedu?isLocalized=true (accessed on 9 June 2020).
13. Kahoot DragonBox. DragonBox. Available online: https://dragonbox.com/ (accessed on 14 March 2022).
14. Siew, N.M.; Geofrey, J.; Lee, B.N. Students' Algebraic Thinking and Attitudes Towards Algebra: The Effects of Game-Based Learning Using Dragonbox 12 + App. *Res. J. Math. Technol.* **2016**, *5*, 66–79.
15. Miljanovic, M.A.; Bradbury, J.S. A Review of Serious Games for Programming. In *Serious Games*; Göbel, S., Garcia-Agundez, A., Tregel, T., Ma, M., Hauge, J.B., Oliveira, M., Marsh, T., Caserman, P., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11243, pp. 204–216. [CrossRef]
16. Hosein, A. Girls' video gaming behaviour and undergraduate degree selection: A secondary data analysis approach. *Comput. Human. Behav.* **2019**, *91*, 226–235. [CrossRef]
17. Freire, M.; Serrano-Laguna, Á.; Iglesias, B.M.; Martínez-Ortiz, I.; Moreno-Ger, P.; Fernández-Manjón, B. Game Learning Analytics: Learning Analytics for Serious Games. In *Learning, Design, and Technology*; Springer International Publishing: Cham, Switzerland, 2016; pp. 1–29. [CrossRef]
18. Elmoazen, R.; Saqr, M.; Khalil, M.; Wasson, B. Learning analytics in virtual laboratories: A systematic literature review of empirical research. *Smart Learn. Environ.* **2023**, *10*, 23. [CrossRef]
19. Kazimoglu, C. Enhancing Confidence in Using Computational Thinking Skills via Playing a Serious Game: A Case Study to Increase Motivation in Learning Computer Programming. *IEEE Access* **2020**, *8*, 221831–221851. [CrossRef]
20. Beltrán, E.J.G.; Arias, J.C.M. Mi Superpoder es la Programación: A tool for teaching programming to children and youth. *Sci. Comput. Program.* **2025**, *240*, 103198. [CrossRef]
21. Díaz, J.; López, J.A.; Sepúlveda, S.; Villegas, G.M.R.; Ahumada, D.; Moreira, F. Evaluating Aspects of Usability in Video Game-Based Programming Learning Platforms. *Procedia Comput. Sci.* **2021**, *181*, 247–254. [CrossRef]
22. Ouahbi, I.; Kaddari, F.; Darhmaoui, H.; Elachqar, A.; Lahmine, S. Learning Basic Programming Concepts by Creating Games with Scratch Programming Environment. *Procedia Soc. Behav. Sci.* **2015**, *191*, 1479–1482. [CrossRef]
23. Moreno-León, J.; Robles, G. Dr. Scratch: A Web Tool to Automatically Evaluate Scratch Projects. In Proceedings of the Workshop in Primary and Secondary Computing Education, London, UK, 9–11 November 2015; ACM: New York, NY, USA, 2015; pp. 132–133. [CrossRef]
24. Delozier, C.; Shey, J. Using Visual Programming Games to Study Novice Programmers. *Int. J. Serious Games* **2023**, *10*, 115–136. [CrossRef]
25. Gundersen, S.W.; Lampropoulos, G. Using Serious Games and Digital Games to Improve Students' Computational Thinking and Programming Skills in K-12 Education: A Systematic Literature Review. *Technologies* **2025**, *13*, 113. [CrossRef]
26. Tlili, A.; Chang, M. *Data Analytics Approaches in Educational Games and Gamification Systems: Summary, Challenges, and Future Insights*; Springer: Singapore, 2019. [CrossRef]
27. Liu, M.; Kang, J.; Liu, S.; Zou, W.; Hodson, J. Learning Analytics as an Assessment Tool in Serious Games: A Review of Literature. In *Serious Games and Edutainment Applications*; Springer International Publishing: Cham, Switzerland, 2017; pp. 537–563. [CrossRef]
28. Serrano-Laguna, Á.; Martínez-Ortiz, I.; Haag, J.; Regan, D.; Johnson, A.; Fernández-Manjón, B. Applying standards to systematize learning analytics in serious games. *Comput. Stand. Interfaces* **2017**, *50*, 116–123. [CrossRef]
29. Winslow, L.E. Programming pedagogy—A psychological overview. *ACM SIGCSE Bull.* **1996**, *28*, 17–22. [CrossRef]
30. Kwon, K. Novice programmer's misconception of programming reflected on problem-solving plans. *Int. J. Comput. Sci. Educ. Sch.* **2017**, *1*, 14–24. [CrossRef]
31. Robins, A.; Rountree, J.; Rountree, N. Learning and teaching programming: A review and discussion. *Int. J. Phytoremediat.* **2003**, *21*, 137–172. [CrossRef]
32. Shute, V.J.; Sun, C.; Asbell-Clarke, J. Demystifying computational thinking. *Educ. Res. Rev.* **2017**, *22*, 142–158. [CrossRef]
33. Alonso-Fernandez, C.; Perez-Colado, I.J.; Calvo-Morata, A.; Freire, M.; Ortiz, I.M.; Manjon, B.F. Applications of Simva to Simplify Serious Games Validation and Deployment. *IEEE Rev. Iberoam. Tecnol. Aprendiz.* **2020**, *15*, 161–170. [CrossRef]
34. Perez-Colado, I.; Pérez-Colado, V.M.; Martinez-Ortiz, I.; Freire, M.; Fernandez-Manjon, B. Simplifiying Serious Games Authoring and Validation with uAdventure and SIMVA. In Proceedings of the 2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT), Tartu, Estonia, 6–9 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 106–108. [CrossRef]
35. de Ruiter, L.E.; Bers, M.U. The Coding Stages Assessment: Development and validation of an instrument for assessing young children's proficiency in the ScratchJr programming language. *Comput. Sci. Educ.* **2022**, *32*, 388–417. [CrossRef]
36. Feigenspan, J.; Kastner, C.; Liebig, J.; Apel, S.; Hanenberg, S. Measuring programming experience. In Proceedings of the 2012 20th IEEE International Conference on Program Comprehension (ICPC), Passau, Germany, 11–13 June 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 73–82. [CrossRef]
37. Kao, G.Y.M.; Chiang, C.H.; Sun, C.T. Customizing scaffolds for game-based learning in physics: Impacts on knowledge acquisition and game design creativity. *Comput. Educ.* **2017**, *113*, 294–312. [CrossRef]

38. Alonso-Fernández, C.; Freire, M.; Martínez-Ortiz, I.; Fernández-Manjón, B. Improving evidence-based assessment of players using serious games. *Telemat. Inform.* **2021**, *60*, 101583. [CrossRef]
39. Rahimi, S.; Shute, V.; Kuba, R.; Dai, C.P.; Yang, X.; Smith, G.; Fernández, C.A. The use and effects of incentive systems on learning and performance in educational games. *Comput. Educ.* **2021**, *165*, 104135. [CrossRef]
40. Yusuf, A.; Noor, N.M.; Román-González, M. Interaction Patterns During Block-based Programming Activities Predict Computational Thinking: Analysis of the Differences in Gender, Cognitive Load, Spatial Ability, and Programming Proficiency. *AI Comput. Sci. Robot. Technol.* **2024**, *3*, 1–39. [CrossRef]
41. Elshiekh, R.; Butgerit, L. Using Gamification to Teach Students Programming Concepts. *Open Access Libr. J.* **2017**, *4*, 1–7. [CrossRef]
42. Giannakoulas, A.; Xinogalos, S. A pilot study on the effectiveness and acceptance of an educational game for teaching programming concepts to primary school students. *Educ. Inf. Technol.* **2018**, *23*, 2029–2052. [CrossRef]
43. Papadakis, S. Evaluating a game-development approach to teach introductory programming concepts in secondary education. *Int. J. Technol. Enhanc. Learn.* **2020**, *12*, 127. [CrossRef]
44. Mathrani, A.; Christian, S.; Ponder-Sutton, A. PlayIT: Game Based Learning Approach for Teaching Programming Concepts. *J. Educ. Technol. Soc.* **2016**, *19*, 5–17.
45. Holly, M.; Habich, L.; Seiser, M.; Glawogger, F.; Innerebner, K.; Kupsa, S.; Einwallner, P.; Pirker, J. FemQuest—An Interactive Multiplayer Game to Engage Girls in Programming. In Proceedings of the 2024 IEEE Conference on Games (CoG), Milan, Italy, 5–8 August 2024; pp. 1–8.
46. Hainey, T.; Baxter, G. A Serious game for programming in higher education. *Comput. Educ. X Real.* **2024**, *4*, 100061. [CrossRef]
47. Arnedo-Moreno, J.; García-Solórzano, D. Programming Fun(damentals): Using commercial video games to teach basic coding to adult learners. *Entertain. Comput.* **2025**, *52*, 100850. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

computers

MDPI

Article

Generative AI in Higher Education Constituent Relationship Management (CRM): Opportunities, Challenges, and Implementation Strategies

Carrie Marcinkevage * and Akhil Kumar

Smeal College of Business, Pennsylvania State University, University Park, PA 16802, USA; axk41@psu.edu
* Correspondence: carriem@psu.edu

Abstract: This research explores opportunities for generative artificial intelligence (GenAI) in higher education constituent (customer) relationship management (CRM) to address the industry's need for digital transformation driven by demographic shifts, economic challenges, and technological advancements. Using a qualitative research approach grounded in the principles of grounded theory, we conducted semi-structured interviews and an open-ended qualitative data collection instrument with technology vendors, implementation consultants, and HEI professionals that are actively exploring GenAI applications. Our findings highlight six primary types of GenAI—textual analysis and synthesis, data summarization, next-best action recommendations, speech synthesis and translation, code development, and image and video creation—each with applications across student recruitment, advising, alumni engagement, and administrative processes. We propose an evaluative framework with eight readiness criteria to assess institutional preparedness for GenAI adoption. While GenAI offers potential benefits, such as increased efficiency, reduced costs, and improved student engagement, its success depends on data readiness, ethical safeguards, and institutional leadership. By integrating GenAI as a co-intelligence alongside human expertise, HEIs can enhance CRM ecosystems and better support their constituents.

Keywords: generative artificial intelligence; GenAI; customer relationship management; constituent relationship management; CRM; higher education; higher ed; HEI

1. Introduction

Higher education in the United States faces a decline of traditional college-aged students, rising costs alongside decreasing tuition and endowments, and societal attitudes questioning the value of higher education [1–4]. The higher education industry has also experienced significant technological turbulence since the COVID-19 pandemic when it transitioned to entirely virtual operations. These challenges have taken their toll on university employees. Gallup indicates a 35% burnout level for university employees, second only to K-12 educators [5]. Given this backdrop, higher education institutions (HEIs) seek ways to optimize their operations so they can continue to attract and retain learners, alumni, and employees.

Customer (or more broadly, constituent) relationship management (CRM) processes and technologies are used to unify contact profiles and data in order to conduct and track workflows and communications. The industry term "Customer" in CRM is renamed and expanded to incorporate many "Constituents" in the higher education setting. CRM systems are used to orchestrate student recruiting processes, track admissions decisions and statuses, manage student advising and services, conduct university fundraising and

alumni outreach, and many more engagement and communication processes. In this paper, we use the terms customer and constituent interchangeably.

Many higher education leaders view generative artificial intelligence (GenAI) as a way to increase the power of their CRM technology to address the sector's urgent challenges. In private industry, research has already been undertaken on GenAI's potential to aid in customer acquisition, retention, churn avoidance, and upgrading [6–11]. In higher education, these activities translate to recruiting and admissions, student success, and advancement (alumni relations and fundraising). While previous AIs have had the power to conduct big data analyses and predictive analytics, researchers point to the power of unlocking unstructured "dark data" as one of the greatest benefits of today's GenAI capabilities [9,12]. Whereas past CRMs with AI enabled could track college application data, class lists, attendance, and grades and make sophisticated analyses and predictions, today's GenAI can combine those data with image recognition of uploaded transcript images, can summarize advising notes, and can conduct sentiment analysis of discussion posts, all leading to automated care plans, approval processes, alerts, and communications. This data harmonization and activation is key to GenAI's potential for CRM to strengthen constituent engagement, leading to lower student acquisition costs, higher tuition and fundraising revenue, and greater student retention and outcomes.

In addition to potential constituent benefits, GenAI's use for CRM may also help address the industry's staffing and burnout concerns through smart automation and increased productivity. As one manager suggested, "I see automation of mundane tasks as a staff retention tool" [13]. The workload that GenAI could assume would free up employees' time to focus on value-added tasks where direct human contact with constituents is most essential.

CRM vendors including Salesforce, Microsoft, Slate, and Campus Solutions are quickly incorporating GenAI into their higher education CRM offerings as they see considerable potential to add both value and revenue. Other technology companies like Ivy.ai, Brado, and GPTfy, are marketing GenAI integration into the CRM ecosystem. With strategic investment into GenAI applied to CRM use cases, HEIs may make significant strides toward addressing industry challenges and providing greater benefit to their constituents.

This study explores the opportunities and challenges for GenAI in higher education CRM applications and identifies key readiness and implementation factors. It was guided by the following research questions related to use cases for GenAI in higher education, industry challenges and opportunities, technology type classification, and implementation considerations.

RQ1. What are the opportunities and challenges in utilizing GenAI in higher education CRM?
RQ2. What factors guide readiness and successful implementation?

Research into GenAI opportunities for higher education CRM applications is in its infancy. Of the approximately 4000 scholarly articles on CRM and AI published in the last ten years, 70% have been written in the last five. Only 200 of those have focused on the higher education industry, 80% of which were written in the last five years. As we could find only six articles on GenAI related to higher education CRM, we believe our study contributes to a promising new area of research. Of the current six articles, five focus on chatbots and one on predicting student retention based on GenAI image analysis from student identification cards. Our study extends the current research by exploring the potential for integrating additional types of GenAI throughout the student lifecycle and determining readiness criteria and recommendations for successful implementation. This contribution clearly connects GenAI and higher education CRM and opens new lines of inquiry into use cases, evaluation, implementation, and adoption.

Backgound

The roots of CRM systems can be traced back to the advent of database marketing in the 1980s, followed by the emergence of sales force automation (SFA) tools and contact management systems in the late 1980s and 1990s. Contact management software tools like ACT! (launched in 1987, https://www.act.org/) and GoldMine (1989, https://www.goldmine.com/) helped salespeople track customer interactions, appointments, and communications. However, CRM systems as we know them now were introduced in the early 2000s as a way for organizations to better integrate their customers' data into a single database and maintain a complete record of interactions with them through multiple channels like web, email, text, call centers, and social media [14–16]. CRM systems evolved to integrate with enterprise resource planning (ERP) systems and into a cloud-based software-as-a-service (SaaS) model. After 2010, CRM systems began to integrate with social media, mobile, email, and in-app interactions, enabling seamless business–customer interactions across multiple channels. As AI has evolved, it has been incorporated into CRM systems to provide descriptive, predictive, and prescriptive modeling capabilities.

The seminal work related to transformer networks and GenAI technology can be found in Vaswani et al., Devlin et al., and Brown et al. [17–19]. Vaswani et al., produced a landmark paper that introduces the transformer architecture that is used in many generative models due to its efficiency and scalability in handling sequential data without recurrent networks [17]. The work in Devlin et al., introduced bidirectional encoder representations from transformers (BERT) and advanced understanding of the context of words in search queries, thus influencing subsequent generative models [18]. Brown et al., demonstrated the unprecedented capabilities in generative pre-trained transformer 3 (GPT-3) for natural language processing (NLP) tasks using a "few-shot" learning approach [19].

Between 2010 and 2015, HEIs began implementing CRM systems for discrete use cases such as executive education or business school enrollment [20,21]. Following these initial use cases, universities explored CRM at the enterprise level. Today, HEIs use CRM systems for degree and non-degree recruiting and admissions processes, student progress tracking, academic and career advising appointments and tracking, alumni engagement program management and communications, fundraising and donor management, corporate relations, and many other individual processes and workflows. HEIs may have a plethora of CRM implementations at varying levels ranging from large-scale, university-wide applications to individual college- and department-level programs.

Although higher education often lags other industries in technology innovation, with the rise of GenAI, higher education finds itself in the unusual position of incorporating new technology at the same time as other industries. Other industries are finding that AI-powered CRM systems can leverage artificial intelligence to automate tasks, analyze customer sentiment, predict purchasing behavior, and offer virtual assistants for instant support [10,22,23]. The transformer network foundations of GenAI technology enable CRMs to generate content, draft personalized messages, and dynamically respond to customer queries in a conversational manner, greatly enhancing customer engagement and satisfaction [6]. Such AI-driven CRMs automate data entry, lead scoring, and communication while providing hyper-personalized interactions, allowing businesses to cater to customer preferences at scale. CRMs now support sophisticated self-service options such as AI agents that empower customers to resolve many of their issues independently and quickly. In higher education, GenAI may enable significant CRM capabilities such as analyzing and summarizing complex student data, modeling enrollment predictions, recommending interventions to prevent student attrition, individualizing constituent communication, and offering virtual agent assistants to provide student and alumni support

and task automation. HEIs are exploring these capabilities to see how they might further optimize their administrative technologies and address their industry challenges.

2. Methods

This study employed a qualitative research approach based on the principles of grounded theory [24]. Grounded theory is particularly suitable for exploring new and emerging phenomena, such as the implementation of GenAI in higher education CRM applications, as it allows for the development of theories based on data collected from participants directly involved in the field and development of theoretical insights directly from data [25].

Data were collected through semi-structured interviews and an open-ended qualitative instrument, both of which posed the same questions and were designed to elicit unrestricted participant responses without pre-defined answer constraints. Data were collected from 20 participants from three key groups: higher education CRM and GenAI software vendors (4), implementation consulting firms (5), and CRM professionals actively exploring or piloting GenAI in higher education CRM contexts (11). This purposive sampling strategy represents a cross-section of key players in this field, and it ensured that participants possessed relevant expertise and experience to address the research questions effectively [26]. The sample size was appropriate for allowing us to reach theoretical saturation [24,27,28]. The semi-structured interview format provided a flexible structure, allowing participants to share insights and experiences while also enabling the researchers to probe into specific areas as needed [29]. The questionnaires allowed additional exploration and supported an exploratory approach to identifying themes related to GenAI in CRM applications.

Data analysis followed an open-ended coding process guided by the principles of grounded theory [30]. Open-ended coding was used to identify and categorize key concepts and themes emerging from the data [25]. This initial coding phase was followed by thematic analysis, which involved identifying patterns and relationships among the coded data to develop a comprehensive understanding of the phenomena under study [31,32]. A co-researcher reviewed and validated the coding and themes to ensure consistency and reliability. As an innovative approach to enhance trustworthiness, we utilized the GenAI code assistant GPT feature known as the Delve qualitative analysis tool (https://app.delvetool.com). Once the code book was developed to the point of saturation by the researchers, the tool was used to recommend the coding application of a subset of data, followed by researcher review. The researchers retained final coding decisions to ensure methodological dependability and confirmability [33]. This approach ensured that AI-assisted insights were integrated thoughtfully, and that final coding was grounded in human interpretation and oversight, which is essential for maintaining rigor in qualitative research [34].

Ethical guidelines were strictly followed throughout the research process. Informed consent was obtained from all participants, and their anonymity and confidentiality were maintained by using participant codes (e.g., P1, P2). Appendix A provides an overview of the participant roles. The study adhered to the ethical guidelines set forth by the governing institutional review board (IRB). This methodology allowed for a rigorous exploration of GenAI in higher education CRM applications, grounding our findings in the perspectives and experiences of key stakeholders while leveraging innovative tools to enhance the analytical process.

3. Results

We utilized semi-structured interviews and open-ended data collection instruments with CRM and AI software vendors, implementation consulting firms, and HEIs who have already made the decision to invest in GenAI in higher education CRM applications. All participants were knowledgeable professionals evaluating, piloting, and utilizing GenAI for CRM applications themselves or with their clients. De-identified participants are listed in Appendix A. Our questions related to how they saw the current and future state of GenAI in higher education; what challenges and advantages the industry faces in GenAI adoption; potential cost/benefit and return on investment (ROI) considerations; and what skills or resources are needed to make progress in GenAI adoption. From the transcripts of the interview and questionnaire responses, particular topics emerged as salient themes. These themes, identified in Table 1, focused on various characteristics of the GenAI space, from types of GenAI applications to challenges and critical success factors.

Table 1. Summary of major themes from our dataset.

Theme	Frequency	Description
Administrative automation	High	Recognized as a significant area for efficiency gains in non-core tasks.
Chatbots and virtual assistants	High	Frequently cited as an accessible, low-cost AI application.
Data quality and integration	High	Noted as a key operational challenge for AI implementation.
Efficiency gains	High	Expected benefit, especially in streamlining administrative processes.
Financial constraints	High	A recurring barrier, especially for high-cost AI tools.
Student engagement and support	High	Seen as a primary goal for improving CRM interactions and retention.
Ethics and compliance	Medium	Focus on ensuring ethical AI deployment and maintaining data privacy.
Leadership and vision	Medium	Highlighted as critical for successful AI implementation.
Scalability	Medium	Cited as important for long-term viability of AI solutions.

3.1. Types of GenAI in Higher Education CRM

To address RQ1 regarding GenAI opportunities for higher education CRM, it is helpful to understand the types of GenAI that appeared in our data (listed in order of reference frequency in interviews and research):

3.1.1. Textual Analysis and Synthesis

Use cases for textual analysis and synthesis include analyzing surveys and feedback, automating knowledge article creation, and creating personalized communications such as recruiting emails, individual application reminders, student messaging, and personalized alumni newsletters or donor updates. Analyzing and creating text is the core of evolving chatbots into their next evolution as virtual agents. Virtual agents combine several capabilities like textual analysis and synthesis, data and text summarization, and code development. This was by far the most highlighted use case for GenAI in CRM throughout the entire constituent lifecycle and across all participants and research, and it could be become the "killer app" for GenAI in higher ed CRM [7,35,36]. Whereas early chatbots relied on a search and return of standard responses based on significant training and pre-developed utterance responses, today's GenAI-enabled agents allow the analysis of structured and unstructured data to create personalized, conversational responses and workflow automations. For example, a traditional admissions chatbot might ask a prospective student to enter their name and cell phone number and ask if they would like to attend an info session. The chatbot would have to send the user to a separate registration form. A

virtual agent can use contextual cues to anticipate the user's intentions, such as using prior indication of interest in online classes to suggest online information session registration. The agent's higher automation level can fill and submit the form, send a confirmation, and even suggest additional services based on context, such as recommending and scheduling an American sign language (ASL) interpreter based on a prior question about ways for hearing-impaired applicants to participate.

3.1.2. Data and Text Summarization

Generating summaries of multiple interactions and data points has great potential for higher education CRM. This technology is already in use for corporate use cases [37–39]. Early pilots at several schools include applicant interview summaries, transcript summaries, scholarship guideline summaries, academic and career advising summaries, and advancement donor contact reports.

3.1.3. Next Best Action Recommendation

This is an evolution of existing AI predictive analytics paired with GenAI summarization and text synthesis [40,41]. For example, in an enrollment use case, the system might utilize existing prospect data for regression and enrollment likelihood, summarize interactions, then generate conversion action recommendations to drive enrollment based on past analysis and training. For externally facing student use cases, this feature could generate course or internship recommendations. For advancement, it could generate next steps for donor outreach or alumni engagement opportunities.

3.1.4. Speech Synthesis and Translation

Many call centers already use speech synthesis to route incoming calls, and this feature can assist in recruiting as well as for student call centers and internal support services [42,43]. Consider the potential for real-time language translation on an inquiry call from an applicant speaking their native language. A listening bot would translate the question to the university's native language, find the answer, and translate it back to the caller's language in real time.

3.1.5. Code Development

Sometimes overlooked, code development is a crucial part of GenAI capability. This capability not only allows faster CRM system code generation based on prompts, but its power is infinitely exponential [38,41]. For example, GenAI could be used to create repeatable code for adding constituent relationship types to contact profiles based on interactions. Vendors are introducing CRM virtual agents that can be developed and operated by having bots create other bots. The AI itself can create code based on its own self-generated AI prompts. This is also a common practice used in training large language models (LLMs). Through having one model evaluate the results of another model by comparing it to expected results, AI can train another AI. One study participant from a higher education technology vendor with significant GenAI model development experience stated that, in their most recent project, this method of GenAI validation increased model accuracy from 74% to 90% (P17).

3.1.6. Image and Video Creation

Image and video creation was the least explored GenAI use case as its capability and ease of use lags behind LLM text synthesis capability. However, it has significant potential in marketing and communication personalization use cases [7,38]. An example of a higher education CRM application is the creation of GenAI images or videos personalized to indi-

vidual recipients in recruiting materials promoting new programs or alumni engagement or welcoming newly admitted students.

3.2. GenAI in CRM Across the Student Lifecycle

The student lifecycle encompasses the range of activities in which a student engages, from their application through their time as a university student and into their post-university alumni experience. HEIs anticipate that the true potential power of these GenAI use case types lies not only in their individual CRM applications, but also in their systematic combination across the constituent lifecycle. These applications showcase the versatility of GenAI in enhancing various aspects of higher education CRM systems, from improving recruiting and admissions processes to supporting student success and alumni and donor engagement. Each type of GenAI can be applied to CRM use cases across the student lifecycle, as shown in Table 2.

Table 2. Examples of GenAI types in CRM applications across the higher education lifecycle.

	GenAI Type	Recruiting and Admissions	Student Success	Alumni/ Development	Other
	Textual analysis and synthesis	Bots/virtual agents for admissions, application, financial aid questions Automated essay scoring	Bots/virtual agents for student wayfinding and cases Automating case routing and resolution	Bots/virtual Agents for alumni engagement and donor services Crafting personalized messages for alumni engagement and fundraising campaigns	Automating creation of program materials Analyze survey responses for sentiment and improvements
	Data and text Summarization	Summarizing applicant profiles to streamline the admissions process	Summarizing student data for advising or to identify at-risk students	Summarizing donor outreach and history for targeted outreach	Generating knowledge articles from case resolution notes
	Next best action recommendation	Recommending next steps for applicants based on their profiles and interaction	Suggesting student personalized advising, interventions, care plans, mentor matching, and resources	Recommending alumni engagement strategies or donor proposals based on their interests and interaction patterns	Providing recommendations for service and program improvements Recommend inquiry responses based on past engagement and interactions
	Speech synthesis and translation	AI recruiters handling large prospect volumes and multi-lingual support	Real-time advising translation Text to speech for student requests	Personalized messages in donor-preferred languages	Enhancing accessibility with text-to-speech for visually impaired students
	Code development	Developing automated CRM workflows for processing applications or generating custom reports	Creating tools to track and improve student engagement and performance. Code small apps for advising and program recommendations based on students' data and preferences	Create custom scripts to segment and engage alumni based on specific interests	Developing CRM extensions to automate repetitive tasks across departments, enhancing operational efficiency

Table 2. *Cont.*

	GenAI Type	Recruiting and Admissions	Student Success	Alumni/ Development	Other
	Image and video creation	Creating engaging visual content for recruitment campaigns Generating personalized welcome videos for admitted students using student data	Producing instructional university process videos and interactive content	Designing visually appealing materials for fundraising events Producing alumni spotlight videos using images and achievements	Generating dynamic visual content for campus announcements or event promotions based on CRM interaction data

3.2.1. Recruiting and Admissions

Figure 1 shows the process for enrollment starting with prospective applicant marketing and outreach and finishing with successful student enrollment and matriculation. Various GenAI technologies are employed at different steps of this process. For example, during the marketing and outreach phases, summarization, text synthesis, and image creation technologies may all be deployed to create a more personalized applicant experience and raise their chances of submitting an application. In the application decision phase, the focus is on data summarization, textual analysis, and recommended next best actions to evaluate and convert candidates to students.

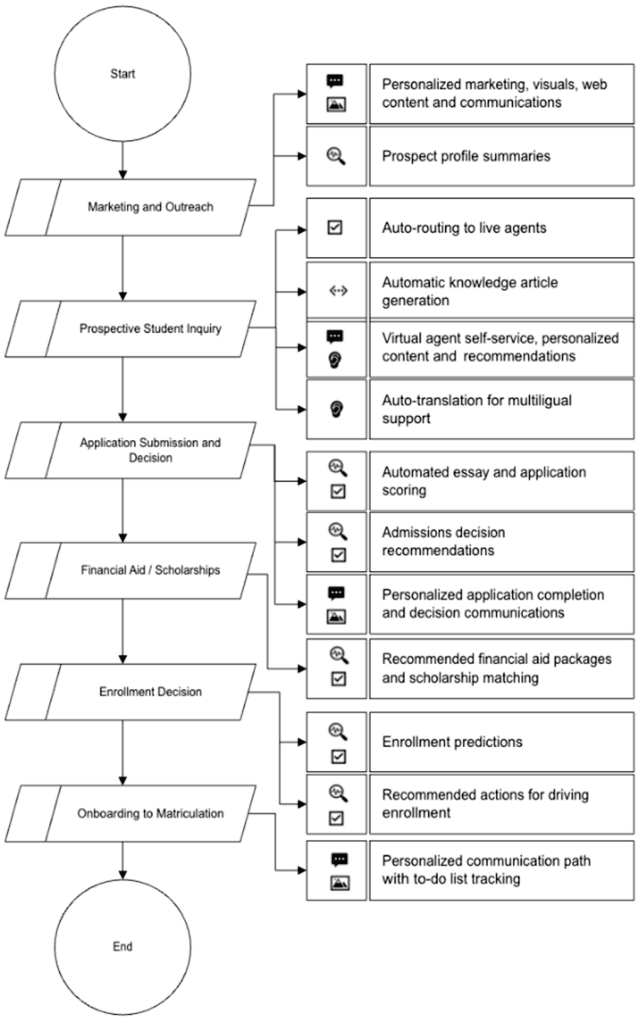


Figure 1. Student recruiting and admissions process enriched with GenAI-CRM application touchpoints.

3.2.2. Student Success

Figure 2 shows a simplified version of the student success process. Many of these activities will occur repeatedly and concurrently. Virtual agents (ideally connected) could play a crucial role at each stage of this lifecycle. Advising and support for academics and career planning can benefit significantly from virtual agents for 24/7 student service and automatic routing of student questions and cases to the correct support queue. Summarization from multiple data sources will be important for building holistic student profiles and tracking graduation requirements and status. The recommended next best action AI can be used for matching organizations, events, mentors, and even internships or jobs based on students' interests, skills, and goals.

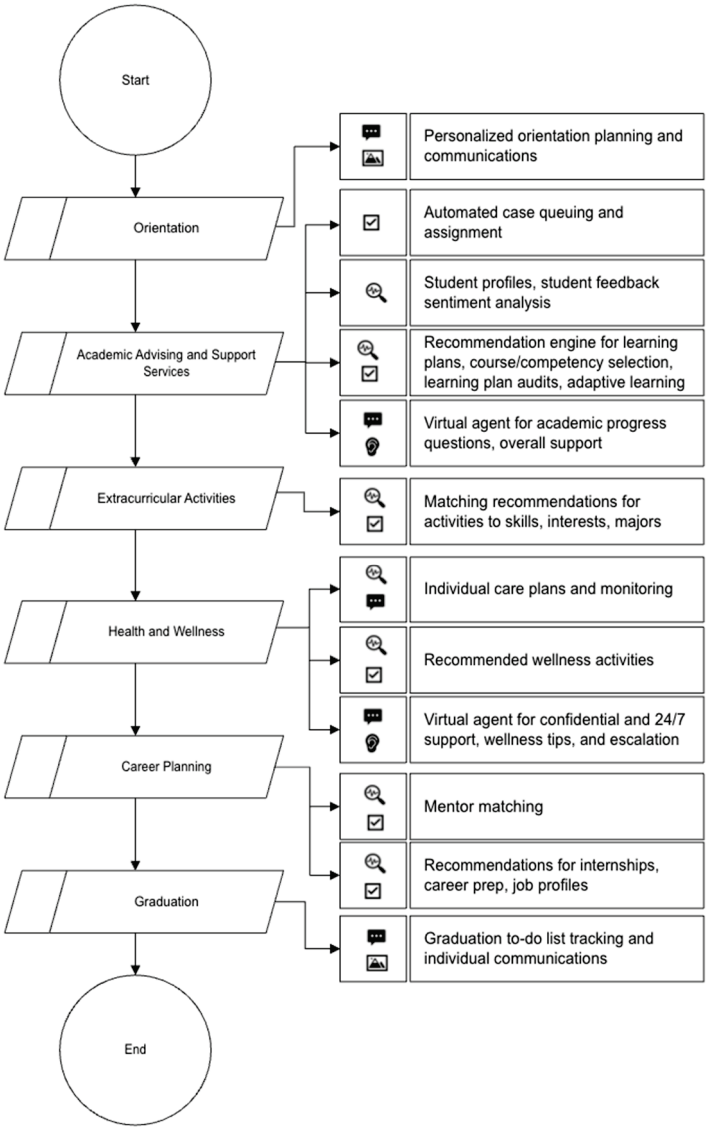


Figure 2. Student success process enriched with GenAI-CRM application touchpoints.

Participants suggested that GenAI may offer significant value in addressing the ongoing mental health crisis. According to *Inside Higher Ed*'s 2024 Student Voice survey, which included responses from over 5000 undergraduates, significant numbers of students report that poor mental health impacts their academic performance [44]. GenAI virtual agents could offer 24/7 confidential support for students at times and through a channel they may likely utilize most with less inhibition than with other channels. Additionally, by

predicting student wellness needs and recommending wellness care plans, it could help prevent student mental health crises.

3.2.3. Advancement (Alumni Engagement and Fundraising)

Figure 3 shows engagement with students after they graduate and become alumni (a status they hold for the rest of their lives). GenAI could play a significant role in tracking and promoting alumni engagement through personalized lifelong learning activities and targeted potential giving opportunities. Several studies suggest that personalized communication needs to start early to keep alumni connected, and that trying to win alumni back after communication lapses is extremely difficult [45,46]. Therefore, the text synthesis and personalization afforded by GenAI may substantially benefit this use case.

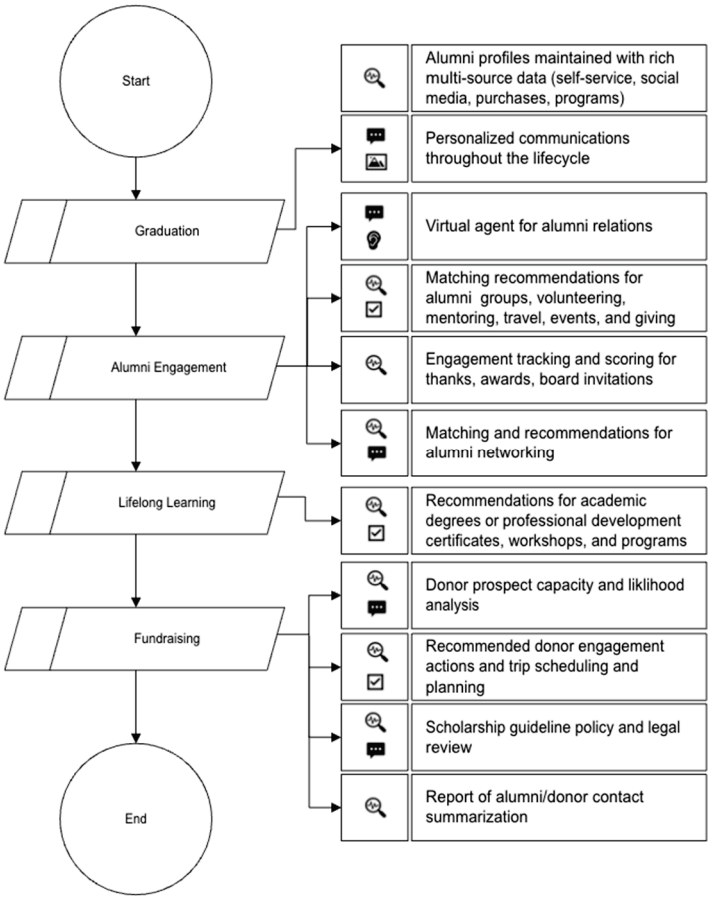


Figure 3. Alumni engagement process enriched with GenAI-CRM application touchpoints.

3.3. Higher Education Industry Challenges and Strengths

In assessing these potential applications, participants and HEIs specifically discussed challenges and strengths of the higher education environment related to RQ1.

> "You've got the innovators. You've got the people that want to be the first ones out there to change. And then you've got the people that are more comfortable and set in their ways, who have been doing this forever and ever and been hugely successful. So why would they want to change? And you have to somehow figure out how to bring these two groups together, because you don't want to be too fast and put the university and students at risk. But you don't want to be too slow and be left behind, because clearly this is moving forward, and this is having an impact on universities, on work, and on society as a whole." (P2)

Higher education is simultaneously an environment of stimulating innovation and inertial stagnation. It houses the most forward-thinking minds alongside the most archaic of systems and processes. The GenAI space is crowded and fast-moving, at odds with higher education's capacity for analysis and change. There are so many types of GenAI and so many potential applications within and outside of CRM that one participant noted, "There's just a lot of competing technologies that make the space noisy and confusing to Higher Ed" (P6). Study participants highlighted challenges they perceived as unique to the higher education environment, particularly its fragmented and siloed data environment, the high need for and risks involved in ensuring ultra-tight security and privacy; uncertain cost/benefit and ROI; and the challenge of promoting employee adoption. They also identified industry strengths born of the nature of working in higher education, such as a strong commitment to student development, significant knowledge sharing within the institution and with other HEIs, and their growing experience with technological change. Highlights of these factors can be seen in Table 3.

Table 3. Strengths and challenges of the higher education industry related to GenAI in CRM.

Industry Factor	Description	Illustrative Quote
Strengths		
Knowledge Sharing	Information and best practices shared inside the intuition as well as with other HEIs industry-wide	"The ability to share is a superpower. And then, maybe because AI comes more out of the academic world than other kinds of technology. Maybe this is the moment when we could bridge our kind of 'researchy' superpowers on one side with how we run our institutions." (P4)
Commitment to students	Dedication to student success as a vocation	"We have a vested interest in providing for the students. What can we do to meet the students' needs? What can we do to make this a better experience? That breeds a curiosity." (P6)
Technology pivot experience	COVID-19 response developed technology change skills	"Who ever thought higher ed could change technology that fast? But we did it, and now we know how." (P19)
Challenges		
Organizational silos	Decentralized structures and budgets lead to disconnected technologies and conflicting goals	"Subject matter experts in the functional areas of the institution often have nothing to do with the content on the website." (P6)
Slow processes and decisions	Committees and competing priorities create a slow pace for decisions and execution	"I would even say it's slow to adopt. I think it hinders the progress of a lot of institutions. It's the whole, you know, overanalyzing. And so you just are paralyzed by all of the different people involved in the conversation and the opinions and the concerns." (P8)
Poor data environment	Data housed across units and in many disconnected systems with no collective ownership points	"Data's a mess. And it's not an isolated issue like, 'Wow, we've really found that schools in the Northeast have it.' It is a very wide issue." (P6)
Highly regulated data environment	Security and privacy are highly critical and regulated	"We have to protect them. If students use the tools <unsecure GPTs and apps> on their own, they're happily signing away their privacy." (P7)

Table 3. *Cont.*

Industry Factor	Description	Illustrative Quote
Difficult cost/benefit and ROI calculation	Structure and lack of metrics focus make measuring success challenging	"Oh, that's great! The student affairs cost went down. But if you actually look, it might not be a positive ROI if you didn't account for the IT costs. Cost goes down in one place, up in another. And are you comparing all together for ROI?" (P6)
Adoption	Overworked staff may resist what they perceive as more work and a job-threatening technology	"You've got the people that are kind of more comfortable and set in their ways, who have been doing this forever and ever, and been hugely successful. So why would they want to change?" (P2)

4. Discussion

The research findings suggest a wide array of possible GenAI use cases for CRM applications across the higher education constituent lifecycle. A gap in current research lies in the evaluation criteria by which to assess readiness for implementing GenAI CRM applications as well as implementation considerations once a project is selected. To address this gap and RQ2, we present a table of readiness evaluation factors and scoring to identify the most valuable and viable GenAI CRM opportunities for an HEI and offer a discussion of considerations for implementation strategy and adoption.

4.1. Evaluating Readiness for GenAI in CRM

Study participants indicated that GenAI CRM application ideas come from both business as well as from IT, and that the project selection process can be challenging, even for experimentation. One implementation consulting firm cautioned, "Be selective and start strategically. . .he wants to use it for teaching assistants, and they need it for their help desk, and you're having an issue with onboarding. And so now we're trying to solve everybody's problem in a very complex institution where there's different organizations within the institution. And all of a sudden nothing gets fixed. None of it happens" (P6).

The ideal GenAI use case becomes the killer app for driving greater CRM adoption and value. When AI teams at Google [47] consider project starting points, they look for "AI-shaped problems" with criteria that make them suitable for AI such as "huge combinatorial search spaces, large amounts of data, and a clear objective function to benchmark performance against," and save for later projects those that seem viable but lack key criteria like connected data. To select potential starting points in higher education, it is helpful to consider a set of readiness criteria for each potential use case. Based on our findings and current scholarly literature, Table 4 provides readiness criteria that may aid in opportunity evaluation. Each criterion topic is framed with a description and key questions to consider.

In considering these readiness criteria, HEIs can evaluate individual use cases in their own institutions. The criteria could be scored from low (1) to high (5) for the fit and readiness of each individual use case given the current environment. For example, if a university's top priority is increasing enrollment, a high "Strategic Alignment" score might be assigned to an admissions virtual agent. A high "Data Readiness" score for that use case would mean that the good quality data needed for this use case are available, including knowledge articles, university policy documents, and access to prospective student information. To compare scoring for use cases, a column could be added representing the total score across criteria assuming equal weighting. Weights could be added to each criterion based on organizational decision-making criteria. Figure 4 represents an example of a sample scoring table.

Table 4. Readiness criteria for selecting GenAI use cases in higher education CRM.

Readiness Criterion	Description	Key Questions
Strategic alignment	The degree to which this application fits with the HEI's current strategy and mission	How well does this application fit our strategy? Would this application make it to the top of our leaders' priority lists?
ROI	Evaluation of the actual ROI or organizational benefits compared with the costs of implementation and operation	With what revenue-based ROI is this associated? How will it be measured? What are the actual costs and benefits across units, and who will accrue them? Will our budget structure allow us to measure this well enough?
Feasibility and implementation speed	Analysis of the knowledge, time, integration complexity, LLM model availability, and all other factors for assessing viability and time to implementation	Can we do this ourselves or do we need an implementation consulting firm? How fast? Do we need an open or closed LLM? If closed, how will it be built, or does it already exist? What systems, people, or other factors have to come together?
Data readiness	The quality, availability, and interoperability of the data upon which the GenAI will rely—the 4Vs of big data: volume, velocity, variety, and value	How clean are our data? How integrated are our data? Do we know source systems, order of operations and integrations, and data survivorship rules? How much structured versus unstructured data are there? How is our system interoperability?
Ethics, privacy, and security ease	The levels of ethical considerations, privacy and security compliance concerns, and regulatory oversight of the involved data	What are the ethical implications of this use case? What regulations apply to this data? What is our security posture? How much risk exposure does this application present?
User adoption likelihood	Willingness, speed, and continuity of user adoption	Why might users adopt/not adopt? How easy is resistance to overcome?
Scalability and flexibility	The degree to which this application can be scaled and/or easily tailored for other uses	How easy is this to scale? Can we use this for other purposes easily?
Measurability and monitoring	The ability to develop and measure key performance indicators (KPIs)	Can we measure performance? Do we have baselines upon which to compare performance? Who will track metrics?

The scoring could also be used to compare readiness criteria across multiple use cases. For example, some prospective student use cases could score better on Data Readiness because they require less data effort (there is simply less data and fewer integrations), and these data are not as heavily bound by regulatory and other privacy and security constraints such as FERPA. In some institutions, an undergraduate program virtual agent might score higher on Feasibility and Implementation Speed than a graduate program agent, because graduate admissions criteria and processes differ widely across programs and are harder for an agent to handle. In the realm of student success, mentor matching may score lower than imagined on Data Readiness, because it requires significant data on both students and alumni, two different populations that are often 'owned' by entirely different institutional units with disconnected processes and systems.

Icon	Use Case	Strategic Alignment	ROI	Feasibility & Implementation Speed	Data Readiness	Ethics, Privacy, & Security Ease	User Adoption Liklihood	Scalability & Flexibility	Measurability & Monitoring	Totals (40)
Enrollment Examples										
	Enrollment Chatbot/Agent	5	5	3	4	4	5	5	5	36
	Automated Essay Scoring	3	3	2	4	3	3	4	3	25
Student Success Examples										
	Live Agent Auto Case Queue Assignment	4	4	3	2	2	4	5	5	29
	Automated Student Care Plan Assignments	4	3	3	3	2	3	2	3	23
Advancement Examples										
	Alumni Engagement and Donor Prediction	3	4	3	3	2	3	3	5	26
	Donor Contact Report Summaries	3	3	4	4	3	3	4	4	28

Figure 4. Sample use case readiness evaluation for GenAI in CRM applications across the student lifecycle. Note: Scoring—1 = Low; 5 = High.

4.2. Implementation Strategy Considerations

> "Don't overthink it. I think I would even say to the slow to adopt: I think it hinders the progress at of a lot of institutions. It's the whole overanalyzing. You are paralyzed by all of the different people involved in the conversation and the opinions and the concerns. . .Don't try to boil the ocean in higher ed. We're at a place where AI needs to be adopted in in one application or use case or another, and you've got to dip your toe in or you'll get left behind." (P7)

Study participants overwhelmingly believed that HEIs should consider where GenAI might best benefit their CRM applications and pilot possibilities quickly, learning and adapting as they find solutions. While HEIs should consider GenAI implementation in alignment with their strategic missions, adoption readiness, data environment, and financial capacity, participants warn against analysis paralysis, asserting that fast experimentation is vital for ramping up on capability and results in GenAI. Based on our findings, we offer the following implementation considerations.

4.2.1. Do Data First

> "I'm afraid of the people who think that in the mediocre data environment that I find, I'm gonna have wonderful outcomes with AI. . .Here's an opportunity to create bad information because the inputs were wrong, and we can do it at a blazing pace in creative new ways." (P4)

Considering the 3V's of big data: volume, velocity, and variety, HEIs are rich in all three [9]. However, the most frequently identified challenge by study participants was the poor state of their institutional data, robbing them of a potential fourth V for value [11]. Simply put by one implementation consulting firm, "Data's a mess" (P6). This issue is critical because vendors, consultants, and researchers consistently identify it as a key factor for successful GenAI optimization.

Data quality, integration, and ownership are notoriously challenging in higher education. HEIs still have many on-premises solutions, where the data and application are hosted on individual systems within the institution rather than cloud-based systems where data are more easily shared and applications more easily updated. Data are also fragmented across different units of the institution, who create micro-solutions to suit their individual needs and become artificial gatekeepers [48]. This leads to replication of data in multiple places with different formats and uses. Local information is often outdated and incomplete; central systems are disconnected and inaccessible.

Before undertaking an implementation, HEIs should inventory the data needed to feed the GenAI and evaluate the quality of the required data environment. Then they might choose use cases where the data are most ready while shoring up other data for more complex applications.

HEIs must find ways to collect, streamline, integrate, and share data to unlock its potential. Some study participants suggested that data orchestration could be undertaken in many ways, whether in multiple data repositories or few, provided the GenAI tools have adequate access and there are no conflicting data or there are rules for resolving conflicts. New data unification platforms are making the job of translation among systems easier, with less need for integration from system to system [49]. For several HEI study participants, their pilot programs focus first on data cleaning and aggregation to ensure readiness for GenAI applications, to "get the data to a place where it could use GenAI" (P1).

Implementation consulting firms and technology vendors included in our study stress the enormity of the data work required to make GenAI useful and safe for higher education CRM—to create guardrails for answers and recommendations, elicit quality responses, and provide accurate summaries. For externally facing GenAI uses like virtual agents, these guardrails will be especially critical. They note that LLMs have nuances and uniquenesses and that the higher education content on websites and FAQs was never written for LLMs and chatbots. It takes time and extensive work to curate content and test the AI in order to provide the best possible experience for constituents and to protect the institution.

4.2.2. Prioritize Credibility, Security, and Privacy

> "We'll need to do more than keep the human in the loop. We'll need the human at the helm." (P13)

HEIs are stewards of sensitive information about many constituent types, from demographics to grades to health data to credit cards, and the responses GenAI provides will have significant potential impact. This data environment requires a sophisticated and highly protective security posture [50]. Therefore, HEIs should be judicious about their entry into GenAI, and carefully manage the tension between caution and speed.

HEIs require adherence to significant data privacy regulations, and participants raised the risk of leaked or hacked sensitive data. They referenced HEIs' responsibility to protect sensitive information and evaluate how much risk they are willing to assume. HEIs would not want a virtual assistant to mistake one student for another and reveal personally identifiable information (PII), grades, or financial status (P5). However, they warned against too much caution causing delay, for they believe that if HEIs do not build suitable GenAI fast enough, students will simply use open GPTs: "if students use the tools on their own, they're happily signing away their privacy" (P7).

The demands of AI-specific security require protection of AI models from tampering and unauthorized access. It is also necessary to ensure that inputs to AI models are sanitized to prevent adversarial attacks. Further, it should not be possible to reverse engineer training data to identify individuals. Models should be trained on decentralized data without transferring sensitive information. There is also a need to implement mechanisms to detect and mitigate biases in AI models that could lead to unfair outcomes.

Finally, HEIs must ensure that the GenAI applications in CRM offer reliable output and avoid "algorithmic opacity" [51]. Early GPTs trained on public data suffered from hallucinations and other challenges to their credibility. As they improve, HEIs will need to vigilantly train and check output to ensure trustworthiness and value. Trust in the responses GenAI provides, whether through an agent, summary, sentiment analysis, or code, will be paramount to adoption.

4.2.3. Measure Cross-Organizational ROI

"You don't want to do AI for AI's sake. There's really no point to that." (P2)

Each GenAI use case, and experimentation with it, needs to have clearly stated and measured value to the organization. Two potential GenAI benefits that participants and research highlight are staff time savings and improved service leading to positive constituent outcomes [6,10,52]. While all see the potential for these outcomes, most HEIs are still in the exploration and pilot phases of GenAI in CRM applications, and they struggle with measuring return on investment (ROI).

Realizing ROI could pose challenges for higher education on the grounds of both cost and measurement. Participants expressed concerns about the high potential cost of GenAI development, noting that "this is actually pretty expensive at scale" (P5). GenAI is likely to be a costly technology considering the preparation of all of the data, the creation of the LLMs, the training of the models, the staffing, the organizational adoption strategies, and the need to keep up with the pace of technological change (P6). In purchasing external GenAI products, HEIs will find a landscape of expensive technologies that do not understand the higher education budget landscape. One participant stated that, "a top executive on campus shared the prohibitively expensive licensing cost to acquiring ChatGPT (https://openai.com/index/chatgpt/, accessed on 4 February 2025) for educational institution and that it was a non-starter" (P11). Not only are they expensive, but pricing is a moving target. What was originally seat licensing might suddenly change to a consumption model as GenAI technology vendors wrestle with how to best monetize this new technology [53–55]. With most HEIs still recovering from COVID-19 era budget challenges and facing additional financial strain, investing in new GenAI technologies will be a challenge. HEIs will need to utilize evaluation and readiness criteria to determine the likelihood of success balanced with potential ROI on each use case investment.

Measuring ROI is critical yet challenging based on higher education organizational accounting structures. Much research cites cost reduction from staff time savings as a major potential benefit of GenAI in CRM and other use cases [3,6,35,38,56]. Costs may be saved in the units providing service to their constituents, though costs in technical units are likely to increase. For example, one technology vendor observed, "The organizational structure is different at every institution. So who owns the budget? It's a mix, right?" (P6). If HEIs cannot aggregate the savings and expenditures on account of GenAI technology across units, true ROI cannot be measured.

One way to justify anticipated GenAI-based staff savings comes from an Accenture 2024 report analyzing all job categories in the economy. By analyzing data from the Occupational Information Network (O*NET) of the US Bureau of Labor Statistics, Accenture estimates that GenAI will impact about 40% of the work we do today through automation or augmentation [41]. If staff time savings are to be realized, HEIs will need to be able to measure those savings through productivity metrics or outcomes.

4.2.4. Focus on Adoption

"AI in Education can only grow at the speed of trust." [57]

GenAI has entered the workplace quickly, with little preparation and less training. University staff may not only be reluctant and distrustful but also fatigued while resetting from major COVID-19 era disruption. Focusing on overcoming adoption challenges may aid the uptake of new technology. While our research suggests that the key factors contributing to the success of CRM-related GenAI applications are mostly like factors in other technology projects, some adoption challenges are unique and require additional consideration [58,59].

An initial adoption challenge is trust in the GenAI itself—its models and explainability [60]. In open commercial GPTs, the output process is hidden behind the models and unknown training data. Even in closed, internally built and tuned models, explainability is challenging. One participant noted, "Even with the same prompt, because of the stochastic nature, you can ask the same question and get a different answer every single time" (P7). While both open and closed models are getting better, one study suggests that people trust their intuition more than AI and are more likely to hold on to their "sticky routines", thus resisting adoption of AI [51]. A 2024 *New York Times* article reported on a study in which ChatGPT alone outperformed both doctors who used and who did not use GenAI, in part due to these routines [61]. Lack of training and overconfidence in their own experience led them to trust their own diagnoses even when the GPT offered conflicting evidence.

Not only do we often trust our own knowledge and experience over AI, but we also believe that human emotional capabilities exceed those of GenAI. However, new research may challenge this assumption. A recent study has shown that AI can be better at dealing with emotions than humans [43]. It has been found that, when compared with humans, AI demonstrates superior discipline in offering emotional support, an element crucial for student support services, while avoiding excessive practical suggestions, which may be less effective in achieving this goal. Even a recent Wall Street Journal article cited an Allstate Insurance discovery that their customers found their GenAI communications more empathetic than live agents [62]. If this capability can be harnessed appropriately, then constituent-facing GenAI applications could be highly suitable and valuable candidate CRM use cases, despite the initial reluctance to accept that an AI is capable of offering empathy.

The growing availability of GenAI in other enterprise technologies may help address some of these adoption challenges. University staff who would be engaged in CRM GenAI applications are likely already experimenting with GenAI in other applications. Most major technology platforms are adding GenAI to their products in ways that encourage use and increase familiarity. Microsoft Office and Google offer GenAI for writing emails and designing presentations. Microsoft Teams and Zoom offer AI-generated transcripts and meeting notes with action items that are widely used in higher education. These applications are doing the heavy lifting of overcoming GenAI resistance and fostering experimentation and adoption in secure, low-risk environments. Additionally, we are all customers experiencing experimentation in GenAI by commercial industries such as Allstate's insurance agents, wireless carriers, and other service providers. As these technologies improve, staff members' lived experiences may aid their acceptance of GenAI in their own work. In early GenAI projects, staff could be selected for their technological openness in addition to their functional capabilities. Once GenAI experimentation has started, the projects themselves can spark innovation and optimism, which Popa et al. suggest are operative competencies needed for GenAI adoption [52]. From there, ongoing training updated at the speed of AI's rapid evolution will be of paramount importance.

Another major adoption challenge for GenAI in higher education CRM is the view of GenAI as a threat to staff jobs. With popular media predicting that significant portions of human work can be replaced with GenAI, employees may seek to protect their jobs by finding rationalizations for not implementing the new technology [41]. This is particularly true for staff in front-facing applicant or student service roles. Staff may be concerned that their performance will now be measured against GenAI, which they could soon expect to outperform them [63]. Addressing this challenge requires strong leadership support, which most current technology and CRM implementation literature identifies as a key factor for implementation success [64–67]. Good leadership implies clarity of intentions and measurement, clear communications of outcomes, and actions consistent

with stated intentions. This includes leaders specifying ROI metrics for their projects and communicating how they will address the meeting or missing of those metrics. For example, if staff time savings is a project objective, communicating what is intended for those saved hours—repurposing toward different work, reducing roles by retraining or moving, or even acknowledging that current staff workload is already over expected hours, with the initiative intended to bring it to sustainable levels to reduce staff overload and burnout.

4.2.5. Embrace GenAI as a Co-Intelligence

"Always invite AI to the table." [68]

Participants and research suggested that the most beneficial way to overcome GenAI adoption resistance is by approaching GenAI as a co-worker and partner in performance—what GenAI researcher Ethan Mollick describes as a "co-intelligence", or a knowledgeable colleague, co-worker, and coach balancing and benefitting everyone's performance [68]. Rather than answering questions about admissions or add/drop deadlines, staff can focus on conversations that utilize their higher-level skill sets. A Deloitte white paper suggests that, while AI benefits tasks requiring high accuracy and high creative difficulty, humans outperform AI in tasks with high context variability in the social relationship realm requiring emotional intelligence and complex decision-making within ambiguity [69]. These are the advising conversations in which admissions, academic, and career advisors can navigate complex organizational systems and foster positive student experiences—even help change lives.

We might also utilize GenAI as a co-intelligence to reduce human error and strengthen new hire competence. Some experienced student advisors and counselors may miss opportunities or concerns based on their own biases or mistakes. As shown in the "doctors versus AI" study, humans often make decisions based on intuition and experience, making their thought processes singularly unique; they might miss some detail or make a mistake [61]. Other staff are new to their jobs and do not have significant experience to bring to bear on the role, especially in departments with high turnover like admissions and academic advising. This would be an appropriate opportunity for a GenAI that is trained by experienced staff to double-check care plans or to help support new staff members with training, role play, quick answers, and ongoing coaching. With appropriate staff training and openness, GenAI has the potential to improve overall performance. It could even expand the thinking of experienced staff, and it could offer experience and ideas to newer staff without taking time away from other employees.

4.3. Limitations and Future Research

Our study reflects only the current state of GenAI, and the rapid advancement of the technology means that this research is limited to its early development. Future research would be able to evaluate industry experience as technology advances. The sample size for this research was also limited by the early stage of industry experience with GenAI in CRM. It may not reflect the diversity of experiences and perspectives of CRM and GenAI technology vendors, implementation consulting firms, or universities. Once the industry has more experience in this space, future research could incorporate a larger sample of perspectives, including the students, staff, and other relevant constituents.

Another limitation of our study is selection bias due to the participants' inherently positive disposition towards GenAI in CRM. The sample for the study intentionally selected technology vendors, implementation consulting firms, and universities who were already investing in and experimenting with GenAI in CRM. This sample was chosen to ensure that the respondents would have experience with the technology and would be able to answer the study questions. In general, these participants had an optimistic view of GenAI

and its potential benefits and this could introduce a bias in our study. A future study could be conducted with a random population sample to explore overall experience with and sentiment toward GenAI use in CRM applications.

This study utilized an exploratory methodology appropriate for the state of HEIs' GenAI and CRM experience. No institution has at this time presented outcomes of implementations or even pilot projects. We were unable to evaluate GenAI outcomes on specific prototypes, as few HEIs have measurable results at this stage. Future studies could share measurable outcomes and learnings from implementations. To increase validity, additional research could be conducted related to the growing experimentation of GenAI use cases to identify results and best practices for future consideration.

Lastly, additional research opportunities would include a quantitative validation study of use case readiness evaluation criteria as the foundation for a critical success factor assessment as well as further work to understand the ethical implications of GenAI in higher education CRM.

5. Conclusions

This research has explored opportunities for GenAI in higher education CRM applications based on a qualitative study of technology suppliers, implementation consulting partners, and HEIs who are pursuing GenAI in CRM use cases. Our study identified the five types of GenAI with the most potential for CRM applications and showed their application through possible use cases within the higher education constituent lifecycle, including the highest prevalence of virtual agents. To assess readiness for GenAI in CRM projects, we have presented eight criteria of technical, organizational, and ethical factors for individual and comparative evaluation. Based on the industry's unique strengths and challenges, we find that successful implementation strategies for GenAI projects must focus on technological elements like data readiness, data and model integrity, and fast experimentation methods, as well as on organizational elements like success measurement and staff adoption.

Our research suggests that technology vendors, implementation consulting partners, and HEIs who are exploring this space believe that GenAI applications within CRM use cases have considerable potential to alleviate higher education staff burnout and shortage concerns, reduce the friction of doing business with universities, increase university enrollment and funding, and allow learners to find their best possible fit for an educational experience and a lifelong learner relationship. At the beginning of the journey, HEIs must remember to put the human at the helm, taking calibrated experimental steps into GenAI for managing their many constituent relationships and then embracing those which provide the best benefit for the university and their constituents.

> "As leaders, we are lucky if we have one opportunity in our careers to identify a genuine catalyst for monumental change. Gen AI is that opportunity." [70]

Author Contributions: Conceptualization, C.M. and A.K.; methodology, C.M.; software, C.M.; validation, C.M. and A.K.; formal analysis, C.M. and A.K.; investigation, C.M. and A.K.; resources, C.M. and A.K.; data curation, C.M. and A.K.; writing—original draft preparation, C.M. and A.K.; writing—review and editing, C.M. and A.K.; visualization, C.M. and A.K.; supervision, A.K.; project administration, C.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
CRM	Constituent (customer) relationship management
GenAI	Generative artificial intelligence
HEI	Higher education institution
LLM	Large language model

Appendix A

Table A1. Participant descriptions.

Participant Code	Type	Description
P1	Partner	Large consulting firm; AI, CRM, and HEI specialist
P2	Partner	Large consulting firm, AI specialist
P3	HEI	Mid-size university
P4	Partner	Mid-size consulting firm, CRM and HEI specialist
P5	Partner	Large consulting and technology firm, AI specialist
P5	Vendor	Large consulting and technology firm, AI specialist
P6	Vendor	CRM and AI technology supplier
P7	HEI	Mid-size university
P8	Vendor	CRM and AI technology supplier
P9	HEI	Mid-size university
P10	HEI	Small private university
P11	HEI	Small public university
P12	Partner	Mid-size consulting firm, CRM and HEI specialist
P13	HEI	Mid-size private university
P14	HEI	Mid-size public university
P15	HEI	Mid-size private university
P16	HEI	Mid-size public university
P17	Vendor	CRM and AI technology supplier
P18	HEI	Large private university
P19	HEI	Large public university

References

1. Bauman, D. Colleges Were Already Bracing for an 'Enrollment Cliff'. Now There Might Be a Second One. Available online: https://www.chronicle.com/article/colleges-were-already-bracing-for-an-enrollment-cliff-now-there-might-be-a-second-one (accessed on 10 November 2024).
2. Carlson, S. Will AI Make College Admissions and Advising Better—Or Worse? *Chron. High. Educ.* Available online: https://www.chronicle.com/article/will-ai-make-college-admissions-and-advising-better-or-worse? (accessed on 1 November 2024).
3. Hannan, E.; Liu, S. AI: New Source of Competitiveness in Higher Education. *Compet. Rev. Int. Bus. J.* **2023**, *33*, 265–279. [CrossRef]
4. Kisling, R.; Peterson, A.; Nisbet, R. The Coming of Age of Data Analytics in Higher Education. *Strateg. Enroll. Manag. Q.* **2021**, *9*, 23.
5. Marken, S.; Agrawal, S. K-12 Workers Have Highest Burnout Rate in U.S. Available online: https://news.gallup.com/poll/393500/workers-highest-burnout-rate.aspx (accessed on 10 November 2024).
6. Das, D. Understanding the Choice of Human Resource and the Artificial Intelligence: "Strategic Behavior" and the Existence of Industry Equilibrium. *J. Econ. Stud.* **2023**, *50*, 234–267. [CrossRef]
7. Ledro, C.; Nosella, A.; Vinelli, A. Artificial Intelligence in Customer Relationship Management: Literature Review and Future Research Directions. *J. Bus. Ind. Mark.* **2022**, *37*, 48–63. [CrossRef]
8. Li, L.; Lin, J.; Luo, W. Investigating the Effect of Artificial Intelligence on Customer Relationship Management Performance in E-Commerce Enterprises. *J. Electron. Commer. Res.* **2023**, *24*, 68–83.
9. Libai, B.; Bart, Y.; Gensler, S.; Hofacker, C.F.; Kaplan, A.; Kötterheinrich, K.; Kroll, E.B. Brave New World? On AI and the Management of Customer Relationships. *J. Interact. Mark.* **2020**, *51*, 44–56. [CrossRef]

10. Ozay, D.; Jahanbakht, M.; Shoomal, A.; Wang, S. Artificial Intelligence (AI)-Based Customer Relationship Management (CRM): A Comprehensive Bibliometric and Systematic Literature Review with Outlook on Future Research. *Enterp. Inf. Syst.* **2024**, *18*, 2351869. [CrossRef]
11. Singh, J.; Joesph, M.H.; Jabbar, K.B.A. Rule-Based Chatbot for Student Enquiries. *J. Phys. Conf. Ser.* **2019**, *1228*, 012060. [CrossRef]
12. Lokuge, S.; Sedera, D.; Ariyachandra, T.; Kumar, S.; Ravi, V. The next Wave of CRM Innovation: Implications for Research, Teaching, and Practice. *Commun. Assoc. Inf. Syst.* **2020**, *46*, 23. [CrossRef]
13. Salesforce New Horizons with AI. Available online: https://www.salesforce.com/content/dam/web/en_us/www/documents/industries/education/new-horizons-with-ai-ebook.pdf (accessed on 15 October 2024).
14. Chen, I.J.; Popovich, K. Understanding Customer Relationship Management (CRM): People, Process and Technology. *Bus. Process Manag. J.* **2003**, *9*, 672–688. [CrossRef]
15. Kumar, V.; Reinartz, W. *Customer Relationship Management*; Springer Texts in Business and Economics; Springer: Berlin/Heidelberg, Germany, 2018; ISBN 978-3-662-55380-0.
16. Sin, L.Y.M.; Tse, A.C.B.; Yim, F.H.K. CRM: Conceptualization and Scale Development. *Eur. J. Mark.* **2005**, *39*, 1264–1290. [CrossRef]
17. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2 August 2017; pp. 1–11.
18. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 24 May 2019; Volume 1, pp. 4171–4186.
19. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901. [CrossRef]
20. Octavia, A.; Jovanka, D.R.; Alqahtani, T.M.; Tanu Wijaya, T.; Habibi, A. Key Factors of Educational CRM Success and Institution Performance: A SEM Analysis. *Cogent Bus. Manag.* **2023**, *10*, 2196786. [CrossRef]
21. Oseni, T.; Chadhar, M.; Ivkovic, S.; Firmin, S. Organisational Learning with SaaS CRM—A Case Study of Higher Education. In *Proceedings of the ACIS 2018 Proceedings*; University of Technology: Sydney, Australia, 2018.
22. Chatterjee, S.; Mikalef, P.; Khorana, S.; Kizgin, H. Assessing the Implementation of AI Integrated CRM System for B2C Relationship Management: Integrating Contingency Theory and Dynamic Capability View Theory. *Inf. Syst. Front.* **2024**, *26*, 967–985. [CrossRef]
23. Sardjono, W.; Cholidin, A. Johan Implementation of Artificial Intelligence-Based Customer Relationship Management for Telecommunication Companies. *E3S Web Conf.* **2023**, *388*, 03015. [CrossRef]
24. Glaser, B.G.; Strauss, A.L. *The Discovery of Grounded Theory: Strategies for Qualitative Research*; Observations; Aldine de Gruyter: Chicago, IL, USA, 1967; ISBN 978-0-202-30028-3.
25. Charmaz, K. *Constructing Grounded Theory (Introducing Qualitative Methods)*, 2nd ed.; Sage Publications, Inc.: London, UK; Thousand Oaks, CA, USA, 2014; ISBN 978-0-85702-913-3.
26. Patton, M.Q. *Qualitative Research & Evaluation Methods: Integrating Theory and Practice*, 4th ed.; Sage Publications, Inc.: Thousand Oaks, CA, USA, 2015; ISBN 978-1-4129-7212-3.
27. Creswell, J.W.; Poth, C.N. *Qualitative Inquiry & Research Design: Choosing among Five Approaches*, 4th ed.; Sage Publications, Inc.: Thousand Oaks, CA, USA, 2018; ISBN 978-1-5063-3020-4.
28. Wutich, A.; Beresford, M.; Bernard, H.R. Sample Sizes for 10 Types of Qualitative Data Analysis: An Integrative Review, Empirical Guidance, and next Steps. *Int. J. Qual. Methods* **2024**, *23*, 16094069241296206. [CrossRef]
29. Kallio, H.; Pietilä, A.-M.; Johnson, M.; Kangasniemi, M. Systematic Methodological Review: Developing a Framework for a Qualitative Semi-Structured Interview Guide. *J. Adv. Nurs.* **2016**, *72*, 2954–2965. [CrossRef]
30. Corbin, J.M.; Strauss, A.L. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*, 4th ed.; Sage Publications, Inc.: Los Angeles, CA, USA, 2015; ISBN 978-1-4129-9746-1.
31. Corley, K.G.; Gioia, D.A. Building Theory about Theory Building: What Constitutes a Theoretical Contribution? *Acad. Manag. Rev.* **2011**, *36*, 12–32. [CrossRef]
32. Rheinhardt, A.; Kreiner, G.; Gioia, D.; Corley, K. Conducting and Publishing Rigorous Qualitative Research. In *The Sage Handbook of Qualitative Business and Management Research Methods*; Cassell, C., Grandy, G., Cunliffe, A., Eds.; Sage Publications Ltd.: London, UK; Thousand Oaks, CA, USA, 2018; pp. 1–31.
33. Lincoln, Y.S.; Guba, E.G. Establishing Trustworthiness. In *Naturalistic Inquiry*; Sage Publications, Inc.: Newbury Park, CA, USA, 1985; pp. 289–331. ISBN 978-0-8039-2431-4.
34. Saldaña, J. *The Coding Manual for Qualitative Researchers*, 4th ed.; Sage Publications, Inc.: London, UK; Thousand Oaks, CA, USA, 2021; ISBN 978-1-5297-3175-0.
35. Charles, V.; Rana, N.P.; Pappas, I.O.; Kamphaug, M.; Siau, K.; Engø-Monsen, K. The next 'Deep' Thing in X to Z Marketing: An Artificial Intelligence-Driven Approach. *Inf. Syst. Front.* **2024**, *26*, 851–856. [CrossRef]

36. Youn, S.; Jin, S.V. "In AI We Trust?" The Effects of Parasocial Interaction and Technopian versus Luddite Ideological Views on Chatbot-Based Customer Relationship Management in the Emerging "Feeling Economy". *Comput. Hum. Behav.* **2021**, *119*, 106721. [CrossRef]
37. Cao, Y.; Zhai, J. Bridging the Gap—The Impact of ChatGPT on Financial Research. *J. Chin. Econ. Bus. Stud.* **2023**, *21*, 177–191. [CrossRef]
38. Feuerriegel, S.; Hartmann, J.; Janiesch, C.; Zschech, P. Generative AI. *Bus. Inf. Syst. Eng.* **2024**, *66*, 111–126. [CrossRef]
39. Robbert, K.; Penn, C.; Wall, J. Use Cases of Large Language Models in Marketing Analytics. *Appl. Mark. Anal.* **2023**, *9*, 249. [CrossRef]
40. İşgüzar, S.; Fendoğlu, E.; Şimşek, A.I. Innovative Applications in Businesses: An Evaluation on Generative Artificial Intelligence. *Amfiteatru Econ.* **2024**, *26*, 511. [CrossRef]
41. Unkefer, H.; Goutham, K. Accenture Helps Organizations Personalize Experiences on Salesforce Using Data and AI. Available online: https://newsroom.accenture.com/news/2024/accenture-helps-organizations-personalize-experiences-on-salesforce-using-data-and-ai (accessed on 21 October 2024).
42. Rahman, M.S.; Bag, S.; Gupta, S.; Sivarajah, U. Technology Readiness of B2B Firms and AI-Based Customer Relationship Management Capability for Enhancing Social Sustainability Performance. *J. Bus. Res.* **2023**, *156*, 113525. [CrossRef]
43. Yin, Y.; Jia, N.; Wakslak, C.J. AI Can Help People Feel Heard, but an AI Label Diminishes This Impact. *Proc. Natl. Acad. Sci. USA* **2024**, *121*, e2319112121. [CrossRef]
44. Mowreader, A. *Inside Higher Ed.* Available online: https://www.insidehighered.com/news/student-success/health-wellness/2024/08/19/experts-weigh-college-student-mental-health-crisis (accessed on 28 December 2024).
45. Miake, A.; Carvalho, R.; Pinto, M.; Graeml, A. Customer Knowledge Management (CKM): Model Proposal and Evaluation in a Large Brazilian Higher Education Private Group. *Braz. Bus. Rev.* **2018**, *15*, 135–151. [CrossRef]
46. Sanchez-Gutierrez, J.; Gonzalez-Uribe, E.G.; Ramirez-Delgadillo, K.P.; Gonzalez-Alvarado, T.E. CRM as an Outreach and Communication Strategy with Graduates of the Master's Degree in Marketing Management. *Compet. Forum* **2016**, *14*, 339–346.
47. Griffin, C.; Wallace, D.; Mateos-Garcia, J.; Schieve, H.; Kohli, P. A New Golden Age of Discovery. Available online: https://deepmind.google/public-policy/ai-for-science/ (accessed on 3 December 2024).
48. Alenezi, M.; Akour, M. Digital Transformation Blueprint in Higher Education: A Case Study of PSU. *Sustainability* **2023**, *15*, 8204. [CrossRef]
49. Gupta, N.; Yip, J. *Databricks Data Intelligence Platform: Unlocking the GenAI Revolution*; Apress: Berkeley, CA, USA, 2024; ISBN 979-8868804434.
50. *Anthology Enhancing Higher Education with Generative AI: A Responsible Approach*; MIT SMR Connections; Massachusetts Institute of Technology: Cambridge, MA, USA, 2024; pp. 1–9.
51. Ångström, R.C.; Björn, M.; Dahlander, L.; Mähring, M.; Wallin, M.W. Getting AI Implementation Right: Insights from a Global Study. *Calif. Manag. Rev.* **2023**, *66*, 5–22. [CrossRef]
52. Popa, I.; Cioc, M.M.; Breazu, A.; Popa, C.F. Identifying Sufficient and Necessary Competencies in the Effective Use of Artificial Intelligence Technologies. *Amfiteatru Econ.* **2024**, *26*, 33–52. [CrossRef]
53. Alfirević, N.; Praničević, D.G.; Mabić, M. Custom-Trained Large Language Models as Open Educational Resources: An Exploratory Research of a Business Management Educational Chatbot in Croatia and Bosnia and Herzegovina. *Sustainability* **2024**, *16*, 4929. [CrossRef]
54. Borsci, S.; Malizia, A.; Schmettow, M.; Van Der Velde, F.; Tariverdiyeva, G.; Balaji, D.; Chamberlain, A. The Chatbot Usability Scale: The Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents. *Pers. Ubiquitous Comput.* **2022**, *26*, 95–119. [CrossRef]
55. Rosenbush, S. AI Agents Can Do More than Answer Queries. That Raises a Few Questions. *Wall Str. J.* Available online: https://www.wsj.com/articles/ai-agents-can-do-more-than-answer-queries-that-raises-a-few-questions-15009853 (accessed on 22 October 2024).
56. Castillo, D.; Canhoto, A.I.; Said, E. The Dark Side of AI-Powered Service Interactions: Exploring the Process of Co-Destruction from the Customer Perspective. *Serv. Ind. J.* **2021**, *41*, 900–925. [CrossRef]
57. Cardona, M.A.; Rodríguez, R.J.; Ishmael, K. *Artificial Intelligence and the Future of Teaching and Learning*; Department of Education Office of Educational Technology: Washington, DC, USA, 2023; pp. 1–67. Available online: https://www.ed.gov/sites/ed/files/documents/ai-report/ai-report.pdf (accessed on 19 November 2024).
58. Chatterjee, S.; Ghosh, S.K.; Chaudhuri, R. Knowledge Management in Improving Business Process: An Interpretative Framework for Successful Implementation of AI–CRM–KM System in Organizations. *Bus. Process Manag. J.* **2020**, *26*, 1261–1281. [CrossRef]
59. Dastjerdi, M.; Keramati, A.; Keramati, N. A Novel Framework for Investigating Organizational Adoption of AI-Integrated CRM Systems in the Healthcare Sector; Using a Hybrid Fuzzy Decision-Making Approach. *Telemat. Inform. Rep.* **2023**, *11*, 100078. [CrossRef]

60. Gaczek, P.; Leszczyński, G.; Mouakher, A. Collaboration with Machines in B2B Marketing: Overcoming Managers' Aversion to AI-CRM with Explainability. *Ind. Mark. Manag.* **2023**, *115*, 127–142. [CrossRef]
61. Kolata, G.A.I. Chatbots Defeated Doctors at Diagnosing Illness. *N. Y. Times*. Available online: https://www.nytimes.com/2024/11/17/health/chatgpt-ai-doctors-diagnosis.html (accessed on 23 November 2024).
62. Bousquette, I. Turns out AI Is More Empathetic than Allstate's Insurance Reps. *Wall Str. J.* Available online: https://www.wsj.com/articles/turns-out-ai-is-more-empathetic-than-allstates-insurance-reps-cf5f7c98 (accessed on 10 February 2025).
63. Bell, S.A.; Korinek, A. AI's Economic Peril. *J. Democr.* **2023**, *34*, 151–161. [CrossRef]
64. Burrell, S. *An Assessment of Salesforce^{TM} Implementation as Student Retention Management System*; Northern Arizona University: Flagstaff, AZ, USA, 2019; p. 63.
65. Marcinkevage, C. Critical Success Factors of Constituent Relationship Management (CRM) Strategy in a Higher Education Institution. Ph.D. Dissertation, Pennsylvania State University, University Park, PA, USA, 2020.
66. Rahimi, I.; Berman, U. Building a CSF Framework for CRM Implementation. *J. Database Mark. Cust. Strategy Manag.* **2009**, *16*, 253–265. [CrossRef]
67. Rigo, G.-E.; Pedron, C.D.; Caldeira, M.; Araújo, C.C.S.D. CRM Adoption in a Higher Education Institution. *J. Inf. Syst. Technol. Manag.* **2016**, *13*, 45–60. [CrossRef]
68. Mollick, E. *Co-Intelligence: Living and Working with AI*; Portfolio: New York, NY, USA, 2024; ISBN 0-593-71671-X.
69. Askew, T.; Fishman, T.; Caron, B.; Mathew, R.; Kunkel, D. How Higher Education Can Realize the Potential of Generative AI. *Deloitte Insights*. Available online: https://www2.deloitte.com/us/en/insights/industry/public-sector/generative-ai-higher-education.html (accessed on 15 October 2024).
70. Shook, E.; Daugherty, P. Work, Workers, Workforce: Reinvented in the Age of AI. Available online: https://www.accenture.com/content/dam/accenture/final/accenture-com/document-2/Accenture-Work-Can-Become-Era-Generative-AI.pdf (accessed on 3 November 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

computers

MDPI

Article

Educational Data Mining and Predictive Modeling in the Age of Artificial Intelligence: An In-Depth Analysis of Research Dynamics

Eloy López-Meneses [1], Pedro C. Mellado-Moreno [2,*], Celia Gallardo Herrerías [3] and Noelia Pelícano-Piris [4]

[1] Department of Education and Social Psychology, Pablo de Olavide University, 41013 Sevilla, Spain; elopmen@upo.es
[2] Department of Education Sciences, Language, Culture and Arts, Rey Juan Carlos University, Paseo Artilleros s/n, 28032 Madrid, Spain
[3] Department of Education, University of Almería, 04120 Almería, Spain; cgh188@ual.es
[4] Faculty of Education, International University of La Rioja, Av. de la Paz, 137, 26006 Logroño, Spain; noelia.pelicano@unir.net
* Correspondence: pedro.mellado@urjc.es

Abstract: This article provides a comprehensive analysis of the research dynamics on the use of Educational Data Mining (EDM) and predictive modeling (PM) in the era of Artificial Intelligence (AI) based on the review of 793 articles published between 2000 and 2024 in the Scopus database. The study employs bibliometric analysis and systematic literature review to identify emerging trends, methodologies, and applications in these fields. The main objective of the study is to examine the primary methodologies and innovations within AI, especially in the context of EDM and PM. It highlights how these technologies can optimize the prediction of student performance, support personalized learning, and enable timely interventions through the analysis of student data. The study also examines the role of AI in improving teaching practices, ensuring that educators maintain control over the system and minimize potential biases. Furthermore, the article addresses the ethical implications of AI implementation in education, such as privacy protection, algorithm transparency, and equity in access to learning. The findings suggest that AI has the potential to significantly improve educational outcomes and optimize student tracking, resource allocation, and the overall effectiveness of educational institutions. The responsible implementation of AI in education is emphasized to ensure inclusive and fair environments for all students.

Keywords: educational data mining; artificial intelligence; predictive modeling; learning analytics; scientific production

1. Introduction

In the last decade, Artificial Intelligence (AI) has burst into various areas of society, profoundly changing sectors such as healthcare, industry, the economy, and, of course, education. In this context, the use of advanced data mining and predictive modeling (PM) tools has gained unprecedented relevance, offering innovative solutions to address current educational challenges. These advances allow for comprehensive, real-time analysis of educational data, allowing for predicting behavior, improving learning, and, in general, transforming educational management at a global level [1,2].

Educational Data Mining (EDM) refers to the application of data mining techniques in the educational field, with the aim of discovering hidden patterns that help to improve teaching-learning processes. PM, on the other hand, is a technique that uses historical

data to predict future events or behaviors. These tools, when implemented with the support of AI, allow for accurate predictions based on large volumes of data from students, teachers, and other stakeholders in the educational environment [3]. The combination of both disciplines is beginning to redefine the educational landscape, optimizing student performance, identifying the factors that influence their success or failure, and personalizing teaching processes to better fit the needs of each student [4].

The convergence of data mining, PM, and AI has the potential to transform education into a more efficient, inclusive, and accessible system. As educational institutions adopt these technologies, new possibilities open up to personalize learning, more effectively track student progress, and improve data-driven decision-making. However, this transformation is not without its challenges [5,6]. Despite the progress made, there are technical, ethical, and social obstacles that must be overcome to ensure that the use of AI in education is fair, accessible, and effective [7].

The purpose of this paper is to provide a comprehensive analysis of the current state of research in the area of EDM and PM, especially in the context of AI. This paper examines how these technologies have been used in education and what their most notable impacts are, as well as the challenges and opportunities they present for future research. Furthermore, it seeks to provide a critical overview of the research dynamics in these fields, highlighting emerging trends and innovative applications that can significantly improve performance and equity in educational systems.

Throughout this paper, several issues related to EDM and PM will be addressed, starting with a theoretical introduction of each of these concepts, followed by a discussion of how they have been implemented at different levels of the educational system. Concrete examples of how these tools are being used in practice to predict student academic performance, improve classroom management, detect behavioral patterns that can influence learning, and generally transform traditional pedagogy will be analyzed. Furthermore, it will be examined how AI is enabling deeper analysis of educational data, making these processes more effective and efficient.

The analysis of research dynamics in this field reveals several trends. Firstly, it is possible to observe an increase in the number of studies examining the relationship between the use of EDM and the improvement of academic performance. This includes the implementation of PM to identify students at risk of failure, the personalization of learning according to the style and pace of each student, and the optimization of assessment methods. Secondly, the use of PM has been a key approach to decision-making in education as it allows predicting academic outcomes and providing early interventions to students in need. This improves the efficiency of the educational system and promotes more inclusive education, helping to reduce dropout rates and improve equity in access to educational opportunities [8,9].

The article will also discuss the challenges associated with the use of these technological tools in education. One of the main issues is data quality and privacy as EDM and PM systems rely on large volumes of information, often containing sensitive data about students. Furthermore, the use of AI raises ethical questions related to algorithmic bias, transparency, and accountability in data-driven decision-making. These issues are critical to ensure that the use of these technologies is beneficial to all educational actors and does not perpetuate inequalities or exclude certain groups of students.

The main objective of this article is to provide a comprehensive overview of current research in these fields, highlighting not only their applications but also the challenges that need to be overcome for the use of EDM and PM to be truly transformative for the education system. It also aims to offer recommendations for future lines of research, considering the

impact of AI in education and how this technology can be used ethically and effectively to improve student learning and academic performance.

This study represents a significant contribution to the field of educational research, offering a detailed review of the current dynamics in data mining and PM and how these disciplines are being revolutionized by AI. Through a critical analysis, it aims to offer a solid foundation for understanding how these technologies can be used to improve the education system and ultimately benefit students and educational institutions globally.

In this way, this research focuses on how advanced modeling techniques can be applied to analyze and improve learning personalization and academic performance in various educational settings.

A key contribution of this work is the identification of patterns and trends in the use of predictive technologies in the educational field through a bibliometric analysis using SciMAT and a systematic review of the literature. This methodological approach enables the mapping of research evolution, identifying the main emerging technologies, the most relevant approaches, and key advances in data mining and PM applied to education. The contribution of this study is based on the structured analysis of 793 publications from 2000 to 2024, allowing us to understand how these technologies can transform teaching methods and support informed decision-making to improve student performance. Furthermore, the impact of these tools on enhancing educational equity has been assessed by analyzing the distribution of research trends and the extent to which personalized solutions are tailored to students' individual needs. The results of this study are grounded in the co-word analysis and thematic evolution of the field, ensuring a robust and data-driven contribution to literature [10].

As a major innovation, this research offers a renewed perspective on the potential impact of AI on learning personalization, suggesting that the integration of PM can not only optimize academic performance but also promote greater equity in access to educational resources. Furthermore, we have highlighted how AI can facilitate the design of more dynamic learning environments where students receive the necessary support according to their individual characteristics. Finally, this study contributes new lines of research that focus on examining the future applications of AI and PM in education, suggesting paths for future research that seek to maximize the benefits of personalized education and expand its reach globally.

The article is structured as follows: Section 2 provides a comprehensive literature review, examining the key concepts of EDM and PM, with a particular focus on the role of AI in these areas. Section 3 details the methodology, combining a bibliometric analysis of 793 articles from the Scopus database with a systematic review of the literature. Section 4 presents both quantitative and qualitative results, highlighting trends in the use of AI and PM in educational research. Section 5 discusses the findings in the context of previous studies, analyzing the implications for educational practices and technologies. Lastly, Section 6 outlines the main conclusions and suggests directions for future research in the field.

2. Theoretical Framework

This section provides an in-depth exploration of the theoretical foundations underpinning the application of EDM and PM in the context of AI within educational systems. The theoretical framework offers a comprehensive analysis of these evolving fields and presents an overview of their current state, emphasizing the major theoretical concepts and research dynamics that have shaped these domains. The purpose of this section is to contextualize the integration of AI-driven data analytics in education, identify the key

challenges involved, and examine the transformative potential of these technologies in educational practices.

2.1. Educational Data Mining

EDM is an interdisciplinary field at the intersection of education and data science, focused on the application of data mining techniques to educational data to uncover patterns and trends that can improve teaching, learning, and educational management. EDM involves the use of computational tools, algorithms, and statistical methods to analyze vast amounts of student data, such as academic performance, behavior patterns, interactions with digital content, and engagement with learning platforms [11,12].

One of the primary goals of EDM is to enhance the personalization of learning. Traditional educational models often treat students as a homogeneous group, applying the same teaching methods and content to all students. In contrast, EDM allows for a more individualized approach, tailoring content and learning strategies to the needs, abilities, and progress of each student. For example, by analyzing a student's past performance data, EDM systems can recommend appropriate learning resources and predict potential areas of difficulty, enabling early interventions.

In recent years, EDM has shown significant potential in various applications, such as predicting student performance, identifying at-risk students, and improving learning outcomes. For instance, PM developed through EDM can be used to forecast whether a student is likely to pass or fail a particular course, which can prompt timely interventions by educators or academic advisors [13,14]. These PM leverage historical data to identify patterns and make accurate predictions, thus improving decision-making processes in educational settings.

Furthermore, EDM is also useful in providing insights into broader educational trends. By analyzing aggregated data from multiple students, educators and researchers can identify trends related to curriculum effectiveness, teaching methods, and student engagement [15–17]. This macro-level analysis helps educational institutions to assess the overall impact of their educational practices and adjust their strategies accordingly to meet the needs of their students.

2.2. Predictive Modeling in Education

PM is a statistical and computational technique used to forecast future outcomes based on historical data. In the context of education, PM has become a powerful tool for assessing student performance, engagement, and progression, as well as predicting potential challenges such as student dropout or failure [18]. PM rely on sophisticated algorithms and machine learning techniques to generate insights from data, which can then be used to inform educational practices.

The application of PM in education can take many forms. One common use is in predicting student academic achievement. By analyzing a range of factors (such as past grades, attendance, socioeconomic background, and learning behaviors), PM can generate accurate forecasts of students' future performance. These models provide valuable insights into which students may be at risk of falling behind, enabling early interventions such as personalized tutoring, counseling, or additional academic support. PM also helps educators to identify students who may benefit from more challenging coursework or advanced learning materials, ensuring that each student receives an educational experience that is tailored to their needs [19,20].

Another important application of PM is its role in course planning and curriculum design. By analyzing large datasets, PM can identify patterns in student learning preferences, content consumption, and engagement with educational materials. This information

can be used to redesign curricula, develop more effective teaching strategies, and enhance overall learning outcomes. PM also enables instructors to adjust the pacing of lessons and allocate resources more efficiently, optimizing both teaching and learning experiences [20].

Additionally, PM has a key role to play in identifying at-risk students and preventing school dropout. Data collected from various student interactions, such as online quizzes, assignments, forum participation, and attendance, can be analyzed to detect signs of disengagement or academic struggles [21]. Early warning systems powered by PM can alert educators to students who may be at risk of failing or leaving school, allowing them to provide timely support and prevent negative outcomes.

While PM offers significant advantages, it is important to note that the accuracy and reliability of these models depend on the quality and comprehensiveness of the data used. Inaccurate or incomplete data may lead to incorrect predictions, which could affect the educational interventions implemented [22]. Therefore, it is key for educational institutions to ensure that the data used in PM is both accurate and representative of the student population.

2.3. AI and Its Application in Education

AI has become an integral part of the digital transformation in education, bringing new opportunities to enhance learning experiences, personalize education, and improve teaching practices. AI in education encompasses a wide range of technologies, including machine learning, natural language processing, and deep learning algorithms, all of which work together to analyze data, automate processes, and adapt learning experiences to meet the unique needs of individual students [23,24].

One of the most significant ways AI is transforming education is through adaptive learning systems. These systems use AI algorithms to analyze student data and dynamically adjust the content, difficulty, and delivery of educational materials in real time. For example, an AI-driven learning platform can assess a student's progress on a particular subject and modify the learning path accordingly [25]. If a student is struggling with a specific topic, the system may offer additional resources or practice exercises to help them to master the material before moving on to more advanced topics.

AI also offers personalized learning experiences that are tailored to the unique needs and learning styles of individual students. By analyzing data such as previous academic performance, learning speed, and preferred learning methods, AI systems can deliver customized content that caters to each student's strengths and weaknesses. This personalization helps students to learn more effectively, reducing the frustration and disengagement that can arise from a one-size-fits-all approach to education [26,27].

Furthermore, AI technologies have the potential to automate various administrative tasks that currently consume valuable time and resources. Grading assignments, assessing student progress, and generating reports are some of the tasks that can be automated using AI, allowing educators to focus more on teaching and engaging with students [28]. In addition, AI systems can provide real-time feedback to students, helping them to understand their mistakes and improve their performance more quickly.

AI can also assist in teacher training and professional development. By using AI to analyze classroom interactions and instructional practices, educators can receive personalized recommendations for improving their teaching techniques. AI-powered platforms can provide teachers with insights into their students' learning progress, suggest new pedagogical approaches, and offer resources for professional development [29,30].

Despite the many advantages AI brings to education, its implementation is not without challenges. One of the major concerns is the potential for bias in AI algorithms. AI systems are trained on data, and if the data used to train the models contains biases or reflects

existing inequalities, the AI algorithms may perpetuate or amplify these biases [31,32]. For example, if an AI model is trained on data from a particular demographic group, it may not perform as effectively for students from different backgrounds. Ensuring that AI systems are fair and unbiased is critical to their success in education.

2.4. Ethical and Technical Challenges in Implementing AI in Education

While AI holds great promises for transforming education, its implementation raises significant ethical and technical challenges that need to be addressed to ensure its effective and responsible use. The ethical challenges of AI in education are multifaceted and include concerns related to data privacy, algorithmic transparency, and fairness.

- Data Privacy and Security: One of the primary ethical concerns related to AI in education is the collection and use of student data. AI systems rely on vast amounts of data to function effectively, including sensitive personal information about students. These data may include academic performance, behavioral data, learning habits, and demographic information [33,34]. Protecting these data and ensuring that they are used responsibly is essential. Educational institutions must adhere to strict data privacy regulations, such as the General Data Protection Regulation (GDPR) in Europe, to ensure that students' personal information is not exploited or misused [35].
- Algorithmic Transparency and Accountability: Another ethical issue is the opacity of AI algorithms. Many AI systems, particularly machine learning models, operate as "black boxes", meaning that it is difficult to understand how they make decisions. In educational settings, where AI-driven decisions can significantly impact students' futures, this lack of transparency is problematic [36]. For instance, if an AI system recommends a student for remedial support or suggests a particular learning path, it is key for educators and students to understand how the decision was made. Ensuring the transparency and explainability of AI algorithms is necessary to build trust in these systems and ensure that they operate fairly [37].
- Bias and Fairness: AI systems can perpetuate existing biases if they are trained on biased data. For example, if an AI model is trained on data that reflect societal inequalities, it may reinforce those biases and result in discriminatory outcomes. In education, this could mean that certain groups of students may receive inaccurate predictions or less favorable treatment based on their gender, race, or socioeconomic status [38]. To address this challenge, it is essential to ensure that the data used to train AI models are diverse and representative of all students and that the algorithms are regularly evaluated for fairness [39].
- Technical Challenges: In addition to the ethical concerns, there are several technical challenges associated with implementing AI in education. One of the primary challenges is the quality and availability of data. AI systems rely on accurate, complete, and well-organized data to function effectively [40]. However, many educational institutions face challenges related to data quality, such as missing or incomplete information, inaccurate data, and inconsistent formats. Educational institutions must invest in data infrastructure and ensure that their data are of high quality and properly structured for AI applications [41].

Another technical challenge is interoperability. Many educational institutions use a variety of digital platforms, learning management systems (LMSs), and other educational technologies that may not be designed to work together seamlessly. This lack of interoperability can make it difficult to integrate AI systems across different platforms and share data between them, limiting the potential impact of AI on educational practices. Ensuring that AI systems can work across diverse educational technologies will be essential for their widespread adoption [42,43].

2.5. The Transformative Potential of AI in Education

Despite the challenges, AI holds transformative potential for the future of education. The ability to personalize learning, predict student outcomes, and enhance teaching practices could revolutionize educational systems around the world. By leveraging AI technologies, educational institutions can provide more individualized learning experiences, improve student outcomes, and create more efficient and effective learning environments [44,45].

AI can help to level the playing field by providing personalized support to students, particularly those from disadvantaged backgrounds or those with learning disabilities. Through adaptive learning technologies and PM, AI can identify students who may need additional support and provide them with the resources they need to succeed. Additionally, AI can help to reduce educational inequalities by providing equitable access to high-quality learning resources, regardless of a student's geographic location or socioeconomic status [46].

As AI continues to evolve, its potential to transform education will only increase. The integration of AI into educational systems offers a unique opportunity to reimagine the way we teach and learn, creating more flexible, efficient, and inclusive learning environments [47,48].

3. Materials and Methods

The study focuses on analyzing the dynamics of research related to EDM and PM in the context of AI. The aim is to provide a comprehensive overview of the trends shaping the development of this field from the earliest available studies in 2000 to the most recent publications in 2024. This analysis combines a systematic qualitative review with advanced bibliometric methods, enabling the identification of key research areas and potential future directions. The research is particularly relevant due to the increasing integration of AI in educational contexts, where data mining and predictive techniques are employed to optimize teaching, learning, and educational outcomes.

To conduct this study, a set of 793 academic articles was collected by performing a keyword-based search in the Scopus database. The selected terms ("education", "teaching", "learning", "artificial intelligence", "data mining", and "predictive modeling") were chosen to ensure a broad yet focused representation of the field. The initial search aimed to gather all relevant publications within the defined timeframe. Subsequently, a data cleaning process was undertaken to remove duplicates and inconsistencies, such as variations in author names or abbreviations. This was fundamental to enhance the reliability of bibliometric analysis.

To analyze the dataset, the Science Mapping Analysis Tool (SciMAT) was employed, providing a detailed review of thematic areas and term co-occurrence networks. Additionally, VOSviewer (version 1.6.20, Leiden University, Leiden, The Netherlands) was used to create visual network maps, representing the frequency of key terms and their connections across the body of literature [49]. The co-occurrence analysis enabled the identification of dominant and emerging research trends, illustrating how the use of AI has evolved in relation to EDM and PM.

Through the bibliometric review, the most frequently occurring keywords were identified, offering insights into the thematic focus of the field. These terms were ranked using relevance scores based on their frequency in titles, abstracts, and keywords. The results revealed a strong focus on topics such as adaptive learning systems, personalized education, ML algorithms, and educational analytics. The analysis also highlighted the growing significance of ethical and practical challenges, including data privacy, algorithmic bias, and the equitable implementation of AI technologies in education [50,51].

The network maps generated by VOSviewer illustrated the evolution of research themes over time. Clusters of related terms were identified, reflecting interconnected areas of study. For example, one cluster might focus on predictive analytics and its applications in identifying at-risk students, while another might center on the integration of AI-driven adaptive learning platforms. These clusters provided a nuanced understanding of how various subfields interact and how they contribute to the overarching goals of improving educational practices through technology [52,53].

This study further examines the methodological approaches used in the selected articles. A considerable proportion of the research employs ML models such as decision trees, neural networks, and support vector machines to predict educational outcomes. These models are often trained on large datasets derived from LMS, online platforms, or standardized assessments. The findings suggest a steady improvement in predictive accuracy and the scalability of these models, enabling their application in diverse educational contexts.

The bibliometric review also sheds light on the temporal dynamics of research activity in this area. The early 2000s saw a modest number of studies focused on basic data mining techniques applied to educational datasets. However, the mid-2010s marked a significant shift, with the advent of more sophisticated AI algorithms and increased computational capacity. This period coincides with a surge in publications examining deep learning and its potential to revolutionize personalized education. More recently, the COVID-19 pandemic acted as a catalyst for research on AI-driven tools to address challenges in remote and hybrid learning environments.

Despite these advancements, the study identifies several limitations and challenges that remain unresolved. For instance, while bibliometric analysis provides a quantitative overview of research trends, it does not fully capture the qualitative dimensions of how AI and PM impact teaching and learning in real-world settings. Factors such as teacher adaptation, student engagement, and the contextual nuances of educational environments are often underexplored. Future research could address these gaps by incorporating qualitative methods, such as case studies or interviews with educators and students.

Another limitation of the current analysis is its reliance on articles indexed in Scopus, which excludes other potentially relevant sources, such as conference proceedings, technical reports, and book chapters. Expanding the scope to include these types of publications could offer a more holistic view of emerging trends and applications. Additionally, ethical concerns related to AI, such as issues of equity, transparency, and accountability, are often treated as peripheral topics rather than integral components of educational research. Addressing these concerns will require interdisciplinary collaboration between technologists, educators, and policymakers.

The study also recognizes that the evolution of research in this field is influenced by external factors. Changes in educational policies, technological advances, and global events, such as the pandemic, have shaped research priorities and applications. For example, the urgent need for scalable, AI-driven solutions during the pandemic accelerated innovations in remote learning and student performance tracking. However, the long-term implications of these developments remain unclear, highlighting the need for ongoing investigation.

This study provides a comprehensive analysis of research dynamics in EDM and PM within the context of AI. By combining systematic review and bibliometric techniques, it identifies key trends, challenges, and opportunities in the field. The findings underscore the transformative potential of AI in education while also highlighting the need for ethical and context-sensitive approaches. As the field continues to evolve, future research should aim to bridge the gap between technological innovation and its practical, equitable implementation in diverse educational settings.

The literature segmentation process and the application of the analysis model used in the study are described. For this purpose, the Science Mapping Analysis Tool (SciMAT) was used, which allowed a detailed bibliometric analysis of the selected articles to be carried out [54]. SciMAT was used to process the literature and perform a visual mapping of the main trends and relationships between the different studies on EDM, PM, and AI in the educational field.

To ensure the relevance and quality of the included articles, inclusion and exclusion criteria were established. The selected articles had to be directly related to the use of EDM techniques and PM and their implementation in education with the support of AI. Studies that did not explicitly address the application of these technologies or that lacked results applicable to improving academic performance or personalizing learning were excluded. In addition, a visualization of the methodological process has been included, showing how the articles were segmented and grouped according to the key themes identified, allowing a more structured analysis of the emerging trends in this field. This segmentation and visualization were essential to understanding the main areas of research and the most-used methodologies in the reviewed studies.

4. Results

This section presents the temporal distribution of publications in the field of EDM and PM from 2000 to 2024. It includes an analysis of the evolution of keywords and research lines during this period, shedding light on the thematic focus and technological advancements that have shaped the domain. Additionally, future research directions are identified using score relevance, offering insights into emerging trends and priorities for advancing personalized learning and optimizing teaching methodologies through AI-driven approaches.

4.1. Assessment of the Evolution of Scientific Output

The evolution of scientific production in the field of EDM and PM between 2000 and 2024 (Figure 1) reflects a panorama of transformation, consolidation, and boom, with differentiated patterns that can be classified into three main phases: the beginning and initial growth (2000–2013), consolidation and expansion (2014–2019), and boom and maturity (2020–2024). This analysis, based on the data provided, reveals both the magnitude and the dynamics of the development of this interdisciplinary field, in which AI, data analytics, and educational sciences converge.

During the first phase, between 2000 and 2013, scientific production was limited and characterized by slow initial growth. In this period, only 33 cumulative articles were published, representing barely 9% of the total produced until 2024. This initial performance can be attributed to the fact that the technologies necessary to apply data mining and PM techniques in the educational field were still in development, as were the technological infrastructures and databases that make data mining possible. Annual production in this period ranged between 1 and 12 articles per year, reflecting an incipient interest in the area. However, it is important to note that, although the volume of publications was low, the annual percentage growth rates in some years were notable. For example, in 2000 and 2001, the growth was 200% and 133%, respectively, although these figures reflect increases on very small numerical bases. Beginning in 2013, the field experienced a significant leap, marked by a 371% variation in the number of articles published. This growth can be explained by the rise of LMS and the increasing access to large volumes of educational data, which laid the foundations for the future expansion of this discipline [55].

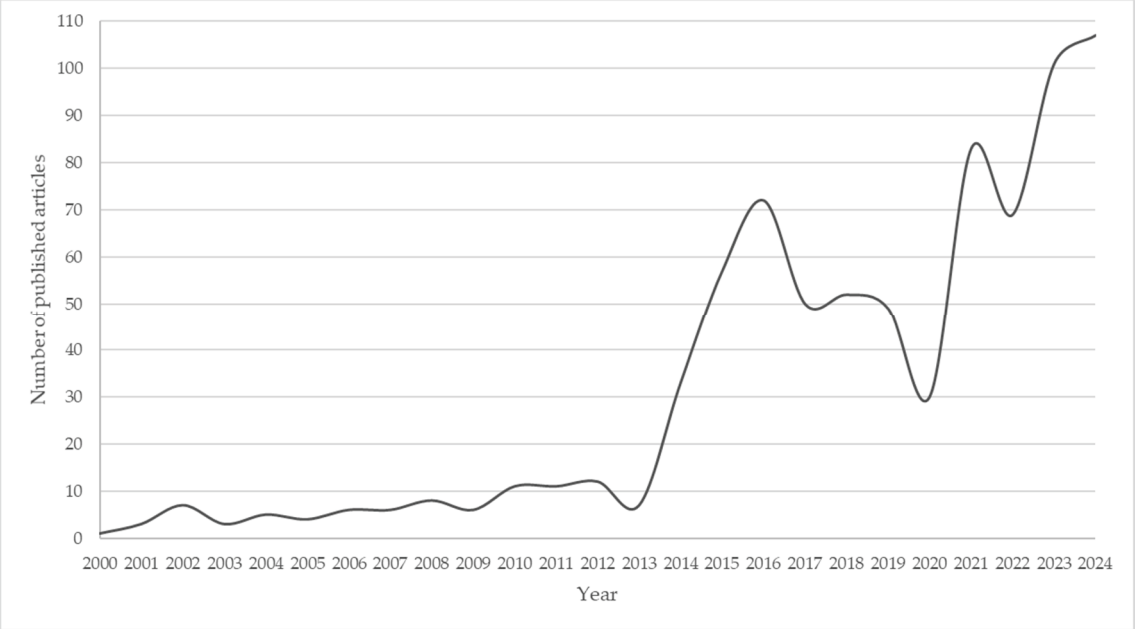


Figure 1. Dynamics of research publications over time (2000–2024).

The second phase, between 2014 and 2019, marks the period of consolidation and the expansion of research in EDM and PM. During this period, there was a notable increase in the number of annual publications, reaching an average of 52 articles per year. By 2019, the cumulative number of articles stood at 299, representing approximately 76% of the total accumulated to that date. This sustained growth suggests a consolidation of the area as an established field of research, driven by the convergence of technological advances in AI, the availability of big data in education, and the interest in improving the personalization and effectiveness of learning through analytical tools. Data mining techniques began to be integrated with pedagogical methodologies, and research examined practical applications in predicting academic performance, identifying dropout patterns, and personalizing learning trajectories [56,57]. However, despite this boom, growth during this phase was not uniform. In 2016, for example, a 31% drop in the annual variation of publications was observed, possibly due to changes in funding priorities or the saturation of certain lines of research. Similarly, in 2019, a 39% drop was recorded, which may indicate that some lines of research were reaching a maturity that required reorientation towards new challenges and research questions.

The third phase, corresponding to the period from 2020 to 2024, is characterized by the rise and maturity of scientific production in EDM and PM, largely due to the impact of AI as a driver of innovation in the educational field. In 2020, there was a significant jump in annual growth, with a variation of 177%, leading to an accumulation of 390 articles. This boom is intricately linked to the increase in the adoption of AI technologies during the COVID-19 pandemic, which accelerated the transition to digital learning environments and encouraged the implementation of PM to adapt learning experiences to the individual needs of students. From this point on, scientific production continued to expand at a remarkable pace, reaching a maximum of 107 publications in 2024, representing 30.23% of the total accumulated to that date. This sustained growth underlines the relevance of the field in the current context of digital education, where data mining and PM tools are playing a key role in informed decision-making, both at the institutional level and in personalizing learning [58,59].

A highlight of this period is the increasing diversification of research lines, which now address topics such as ethics in the use of educational data, equity in automated learning

systems, and the integration of PM with inclusive pedagogical approaches. Furthermore, advances in ML techniques, such as deep learning models and adaptive learning systems, have made it possible to examine new applications in formative assessment, the early detection of learning problems, and the improvement of educational outcomes. In this sense, research is not only focused on the development of more accurate models but also on their practical implementation in diverse educational contexts, fostering a closer connection between theory and practice [60,61].

In terms of growth dynamics, the evolution of the annual percentage change over this period shows fluctuations, reflecting both the challenges and opportunities inherent to the field. For example, the 17% decrease in 2021 suggests a slowdown after the initial pandemic-driven boom, while the 46% increase in 2022 indicates a revival of interest in new applications and approaches. These variations highlight the dynamic and evolving nature of EDM and PM research, which continues to adapt to the emerging needs of an increasingly digitalized and data-centric global education system.

Overall, the evolution of scientific production in this field reflects a journey of maturation, from its experimental beginnings to its consolidation as a key discipline to address contemporary challenges in education. As it progresses, the field continues to expand in innovative directions, guided by the intersection of AI, data analytics, and educational sciences. This historical analysis not only provides insight into past developments but also identifies emerging areas that promise to further transform the way education is understood and practiced in the digital age [62,63].

4.2. Keyword Dynamics in Scholarly Output

In recent years, the intersection of EDM and AI has become a central focus in the evolution of teaching and learning processes. The advancements in AI technologies, particularly ML and deep learning, have revolutionized how educational systems analyze and predict student performance. In this context, PM emerges as a powerful tool to optimize educational practices by forecasting learning outcomes and personalizing the learning experience.

This study delves into the research dynamics of EDM and PM, drawing insights from a comprehensive analysis of key research areas. Using co-occurrence analysis, we identified the most influential keywords and their corresponding clusters, which reflect the thematic evolution of this field. The data, extracted using VOSviewer 1.6.20 software, categorize the keywords into five distinct clusters, namely, (1) Deep Learning and Natural Language Processing in Education and Technology; (2) ML and AI Applications in Security and Engineering; (3) Feature Extraction and Pattern Recognition in ML and AI Systems; (4) Intelligent Systems and Data Analysis in Education; and (5) Knowledge Management Systems and Big Data Analytics. Each cluster highlights a key area of focus within the intersection of AI and education, providing a clearer overview of the evolving trends and research priorities in this field.

The keywords presented in Table 1 reveal the rapid integration of AI in education. Concepts such as “machine learning”, “deep learning”, and “neural networks” underscore the pivotal role of AI algorithms in enhancing learning personalization and prediction of academic success. Furthermore, keywords like “educational data mining”, “data science”, and “predictive analysis” are instrumental in the analysis of large datasets, enabling educators to identify performance trends and make informed decisions. Keywords related to “big data”, “e-learning”, and “online learning” emphasize the ongoing digital transformation of education, fostering more flexible, accessible, and tailored learning experiences.

Table 1. 30 key terms (2000–2024).

R	Keyword	N	Cluster	Links	TLS	R	Keyword	N	Cluster	Links	TLS
1	machine learning	286	2	231	1085	16	semantics	33	5	87	184
2	deep learning	105	1	156	447	17	extraction	31	1	92	157
3	forecasting	73	2	131	322	18	data science	30	1	77	163
4	neural networks	61	2	143	286	19	natural languages	30	1	79	203
5	feature extraction	61	3	123	292	20	pattern recognition	30	3	67	106
6	big data	48	5	102	188	21	information retrieval	29	5	88	164
7	natural language processing	47	1	106	307	22	feature selection	28	3	81	163
8	decision making	46	4	98	187	23	text processing	27	1	70	162
9	sentiment analysis	43	1	90	220	24	machine learning techniques	25	2	71	118
10	social networking (online)	39	1	95	197	25	internet of things	23	1	54	99
11	text mining	35	1	77	165	26	supervised learning	21	3	57	88
12	data handling	35	2	103	180	27	visualization	21	4	50	81
13	e-learning	34	4	80	147	28	data visualization	20	4	55	84
14	nearest neighbor search	33	2	91	161	29	education computing	20	4	40	75
15	educational data mining	33	4	54	103	30	computer vision	19	3	56	88

R: rank, N: occurrences, and TLS: total link strength.

By examining these trends, this analysis provides a comprehensive understanding of how AI technologies are reshaping the educational landscape, offering new opportunities for innovation, and improving learning outcomes.

Each of these clusters has a significant interrelationship in the educational field, collaborating to improve the personalization of learning, optimize educational management, and guarantee a more accessible, equitable, and quality education. A co-occurrence analysis of the keywords was carried out in the sample articles related to the research topic. Figure 2 presents the network representation of these keywords. From this analysis, it was identified that the keywords were grouped into five clusters, each with homogeneous characteristics, reflecting specific thematic areas within the field of EDM and PM.

- Cluster 1: Deep Learning and Natural Language Processing in Education and Technology

This cluster (pink) encompasses keywords related to deep learning and natural language processing and their application in various fields, particularly in education and technology. It includes topics such as sentiment analysis, social networking (online), data science, and text mining, which highlight the integration of advanced AI techniques for personalizing learning and predicting academic behaviors. The use of complex algorithms, deep neural networks, and data mining in this cluster reflects the focus on analyzing large volumes of unstructured data, such as text and interactions on digital platforms, to improve decision-making and optimize educational management [14].

This field of research is oriented toward developing intelligent systems that can adapt to the individual needs of students, using ML models that analyze complex data to detect performance patterns, personalize content, and predict outcomes. In addition, applications of emerging technologies, such as cloud computing, chatbots, emotion recognition, and face recognition, highlight the potential of AI in creating more interactive and dynamic educational environments [64].

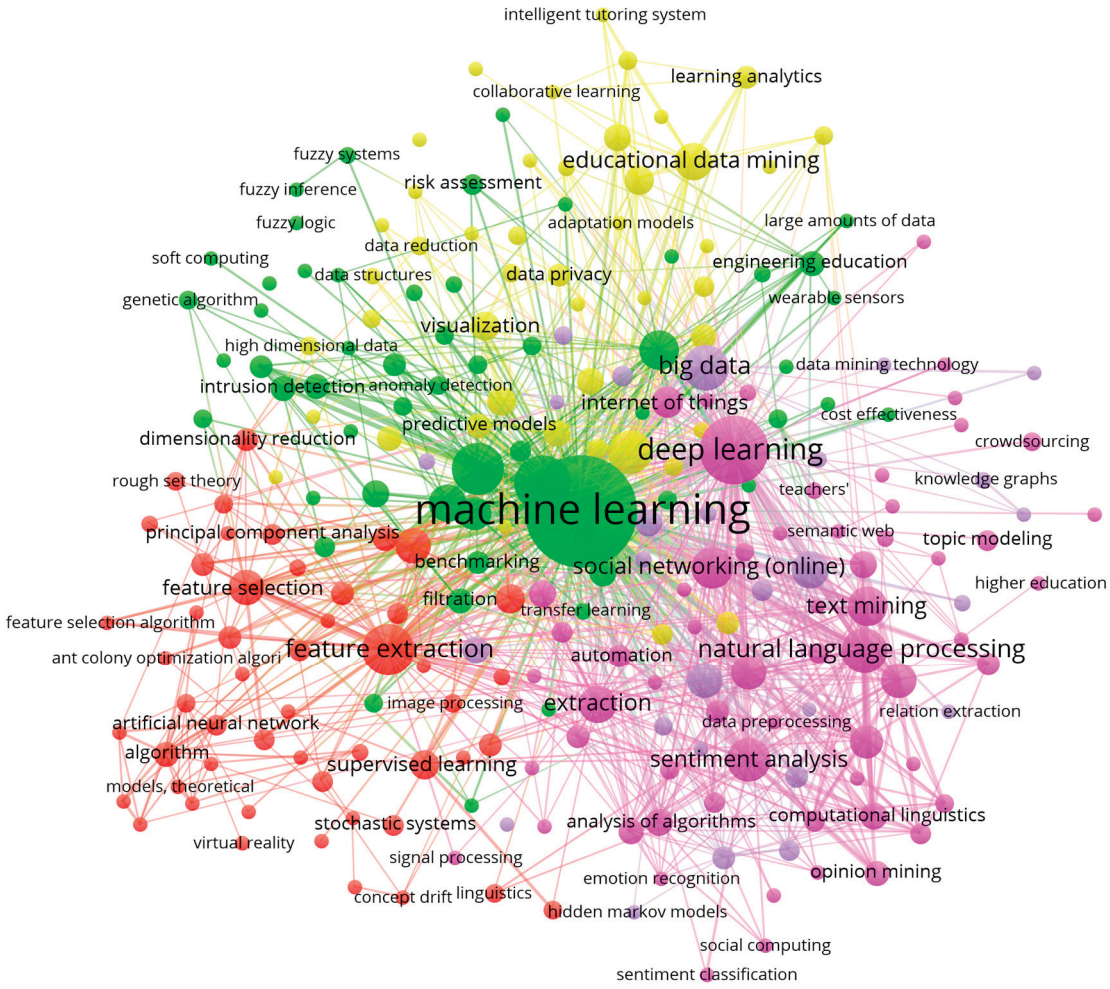


Figure 2. Keyword Co-occurrence Network (2000–2024).

- Cluster 2: ML and AI Applications in Security and Engineering

This cluster (green) is composed of key terms related to ML techniques, neural networks, and AI applied to various fields, with a particular focus on data security and engineering. Topics include forecasting, anomaly detection, ML algorithms, support vector machine, and risk assessment, which highlight the application of predictive and diagnostic models in areas such as intrusion detection and risk analysis. Also highlighted are terms such as decision tree, neural network, and swarm intelligence, which reflect the use of advanced AI algorithms to optimize processes and improve decision-making in complex environments [65].

This field of research focuses on the use of ML models to address problems such as fault detection, big data classification, and improving computer security through anomaly detection algorithms. In addition, the integration of techniques such as fuzzy logic, genetic algorithms, and radial basis function networks shows the interest in adaptive and accurate solutions in situations of high uncertainty, such as data protection and the optimization of industrial systems. The research also examines applications in areas such as engineering education and computer-aided diagnosis, highlighting the transformative potential of ML in education and medicine [66].

- Cluster 3: Feature Extraction and Pattern Recognition in ML and AI Systems

This cluster (red) focuses on advanced ML and data processing techniques, especially feature extraction and selection, pattern recognition, and the use of AI algorithms to improve the accuracy and efficiency of PM. Key terms such as feature extraction, pattern

recognition, supervised learning, and unsupervised learning indicate an intensive focus on identifying and classifying patterns within large volumes of data, which is essential for tasks such as computer vision and computer simulation [67].

One of the main contributions of this cluster lies in the use of optimization algorithms and statistical techniques to improve the performance of models. Terms such as evolutionary algorithms, reinforcement learning, and stochastic systems reflect the use of adaptive and self-adjusting approaches in solving complex problems. In addition, the use of Principal Component Analysis (PCA) and dimensionality reduction underscores the importance of reducing data complexity to improve the interpretability and performance of learning models.

This cluster also addresses the field of bioinformatics and the use of algorithms such as Ant Colony Optimization for the optimization of biological processes, which could have implications in drug discovery or genetic sequencing. The incorporation of ensemble learning and adaptive boosting indicates an emphasis on improving the accuracy of models by combining multiple algorithms. Furthermore, the exploration of decision support systems and feedback shows the potential of these models for practical applications, such as decision-making in complex environments [68].

This line of research stands out for its ability to integrate supervised and unsupervised learning techniques, together with advanced optimization and dimensionality reduction methods, in applications ranging from computer vision to stochastic systems analysis in various fields, including the health and engineering sectors.

- Cluster 4: Intelligent Systems and Data Analysis in Education

This cluster (yellow) focuses on the integration of advanced technologies, such as AI, ML, intelligent systems, and data analysis, with the aim of optimizing educational processes. It examines how digital tools, PM, and recommendation systems can improve educational efficiency, personalized learning, and the assessment of student performance.

One of the main contributions of this line is personalized education since the use of PM and ML allows content and pedagogical strategies to be adapted to the individual needs of students, favoring a more efficient learning experience. In addition, through the analysis of educational data and the use of techniques such as data mining, data visualization, and ML, the prediction of students' academic performance and the early identification of problems are facilitated, allowing for more effective interventions [69].

Another relevant aspect is the optimization of decision-making within the classroom. The incorporation of intelligent systems, such as intelligent tutoring systems or intelligent agents, helps teachers to make informed decisions about content and pedagogical approach, adapting them to the changing needs of students. Furthermore, with the increase in the use of educational data, privacy protection and information security have become key areas, driving research into technologies such as blockchain and cryptography to preserve user privacy.

The use of data visualization and visual analytics also plays an essential role in facilitating the interpretation of large volumes of educational data, helping teachers to make more informed decisions based on student behaviors and outcomes [70]. Finally, this line of research contributes to innovation in curricular design, using PM and multilayer neural networks to adapt educational curricula to the emerging needs of the technological and educational environment. Together, these investigations seek to improve teaching and learning methods, focusing on educational technology, automatic adaptation to the student, and the predictive analysis of performance.

- Cluster 5: Knowledge Management Systems and Big Data Analytics

This cluster (cyan) focuses on the integration of advanced technologies for the management, analysis, and extraction of information from large volumes of data, known as big data. Approaches such as ML, data mining, information retrieval, and knowledge discovery are examined to improve the understanding and use of data in various areas.

One of the main contributions of this line is the development of knowledge-based systems, such as information retrieval systems and knowledge-based systems, which allow large amounts of unstructured data to be processed and transformed into useful information. Techniques such as information extraction, named entity recognition, and association rule mining stand out, which allow patterns and relationships to be identified within data, facilitating decision-making in different contexts, including the criminal field, where it can be applied to predict crimes [71,72].

In addition, knowledge graphs and data warehouses are fundamental tools for organizing and representing information efficiently, which improves knowledge management within organizations. Big data analytics also plays a key role in making informed decisions and identifying trends from large data sets. Research in hidden Markov models and image segmentation also contributes to advancement in areas such as character recognition and image analysis, which have significant applications in automation and AI.

This line of research stands out for its ability to improve information management in complex contexts where semantic analysis and relation extraction are essential to understanding data and obtaining actionable insights, improving efficiency in problem solving in sectors such as information search, knowledge management, and large-scale data analysis [73].

4.3. Emerging Themes in Future Research

Table 2 presents a detailed overview of the emerging research directions identified through the cluster analysis based on the relevance scores assigned to key terms. This table highlights the most significant topics currently shaping the field of EDM and PM in the context of big data and knowledge management. The terms listed in the table represent the latest advancements and areas of focus within this rapidly evolving discipline, offering a clear indication of the trends and challenges that researchers are addressing.

Table 2. Research horizons identified through relevance scoring methods.

Rank	Future Line of Research	Relevance Score
1	Hierarchical Active Learning	10.668
2	Comprehensive Understanding	10.638
3	Vulnerability Detection Method	10.149
4	Collaborative Consortium	10.149
5	Ant Colony Optimization	10.133

Each term in Table 2 is linked to a specific cluster, providing insights into how these concepts are interrelated and contributing to the development of new methodologies in educational research. The table categorizes the terms according to their relevance and frequency of occurrence, making it possible to trace the shifts in focus and to pinpoint the most innovative areas for future exploration.

The detailed description that follows in the next sections expands on the clusters and research directions identified in Table 2, offering a comprehensive analysis of how these emerging topics are shaping the future of AI applications in education, data-driven decision-making, and knowledge discovery.

- Hierarchical Active Learning

Hierarchical Active Learning (HAL) is a line of research aimed at improving efficiency and accuracy in EDM and PM. This approach integrates the principles of active learning with hierarchical structures to optimize data selection and labeling, strategically addressing the heterogeneity and multidimensionality characteristic of educational data, such as academic records, interactions on digital platforms, and student behavior patterns [74,75].

The advancement of knowledge in this area focuses on the development of systems that reduce the dependence on massive manual labeling, prioritizing data that provide greater value to model learning. The incorporation of HAL in combination with AI can facilitate the early detection of at-risk students, improve learning personalization, and increase the accuracy of academic predictions.

The hierarchical approach allows data to be analyzed at different levels, from individual patterns to institutional trends, offering a comprehensive view that promotes more informed data-driven decisions. This opens up possibilities for significant innovations in pedagogical design and academic management, responding to the challenges of modern education [76,77].

HAL research can contribute to the advancement of knowledge in areas such as the optimization of learning algorithms, adaptability to diverse educational contexts, and the creation of scalable solutions that support the transformation of educational processes through AI.

- Comprehensive Understanding

The Comprehensive Understanding research line seeks to address the need for comprehensive analysis that connects data, models, and educational contexts. This approach focuses on going beyond isolated predictions, developing methodological frameworks that allow interpreting and contextualizing results in terms of pedagogical impact, equity in learning, and the personalization of educational experiences.

The advancement of knowledge in this area lies in the ability to integrate multiple dimensions of educational data, such as sociodemographic factors, digital interaction patterns, and academic results, with AI techniques that are not only predictive but also explanatory. This would allow identifying not only what will happen but also why it happens, offering educators and academic managers more robust tools to make informed decisions [78,79].

In addition, this approach can help to bridge the gap between technical analysis and practical applications, ensuring that PM is accessible and understandable to a wider audience, including teachers and educational policymakers. The pipeline also has the potential to address ethical and transparency issues in the use of AI in education, promoting responsible and equitable practices.

This research can advance knowledge by providing a deeper and more holistic view of how educational data can be used not only for prediction but also to improve the quality of teaching, design more effective interventions, and personalize learning processes for the benefit of students [80,81].

- Vulnerability Detection Method

The Vulnerability Detection Method research line focuses on developing advanced methods to identify critical points or vulnerabilities in educational systems through data analysis. This approach seeks to detect early risk factors that may affect student performance, equity in access to learning, or the effectiveness of pedagogical strategies [82].

The advancement of knowledge in this area lies in the creation of models that combine AI techniques and statistical analysis with a deep understanding of educational contexts. This would allow the identification of hidden patterns that may indicate structural vulnera-

bilities, such as the exclusion of specific groups of students, the ineffectiveness of certain teaching methods, or the lack of resources in critical areas.

In addition, vulnerability detection methods can help to predict and prevent problems before they manifest, providing educators and academic managers with key information to design specific and targeted interventions. This approach has the potential to transform the way educational institutions manage risks by allowing them to act proactively rather than reactively [83]. This line of research can also contribute to ensuring the sustainability of AI applications in education by identifying limitations in existing models and proposing solutions that promote their adaptability and robustness. Ultimately, research in this field will advance knowledge by improving the resilience of educational systems, fostering equity, and personalizing learning for the benefit of all students.

- Collaborative Consortium

The Collaborative Consortium research line proposes the development of interdisciplinary collaboration networks that integrate researchers, educators, technological developers, and educational policymakers. This approach seeks to enhance the exchange of knowledge, resources, and data to jointly address the complex challenges in the application of AI in education [84].

The advancement of knowledge in this area lies in the ability to bring together diverse and complementary perspectives to design PM that is more inclusive, robust, and adapted to different educational contexts. Collaborative consortia can facilitate the collection and analysis of educational data from different regions and educational levels, promoting a global approach that identifies patterns and solutions applicable on a large scale.

In addition, this approach encourages the creation of common standards for the use of educational data, guaranteeing interoperability between systems, data privacy, and transparency in the development of PM. Consortia can also promote the training of educators and academic managers, facilitating access to AI tools and strengthening their capacity to interpret and apply results in practice [85].

Research has the potential to advance knowledge by creating a shared framework that maximizes the impact of scientific findings, accelerates technological innovation, and fosters more equitable and sustainable educational solutions. This approach can be key to addressing global challenges, such as inequality in access to education and adapting to the changing needs of students in the digital age.

- Ant Colony Optimization

The line of research on Ant Colony Optimization (ACO) applied to EDM and PM focuses on the use of algorithms inspired by the behavior of ant colonies to solve complex problems related to optimization and decision-making in educational environments. This approach takes advantage of ACO's ability to find optimal solutions by simulating adaptive and collaborative processes, which is especially useful in educational systems that generate large volumes of heterogeneous data [86].

The advancement of knowledge in this area is oriented towards improving the precision and efficiency of tasks such as grouping students according to their performance, identifying patterns in learning data, and optimizing educational resources. ACO can be adapted to address specific problems in the educational field, such as personalizing learning, predicting academic trajectories, and efficiently allocating resources in online learning platforms [87].

Furthermore, this methodology has the potential to improve educational recommendation systems by identifying optimal routes for students based on their individual needs and goals. It can also contribute to improving equity in educational systems by ensuring that data-driven decisions are optimized for all students, regardless of their background.

The research can advance knowledge by providing new tools to address optimization problems in the educational context, foster innovation in the design of PM, and promote adaptive solutions that respond to the dynamic challenges of education in the AI era [86]. Some adaptive solutions that could respond to the dynamic challenges of education in the age of AI include personalized learning systems, which adjust content and pedagogical strategies according to each student's needs, interests, and abilities. In addition, predictive analytics make it possible to anticipate student performance and detect potential learning problems before they become serious. AI-based continuous assessment facilitates constant feedback, helping teachers to adapt their approaches in real time [1,88]. Adaptive learning platforms adjust the pace and difficulty level of content, optimizing the educational experience for each individual, while intelligent tutoring systems act as personalized tutors, offering support to students outside the classroom and guiding them in their learning process. These solutions not only improve personalization but also address the diversity of needs in an increasingly digitalized educational environment.

5. Discussion

The discussion of the results obtained in this study on EDM and PM in the AI era should be framed in the context of previous advances in these areas, particularly in the context of how these technologies are reshaping educational processes. The findings presented in this study reinforce and extend the ideas and conclusions of previous research while proposing new lines of research that address some of the current challenges faced by educational systems when integrating AI.

This study highlights that the combination of data mining and PM, supported by AI, has enormous potential to improve education by personalizing learning processes and optimizing the prediction of students' academic performance. This finding is consistent with previous studies, which highlighted that EDM can be a powerful tool to better understand learning and improve pedagogical intervention [89–91]. By applying advanced PM techniques, this study confirms that AI is capable of analyzing large volumes of educational data and can generate accurate predictions about student performance, identify those at risk of dropping out of school, and, even more relevantly, offer personalized interventions [6].

The study also highlights how these technologies can predict future student behaviors more efficiently than traditional interventions, allowing teachers and administrators of educational institutions to make decisions based on more accurate data. This advance is particularly significant because it aligns with the theory of personalized learning, which advocates adapting the educational process to the individual needs and characteristics of each student [92,93]. Data mining and PM open the door to more effective personalization, allowing students to receive specific educational support based on their behavior and performance patterns. This early intervention capability has been highlighted in research, which suggests that data mining could improve student retention by detecting academic difficulties early [12].

The use of AI has also been shown to be a key factor in personalizing learning, allowing systems to tailor recommendations and resources according to each student's needs. This approach contributes to improved academic performance and increases student motivation and engagement. The results obtained in this study reflect this trend and underline how AI and PM can improve the educational experience by offering materials and activities that fit each student's pace and learning style. In fact, several studies had already demonstrated the effectiveness of AI-based platforms in improving learning, reinforcing these findings by highlighting the direct relationship between personalized learning and improved academic outcomes [94,95].

In addition to educational outcomes, this study also addresses the importance of equity in access to education. As data mining and AI are integrated into educational systems, it is essential to ensure that these technologies benefit students who are already in a favorable position and help to reduce disparities in access to educational resources. This study suggests that data mining can identify not only students at academic risk but also patterns of inequality in access to resources and support [16,96]. This approach is supported by various studies, which indicate that data mining can help detect gaps in education and promote greater equity [6,92]. For example, if a student is in a situation of vulnerability or lack of resources, predictive systems can quickly identify it, allowing for early intervention that provides the necessary support to prevent them from falling into academic failure.

One of the most interesting findings in the present study is the use of PM to intervene in a personalized way. Predictive systems allow for the identification of students at higher risk and can generate adaptive interventions. This concept of personalized intervention, driven by PM, holds immense potential to revolutionize how educational institutions respond to the diverse needs of their students. By leveraging data-driven insights, institutions can tailor educational experiences to individual learning styles, pacing, and specific challenges, ensuring that each student receives the support they need to succeed [97]. PM can anticipate student struggles, enabling early interventions and targeted strategies that address academic performance, engagement, and emotional well-being. This shift towards personalized learning also empowers educators to make more informed decisions, enhancing their ability to adjust teaching methods, materials, and support mechanisms in real time. As a result, personalized interventions not only improve student outcomes but also foster a more inclusive, equitable, and effective educational environment [98].

The predictive model predicts academic outcomes and can be used to recommend teaching resources and strategies tailored to students' needs, which aligns with trends observed in other studies, which argue that personalized interventions can significantly improve teaching effectiveness [99].

In terms of future research directions, one of the key areas emerging from this study is the need to improve the accuracy and capacity of PM to address the complexity of student behavior and performance. While the results obtained are promising, current models still face challenges in terms of the accuracy of their predictions. PM systems used in EDM are generally based on structured data, such as grades and questionnaire responses, but academic performance cannot always be reduced to these factors [100]. More sophisticated models that include unstructured data, such as the analysis of students' online behavior, their interactions on digital platforms, or even their emotional well-being, could offer a more accurate and complete prediction of performance. This integrative approach has been suggested by several researchers, who argue that incorporating more types of data could improve the effectiveness of predictive systems [101,102].

Furthermore, the study highlights the importance of addressing ethical issues in the use of AI in education. As predictive systems become more deeply integrated into educational settings, issues related to data privacy, algorithm transparency, and fairness in their application must be considered. Several studies highlight the need for EDM systems to be transparent and respect student rights, which should be a priority in future research and development [103,104]. Predictive systems should be designed in a way that minimizes bias and is used fairly, without discriminating against any group of students.

In terms of future research, this study suggests that research should focus on creating more inclusive models, which can be adapted to a greater diversity of educational contexts and student populations. This includes examining the applicability of PM at different educational levels and in varied cultural contexts, which will allow for a better under-

standing of how these systems can be effectively used in different settings. Furthermore, collaboration between educational institutions, researchers, and technology developers will be key to further advance the creation of AI-based educational tools that are accessible and useful for all students.

This study confirms that EDM and PM are powerful tools to improve educational processes, allowing for personalized learning and early interventions. However, it also highlights that there are significant challenges that need to be addressed, especially in terms of accuracy, equity, and ethics. As AI continues to be a principal component of education, it is critical that future research focuses on improving the effectiveness of PM and ensuring that these benefits are equitably distributed across all students. Only then can we fully harness the potential of these technologies to transform education in positive and lasting ways.

6. Conclusions

This work has examined the impact of EDM and PM in the educational domain, especially in the context of AI, and how these tools can transform teaching and learning processes. Throughout the study, the enormous potential of these technologies to personalize learning, improve educational intervention, and offer accurate predictions on student performance has been highlighted, contributing significantly to the optimization of educational systems.

Between 2000 and 2024, scientific production in EDM and PM has evolved in three phases: initial growth (2000–2013), consolidation (2014–2019), and boom (2020–2024). The use of AI and access to large volumes of data have driven the expansion, highlighting applications in personalizing learning and improving educational outcomes.

One of the key findings of this work is that data mining and PM, supported by AI, can provide deeper insights into student behavior and needs, allowing educators to intervene early and effectively. The ability to accurately predict academic performance and dropout risk is a significant advantage of these technologies as it allows for a more detailed customization of teaching programs. By identifying behavioral patterns, the system can suggest educational interventions tailored to each student, increasing the chances of academic success, and reducing the dropout rate. This personalization, supported by predictive analytics, improves the educational experience for both students and teachers, allowing for a more student-centered approach.

This study has also shown how AI, by analyzing large volumes of data generated during the educational process, can provide teachers with valuable information about students' progress and needs. Through PM, it is possible to anticipate difficulties and obstacles before they become serious problems, allowing educational institutions to design more effective intervention strategies. The results have confirmed that the integration of AI in education improves the ability of systems to adapt to the individual needs of students, creating a more inclusive and accessible environment for all.

A fundamental aspect emphasized in this study is the importance of ensuring equitable access to AI-based educational tools, such as PM systems and data mining platforms, which are increasingly used to personalize learning and predict student performance. Ensuring that these tools are accessible to all students, regardless of socioeconomic background, is critical for fostering an inclusive educational environment. Although data mining and PM have great potential to improve academic performance, it is key that their implementation does not reproduce or amplify existing inequalities in access to education. This work highlights that educational systems must ensure that all students, regardless of their socioeconomic or cultural background, have access to these technologies. AI and data mining can help to identify gaps in access to educational resources, which could contribute to the

promotion of greater equity in education. However, for this to become a reality, a conscious and equitable approach is needed in the design and implementation of these tools.

The integration of AI in education also poses significant ethical challenges, which this work has highlighted as a key area for future research. One of the main challenges is to ensure that the use of these technologies is transparent, fair, and respectful of students' privacy. The algorithms that feed the PM must be designed in a way that minimizes bias and prevents discrimination. Transparency in decision-making processes is essential to ensure student and educator trust in these systems. Furthermore, policies need to be implemented to ensure that students' personal data are treated with the utmost respect and in compliance with privacy and data protection regulations. The results of this study suggest that future research should address these ethical challenges, ensuring that the implementation of AI in education is beneficial for all without compromising students' rights.

Despite the advances that AI has enabled in the educational field, this work also reveals that there are areas that still require further research and development. First, PM used in EDM can improve their accuracy if more types of data are integrated. Currently, systems tend to focus on structured data, such as students' grades and performance on standardized tests, but academic performance is the result of multiple factors, both academic and personal. The inclusion of unstructured data, such as the analysis of interactions on digital platforms, online behavior, or students' emotional well-being, could make PM more comprehensive and accurate, allowing for more effective interventions. The need to develop models that take these additional factors into account is a key direction for future research.

Another critical area highlighted in this work is the need to improve the adaptability of predictive systems to diverse educational contexts and student populations. As AI-based educational technology expands, it is critical that data mining models can be effectively used in different educational contexts, from primary to university education, and in diverse cultural and social settings. The study has underlined that predictive systems must be flexible enough to adjust to the particularities of each context, taking into account the different educational realities and needs of students. This will allow for even more precise personalization and greater effectiveness of interventions.

Despite these challenges, the study concludes that the integration of data mining and PM in education holds great potential to positively transform the education sector. By leveraging these advanced technologies, educational institutions can optimize teaching strategies, enhance student outcomes, and foster a more personalized, data-driven learning environment. As AI technologies continue to advance, new opportunities open up to improve teaching and learning. Predictive systems can improve retention and academic success and offer a more dynamic and flexible approach to learning, which adapts to the individual needs of students. These advances, however, must be implemented with an ethical, inclusive, and equitable vision so that the benefits of AI reach all students fairly.

Hence, this work has shown that EDM and PM are powerful tools that, if properly implemented, can significantly improve educational quality. However, it has also become clear that the development of these technologies must be accompanied by an ethical approach that ensures that all students benefit equitably and that potential biases or discrimination are minimized. The future of education, driven by AI, is promising, but it will only be successful if these challenges are addressed responsibly and consciously. The potential of AI in education is immense, and, with the right approach, it can profoundly transform the way we learn and teach.

Author Contributions: Conceptualization, methodology, software, formal analysis, resources, data curation, and writing—original draft preparation: E.L.-M. and P.C.M.-M.; investigation, validation, writing—review and editing, visualization, supervision, and project administration: E.L.-M., P.C.M.-M., C.G.H. and N.P.-P.; funding acquisition, E.L.-M. and P.C.M.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data were obtained from Elsevier's Scopus database (https://www.scopus.com/) accessed on 1 November 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chen, Z.; Zhang, J.; Jiang, X.; Hu, Z.; Han, X.; Xu, M.; Savitha, V.; Vivekananda, G.N. Education 4.0 using artificial intelligence for students performance analysis. *Intel. Artif.* **2020**, *23*, 124–137. [CrossRef]
2. van Leeuwen, A.; Goudriaan, M.; Aksu, Ü. How to responsibly deploy a predictive modelling dashboard for study advisors? A use case illustrating various stakeholder perspectives. *Comput. Educ. Artif. Intell.* **2024**, *7*, 100304. [CrossRef]
3. Castro, G.P.B.; Chiappe, A.; Rodríguez, D.F.B.; Sepulveda, F.G. Harnessing AI for Education 4.0: Drivers of Personalized Learning. *Electron. J. E-Learn.* **2024**, *22*, 1–14. [CrossRef]
4. Chavarro, D.; Perez-Taborda, J.A.; Ávila, A. Connecting brain and heart: Artificial intelligence for sustainable development. *Scientometrics* **2022**, *127*, 7041–7060. [CrossRef]
5. Lee, D.; Arnold, M.; Srivastava, A.; Plastow, K.; Strelan, P.; Ploeckl, F.; Lekkas, D.; Palmer, E. The impact of generative AI on higher education learning and teaching: A study of educators' perspectives. *Comput. Educ. Artif. Intell.* **2024**, *6*, 100221. [CrossRef]
6. Almiman, A.; Othman, M.T.B. Predictive Analysis of Computer Science Student Performance: An ACM2013 Knowledge Area Approach. *Ing. Des Syst. D'information* **2024**, *29*, 169–189. [CrossRef]
7. Flores-Castañeda, R.O.; Olaya-Cotera, S.; Iparraguirre-Villanueva, O. Benefits of Metaverse Application in Education: A Systematic Review. *Int. J. Eng. Pedagog.* **2024**, *14*, 61–81. [CrossRef]
8. Martín-García, A.V.; Martínez-Abad, F.; Reyes-González, D. TAM and stages of adoption of blended learning in higher education by application of data mining techniques. *Br. J. Educ. Technol.* **2019**, *50*, 2484–2500. [CrossRef]
9. Aunimo, L.; Kauttonen, J.; Vahtola, M.; Huttunen, S. Combining data from a LMS and a student register for exploring factors affecting study duration. *J. Comput. High. Educ.* **2024**, 1–24. [CrossRef]
10. López-Belmonte, J.; Parra-González, M.E.; Segura-Robles, A.; Pozo-Sánchez, S. Scientific mapping of gamification in web of science. *Eur. J. Investig. Health Psychol. Educ.* **2020**, *10*, 832–847. [CrossRef] [PubMed]
11. Villegas, C.W.; García-Ortiz, J.; Román-Cañizares, M.; Sánchez-Viteri, S. Proposal of a remote education model with the integration of an ICT architecture to improve learning management. *PeerJ Comput. Sci.* **2021**, *7*, e781. [CrossRef] [PubMed]
12. Dad, I.; He, J.; Noor, W.; Samad, A.; Ullah, I.; Ara, S. Cross Classification Matrix to Evaluate the Performance of Machine Learning Algorithms in Predicting Students Performance of Developing Regions. *SN Comput. Sci.* **2024**, *5*, 621. [CrossRef]
13. Ahmed, E. Student Performance Prediction Using Machine Learning Algorithms. *Appl. Comput. Intell. Soft Comput.* **2024**, *2024*, 4067721. [CrossRef]
14. Abdullah Sharadgah, T.; Abdulatif Sa'di, R. A Systematic Review of Research on the Use of Artificial Intelligence in English Language Teaching and Learning (2015–2021): What are the Current Effects? *J. Inf. Technol. Educ. Res.* **2022**, *21*, 337–377. [CrossRef]
15. Silva Filho, R.L.C.; Brito, K.; Adeodato, P.J.L. A data mining framework for reporting trends in the predictive contribution of factors related to educational achievement. *Expert Syst. Appl.* **2023**, *221*, 119729. [CrossRef]
16. Guo, S.; Zheng, Y.; Zhai, X. Artificial intelligence in education research during 2013–2023: A review based on bibliometric analysis. *Educ. Inf. Technol.* **2024**, *29*, 16387–16409. [CrossRef]
17. Bittencourt, I.I.; Chalco, G.; Santos, J.; Fernandes, S.; Silva, J.; Batista, N.; Hutz, C.; Isotani, S. Positive Artificial Intelligence in Education (P-AIED): A Roadmap. *Int. J. Artif. Intell. Educ.* **2024**, *34*, 732–792. [CrossRef]
18. Bañeres, D.; Rodríguez-González, M.E.; Guerrero-Roldán, A.-E.; Cortadas, P. An early warning system to identify and intervene online dropout learners. *Int. J. Educ. Technol. High. Educ.* **2023**, *20*, 3. [CrossRef]
19. Zhu, J. Strategies for the implementation of aesthetic education in music teaching in colleges and universities under the threshold of big data. *Appl. Math. Nonlinear Sci.* **2024**, *9*, 1–16. [CrossRef]
20. Juan, L.X.; Bin, Z. Assessment of the Effect of Attitudes of Information Engineering Students on Academic Misconduct. *Int. J. Eng. Educ.* **2024**, *40*, 776–787.
21. Rouhi, A.H. Applying Markov-based forecasting in enrolment planning. *J. Institutional Res. South East Asia* **2019**, *17*, 133–154.

22. Walker, E.; Ogan, A. We're in this Together: Intentional Design of Social Relationships with AIED Systems. *Int. J. Artif. Intell. Educ.* **2016**, *26*, 713–729. [CrossRef]
23. Ishak, N.A.; Jiang, C.X. Research on the Relationship between the Development of Artificial Intelligence Technology and the Change of Education Goals. *Int. J. Intell. Syst. Appl. Eng.* **2022**, *10*, 329–335.
24. Areed, M.F.; Amasha, M.A.; Abougalala, R.A.; Alkhalaf, S.; Khairy, D. Developing gamification e-quizzes based on an android app: The impact of asynchronous form. *Educ. Inf. Technol.* **2021**, *26*, 4857–4878. [CrossRef] [PubMed]
25. Ma, Z.-G.; Ni, R.-Y.; Yu, K.-H. Recent advances, key techniques and future challenges of knowledge graph. *Chin. J. Eng.* **2020**, *42*, 1254–1266. [CrossRef]
26. Tolentino, R.; Rodriguez, C.; Hersson-Edery, F.; Lane, J.; Abbasgholizadeh Rahimi, S. Perspectives on virtual interviews and emerging technologies integration in family medicine residency programs: A cross-sectional survey study. *BMC Med. Educ.* **2024**, *24*, 975. [CrossRef] [PubMed]
27. Chen, B.; Hwang, G.-H.; Wang, S.-H. Gender Differences in Cognitive Load when Applying Game Based Learning with Intelligent Robots. *Educ. Technol. Soc.* **2021**, *24*, 102–115.
28. Lokare, V.T.; Jadhav, P.M. A Comprehensive Pedagogical Framework for Teaching IoS App Design Course. *J. Eng. Educ. Transform.* **2024**, *37*, 491–499. [CrossRef]
29. Bai, X.; Stede, M. A Survey of Current Machine Learning Approaches to Student Free-Text Evaluation for Intelligent Tutoring. *Int. J. Artif. Intell. Educ.* **2023**, *33*, 992–1030. [CrossRef]
30. Yu, J.H.; Chauhan, D. Trends in NLP for personalized learning: LDA and sentiment analysis insights. *Educ. Inf. Technol.* **2024**, 1–42. [CrossRef]
31. AlAfnan, M.A. Taxonomy of Educational Objectives: Teaching, Learning, and Assessing in the Information and Artificial Intelligence Era. *J. Curric. Teach.* **2024**, *13*, 173–191. [CrossRef]
32. Ayanwale, M.A.; Molefi, R.R. Exploring intention of undergraduate students to embrace chatbots: From the vantage point of Lesotho. *Int. J. Educ. Technol. High. Educ.* **2024**, *21*, 20. [CrossRef]
33. Li, H.; Han, D. EduRSS: A Blockchain-Based Educational Records Secure Storage and Sharing Scheme. *IEEE Access* **2019**, *7*, 179273–179289. [CrossRef]
34. Sookhak, M.; Tang, H.; He, Y.; Yu, F.R. Security and Privacy of Smart Cities: A Survey, Research Issues and Challenges. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 1718–1743. [CrossRef]
35. Onik, M.M.H.; Kim, C.-S.; Lee, N.-Y.; Yang, J. Privacy-aware blockchain for personal data sharing and tracking. *Open Comput. Sci.* **2019**, *9*, 80–91. [CrossRef]
36. Adegboye, M. Impact of artificial intelligence on health information literacy: Guidance for healthcare professionals. *Libr. Hi Tech News* **2024**, *41*, 1–4. [CrossRef]
37. Craig, J. Computer vision for visual arts collections: Looking at algorithmic bias, transparency, and labor. *Art Doc.* **2021**, *40*, 198–208. [CrossRef]
38. Prinsloo, P.; Khalil, M.; Slade, S. Vulnerable student digital well-being in AI-powered educational decision support systems (AI-EDSS) in higher education. *Br. J. Educ. Technol.* **2024**, *55*, 2075–2092. [CrossRef]
39. Vega-Gonzalo, M.; Christidis, P. Fair Models for Impartial Policies: Controlling Algorithmic Bias in Transport Behavioural Modelling. *Sustainability* **2022**, *14*, 8416. [CrossRef]
40. Ng, D.T.K.; Leung, J.K.L.; Su, J.; Ng, R.C.W.; Chu, S.K.W. Teachers' AI digital competencies and twenty-first century skills in the post-pandemic world. *Educ. Technol. Res. Dev.* **2023**, *71*, 137–161. [CrossRef] [PubMed]
41. Nikonova, E.; Yakhyaeva, K.; Pivkina, N.; Schetinina, A. Using Artificial Intelligence Tools in Teaching a Foreign Language in Higher Technical Institutions. *Eur. J. Contemp. Educ.* **2023**, *12*, 578–589. [CrossRef]
42. Mondragon-Estrada, E.; Kirschning, I.; Nolazco-Flores, J.A.; Camacho-Zuñiga, C. Fostering digital transformation in education: Technology enhanced learning from professors' experiences in emergency remote teaching. *Front. Educ.* **2023**, *8*, 1250461. [CrossRef]
43. Abdullahi, I.; Longo, S.; Samie, M. Towards a Distributed Digital Twin Framework for Predictive Maintenance in Industrial Internet of Things (IIoT). *Sensors* **2024**, *24*, 2663. [CrossRef] [PubMed]
44. Squalli Houssaini, M.; Aboutajeddine, A.; Toughrai, I.; Ibrahimi, A. Development of a design course for medical curriculum: Using design thinking as an instructional design method empowered by constructive alignment and generative AI. *Think. Ski. Creat.* **2024**, *52*, 101491. [CrossRef]
45. Talaat, F.M. The effect of consanguineous marriage on reading disability based on deep neural networks. *Multimed. Tools Appl.* **2024**, *83*, 51787–51807. [CrossRef]
46. Namboothiri, S.; Varghese, T.; Jacob, M.; Job, S.; Cyriac, J. Integrating Artificial Intelligence with NHEQF Descriptors for Pedagogical Excellence. *High. Educ. Future* **2025**, *12*, 27–50. [CrossRef]
47. Khemka, M.; Houck, B. Toward Effective AI Support for Developers. *Queue* **2024**, *22*, 53–78. [CrossRef]

48. Ukeje, I.O.; Elom, C.O.; Ayanwale, M.A.; Umoke, C.C.; Nwangbo, S.O. Exploring an Innovative Educational Governance Framework: Leveraging Artificial Intelligence in a Stakeholder-Driven «Open Campus Model» in South East Nigerian Universities. *Int. J. Learn. Teach. Educ. Res.* **2024**, *23*, 416–440. [CrossRef]
49. van Eck, N.J.; Waltman, L. Visualizing Bibliometric Networks. In *Measuring Scholarly Impact*; Springer International Publishing: Cham, Switzerland, 2014; pp. 285–320.
50. van Eck, N.J.; Waltman, L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* **2010**, *84*, 523–538. [CrossRef] [PubMed]
51. Van Eck, N.J.; Waltman, L. Citation-based clustering of publications using CitNetExplorer and VOSviewer. *Scientometrics* **2017**, *111*, 1053–1070. [CrossRef] [PubMed]
52. González-Zamar, M.-D.; Abad-Segura, E. La Realidad Aumentada como recurso creativo en la educación: Una revisión global. *Rev. Creat. Y Soc.* **2020**, *32*, 164–190. [CrossRef]
53. González-Zamar, M.D.; Abad-Segura, E. Analysing the teaching-learning process in the university virtual classroom at the beginning of the COVID-19 pandemic. *Cult. Educ.* **2024**, *36*, 190–217. [CrossRef]
54. Gutiérrez-Salcedo, M.; Martínez, M.Á.; Moral-Munoz, J.A.; Herrera-Viedma, E.; Cobo, M.J. Some bibliometric procedures for analyzing and evaluating research fields. *Appl. Intell.* **2018**, *48*, 1275–1287. [CrossRef]
55. Wang, S. Smart data mining algorithm for intelligent education. *J. Intell. Fuzzy Syst.* **2019**, *37*, 9–16. [CrossRef]
56. Soomro, K.A.; Zai, S.Y.; Jafri, I.H. Competence and usage of Web 2.0 technologies by higher education faculty. *Educ. Media Int.* **2015**, *52*, 284–295. [CrossRef]
57. Marin-Gomez, F.X.; Garcia-Moreno Marchán, R.; Mayos-Fernandez, A.; Flores-Mateo, G.; Granado-Font, E.; Barrera Uriarte, M.L.; Duch, J.; Rey-Reñones, C. Exploring Efficacy of a Serious Game (Tobbstop) for Smoking Cessation During Pregnancy: Randomized Controlled Trial. *JMIR Serious Games* **2019**, *7*, e12835. [CrossRef]
58. Villarreal-Torres, H.; Ángeles-Morales, J.; Cano-Mejía, J.; Mejía-Murillo, C.; Flores-Reyes, G.; Palomino-Márquez, M.; Marín-Rodriguez, W.; Andrade-Girón, D. Classification model for student dropouts using machine learning: A case study. *EAI Endorsed Trans. Scalable Inf. Syst.* **2023**, *10*, 1–12. [CrossRef]
59. Griffiths, P.D.R.; Baudier, P. Enabling Responsible Banking through the Application of Blockchain. *J. Innov. Econ. Manag.* **2023**, *41*, 17–49. [CrossRef]
60. Reis-Andersson, J. Leading the digitalization in K–12 education at the municipality level. *Cogent Educ.* **2024**, *11*, 2368997. [CrossRef]
61. Vinet, L.; Zhedanov, A. A «missing» family of classical orthogonal polynomials. *Antimicrob. Agents Chemother.* **2010**, *58*, 7250–7257. [CrossRef]
62. Mbambo, G.P.; Du Plessis, E.C. Impact of Artificial Intelligence on Teacher Training in Open Distance and Electronic Learning. *Int. J. Learn. Teach. Educ. Res.* **2024**, *23*, 370–386. [CrossRef]
63. Subedi, K.R.; Sharma, S.; Bista, K. Academic Identity Development of Doctoral Scholars in an Online Writing Group. *Int. J. Dr. Stud.* **2022**, *17*, 279–300. [CrossRef]
64. Wang, S.; Wang, H.; Jiang, Y.; Li, P.; Yang, W. Understanding students' participation of intelligent teaching: An empirical study considering artificial intelligence usefulness, interactive reward, satisfaction, university support and enjoyment. *Interact. Learn. Environ.* **2023**, *31*, 5633–5649. [CrossRef]
65. Wazid, M.; Das, A.K.; Park, Y. Blockchain-Envisioned Secure Authentication Approach in AIoT: Applications, Challenges, and Future Research. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 3866006. [CrossRef]
66. Zhang, P.; Xu, X.; Qin, X.; Liu, Y.; Ma, N.; Han, S. Evolution Toward Artificial Intelligence of Things Under 6G Ubiquitous-X. *J. Harbin Inst. Technol.* **2020**, *27*, 116–135. [CrossRef]
67. Upreti, K.; Dadhich, P.; Gautam, A.; Singh, J.; Kushwah, V.S.; Parashar, J.; Gangwar, D.; Ghosh, S. Investigation on the analysis of integration of IoT and AI technologies with information security for advanced education 4.0. *J. Discret. Math. Sci. Cryptogr.* **2024**, *27*, 1445–1454. [CrossRef]
68. Kondoyanni, M.; Loukatos, D.; Arvanitis, K.G.; Lygkoura, K.-A.; Symeonaki, E.; Maraveas, C. Adding Machine-Learning Functionality to Real Equipment for Water Preservation: An Evaluation Case Study in Higher Education. *Sustainability* **2024**, *16*, 3261. [CrossRef]
69. Ahajjam, T.; Moutaib, M.; Aissa, H.; Azrour, M.; Farhaoui, Y.; Fattah, M. Predicting Students' Final Performance Using Artificial Neural Networks. *Big Data Min. Anal.* **2022**, *5*, 294–301. [CrossRef]
70. Brand, C.E.J.; Gabriel Ramirez, V.M.; Diaz, J.; Moreira, F. Toward Educational Sustainability: An AI System for Identifying and Preventing Student Dropout. *Rev. Iberoam. Tecnol. Aprendiz.* **2024**, *19*, 100–110. [CrossRef]
71. Sallan, G.; Behal, S. Prediction of student dropout using enhanced machine learning algorithm. *Adv. Math. Sci. J.* **2020**, *9*, 3821–3826. [CrossRef]
72. Yao, K.-C.; Huang, W.-T.; Hsu, L.-C.; Yang, C.-K.; Lai, J.-Y. Evaluation of the Established IoT Smart Home Robot Teaching Module Based on Embedded Thematic-Approach Strategy. *Math. Probl. Eng.* **2020**, *2020*, 1–10. [CrossRef]

73. Mouta, A.; Torrecilla-Sánchez, E.M.; Pinto-Llorente, A.M. Design of a future scenarios toolkit for an ethical implementation of artificial intelligence in education. *Educ. Inf. Technol.* **2024**, *29*, 10473–10498. [CrossRef]
74. Wang, M.; Zhang, L. Teacher leadership in professional learning communities in early childhood education: A Chinese perspective. *Educ. Manag. Adm. Leadersh.* **2024**, *2024*, 17411432241240604. [CrossRef]
75. Montie, J.E.; Xiang, Z.; Schweinhart, L.J. Preschool experience in 10 countries: Cognitive and language performance at age 7. *Early Child. Res. Q.* **2006**, *21*, 313–331. [CrossRef]
76. Dassanayaka, D.M.C.; Weerasinghe, I.M.S.; Dahanayaka, S.N.S.; Dedunu, H.H.; Athurupana, K.A.H.M.K.; Muthuweera, N.K.G. Academics' attitudes towards online education amidst the Covid-19 outbreak. *Int. J. Educ. Manag.* **2022**, *36*, 661–677. [CrossRef]
77. Ebrahim, E.A.; Cengiz, M.A. Predicting Verbal Learning and Memory Assessments of Older Adults Using Bayesian Hierarchical Models. *Front. Psychol.* **2022**, *13*, 855379. [CrossRef] [PubMed]
78. Bakhov, I.; Bilous, N.; Saiko, M.; Isaienko, S.; Hurinchuk, S.; Nozhovnik, O. Beyond the Dictionary: Redefining Translation Education with Artificial Intelligence-Assisted App Design and Training. *Int. J. Learn. Teach. Educ. Res.* **2024**, *23*, 118–140. [CrossRef]
79. Salhab, R. AI Literacy Across Curriculum Design: Investigating College Instructors' Perspectives. *Online Learn. J.* **2024**, *28*, 22–47. [CrossRef]
80. Walter, Y. Embracing the future of Artificial Intelligence in the classroom: The relevance of AI literacy, prompt engineering, and critical thinking in modern education. *Int. J. Educ. Technol. High. Educ.* **2024**, *21*, 1–29. [CrossRef]
81. Churchill, D.; Churchill, N. Educational affordances of PDAs: A study of a teacher's exploration of this technology. *Comput. Educ.* **2008**, *50*, 1439–1450. [CrossRef]
82. Wang, J.; Sun, G.; Gu, Y.; Liu, K. ConGradetect: Blockchain-based detection of code and identity privacy vulnerabilities in crowdsourcing. *J. Syst. Archit.* **2021**, *114*, 101910. [CrossRef]
83. Su, X.-H.; Zheng, W.-N.; Jiang, Y.; Wei, H.-W.; Wan, J.-Y.; Wei, Z.-Y. Research and Progress on Learning-Based Source Code Vulnerability Detection. *Chin. J. Comput.* **2024**, *47*, 337–374. [CrossRef]
84. Yang, J.; Onik, M.; Lee, N.-Y.; Ahmed, M.; Kim, C.-S. Proof-of-Familiarity: A Privacy-Preserved Blockchain Scheme for Collaborative Medical Decision-Making. *Appl. Sci.* **2019**, *9*, 1370. [CrossRef]
85. Zabaniotou, A. Learning outcomes of leader and leadership development: Method of Soft and Female leadership (SFL) empirical framework for driving the gender equality change in Mediterranean Engineering Education. *Open Res. Eur.* **2021**, *1*, 63. [CrossRef]
86. Chen, Y.; Mazlack, L.J.; Minai, A.A.; Lu, L.J. Inferring causal networks using fuzzy cognitive maps and evolutionary algorithms with application to gene regulatory network reconstruction. *Appl. Soft Comput. J.* **2015**, *37*, 667–679. [CrossRef]
87. Salama, K.M.; Abdelbar, A.M. Learning cluster-based classification systems with ant colony optimization algorithms. *Swarm Intell.* **2017**, *11*, 211–242. [CrossRef]
88. Isiaku, L.; Kwala, A.F.; Sambo, K.U.; Isiaku, H.H. Academic Evolution in the Age of ChatGPT: An In-depth Qualitative Exploration of its Influence on Research, Learning, and Ethics in Higher Education. *J. Univ. Teach. Learn. Pract.* **2024**, *21*, 1–25. [CrossRef]
89. Abdalkareem, M.; Min-Allah, N. Explainable Models for Predicting Academic Pathways for High School Students in Saudi Arabia. *IEEE Access* **2024**, *12*, 30604–30626. [CrossRef]
90. Spector, J.M. Emerging educational technologies and research directions. *Educ. Technol. Soc.* **2013**, *16*, 21–30.
91. Costa, E.B.; Fonseca, B.; Santana, M.A.; de Araújo, F.F.; Rego, J. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Comput. Hum. Behav.* **2017**, *73*, 247–256. [CrossRef]
92. Sghir, N.; Adadi, A.; Lahmer, M. Recent advances in Predictive Learning Analytics: A decade systematic review (2012–2022). *Educ. Inf. Technol.* **2023**, *28*, 8299–8333. [CrossRef]
93. Alé-Ruiz, R.; Martínez-Abad, F.; del Moral-Marcos, M.T. Academic engagement and management of personalised active learning in higher education digital ecosystems. *Educ. Inf. Technol.* **2024**, *29*, 12289–12304. [CrossRef]
94. Nanglae, L.; Iam-On, N.; Boongoen, T.; Kaewchay, K.; Mullaney, J. Determining patterns of student graduation using a bi-level learning framework. *Bull. Electr. Eng. Inform.* **2021**, *10*, 2201–2211. [CrossRef]
95. Lescano-Delgado, M. Avances en el uso de inteligencia artificial para la mejora del control y la detección de fraudes en organizaciones. *Rev. Científica Sist. Informática* **2024**, *4*, e671. [CrossRef]
96. Hamal, O.; El Faddouli, N.-E.; Harouni, M.H.A.; Lu, J. Artificial Intelligent in Education. *Sustainability* **2022**, *14*, 2862. [CrossRef]
97. Verma, A.; Singh, A.; Anand, D.; Aljahdali, H.M.; Alsubhi, K.; Khan, B. IoT Inspired Intelligent Monitoring and Reporting Framework for Education 4.0. *IEEE Access* **2021**, *9*, 131286–131305. [CrossRef]
98. Jacinto, C. Redistribution and affectivity as dimensions of social justice. The interventions of the Third Sector in the transitions of the education to the work. *Profr.-Rev. Curric. Form. Profr.* **2017**, *21*, 177–195.
99. Farhood, H.; Joudah, I.; Beheshti, A.; Muller, S. Advancing student outcome predictions through generative adversarial networks. *Comput. Educ. Artif. Intell.* **2024**, *7*, 100293. [CrossRef]

100. Mergen, M.; Meyerheim, M.; Graf, N. Reviewing the current state of virtual reality integration in medical education—A scoping review protocol. *Syst. Rev.* **2023**, *12*, 97. [CrossRef] [PubMed]
101. Nnadi, L.C.; Watanobe, Y.; Rahman, M.M.; John-Otumu, A.M. Prediction of Students' Adaptability Using Explainable AI in Educational Machine Learning Models. *Appl. Sci.* **2024**, *14*, 5141. [CrossRef]
102. Nagy, M.; Molontay, R. Interpretable Dropout Prediction: Towards XAI-Based Personalized Intervention. *Int. J. Artif. Intell. Educ.* **2024**, *34*, 274–300. [CrossRef]
103. Janahi, Y.M.; AlDhaen, E.; Hamdan, A.; Nureldeen, W.A. Emerging technologies for digitalized learning in higher education. *Dev. Learn. Organ.* **2023**, *37*, 29–31. [CrossRef]
104. Rasheed, Z.; Ghwanmeh, S.; Abualkishik, A.Z. Harnessing Artificial Intelligence for Personalized Learning: A Systematic Review. *Data Metadata* **2023**, *2*, 146. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

computers

MDPI

Article

A Proposed Method of Automating Data Processing for Analysing Data Produced from Eye Tracking and Galvanic Skin Response

Javier Sáez-García [1,*], María Consuelo Sáiz-Manzanares [2,*] and Raúl Marticorena-Sánchez [3]

[1] Higher Polytechnic School (Campus Milanera), Universidad de Burgos, 09001 Burgos, Spain
[2] Department of Health Sciences, GIR DATAHES, UIC JCYL No. 348, Universidad de Burgos, 09001 Burgos, Spain
[3] ADMIRABLE Research Group, Department of Computer Engineering, Higher Polytechnic School (Campus Vena), Universidad de Burgos, 09006 Burgos, Spain; rmartico@ubu.es
* Correspondence: jsg1013@alu.ubu.es (J.S.-G.); mcsmanzanares@ubu.es (M.C.S.-M.)

Abstract: The use of eye tracking technology, together with other physiological measurements such as psychogalvanic skin response (GSR) and electroencephalographic (EEG) recordings, provides researchers with information about users' physiological behavioural responses during their learning process in different types of tasks. These devices produce a large volume of data. However, in order to analyse these records, researchers have to process and analyse them using complex statistical and/or machine learning techniques (supervised or unsupervised) that are usually not incorporated into the devices. The objectives of this study were (1) to propose a procedure for processing the extracted data; (2) to address the potential technical challenges and difficulties in processing logs in integrated multichannel technology; and (3) to offer solutions for automating data processing and analysis. A Notebook in Jupyter is proposed with the steps for importing and processing data, as well as for using supervised and unsupervised machine learning algorithms.

Keywords: eye tracking; galvanic skin response; behavioural monitoring; learning process; data processing

1. Introduction

The use of eye tracking technology, together with other physiological biomarker measurements such as psychogalvanic skin response (GSR) and electroencephalographic recordings (EEG), provides researchers with information about users' physiological behavioural responses during their learning processes when performing various types of tasks. These records are obtained through sensors and provide non-subjective information that is put together with other data from the user (learning outcomes, perceived satisfaction with the learning process, perception of the use of metacognitive strategies, etc.). This technology has been applied in a range of fields, each with different objectives. For example, in marketing it has been used to understand client behaviour in order to achieve customer loyalty and increase sales [1–11]. In healthcare sciences, this technology can be applied for diagnosis (e.g., mental illness [12], autism spectrum disorder [13–15], examining medical imaging [16], mental fatigue [17], application in various clinical needs [18], Alzheimer's [19], dysfunctional cerebral networks [20], various disorders [21], Rett's Syndrome [22], and activity with prosthetic limbs [23]). Eye tracking technology has also been used in therapeutic intervention (for example, in Autism Spectrum Disorder [15], cerebral palsy [24], in patients with depression [25,26], in Parkinson's [27], with various addictions [28], and for anorexia nervosa [29]). It has also been used in education in order to understand how learners process information when doing various tasks, with the ultimate aim being to improve students' learning processes. Eye tracking technology can be used in

educational settings to analyse different learning styles [30], for educational rehabilitation from problems related to dyslexia [31], and to examine the effectiveness of educational methodologies such as evidenced-based learning [32].

More specifically in educational settings, this technology is commonly used in higher education, especially with health sciences students [33–39]. One of the specific functions the technology offers is that it can give the teacher or researcher data that, once processed and analysed, will help in the design of explanatory models of how learning occurs in different types of learners [30].

Eye tracking technology is also being used in driving to understand the characteristics of different drivers in pursuit of better driving and fewer accidents [40,41]. It is also being used to test different effects on driving of things such as advertising [42].

Recent studies on human perception indicate that higher levels of attention are related to variables such as gender, age, and educational level. Studies on vigilance indicate that higher levels of self-attention [35] are related to shorter total fixation durations and fewer fixation counts while the user looks at picture stimuli [40]. Also, the use of animated images seems to improve learning outcomes [41]. Similarly, recent studies call for increased research into the effects of self-regulated learning strategy (SRL) training on student motivation, engagement and performance [43]. Specifically, using think-aloud protocols involving self-regulated strategies appears to improve learning outcomes [44]. Similarly, feedback on user performance appears to enhance deep learning in learners [45]. Along these lines, using materials that include self-regulated simulation activities has been shown to be very effective in increasing information acquisition and mitigating the effects of prior knowledge about the task on learning outcomes [46].

In other studies we have addressed the meanings of metrics from applying eye tracking technology integrated with GSR devices and their relationship to the use of cognitive and metacognitive strategies as well as learning profiles [37,47,48]. One reference that may help researchers to prepare metrics is Holmqvist et al. [49], a summary of which can be found in Table 1.

Table 1. Some representative metrics and interpretation in integrated multichannel eye tracking.

Metric	Unit of Measurement	Meaning	Interpretation in the Context of the Human Learning
Fixation count	Count	Number of fixations on a stimulus or part of a stimulus.	A higher number of fixations may be related to a difficulty in processing that information because it is novel or difficult for the learner to process.
Fixation duration	Milliseconds	Duration of fixation.	Refers to the reaction times of the learner. A longer duration may be related to a higher cognitive load in the processing of the stimulus. It may also be related to the use of metacognitive orientation strategies, i.e., searching for information or relating it to previous knowledge.
Saccade count	Count	Refers to the shift of gaze from one part of the stimulus to another.	A higher number of saccades implies the use of metacognitive orientation strategies, i.e., searching for information or relating it to prior knowledge. Likewise, the greater the amplitude of the saccade, the lower the cognitive effort, although this may also be related to information processing problems. Younger learners apply shorter saccades.
Pupil diameter	Millimetres	The mean pupil diameter is collected for all fixations within an AOI during a time interval.	It refers to the interest that a stimulus or part of it has for the learner. A larger pupil diameter may be related to increased cognitive load and/or difficulty in processing a task.
Number of visits		Number of bindings within a stimulus or part of a stimulus (of an area of interest or AOI).	Refers to attention to a stimulus or part of a stimulus.
Scan Path Length or Gaze Point	X and Y position coordinates	Refers to the chain of fixations in order of succession.	It involves a pattern of visual tracking on a stimulus. It gives information about how each learner processes information.

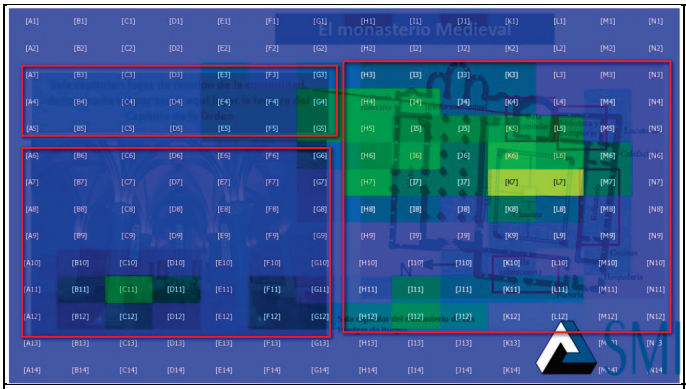
Table 1. *Cont.*

Metric	Unit of Measurement	Meaning	Interpretation in the Context of the Human Learning
SCR count	Count	The number of skin conductance responses (SCRs), for each Interval in Time of Interest.	It provides information about the emotional state of a learner. The SCR count can be used to identify which specific moments of a dynamic stimulus and specific information within a stimulus elicit an emotional response in a learner. A higher SCR count indicates a higher level of emotional arousal.
ER SCR Amplitude	Microsiemens	The amplitude of each event related skin conductance response (ER-SCR), for each Interval in Time of Interest. Time of Interest intervals that do not have an ER-SCR are calculated using filtered GSR data.	When an SCR occurs between 1 and 5 s after an event (ER-SCR), it is considered whether the stimulus elicits an emotional reaction in the learner. This measure can provide information about different emotional states such as: anxiety, stress, frustration, or relaxation.
GSR average	Microsiemens	The mean of the average galvanic skin response (GSR) signal, after filtering for each time of interest.	When environmental conditions are held constant, slow fluctuations in the GSR signal (between seconds and minutes) reflect changes in the participant's emotional arousal. The researcher can use the average GSR metric in different sections of the session to determine whether a learner might be stressed, frustrated, or relaxed during the course of an experiment.

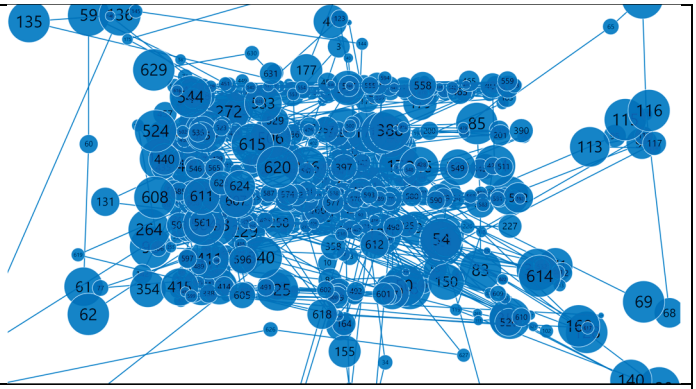
How to Extract Records from Eye Tracking Devices and How to Process and Analyse Records

The devices that apply eye tracking technology include records of static metrics (fixations, saccades, pupil diameter, blinking, etc.). They also apply various measuring parameters to those metrics such as velocity, duration, etc. These metrics are found by applying statistical resources such as frequencies, means, standard deviations, etc. (examples of these measures according to parameter may be found in Table 1).

Eye tracking devices also include measurement of dynamic metrics (these involve the points where the participant is looking on the screen's cartesian co-ordinates), called gaze point or scan path. An example is shown in Figure 1.



Example of gaze point produced using an eye tracking device: iViwe XTM, SMI Experimenter Center 3.0, and SMI BeGazeTM and a monitor with a resolution of 1680 × 1050



Example of a scan path produced by a Tobii Pro Lab v. 1.241.54542 and a monitor with a resolution of 1920 × 1080

Figure 1. Examples of gaze point and scan path.

Nonetheless, more thorough analysis of dynamic metrics requires the application of machine learning techniques. These may be supervised techniques, which include techniques such as regression (about predicting numerical values), including linear regression, Support Vector Machine (SVM), Decision Tree, and Neural Networks. The other supervised techniques are classification (about predicting categories). These include SVM, discriminant analysis, nearest neighbour, and naïve Bayes. Machine learning techniques may also be unsupervised (which are applied without using labels in order to discover hidden patterns).

Unsupervised techniques include clustering algorithms (which help to group data based on similarity, such as: *k*-means, *k*-means++, etc.), and are applied to reduce dimensionality (such as in Principal Component Analysis—PCA) [50,51].

That said, it is important to remember that the data that can be extracted from eye tracking devices are unprocessed. This means that, in order to be able to apply statistical techniques or machine learning algorithms (supervised or unsupervised), they must first be processed. According to García et al. [52], pp. 10–13, the steps in the procedure are:

1. Produce the database.
2. Define the problem.
3. Understand the problem.
4. Process the data, with the following sub-steps:
    - 4.1. Prepare the data: clean the data, which includes operations to correct erroneous data, filter incorrect data from the dataset, and reduce unnecessary data. Other data cleaning tasks involve detecting discrepancies and dirty data (fragments of original data that make no sense)—these tasks are more closely related to understanding the original data and usually need human involvement in the process—and transformation of the data (in this step, the data are converted or consolidated so that the result of the data-mining process can be applied or be more effective). Sub-tasks within transformation include smoothing, construction of characteristics, and data aggregation or data summary; these require human supervision.
    - 4.2. Integrate the data (covers the fusion of data from various data stores). This process is applied to avoid redundancies and inconsistencies in the resulting dataset. Typical operations in data integration are identification and unification of variables and domains, analysis of the correlation of attributes, duplication of tuples, and detection of conflicts in data values from different sources.
    - 4.3. Normalise the data (this refers to the unity of the measure used, as this may affect the data analysis). This means that all of the attributes must be expressed in the same units of measurement and must use a common scale or range. This step is particularly useful in statistical learning methods.
    - 4.4. Impute missing data (this is a form of cleaning data, the aim of which is to fill in missing values using a reasonable estimation via different methods, such as applying a mean value).
    - 4.5. Identify noise (this refers to the act of smoothing in data transformation, the main objective of which is to detect random errors or variance in a measured variable).
5. Apply machine learning techniques.
6. Evaluation.
7. Exploit the results.

In summary, using integrated multichannel eye tracking technology can be of great help to professionals in various fields (marketing, health, education, driving, etc.). However, the processed information these devices provide is very basic. Therefore, we believe that ordinary users (researchers, teachers, professionals, etc.) should be able to get information about the results simply and easily [53]. Current research indicates that machine learning techniques can help make data analysis more effective, as they help people to interpret the data [54–57]. Many concepts in machine learning are related to ideas that were previously addressed in psychology and neuroscience [58].

Therefore, our research field could benefit from proposals for automating the preprocessing and processing [52] of data recorded by devices applying integrated eye tracking techniques (i.e., data from eye tracking and other data such from integrated GSR devices or electroencephalographic response (EEG)). There is a gap between the output of such software and the data analysis that researchers would perform. The present paper seeks to

bridge that gap, aiming to explore this issue in greater depth in order to help researchers in this field.

In view of the above, the objective of the current study was to suggest a procedure for processing data extracted from these types of devices by applying integrated multichannel, eye tracking technology.

2. Materials and Methods

2.1. Participants

A convenience sample was used to illustrate the proposed procedure to readers by example. More specifically, the study used a sample of 17 students (nine women and eight men) in the fourth year of a biomedical engineering degree at the University of Burgos. They were between 21 and 22 years old, with a standard deviation for both genders between 0.7 and 0.9 years. Participation was voluntary and without any financial reward. The students were told about the study objectives and signed their informed consent before taking part. In addition, the study was approved beforehand by the University of Burgos Bioethics Committee (report no. IO 03/2022).

2.2. Instruments

(a) Tobii Pro Lab v. 1.241.54542 [59]. This is a hardware and software platform that is used by researchers to analyse human behaviour. It has a high degree of flexibility and can be used to perform advanced studies on attention and provide deep analysis of cognitive processes. This software was chosen due to it currently being one of the most widely used in research [60], as it makes it easy to integrate other measuring instruments, such as GSR.

(b) Shimmer 3 GSR+ [61]. The Shimmer3 GSR+ (Galvanic Skin Response) unit provides connections and pre-amplification for a Galvanic Skin Response (Electrodermal Resistance Measurement—EDR/Electrodermal Activity (EDA) data acquisition channel). The GSR+ unit is suitable for measuring the electrical characteristics or conductance of the skin. The device is compatible for use in conjunction with Tobii Pro Lab v. 1.241.54542 and the metrics are integrated with those obtained with the Tobii Pro Lab v. 1.241.54542 device.

(c) Virtual Laboratory 8—Eye tracking technology applied to Early Intervention II. A virtual classroom (also called a lab) in the eEarlyCare-T Research Project. This classroom is free and open access after logging into the platform https://www2.ubu.es/eearlycare_t/en/project (accessed on 8 August 2024) [62]. Images from the virtual classroom are provided in Figure A1.

(d) Python libraries: Pandas, sklearn, matplotlib, seaborn.

2.3. Procedure

The objective of this study was to suggest a procedure for processing data extracted from an eye tracking device that incorporated GSR recording. More specifically, the study used a Tobii Pro Lab v. 1.241.54542 [59] with Shimmer 3 GSR+ [61]. It addressed the technical difficulties and record processing challenges that users may have to deal with when using this integrated multichannel technology. Lastly, the study also aimed to propose potential solutions for automating data analysis via the design of a Jupyter Notebook in which machine learning algorithms were applied.

Researchers do not normally work with very large samples of participants because the experimental process for these types of studies is laborious and needs to be tailored. This means spending a lot of time on each user in each experimental phase. Despite the small numbers of participants, each one produces a large amount of diverse records, meaning a huge volume of data. Integrated multichannel eye tracking devices facilitate the processing and analysis of data through frequency analysis—for example, through data extracted using the heat map model, which visualises the most common gaze points recorded for the object of study, be that an image, text, video, or, web page. Figure 2 shows an example.

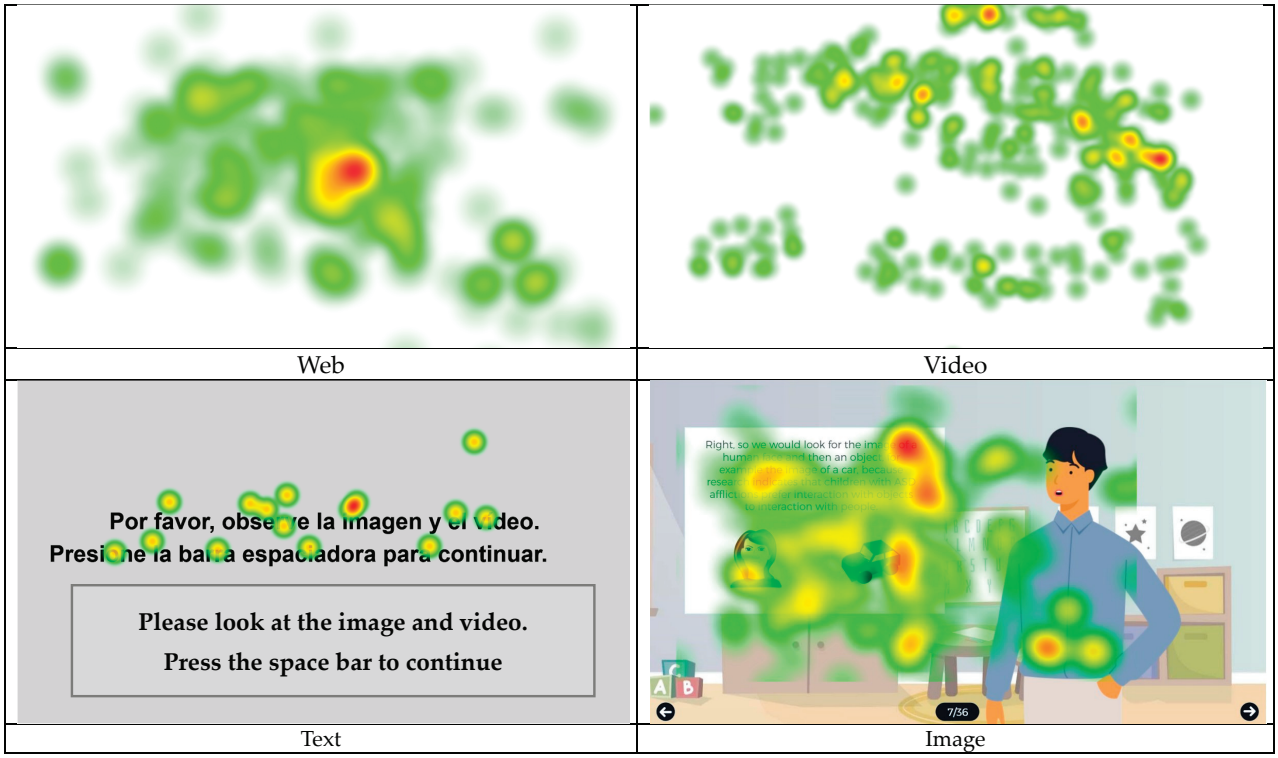


Figure 2. Heat Map for different stimuli (web, video, text, and image).

In addition, integrated multi-channel eye tracking technology provides information about the gaze positioning chain in hierarchical order of appearance. This is a dynamic measure called scan path or gaze point, an example of which is shown in Figure 3.

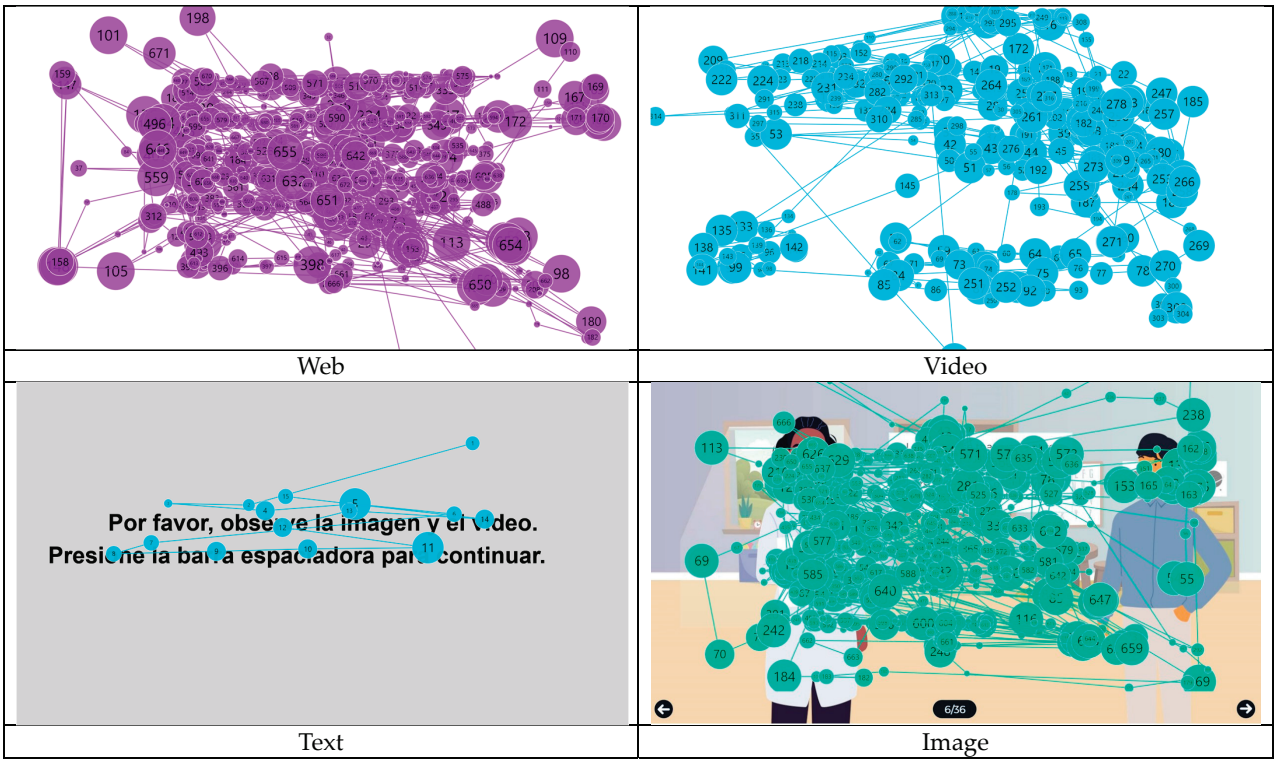


Figure 3. Gaze Point in different stimuli (web, video, text, and image).

Furthermore, a variety of metrics can be extracted for each participant once Areas of Interest (AOIs) have been established. AOIs can be divided into relevant AOIs and non-relevant AOIs. Relevant AOIs refer to the essential information in the learning process. The users should pay more attention to them. Non-relevant AOIs refer to information that the user should not pay attention to. Table 2 gives an example of some of these metrics, the meaning of the record, and the unit of measurement. These metrics refer specifically to the Tobii Pro Lab technology v. 1.241.54542 [59]. In addition, researchers can extract metrics for saccades and fixations about parameters of duration, start, end, mapping, direction, etc. These records can be produces in various formats, including .xlsx, .tsv, .plof, etc.

Table 2. Most representative metrics, meaning, and unit of measurement.

Metric	Meaning	Unit of Measurement
Fixation point	The normalized horizontal and vertical coordinate of the fixation point.	Normalized coordinates (DACS)
Average pupil diameter	The average diameter of the pupil of the fixation. Calculated using the resulting pupil diameter after applying pupil diameter.	Millimeters
Saccade direction	The angle of the straight line between the preceding fixation and succeeding fixation. This can only to applied to whole saccades.	Degrees
Average velocity	The average velocity across all samples belonging to the saccade, even outside the interval.	Degrees/second
Peak velocity	The maximum velocity across all samples belonging to the saccade, even outside the interval.	Degrees/second
Saccade amplitude	The amplitude for whole saccades.	Degrees
Mouse position	The position of the mouse.	Pixels (DACS)
Galvanic skin response (GSR)	The raw galvanic skin response signal of the participant.	Microsiemens
Average GSR	The average galvanic skin response (GSR) signal, after filtering, for an interval.	Microsiemens
Number of SCR	The number of skin conductance responses (SCRs) for an interval.	Count
Amplitude of event related SCR	The amplitude of each event-related skin conductance response (ER-SCR), for an interval. ER-SCRs are calculated using filtered GSR data.	Microsiemens

However, in order to be able to analyse the recorded data, researchers must process and analyse most of these records. Furthermore, once processed, statistical or machine learning techniques (supervised or unsupervised) must be applied in order to analyse and subsequently interpret the data [52]. Figure 4 shows an outline of the process that should be followed. It applies a more complex log analysis procedure than the default analysis offered by the devices.

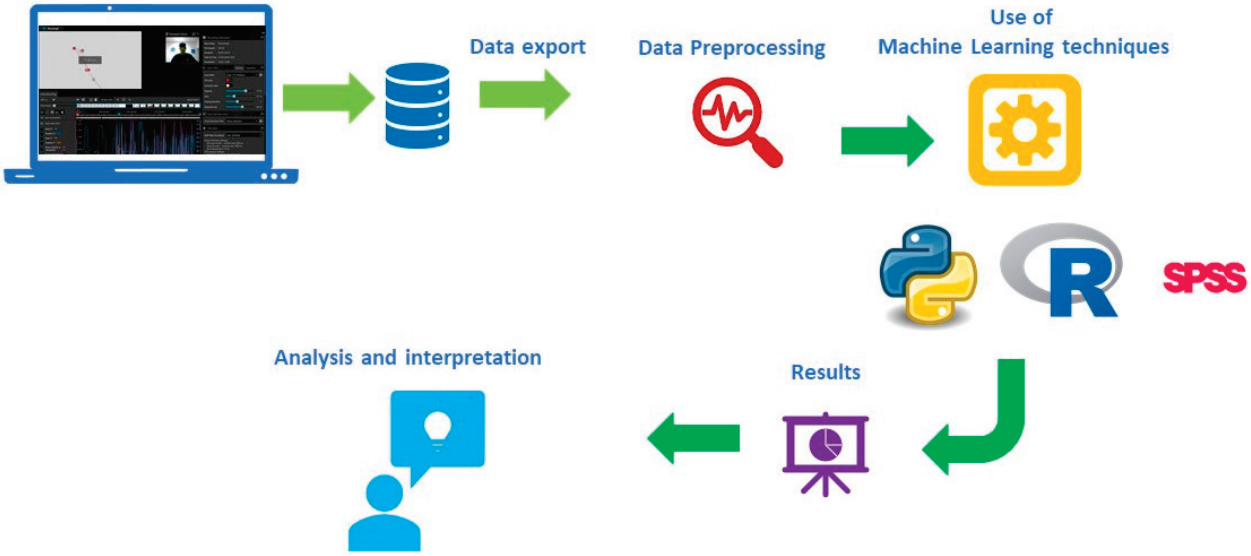


Figure 4. Procedure for analysing records produced with integrated multichannel eye tracking technology.

2.4. Data Analysis

The objective of this study was not to test any research hypotheses, but rather to suggest how to create a Notebook in Jupyter as a way of providing an example of how one might process data recorded via integrated multichannel eye tracking technology (in this case incorporating a GSR device). More specifically, we applied supervised machine learning algorithms (regression—linear regression in this case) and unsupervised machine learning algorithms (clustering: using *k*-means) using the Python libraries: pandas, sklearn, matplotlib, and seaborn.

3. Results

A Notebook was made in Jupyter with the steps for importing and processing the database extracted from Tobii Pro Lab v. 1.241.54542 [59], after a pre-processing phase. In this phase, all detected noise was cleaned. The steps to be followed are described below.

Step 1. Clean and filter the files (one file per participant was exported), in this case, they were extracted in .xlsx format.
Step 2. Import libraries to read the extracted files.

```
Import
import pandas as pd
```

Step 3. Create a function to concatenate the files

Input:

- data_folder: path to the folder containing data files
- extension: file extension (e.g., .xlsx)

Output:

- concatenated_data: A DataFrame formed by concatenating all the files.

Process:

1. **Initialize** files as the list of files in data_folder.
2. **Initialize** an empty list list_dataframes.
3. **For** each file in files:
    - Set full_path as the complete path of the file.
    - **If** full_path is a valid file and its extension matches extension:
        - **If** the extension is .xlsx:
            - Display message: ‘‘Reading file i + 1/len(files): file’’.
            - Read the file into a DataFrame df.
            - Append df to list_dataframes.
        - **Else**: Continue to the next file.
    - **If** an error occurs while reading the file:
        - Display message: ‘‘Error reading file file: error_message’’.
4. **If** list_dataframes is not empty:
    - Concatenate all DataFrames in list_dataframes into concatenated_data.
    - Display message: ‘‘All files have been processed!’’.
    - Return concatenated_data.
5. **Else**:
    - Display message: ‘‘No files were found for processing’’.
    - Return None.

Step 4. Data Preprocessing.

Input:

- data: DataFrame containing the data

- `delete_rows:` dictionary where keys are column names and values are rows to be deleted
- `filter_rows:` dictionary where keys are column names and values are rows of interest
- `select_columns:` list containing the columns to be kept

`Output:`

- `processed_data:` A DataFrame with the processed data.

`Process:`

1. **For** each `column, values` **in** `delete_rows:`
   - Remove rows from data where the column contains any of the values.
2. **For** each column, `values` **in** `filter_rows:`
   - `Filter data to keep only rows where the column contains any of the values.`
3. **Select** columns of interest in `data using select_columns`.
4. **Return** `processed_data`, which is now the processed DataFrame.

Step 5. Data Processing with Filters and Selection.

`Process:`

1. **Create a dictionary** `delete_rows:`
   - Keys: 'Participant name'
   - Values: ['Test XX 2', 'Participant_X3', 'NP_01', 'NP_01_2', 'NP_02']
   - Purpose: To remove rows where the 'Participant name' matches any value in the list.
2. **Create a dictionary** `filter_rows:`
   - Keys: 'Sensor'
   - Values: ['Mouse', 'GSR']
   - Purpose: To filter rows where the 'Sensor' column contains only 'Mouse' and 'GSR'.
3. **Create a list** `select_columns:`
   - ['Participant name', 'Gender', 'Audio', 'Recording duration', 'Eye movement type', 'Galvanic skin response (GSR)']
   - Purpose: To select the columns of interest from the DataFrame.
4. **Applythe processing:**
   - Call `processing_Data` function with the parameters: `data`, `delete_rows`, `filter_rows`, and `select_columns`.
5. **Return the processed data** `as processed_data.`

Step 6. Data grouping and aggregation

`Input`:

- `data:` DataFrame containing participant data

`Output`:

- `grouped:` DataFrame with data grouped by 'Participant name' and aggregated metrics

`Process`:

1. **Group** the `data` by 'Participant name', applying the following aggregation functions:
   - **Gender**: Select the first value (assuming gender is consistent across all rows).
   - **Audio:** Select the first value (assuming the audio characteristic is consistent across all rows).
   - **Recording duration:** Select the first value (assuming test duration is the same across all rows).
   - **Eye movement type:**
     - **If** both 'Fixation' and 'Saccade' exist in the column:

- Compute the ratio of 'Fixation' events to 'Saccade' events.
- **Else:** Return None.

- **Galvanic skin response (GSR):** Compute the mean value for GSR.

2. **Reset the index** of the grouped DataFrame.
3. **Rename** the 'Eye movement type' column to 'Fixation to Saccade Ratio' for a more descriptive name.
4. **Return** the grouped DataFrame.

Step 7. Load data

`Input`:

- `location`: File path of the Excel file

`Output`:

- `data`: Loaded DataFrame from the Excel file

`Process`:

1. **Load** the Excel file from `location` into a DataFrame `data`
2. **Return** the `data`

Step 8. Combine data

`Input`:

- `data1`: First DataFrame
- `data2`: Second DataFrame
- `column`: Column name on which to merge both DataFrames

`Output`:

- `data`: Merged DataFrame based on the common column

`Process`:

1. **Merge** `data1` and `data2` on the column
2. **Return** the merged `data`

An example is presented in Figure 5.

	Participant name	Gender	Visualisation	Recording duration	Fixation to Saccade Ratio	Galvanic skin response (GSR)
0	NP_01	Female	no audio	237413	7.327740	1.827789
1	NP_02	Male	no audio	134613	8.908907	2.833627
2	NP_03	Female	no audio	276957	6.652087	2.613371
3	NP_04	Female	with audio	264543	8.836283	0.582247
4	NP_05	Female	no audio	321678	5.249753	0.858351
5	NP_06	Female	no audio	296533	7.607504	0.575296
6	NP_07	Male	no audio	323631	9.199501	2.142786
7	NP_08	Female	no audio	241836	7.555053	6.657798
8	NP_09	Male	no audio	237535	7.835564	0.044381
9	NP_10	Male	no audio	250432	6.875283	6.913355
10	NP_11	Male	with audio	255166	7.862782	14.114906
11	NP_12	Male	with audio	220215	8.055693	0.013019
12	NP_13	Female	no audio	246682	7.146648	0.181956
13	NP_14	Male	no audio	399585	7.582560	0.184202
14	NP_15	Female	no audio	249501	10.839836	0.082633
15	NP_16	Male	no audio	267439	9.540541	10.012614
16	NP_17	Female	no audio	241825	5.634638	0.197920

Figure 5. DataFrame of the data grouped by participants.

An example of the final integration of the data is presented in Figure 6.

Participant name	Gender	Recording durantion	Fixation Saccade Ratio	Glavanic skin response (GSR)	Visualisation Lab	Visualisation Type	Test success
NP_01	Female	237413	7.327740	1.827789	1	0	2
NP_02	Male	134613	8.908907	2.833627	1	0	5
NP_03	Female	276957	6.652087	2.613371	1	0	4
NP_04	Female	264543	8.836283	0.582247	1	1	0
NP_05	Female	321678	5.249753	0.858351	0	1	0
NP_06	Female	296533	7.607504	0.575296	1	0	3
NP_07	Male	323631	9.199501	2.142786	1	0	2
NP_08	Female	241836	7.555053	6.657798	1	0	1
NP_09	Male	237535	7.835564	0.044381	1	0	3
NP_10	Male	250432	6.875283	6.913355	1	0	4
NP_11	Male	255166	7.862782	14.114906	0	1	0
NP_12	Masculino	220215	8.055693	0.013019	0	1	0
NP_13	Female	246682	7.146648	0.181956	1	1	2
NP_14	Male	399585	7.582560	0.184202	1	0	0
NP_15	Female	249501	10.839836	0.082633	0	0	1
NP_16	Male	267439	9.540541	10.012614	1	0	0
NP_17	Female	241825	5.634638	0.197920	1	0	0

Figure 6. Final data integration.

Step 9. Clustering preparation

`Input:`

- `data`: DataFrame containing the original data

`Output:`

- `data_normalized`: DataFrame with encoded, imputed, and normalized data ready for clustering

`ProcAess`:

1. **Detect Boolean variables:**
   - Identify Boolean columns.
   - Store these column names in `booleans`.
2. **Detect non-Boolean categorical variables:**
   - Identify categorical columns.
   - Store these column names in `categorical`.
3. **Encode categorical variables:**
   - Convert categorical variables into Boolean (binary) variables.
   - Store the result in `data_encoded`.
4. **Convert Boolean columns** to binary values:
   - **For** each column in `booleans`:
     - Convert the column values to binary (`0` or `1`).
5. **Handle missing values:**
   - Apply a **mean imputation** strategy to fill missing values:
     - Impute missing values using mean in `data_encoded`.
     - Store the imputed data in data_filled.
6. **Normalize the data:**
   - Apply **StandardScaler** to normalize the data:

- Scale the imputed data so that all features have the same scale.
- Store the normalized data in `data_normalized`.

7. **Return** the `data_normalized`.

Final Step: Apply Clustering Preparation to Combined Data

1. **Call** `clustering_preparation(data_combined)` to normalize and prepare the data for clustering.
2. **Return** `data_normalized`.

Algorithm: K-Means Clustering (Elbow Method)

Input:

- `data`: DataFrame containing the preprocessed data for clustering
- `k_range`: List of values representing the range of potential cluster numbers

Output:

- `inertia`: List of inertia values for different cluster sizes

Process:

1. **Initialize** an empty list `inertia` to store the inertia values for each K.
2. **For each value of** k in `k_range`:
   - **Create** a KMeans object with the following parameters:
     - `n_clusters` = k: Specifies the number of clusters.
     - `random_state = 42`: Ensures reproducibility.
     - `n_init = 10`: Number of times the KMeans algorithm will be run with different initializations.
   - **Fit** the KMeans model to `data` using the `fit` method.
   - **Append** the `inertia` value of the current KMeans model to the `inertia` list.

Algorithm: K-Means Clustering Visualization

Input:

- `data`: Original DataFrame with all data (including the cluster assignment)
- `normalised_data`: Data that has been preprocessed and normalized for clustering
- `delete_columns`: List of columns to be removed (if necessary) before visualization
- `clusters`: The number of clusters from the KMeans model

Output:

- A plot showing relationships between variables involved in clustering.

Process:

1. Initialize a figure for the plot with size (12, 12).
2. Create a pairplot using Seaborn (`sns.pairplot`):
   - Set `data` as the source DataFrame.
   - Specify the variables to plot using `numerical_cols` (the numerical columns involved in the clustering).
   - Set `hue` to 'Cluster' to color the points by their assigned cluster.
   - Use the `palette='tab10'` to specify a color palette for the clusters.
   - Set `plot_kws` with additional plot options:
     - `alpha = 0.6`: Transparency of the points.
     - `s = 100`: Size of the scatter plot points.
     - `edgecolor = 'w'`: White edge around the points.
3. Add a title to the plot:
   - Title format: "Relationships between variables involved in clustering with KMeans({clusters} clusters)".
   - Adjust the title position with `y=1.02`.
4. Save the plot as a PDF file:

- File name: `Variables_KMeans_{clusters}Cluster.pdf`.

5. Show the plot using `plt.show()`.

   In this case the graph points to the application of k = 2 (see Figure 7).

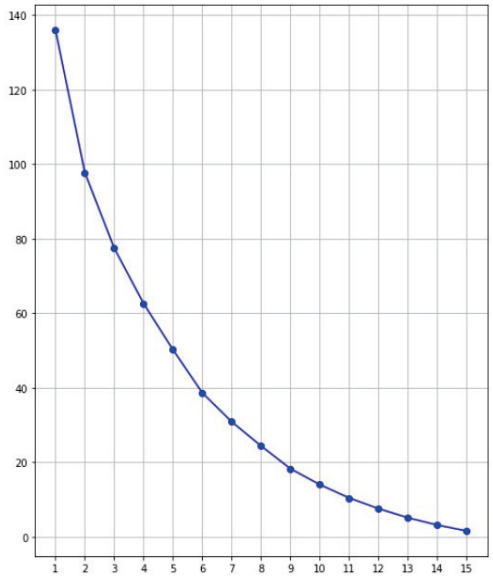


Figure 7. Graph of the elbow method.

An example of the graph obtained is shown in Figure 8.

Step 10. Linear Regression and Feature Impact Visualization

Input:

- `data:` DataFrame containing the feature variables and target variable
- `target_column:` The column name of the target variable to be predicted

Output:

- `df_coefficients:` DataFrame containing the coefficients and *p*-values of each feature

Process:

1. **Split the DataFrame** into features `(X)` and target variable `(y)`:
   - Set `y` as the target variable column.
   - Set `X` as the remaining features.
2. **Scale the features** using `StandardScaler:`
   - Apply `StandardScaler` to scale the feature values to the same scale.
   - Store the scaled data in `X_scaled`.
3. **Initialize the Linear Regression model**:
   - Create a LinearRegression object.
4. **Create empty lists** to store coefficients and *p*-values:
   - `coefficients:` Stores the regression coefficients (weights).
   - `p_values:` Stores the *p*-values for each coefficient, indicating their statistical significance.
5. **Fit the linear regression model for each feature**:
   - **For each feature column** in `X`:
     - Extract the feature column from `X_scaled: X_temporal = X_scaled[:, [i]]`.
     - Fit the model.
     - Append the coefficient `(model.coef_[0])` to `coefficients`.
     - Add a constant column to `X_temporal` for calculating the intercept: `X_temporal sm.add_constant(X_temporal)`.

- Create and fit the OLS model.
- Append the *p*-value (`model_sm.pvalues [1]`) to `p_values`.

6. **Create a DataFrame** with the coefficients and *p*-values:
   - Store the results in `df_coefficients`.
7. **Sort the DataFrame** by the absolute value of the coefficient:
   - Sort `df_coefficients` by coefficient magnitude.
8. **Visualize the regression coefficients**:
   - Create a figure for the bar plot.
   - Use Seaborn to create a bar plot.
   - Set plot titles and axis labels.
   - Save the plot as a PDF.
   - Show the plot.
9. **Return the DataFrame** with the sorted coefficients and *p*-values (`df_coefficients`).

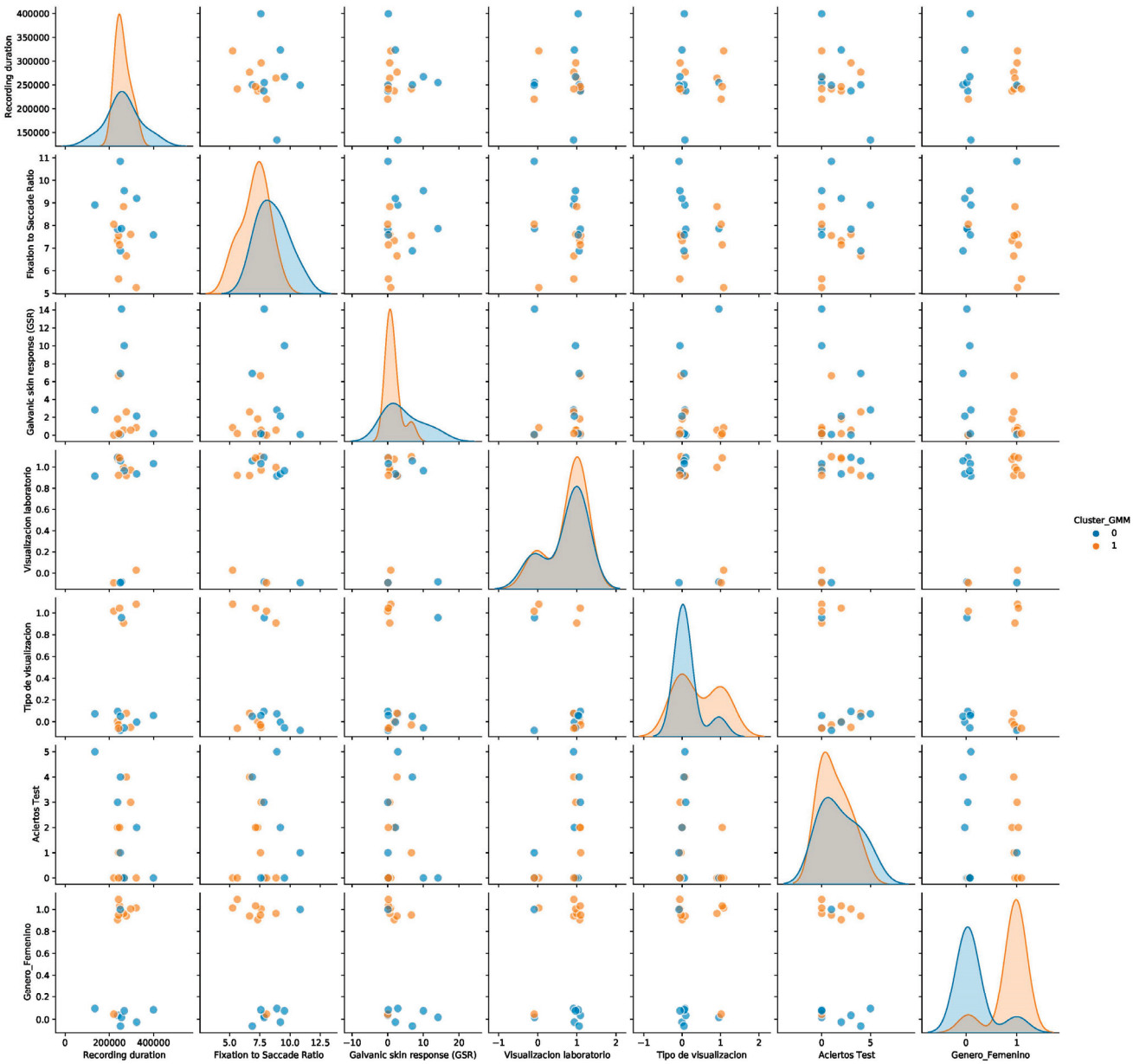


Figure 8. Scatter plot of the relationship between all variables.

Specifically, the regression coefficients and their significance at $\alpha = 0.05$ were found. As Table 3 shows, none of them were significant. There was only a pattern of difference in the variables "Type of presentation" $\alpha = 0.07$ and "Recording duration".

Table 3. Lineal Regression for each of the variables.

Variables (Characteristics)	Standardized Beta Coefficients	Standard Errors of Coefficients	*p*-Value
Type of presentation	0.48	0.23	0.07
Recording duration	−0.45	0.23	0.07
Gender	−0.15	0.25	0.57
GSR	−0.16	0.24	0.73
Fixation saccade ratio	−0.19	0.24	0.52

4. Discussion

The aim of this study was to highlight the need for integrated multichannel eye tracking devices to also offer—in addition to analysis of metrics based on using simple statistical tests (e.g., duration of an interval, number of events, time of each event, total duration of a fixation on an AOI, mean duration of fixations on an AOI, the number of fixations on an AOI, the duration of visits to a stimulus, the number of visits, the number of clicks on an AOI, gaze direction, eye position, mean duration of fixations, minimum duration of fixations, maximum duration of fixations, pupil diameter, etc.)—the possibility of processing records by applying other analytical techniques, such as more complex statistical or machine learning techniques (both supervised and unsupervised). Various studies have made effective use of supervised machine learning techniques (classification and regression) and unsupervised techniques (clustering) to interpret eye tracking data [50,51,54–57]. The objective is for the processing of all of this data to be done as smoothly as possible, automating the phases that researchers currently have to do by hand [52]. This paper provides an example of how data may be extracted and how machine learning algorithms may be automated, specifically in this case with the Python programming language—although the procedure may be used with other languages, such as R and Matlab.

5. Conclusions

Currently, devices applying eye tracking technology give researchers and users simple information for data analysis (frequencies, means, simple analysis of gaze paths for each participant). Extracting this data is relatively painless. However, the devices record much more information than that. To make the best analysis of all of this data, these records need to be pre-processed and processed. This requires an understanding of computation or IT. This is a major limitation on making use of the potential offered by multichannel integrated eye tracking technology. This is a pressing issue, as we believe that this technology should be habitually applied in advertising, health, education, and driving (among other areas).

5.1. Study Limitations

The limitations of this study are related to the small sample of user data. Because of that, future studies will increase the size of the samples in order to produce more versatile Jupyter Notebooks that can be used in different research fields. In addition, this study only applied two machine learning algorithms (supervised: linear regression; unsupervised: *k*-means clustering). Because of that, other studies will expand the use of supervised and unsupervised algorithms.

5.2. Future Lines of Research

The use of integrated multichannel eye tracking technology can offer professionals in marketing, education, and health (among others) real-time analysis of the processing of all of the information the devices record. This functionality will be extremely useful in various fields of application (for example, in education it will be useful for teachers to

have prediction, classification, or clustering data to help them create tailored educational resources. Similarly, doctors and psychologists may find help in individualised differential diagnosis and treatment). In other words, it opens up a wide range of possibilities for practical application. However, the obstacle of data processing means that this technology is not being used to its fullest potential. In this article, we have presented an example of how the phases of data pre-processing and processing may be automated, including the use of supervised and unsupervised machine learning algorithms. We would encourage researchers to continue working along these lines, and those responsible for the technology to incorporate these analytical resources into their multichannel eye tracking devices. Current researchers who avail themselves of machine learning techniques are more effective at analysing eye tracking data. All of this will contribute to this technology being used more habitually, and to improving the conclusions studies can produce, helping to improve research in the fields in which it is applied.

Author Contributions: Conceptualisation, all authors; methodology, all authors; software, all authors; validation, all authors; formal analysis, all authors; investigation, all authors; resources, all authors; data curation, all authors; writing—original draft preparation, M.C.S.-M. and R.M.-S.; writing—review and editing, all authors; visualisation, all authors; supervision, all authors; project administration, all authors; funding acquisition, M.C.S.-M. and R.M.-S. All authors have read and agreed to the published version of the manuscript.

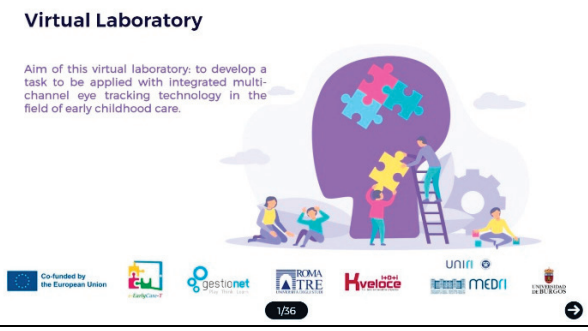
Funding: Project "Voice assistants and artificial intelligence in Moodle: a path towards a smart university" SmartLearnUni. Call 2020 R&D&I Projects—RTI Type B. MINISTRY OF SCIENCE AND INNOVATION AND UNIVERSITIES. STATE RESEARCH AGENCY. Government of Spain, grant number PID2020-117111RB-I00". Specifically, in the part concerning the application of multichannel eye tracking technology with university students and Project "Specialized and updated training on supporting advance technologies for early childhood education and care professionals and graduates" (eEarlyCare-T) grant number 2021-1-ES01-KA220-SCH-000032661 funded by the EUROPEAN COMMISSION. In particular, the funding has enabled the development of the e-learning classroom and educational materials.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Ethics Committee of UNIVERSITY OF BURGOS (protocol code IO 03/2022 and date of approval 7 April 2022).

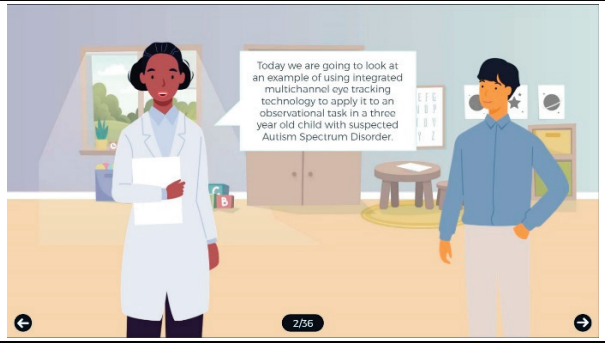
Data Availability Statement: We confirm that the data is not available for ethical reasons, but that it can be requested as long as the person, the institution to which it belongs and the purposes for which the data is used are credited.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A



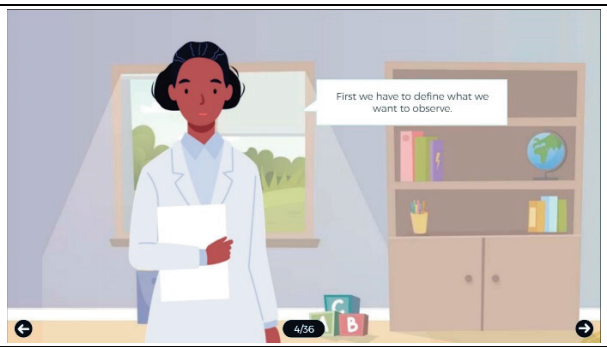
Presentation of the virtual laboratory



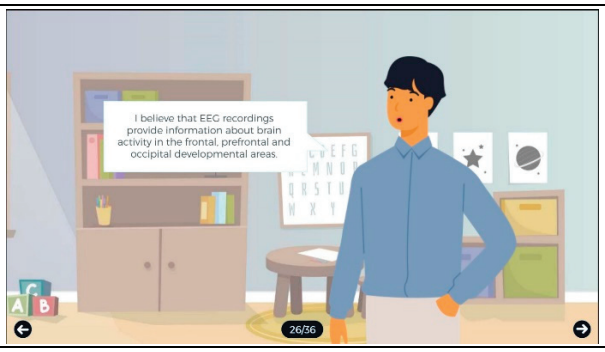
Presentation of the virtual laboratory through avatars



Role of biomedical engineering teacher and role of biomedical engineering student



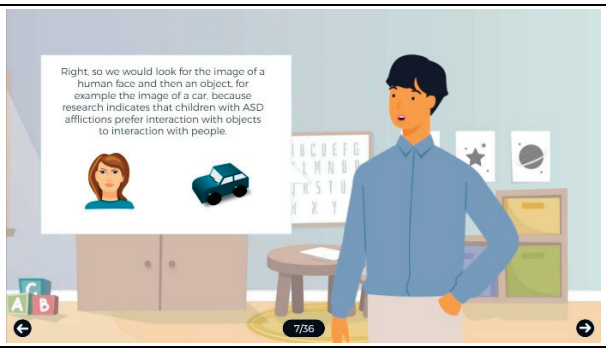
The teacher-role avatar regulates information recall



The student-role avatar remembers some of the information.



The avatar with the role of teacher reinforces the student

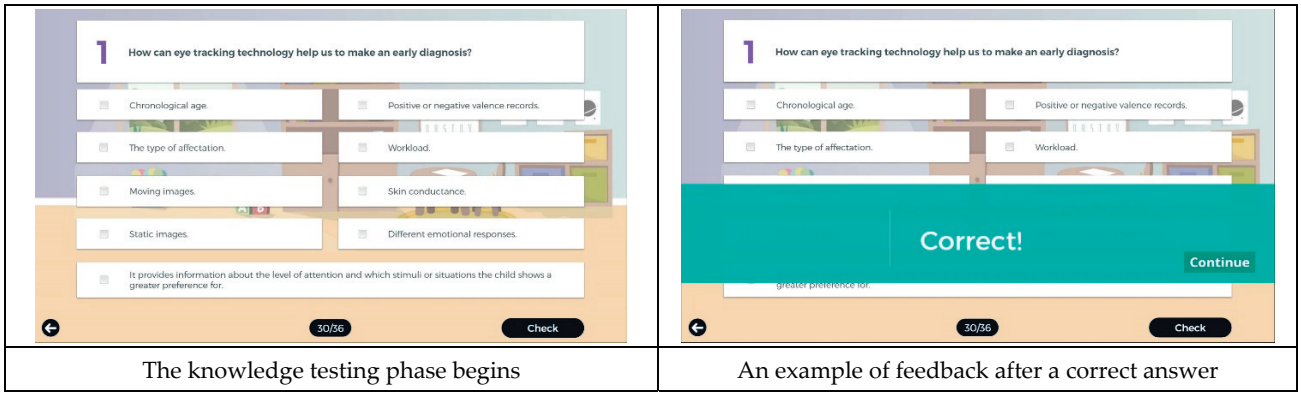


Avatar in the student role ponders aloud the answer



Avatar with teacher role reaffirms the response of the avatar with student role

Figure A1. *Cont.*



The knowledge testing phase begins | An example of feedback after a correct answer

Figure A1. Description of the virtual laboratory.

References

1. Retamosa, M.; Aliagas, I.; Millán, A. Displaying ingredients on healthy snack packaging: A study on visual attention, choice, and purchase intention. *J. Sens. Stud.* **2024**, *39*, e12944. [CrossRef]
2. Bajaj, R.; Ali Syed, A.; Singh, S. Analysing applications of neuromarketing in efficacy of programmatic advertising. *J. Consum. Behav.* **2024**, *23*, 939–958. [CrossRef]
3. Bhardwaj, S.; Thapa, S.B.; Gandhi, A. Advances in neuromarketing and improved understanding of consumer behaviour: Analysing tool possibilities and research trends. *Cogent Bus. Manag.* **2024**, *11*, 2376773. [CrossRef]
4. Calderón-Fajardo, V.; Anaya-Sánchez, R.; Rejón-Guardia, F.; Molinillo, S. Neurotourism Insights: Eye Tracking and Galvanic Analysis of Tourism Destination Brand Logos and AI Visuals. *Tour. Manag. Stud.* **2024**, *20*, 53–78. [CrossRef]
5. Modi, N.; Singh, J. An analysis of perfume packaging designs on consumer's cognitive and emotional behavior using eye gaze tracking. *Multimed. Tools Appl.* **2024**, *83*, 82563–82588. [CrossRef]
6. Thiebaut, R.; Elshourbagi, A. The Effect of Neuromarketing and Subconscious Branding on Business Profitability and Brand Image: A New Business Model Innovation for Startups. In *Fostering Global Entrepreneurship Through Business Model Innovation*; Gupta, V., Ed.; IGI Global: Hershey, PA, USA, 2024; pp. 217–252. [CrossRef]
7. Cong, L.; Luan, S.; Young, E.; Mirosa, M.; Bremer, P.; Torrico, D.D. The Application of Biometric Approaches in Agri-Food Marketing: A Systematic Literature Review. *Foods* **2023**, *12*, 2982. [CrossRef] [PubMed]
8. Al-Nafjan, A.; Aldayel, M.; Kharrat, A. Systematic Review and Future Direction of Neuro-Tourism Research. *Brain Sci.* **2023**, *13*, 682. [CrossRef] [PubMed]
9. Khondakar, F.K.; Sarowar, H.; Chowdhury, M.H.; Majumder, S.; Hossain, A.; Dewan, M.A.A.; Hossain, Q.D. A systematic review on EEG-based neuromarketing: Recent trends and analyzing techniques. *Brain Inform.* **2024**, *11*, 17. [CrossRef]
10. Tavares-Filho, E.R.; Hidalgo, L.G.S.; Lima, L.M.; Spers, E.E.; Pimentel, T.C.; Esmerino, E.A.; Cruz, A.G. Impact of animal origin of milk, processing technology, type of product, and price on the Boursin cheese choice process: Insights of a discrete choice experiment and eye tracking. *J. Food Sci.* **2024**, *89*, 640–655. [CrossRef]
11. Madlenak, R.; Chinoracky, R.; Stalmasekova, N.; Madlenakova, L. Investigating the Effect of Outdoor Advertising on Consumer Decisions: An Eye-Tracking and A/B Testing Study of Car Drivers' Perception. *Appl. Sci.* **2023**, *13*, 6808. [CrossRef]
12. Kim, M.; Lee, J.; Lee, S.Y.; Ha, M.; Park, I.; Jang, J.; Jang, M.; Park, S.; Kwon, J.S. Development of an eye-tracking system based on a deep learning model to assess executive function in patients with mental illnesses. *Sci. Rep.* **2024**, *14*, 18186. [CrossRef] [PubMed]
13. Perkovich, E.; Laakman, A.; Mire, S.; Yoshida, H. Conducting head-mounted eye-tracking research with young children with autism and children with increased likelihood of later autism diagnosis. *J. Neurodev. Disord.* **2024**, *16*, 7. [CrossRef] [PubMed]
14. Amirbay, A.; Mukhanova, A.; Baigabylov, N.; Kudabekov, M.; Mukhambetova, K.; Baigusheva, K.; Baibulova, M.; Ospanova, T. Development of an algorithm for identifying the autism spectrum based on features using deep learning methods. *Int. J. Electr. Comput. Eng. (IJECE)* **2024**, *14*, 5513–5523. [CrossRef]
15. Bent, C.; Glencross, S.; McKinnon, K.; Hudry, K.; Dissanayake, C.; The Victorian ASELCC Team; Vivanti, G. Predictors of Developmental and Adaptive Behaviour Outcomes in Response to Early Intensive Behavioural Intervention and the Early Start Denver Model. *J. Autism Dev. Disord.* **2024**, *54*, 2668–2681. [CrossRef]
16. Ibragimov, B.; Mello-Thoms, C. The Use of Machine Learning in Eye Tracking Studies in Medical Imaging: A Review. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 3597–3612. [CrossRef] [PubMed]
17. Mehmood, I.; Li, H.; Umer, W.; Tariq, S.; Wu, H. Non-invasive detection of mental fatigue in construction equipment operators through geometric measurements of facial features. *J. Saf. Res.* **2024**, *89*, 234–250. [CrossRef]
18. Schmidt, A.; Mohareri, O.; DiMaio, S.; Yip, M.C.; Salcudean, S.E. Tracking and mapping in medical computer vision: A review. *Med. Image Anal.* **2024**, *94*, 103131. [CrossRef] [PubMed]
19. Boujelbane, M.A.; Trabelsi, K.; Salem, A.; Ammar, A.; Glenn, J.M.; Boukhris, O.; AlRashid, M.M.; Jahrami, H.; Chtourou, H. Eye Tracking During Visual Paired-Comparison Tasks: A Systematic Review and Meta-Analysis of the Diagnostic Test Accuracy for Detecting Cognitive Decline. *J. Alzheimer's Dis.* **2024**, *99*, 207–221. [CrossRef]

20. Klotzek, A.; Jemni, M.; Groves, S.J.; Carrick, F.R. Effects of Cervical Spinal Manipulation on Saccadic Eye Movements. *Brain Sci.* **2024**, *14*, 292. [CrossRef]
21. Pauszek, J.R. An introduction to eye tracking in human factors healthcare research and medical device testing. *Hum. Factors Healthc.* **2023**, *3*, 100031. [CrossRef]
22. Passaro, A.; Zullo, A.; Di Gioia, M.; Curcio, E.; Stasolla, F. A Narrative Review on the Use of Eye-Tracking in Rett Syndrome: Implications for Diagnosis and Treatment. *OBM Genet.* **2024**, *8*, 250. [CrossRef]
23. Hill, W.; Lindner, H. Using eye tracking to assess learning of a multifunction prosthetic hand: An exploratory study from a rehabilitation perspective. *J. Neuroeng. Rehabil.* **2024**, *21*, 148. [CrossRef] [PubMed]
24. Pulay, M.Á.; Szabó, É.F. Developing Visual Perceptual Skills with Assistive Technology Supported Application for Children with Cerebral Palsy. *Acta Polytech. Hung.* **2024**, *21*, 25–38. [CrossRef]
25. Feldmann, L.; Zsigo, C.; Mörtl, I.; Bartling, J.; Wachinger, C.; Oort, F.; Schulte-Körne, G.; Greimel, E. Emotion regulation in adolescents with major depression—Evidence from a combined EEG and eye-tracking study. *J. Affect. Disord.* **2023**, *340*, 899–906. [CrossRef] [PubMed]
26. Tao, Z.; Sun, N.; Yuan, Z.; Chen, Z.; Liu, J.; Wang, C.; Li, S.; Ma, X.; Ji, B.; Li, K. Research on a New Intelligent and Rapid Screening Method for Depression Risk in Young People Based on Eye Tracking Technology. *Brain Sci.* **2023**, *13*, 1415. [CrossRef]
27. Brien, D.C.; Riek, H.C.; Yep, R.; Huang, J.; Coe, B.; Areshenkoff, C.; Grimes, D.; Jog, M.; Lang, A.; Marras, C.; et al. Classification and staging of Parkinson's disease using video-based eye tracking. *Park. Relat. Disord.* **2023**, *110*, 105316. [CrossRef] [PubMed]
28. Ghiţă, A.; Hernández-Serrano, O.; Moreno, M.; Monràs, M.; Gual, A.; Maurage, P.; Gacto-Sánchez, M.; Ferrer-García, M.; Porras-García, B.; Gutiérrez-Maldonado, J. Exploring Attentional Bias toward Alcohol Content: Insights from Eye-Movement Activity. *Eur. Addict. Res.* **2024**, *30*, 65–79. [CrossRef]
29. Puttevils, L.; De Bruecker, M.; Allaert, J.; Sanchez-Lopez, A.; De Schryver, N.; Vervaet, M.; Baeken, C.; Vanderhasselt, M.-A. Attentional bias to food during free and instructed viewing in anorexia nervosa: An eye tracking study. *J. Psychiatr. Res.* **2023**, *164*, 468–476. [CrossRef]
30. Guo, X.; Liu, Y.; Tan, Y.; Xia, Z.; Fu, H. Hazard identification performance comparison between virtual reality and traditional construction safety training modes for different learning style individuals. *Saf. Sci.* **2024**, *180*, 106644. [CrossRef]
31. Virlet, L.; Sparrow, L.; Barela, J.; Berquin, P.; Bonnet, C. Proprioceptive intervention improves reading performance in developmental dyslexia: An eye-tracking study. *Res. Dev. Disabil.* **2024**, *153*, 104813. [CrossRef]
32. Liang, Z.; Ga, R.; Bai, H.; Zhao, Q.; Wang, G.; Lai, Q.; Chen, S.; Yu, Q.; Zhou, Z. Teaching expectancy improves video-based learning: Evidence from eye-movement synchronization. *Br. J. Educ. Technol.* **2024**. [CrossRef]
33. Kok, E.M.; Niehorster, D.C.; van der Gijp, A.; Rutgers, D.R.; Auffermann, W.F.; van der Schaaf, M.; Kester, L.; van Gog, T. The effects of gaze-display feedback on medical students' self-monitoring and learning in radiology. *Adv. Health Sci. Educ.* **2024**. *online ahead of print*. [CrossRef] [PubMed]
34. Sáiz-Manzanares, M.C.; Marticorena-Sánchez, R.; Escolar-Llamazares, M.C.; González-Díez, I.; Martín Antón, L.J. Using integrated multimodal technology: A way to personalised learning in Health Sciences and Biomedical engineering Students. *Appl. Sci.* **2024**, *14*, 7017. [CrossRef]
35. Mullen, B. Self-Attention Theory: The Effects of Group Composition on the Individual. In *Theories of Group Behavior*; Springer Series in Social, Psychology; Mullen, B., Goethals, G.R., Eds.; Springer: New York, NY, USA, 1987. [CrossRef]
36. Korteland, R.J.; Kok, E.; Hulshof, C.; van Gog, T. Teaching through their eyes: Effects on optometry teachers' adaptivity and students' learning when teachers see students' gaze. *Adv. Health Sci. Educ.* **2024**. [CrossRef]
37. Sáiz-Manzanares, M.C.; Marticorena-Sánchez, R.; Martín-Antón, L.J.; González-Diez, I.; Carbonero-Martín, I. Using eye tracking technology to analyse cognitive load in multichannel activities in university students. *Int. J. Hum. Comput. Interact.* **2024**, *40*, 3263–3281. [CrossRef]
38. Wronski, P.; Wensing, M.; Ghosh, S.; Gärttner, L.; Müller, W.; Koetsenruijter, J. Use of a quantitative data report in a hypothetical decision scenario for health policymaking: A computer-assisted laboratory study. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 32. [CrossRef]
39. Lee, Y.; de Jong, N.; Donkers, J.; Jarodzka, H.; van Merriënboer, J.G. Measuring Cognitive Load in Virtual Reality Training via Pupillometry. *IEEE Trans. Learn. Technol.* **2024**, *17*, 704–710. [CrossRef]
40. Wang, Y.; Lu, Y.; Shen, C.-Y.; Luo, S.-J.; Zhang, L.-Y. Exploring product style perception: A comparative eye-tracking analysis of users across varying levels of self-monitoring. *Displays* **2024**, *84*, 102790. [CrossRef]
41. Cazes, M.; Noël, A.; Jamet, E. Cognitive effects of humorous drawings on learning: An eye-tracking study. *Appl. Cogn. Psychol.* **2024**, *38*, e4178. [CrossRef]
42. Tarkowski, S.; Caban, J.; Dzieńkowski, M.; Nieoczym, A.; Zarajczyk, J. Driver's distraction and its potential influence on the extension of reaction time. *Arch. Automot. Eng.* **2022**, *98*, 65–78. [CrossRef]
43. Cheng, G.; Di, Z.; Xie, H.; Wang, F.L. Exploring differences in self-regulated learning strategy use between high- and low-performing students in introductory programming: An analysis of eye-tracking and retrospective think-aloud data from program comprehension. *Comput. Educ.* **2024**, *208*, 104948. [CrossRef]
44. Omobolanle, O.; Abiola, A.; Nihar, G.; Mohammad, K.; Abiola, A. Detecting Learning Stages within a Sensor-Based Mixed Reality Learning Environment Using Deep Learning. *J. Comput. Civ. Eng.* **2024**, *37*, 04023011. [CrossRef]
45. Bouwer, R.; Dirkx, K. The eye-mind of processing written feedback for revision. *Learn. Instr.* **2024**, *85*, 101745. [CrossRef]

46. Ferreira, C.P.; Soledad Gonzalez-Gonzalez, C.; Francisca Adamatti, D.; Moreira, F. Analysis Learning Model with Biometric Devices for Business Simulation Games: Brazilian Case Study. *IEEE Access* **2024**, *12*, 95548–95564. [CrossRef]
47. Sáiz-Manzanares, M.C.; Marticorena-Sánchez, R.; Martín-Antón, L.J.; Almeida, L.; Carbonero-Martín, I. Application and challenges of eye tracking technology in Higher Education. *Comunicar* **2023**, *76*, 1–12. [CrossRef]
48. Sáiz-Manzanares, M.C.; Ramos Pérez, I.; Arnaiz-Rodríguez, Á.; Rodríguez-Arribas, S.; Almeida, L.; Martin, C.F. Analysis of the learning process through eye tracking technology and feature selection techniques. *Appl. Sci.* **2021**, *11*, 6157. [CrossRef]
49. Holmqvist, K.L.U.; Nyström, M.L.U.; Andersson, R.L.U.; Dewhurst, R.L.U.; Halszka, J.; van de Weijer, J.L. *Eye Tracking: A Comprehensive Guide to Methods and Measures*; Oxford University Press: Oxford, UK, 2011.
50. Vortmann, L.-M.; Putze, F. Combining Implicit and Explicit Feature Extraction for Eye Tracking: Attention Classification Using a Heterogeneous Input. *Sensors* **2021**, *21*, 8205. [CrossRef]
51. Sáiz Manzanares, M.C.; Rodríguez Diez, J.J.; Marticorena Sánchez, R.; Zaparaín Yáñez, M.J.; Cerezo Menéndez, R. Lifelong Learning from Sustainable Education: An Analysis with Eye Tracking and Data Mining Techniques. *Sustainability* **2020**, *12*, 1970. [CrossRef]
52. García, S.; Luengo, J.; Herrera, F. *Data Preprocessing in Data Mining*; Springer: London, UK, 2015. [CrossRef]
53. Thilderkvist, E.; Dobslaw, F. On current limitations of online eye-tracking to study the visual processing of source code. *Inf. Softw. Technol.* **2024**, *174*, 107502. [CrossRef]
54. Cho, S.-W.; Lim, Y.-H.; Seo, K.-M.; Kim, J. Integration of eye-tracking and object detection in a deep learning system for quality inspection analysis. *J. Comput. Des. Eng.* **2024**, *11*, 158–173. [CrossRef]
55. Liu, Z.; Yeh, W.-C.; Lin, K.-Y.; Lin, C.-S.H.; Chang, C.-Y. Machine learning based approach for exploring online shopping behavior and preferences with eye tracking. *Comput. Sci. Inf. Syst.* **2024**, *21*, 593–623. [CrossRef]
56. Zhang, H.; Wang, C. Integrated neural network-based pupil tracking technology for wearable gaze tracking devices in flight training. *IEEE Access* **2024**, *12*, 133234–133244. [CrossRef]
57. Born, J.; Ram, B.; Ramachandran, N.; Romero Pinto, S.A.; Winkler, S.; Ratnam, R. Multimodal Study of the Effects of Varying Task Load Utilizing EEG, GSR and Eye-Tracking. *bioRxiv* **2019**, 798496. [CrossRef]
58. Lindsay, G.W. Attention in Psychology, Neuroscience, and Machine Learning. *Front. Comput. Neurosci.* **2024**, *14*. [CrossRef]
59. Tobii AB Corp. *Tobii Pro Lab [Computer Software]*, version 1.241.54542; Tobii Corp: Danderyd, Sweden, 2024.
60. Grabinger, L.; Hauser, F.; Wolff, C.; Mottok, J. On Eye Tracking in Software Engineering. *SN Comput. Sci.* **2024**, *5*, 729. [CrossRef]
61. Shimmer3 GSR. *Shimmer [Computer Software]*, version 3; Shimmer Corp.: Dublin, Ireland; Boston, MA, USA, 2024.
62. Sáiz-Manzanares, M.C.; Marticorena-Sánchez, R. *Manual for the Development of Self-Regulated Virtual Laboratories*; Servicio de Publicaciones de la Universidad de Burgos: Burgos, Spain, 2024. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

computers

MDPI

Article

Development of a Children's Educational Dictionary for a Low-Resource Language Using AI Tools

Diana Rakhimova [1,2,*], Aidana Karibayeva [1,2], Vladislav Karyukin [1,2], Assem Turarbek [1,2], Zhansaya Duisenbekkyzy [1] and Rashid Aliyev [1]

[1] Department of Information Systems, Al-Farabi Kazakh National University, Almaty 050040, Kazakhstan; a.s.karibayeva@gmail.com (A.K.); vladislav.karyukin@gmail.com (V.K.); turarbekasem1@gmail.com (A.T.); 1695507zh@gmail.com (Z.D.); rasidaliev53@gmail.com (R.A.)

[2] Institute of Information and Computational Technologies, Almaty 050010, Kazakhstan

* Correspondence: diana.rakhimova@kaznu.edu.kz or di.diva@mail.ru; Tel.: +7-777-242-0775

Abstract: Today, various interactive tools or partially available artificial intelligence applications are actively used in educational processes to solve multiple problems for resource-rich languages, such as English, Spanish, French, etc. Unfortunately, the situation is different and more complex for low-resource languages, like Kazakh, Uzbek, Mongolian, and others, due to the lack of qualitative and accessible resources, morphological complexity, and the semantics of agglutinative languages. This article presents research on early childhood learning resources for the low-resource Kazakh language. Generally, a dictionary for children differs from classical educational dictionaries. The difference between dictionaries for children and adults lies in their purpose and methods of presenting information. A themed dictionary will make learning and remembering new words easier for children because they will be presented in a specific context. This article discusses developing an approach to creating a thematic children's dictionary of the low-resource Kazakh language using artificial intelligence. The proposed approach is based on several important stages: the initial formation of a list of English words with the use of ChatGPT; identification of their semantic weights; generation of phrases and sentences with the use of the list of semantically related words; translation of obtained phrases and sentences from English to Kazakh, dividing them into bigrams and trigrams; and processing with Kazakh language POS pattern tag templates to adapt them for children. When the dictionary was formed, the semantic proximity of words and phrases to the given theme and age restrictions for children were taken into account. The formed dictionary phrases were evaluated using the cosine similarity, Euclidean similarity, and Manhattan distance metrics. Moreover, the dictionary was extended with video and audio data by implementing models like DALL-E 3, Midjourney, and Stable Diffusion to illustrate the dictionary data and TTS (Text to Speech) technology for the Kazakh language for voice synthesis. The developed thematic dictionary approach was tested, and a SUS (System Usability Scale) assessment of the application was conducted. The experimental results demonstrate the proposed approach's high efficiency and its potential for wide use in educational purposes.

Keywords: low-resource language; ChatGPT; children's education; Kazakh language; artificial intelligence

1. Introduction

Linguistic dictionaries are typically designed for adults and older children, often containing complex terms and specialized language. These can be difficult for younger audiences to comprehend. In contrast, children's dictionaries employ simpler, more understandable language appropriate for a specific age group. They often feature bright illustrations to aid in the visual memorization of words. The primary differences between dictionaries for children and adults lie in their purpose, language, and methods of presenting information.

In many countries, preschool education is considered to play a compensatory role for children who have not received support from their immediate environment, i.e., from home [1]. Researchers such as Tsiolkovsky and Goldstein [2] have shown that children with delayed speech development can catch up with their peers and improve their communication skills and literacy if they receive support. This is confirmed by Grabmayer [3] and Lennox [4]. It is also evident that children with extensive reading experience from various media and environments where parents are involved in reading have higher phonological awareness, broader letter knowledge, and more extensive vocabulary than children who engage in reading. Preschool educators in various global studies have reported that many children have poor speech and language development and, therefore, need support [5]. It is known that children with poor communication skills can develop both active and passive vocabulary if they receive the right support [6]. In this case, "right" means individually adapted. Thus, reading aloud can be seen as a tool for improving communication [7]. One of the tasks of preschool educators is to observe children's communication abilities based on each child's capabilities.

In the era of active digital society development, the importance of effective communication becomes apparent from an early age. Every child, regardless of their native language or social background, should have equal opportunities to learn their native language. Preschools and schools play a crucial role in developing language skills and serve as the foundation for children's education. Quantitative differences in children's speech behavior are observed in early preschool years. According to G. Lyamina's research [8], the volume of independent speech utterances during a 30 min play period changes as follows: for children aged 2 to 2.5 years, the average speech activity is 25–28 words; for children from 2.5 to 3 years—70–80 words; for children from 3 to 3.5 years—110–115 words [8]. Speech development delay can be diagnosed during the period when speech should already be developed, namely, at 3–4 years.

Speech development delays can be diagnosed during the period when speech should be well-developed, typically around 3–4 years. Vocabulary development in children is closely tied to their ability to understand words that denote objects in their immediate environment, actions associated with those objects, and their characteristics. In the first two years of life, a word functions as part of a complex influence from adults, including gestures, intonation, and the situation in which the word is spoken. The word then becomes an integrating signal. As children grow, the number of words they use increases rapidly, although the rate of growth slows down. For children aged 3–5 years, mastering the clear referential meaning of words and their specific meanings is central, and by 5–6 years old, they begin to develop a system of so-called everyday concepts, which predominantly have emotional-figurative visual connections. The clarification of the semantic content of words at 6–7 years old begins to gain momentum, related to acquiring new knowledge about the world and developing an aesthetic attitude towards words and speech in general.

This study makes a new contribution to understanding early childhood education by proposing the creation of a topic-based dictionary of the Kazakh language using artificial intelligence (AI). Unlike existing studies focusing on resource-rich languages, this work explores resource-poor languages, such as Kazakh, and addresses the unique challenges associated with their morphological complexity and the lack of available educational resources. The morphological analysis of the generated sentences and word phrases on the specific topic gives an ability to choose the most optimal elements of the dictionary that are well adapted for children of a certain age. Moreover, the dictionary is elaborated for an audio and visualization model that provides a qualitative educational approach for children.

Creating a dictionary for children will help achieve goals such as increasing the availability of educational resources, personalized learning, preserving and developing languages with limited resources, providing support for parents and teachers, and improving the quality of linguistic data.

The development of automated dictionaries for children using AI is a necessary field of research that combines linguistic knowledge with AI capabilities to create educational

tools. These thematic dictionaries can play a crucial role in enhancing preschool children's language learning, cognitive development, and cultural awareness.

Adapting AI technologies to educational resources for children, including in the Kazakh language, can significantly impact the approach to linguistic and cultural education, making it more engaging and effective. Using AI, educators can provide a personalized learning experience that meets the developmental needs of children, promoting deeper learning of their native and additional languages.

Speech disorders in preschool children are a serious issue worldwide. Research in this area is highly relevant as the number of children with delayed speech development increases every year. The number of parents seeking help from special centers and speech therapy specialists is also growing. The development and application of new technologies using natural language processing methods and machine learning are highly demanded to address this issue in the Kazakh language.

The article describes such stages as project planning, language analysis, needs assessment, cultural considerations, community involvement and validation, monitoring and updating, sustainability, and future expansion. Generally, the structure of the paper is organized in the following way. The introduction describes the needs assessment and cultural aspects of creating and researching a children's dictionary. Section 2 provides an extensive literature review where many significantly essential papers related to the development of children's dictionaries, speech recognition, and text generation are described. The literature review will analyze the Kazakh language and its applications in developing children's speech in other languages. Section 3 discusses the methodological approach applied to developing the children's dictionary, its evaluation, and designing the dictionary visualization and audio voice module. The methodology section describes community involvement, how the verification was carried out, and considers the developed dictionary's sustainability. Section 4 fully describes the architecture and process of the children's dictionary formation using ChatGPT and the Kazakh language POS patterns. Section 5 analyzes the obtained dictionary by evaluating the closeness of generated dictionary phrases and sentences to the original word using the cosine similarity, Euclidean similarity, and Manhattan distance metrics. Section 6 is devoted to creating the developed dictionary's visualization and audio voice module. Finally, Section 7 concludes the research by analyzing all the steps taken to develop a children's dictionary for the low-resource Kazakh language using modern AI tools. Future expansion, updating, and monitoring of the created dictionary will be discussed in the conclusion and future work. All these aspects will be discussed in detail in the article.

2. Literature Review

The application of AI to create various methods, tools, and techniques that impact the development of children's speech is one of the first steps in addressing social and communication problems.

M. Verburg's research [9] is intended to understand the methodologies for developing dictionaries tailored to a specific age group of children. D. Nkomo, in his work [10], analyzes the needs of African children for dictionaries, highlighting the critical importance of their accessibility and practicality. Figure 1 presents the key characteristics that a dictionary should have.

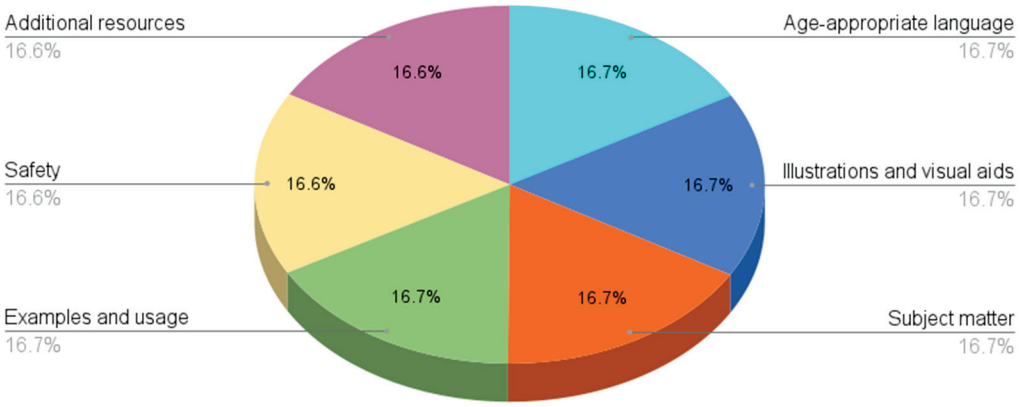


Figure 1. Important parameters for the development of thematic dictionaries for children [9–11].

Based on this diagram, it can be seen that the development of a children's dictionary takes into account many didactic and linguistic aspects that are directly related to the language's characteristics.

The study by Dan Zhao and Hai Xu [12], dedicated to optimizing the development of English–Chinese children's dictionaries, points out that thematic dictionaries intended for an adult audience are often not suitable for children because adult dictionaries use more complex language and terms that may be beyond children's comprehension. Children's cognitive abilities are still developing, so adult dictionaries can be too complex and confusing. In contrast, children's dictionaries are designed with age-appropriate features, including simple language and illustrations, which facilitate learning and understanding. Thus, adult dictionaries are not suitable as educational tools for younger students.

Next, we will compare children's and adult thematic dictionaries, considering the aforementioned dictionary development parameters, as shown in Table 1.

Table 1. Comparison of children's and adult thematic dictionaries: features and applicability [12–14].

Feature	A Thematic Dictionary for Children	A Thematic Dictionary for Adults
Content	Simple, child-friendly vocabulary	A complex and comprehensive vocabulary, including specialized terms
Goal	Improve basic language skills and learning	Provide detailed information on various professional fields
Age	Adapted to the level of understanding of children of different ages	It assumes a high level of literacy and knowledge of the subject
Definitions	The definitions are simple, concise, and often accompanied by examples	The definitions are provided with detailed technical descriptions, without any simplifications
Illustrations	Frequent use of colorful illustrations helps to better understand and remember the text	A minimum of illustrations or their complete absence; more attention is paid to the text
Design and format	Bright, fascinating to interest children Divided by topic to facilitate the study of specific categories	A more formal and dense format designed for familiarization. It is divided alphabetically or by area/topic, which provides systematic access to information

The comparison highlights that dictionaries for adults may be unsuitable for children due to the complexity of content, lack of engaging elements, and inappropriate materials. Terms and explanations in such dictionaries are often too complex and can overwhelm a child's perception, making understanding more difficult. Additionally, adult dictionaries typically lack interactive elements or visual aids, making them less attractive and interesting for children, which may reduce their interest in learning new words.

Research disseminates practices for developing and using applications and mobile technologies related to human communication and interaction, as reported by doctors, for

children with various speech disorders. The development of phonological awareness, i.e., manipulating the language structure, from an early age is essential for acquiring literacy skills. Communication technologies aimed at learning play a key role and can stimulate the interest and thinking of today's children living in a digital society.

Currently, the App Store, Google Play, and Play Market offer a vast number of speech development applications for children. However, few of them meet the standards of effective education from a developmental psychology perspective [15–17]. The increased use of educational applications provides additional opportunities to support language development, for example, by introducing children to new words just like in storybooks, especially when applications are used jointly between a parent and a child. Joint media use between parent and child reduces time costs. However, it should be noted that only 20% of parents report using applications with children aged 2 to 4 most of the time [18], although for children aged 0–2, this figure rises to 49%, and for children aged 0–8, 80% of parents report using applications with their child most or some of the time.

Parents and children can use applications to expand their language vocabulary, especially when learning in the application is embedded in an authentic, meaningful, and purposeful context [15,19]. Several studies conducted to date have shown that children can learn new words through applications [20–23].

A multilingual thematic dictionary for primary school students was created based on design and development methodology [24]. Authors [25] identify the educational impact of AI technologies as a tool to help children develop ethical and critical thinking skills.

Studies by specialists (speech therapists, defectologists) at the Astana Rehabilitation Center have shown that children produce sound reactions from birth [26]. A one-year-old child should be able to pronounce at least 10–15 words, a two-year-old around 300 words, and a child over three years old should communicate freely. A normal three-year-old child should have a vocabulary of 800–1500 words. The absence of two-word phrases at the age of 2 years and older, such as ("ana ber"—mom give it, "tamak zheymin"—want to eat, "dalaga baram"—go outside . . .), and at age 3, a small vocabulary, poor pronunciation, and inability to express their desires and thoughts through speech are considered abnormal [27].

Currently, research in Kazakhstan actively focuses on speech recognition in the Kazakh language, machine translation to Kazakh, and the use of AI in Kazakh speech recognition by scientists such as Sharipbay A.E., Bekmanova G.T., Tukejev U.A., Mansurova M., Amirgaliyev E.N., Musabayev R.R., Mamyrbayev O.Zh., and others [28,29]. These works present research and results on Kazakh speech STT (Speech to Text) and TTS (Text to Speech) tasks based on neural network models using datasets from adult speakers with correct pronunciation. Unfortunately, the analysis and recognition of children's speech have not been considered. The lack of specialized electronic and accessible resources complicates the development of this field.

The study [30] developed a dictionary of Kazakh language synonyms for children with weak speech skills using the Word2Vec method, grouping the dictionary by special word groups. The proposed model can be integrated into various educational and language applications to improve children's vocabulary and speech skills. This approach depends on input resources for training and requires significant computational resources.

When examining scholars' work, it can be seen that many use neural networks and machine learning methods to achieve their goals and solve various problems. In his work, Lukashik [31] demonstrates the application of corpus linguistics and generative AI tools for terminology development in the Kashubian language. The study proved that ChatGPT offers a faster and more efficient method for terminology extraction compared to traditional corpus linguistics, although it requires human oversight to ensure data accuracy. R. Ponnusami's research [32] focuses on using the ChatGPT-3 model for text generation in the Tamil language. The results showed that the model successfully performs text generation tasks and is applicable for educational and entertainment purposes.

The use of AI in education, especially through chatbots like ChatGPT, has significantly expanded and enhanced learning opportunities from early childhood to higher education.

AI, especially in the form of chatbots, has been effectively integrated into educational institutions, assisting children with tasks such as pronunciation and spelling. While AI offers benefits such as personalized learning and constant availability, it is emphasized that it should complement rather than replace human educators, ensuring a comprehensive and effective educational process [33].

AI-driven thematic dictionaries also demonstrate the potential to support children with special educational needs. For example, I.A. Pisarev's ontological approach [34] includes thematic dictionaries that use computational linguistics and lexical-semantic field analysis, as well as AI for adapted learning for children.

Analyzing researchers' works focused on the Kazakh language, we can highlight Gaziza Yelibayeva's research [35], which discusses the search for phrases in the Kazakh language based on ontology, illustrating the technical capabilities for developing educational tools based on AI that could support dictionary creation. These tools can help systematize vocabulary and facilitate effective language learning through intuitive search processes and semantic networks. Table 2 presents a comprehensive description of research in this area.

Table 2. Comparative analysis of research in the field of language innovations and educational resources.

Researcher	Year	Language	Methodology	Age Group	Developed
Marek Lukasik	2024	Kashubian	Corpus Linguistics	for all age groups	Terminology in the Kashubian language
Mubashir Munaf, Hammad Afzal, Naima Iltaf, Khawir Mahmood	2023	Urdu	mBERT, mT5	for all age groups	Methods of Summarizing Texts
R. Ponnusami	2023	Tamil	ChatGPT-3	for all age groups	Generating texts in Tamil
Prabhi Siddhesh Kadam	2021	English	Research in the form of a questionnaire	2–5 years old	The impact of multimedia on learning skills
Lou Parmavati et al.	2022	English–Indonesian–Balinese	DnD	primary school	Multilingual thematic digital dictionary
Gaziza Yelibayeva et al.	2021	Kazakh	Ontology	for all age groups	Dictionary phrases in the Kazakh language

Based on the analysis of the studies presented in Table 2, it can be noted that the development of a dictionary for preschool children for cognitive languages is not yet so developed. After all, creating and preparing a dictionary for the Kazakh language using AI is a very difficult and new approach that can significantly contribute to this area's development and help many children in speech development.

The integration of visualization and voiceover modules in educational tools has been a significant area of research. Various studies have highlighted the importance of visual and auditory aids in enhancing language learning, particularly for young children. According to Piaget's theory of cognitive development [36], children in the preoperational stage (ages 2–7) benefit greatly from visual aids as they rely heavily on visual thinking and imagery. Similarly, Vygotsky's work [37] emphasizes the role of social interaction and language in cognitive development, supporting the use of voiceover to facilitate learning.

Studies have shown that visual aids, such as images and videos, can significantly enhance vocabulary acquisition in young learners. For instance, a study by Verhallen and Bus [38] demonstrated that children who were exposed to multimedia storybooks showed improved vocabulary and comprehension skills compared to those who were only given traditional books. Using images and animations helps contextualize new words, making them easier for children to understand and remember.

The application of AI in educational tools has further enhanced the effectiveness of visualization and voiceover modules. AI-driven tools can provide personalized learning experiences by adapting to each child's individual needs. For example, AI algorithms can analyze a child's progress and adjust the difficulty level of the content accordingly. Studies have shown that personalized learning can significantly improve educational outcomes [39].

The use of AI for text-to-speech (TTS) technology has also advanced significantly, making it possible to generate high-quality, natural-sounding voiceovers. This technology can be particularly beneficial for children learning to read, as it allows them to hear the correct pronunciation of words and phrases. Research has shown that TTS can support reading development by providing auditory reinforcement and aiding in recognizing written words [40].

Visualization and audio support modules significantly enhance children's learning experience by making the educational process more engaging and effective. According to Professor Prem Sunder in "The Effectiveness of Audio-Visual Aids in Teaching-Learning Process," [41] audio-visual aids stimulate students' interest and curiosity, providing intrinsic motivation and making learning enjoyable. This is achieved by transforming traditional verbal instruction, which can often be monotonous, into a dynamic and interactive experience that captures students' attention. Sunder's research involving 10th-grade students and teachers in Gohana City demonstrated that those exposed to audio-visual aids performed significantly better than those who received traditional instruction, highlighting the effectiveness of these tools in enhancing educational outcomes.

On the other hand, the longitudinal study by Sundqvist et al. on "The relationship between children's exposure to screen media and vocabulary development" [42] suggests a nuanced perspective. While the study found a negative relationship between excessive screen media use and vocabulary development, it underscores the importance of balancing screen time with other forms of learning. The study tracked children from infancy to early childhood, revealing that high-screen media exposure could impede language development. This suggests that while audio-visual aids are beneficial, they must be moderated and integrated thoughtfully within a broader educational strategy to prevent potential adverse effects on language skills.

In summary, visualization and audio support modules are valuable in making learning more effective and engaging for children, as demonstrated by the positive impacts on motivation and performance in Sunder's study. However, as indicated by Sundqvist et al., it is crucial to monitor and balance the use of these tools to ensure they complement rather than hinder overall cognitive development. Integrating audio-visual aids in a balanced manner can foster an enriching learning environment that supports the all-round development of children.

Moreover, visualization and voiceover can help children develop their cognitive skills, including analyzing, synthesizing, and making decisions. This can be especially useful when learning new or complex topics.

Over the past ten years, digital media have become more accessible and increasingly used among preschoolers [43], largely due to intuitive design. According to the UK Occam report, by 2018, 58% of children aged 3 and 4 used tablets; 19% of them had their own tablets. Thus, young children are surrounded by educational technologies and are native users of smartphones, tablets, and computers; therefore, they have been called "digits", a term emphasizing the fact that digital media have been present in their lives since birth.

In conclusion, it can be noted that if you create an application that uses a correctly compiled children's thematic dictionary, it can significantly improve the development of speech in children.

3. Methodology

This section presents the full methodology of the proposed approaches to building the dictionary adapted for children aged 3 to 5 with a sufficient level of speech development and language skills to be able to independently express words and phrases in the Kazakh

language, as well as the ability to maintain attention and concentration during the process of recording and reproducing speeches in the Kazakh language that were selected for recording. During the study, with the permission of the parents, an experiment was conducted with children aged 3–5 years who live in Almaty, Kazakhstan. The experiment was conducted the following way: the group of educators was given an application for teaching the spoken Kazakh language to children, and the educators used it to teach the children. The children are ordinary children of the Kazakh group, mostly boys.

Children begin to develop speech at different ages. For example, at 2–3 years of age, a child starts to speak distortedly: words he or she made up, words with mistakes, and different sounds. Then, by 3–5 years, the child's speech becomes more intelligible, he or she begins to use verbs in speech, and the number of pronouns, adverbs, and adjectives increases. That is, he or she begins to speak correctly. This age is considered the most correct time for studying the development of a child's speech [44].

The children were selected for the study based on certain criteria, including their age (2 years 11 months–5 years 11 months) and their level of literacy development in [45]. In addition, the child's ability to control their behavior and attention could be an important criterion since these were the skills that were studied in the study. Thus, the children included in the sample met the specified parameters and were ideal candidates for studying the influence of inhibitory control on the development of literacy skills.

In this article, we aimed to investigate the speech development of Kazakh-speaking children in this age range to understand any common speech difficulties better and provide appropriate interventions where needed. By focusing on this specific age group, we were able to gather valuable information about the typical speech patterns and potential problems faced by Kazakh speakers aged 3 to 5 years.

It is based on methodological rigor, which is a key concept to ensuring the validity and reliability of scientific research. The principles include the general description of the development of the children's dictionary. The subsequent steps provide thorough descriptions of each phase regarding selecting research methods and tools, including the adequacy and validity of the methods used in the context of the tasks at hand. The lowest level provides smaller details, such as the correctness and presentation of results, which, although less critical, are also important for the overall perception and evaluation of the work. By following these levels of rigor, the methodological section ensures the high quality and integrity of the research conducted with the most prominent obtained results.

The methodological part of building a qualitative dictionary for children encompasses several important stages that contribute to thoroughly describing its design, evaluation, and enhancement. The development of the children's dictionary includes the following parts:

- The generation of reference words is a significant step that allows getting the basic important words on which the successive generated phrases and sentences rely. The ChatGPT model is effectively employed in this step to generate a comprehensive list of reference words, and the basic initial words serve as a dictionary's foundation and are selected based on the relevance and frequency of their use by children. The variety of reference words is defined by adding support parameters highlighting the necessity of including five nouns, adjectives, verbs, and adverbs in the list of reference words. In addition, after all phrases and sentences are generated, their meaningfulness and adequacy are checked and verified by specially qualified linguists and educators.
- The obtained reference words are then utilized to generate phrases and sentences that form the children's dictionary. These phrases and sentences are formed in English; translating them into Kazakh using a high-quality translation system is also essential. Here, the Google Translate system is used.
- When the phrases and sentences are translated into Kazakh, the filtering phase takes place. In this phase, all phrases and sentences go through morphological analysis to ensure their grammatical correctness. They are split into bigrams [46] and trigrams [47] that present very simplified forms of words. Bigrams present a sequence of two consecutive words in the text. For example, there is a sentence "Natural lan-

guage processing is interesting". The bigrams of this sentence will be [("Natural", "language"), ("language", "processing"), ("processing", "is"), ("is", "interesting")]. At the same time, trigrams include a sequence of three consecutive words. The trigrams of the sentence will be [("Natural", "language", "processing"), ("language", "processing", "is"), ("processing", "is", "interesting")]. The use of bigrams and trigrams helps in text analysis and pattern recognition in natural language processing. Then, specially prepared templates are utilized to filter the content and align it with educational standards. This step proves that the generated phrases are age-appropriate and favorable to learning.

The whole methodological structure is shown in Figure 2.

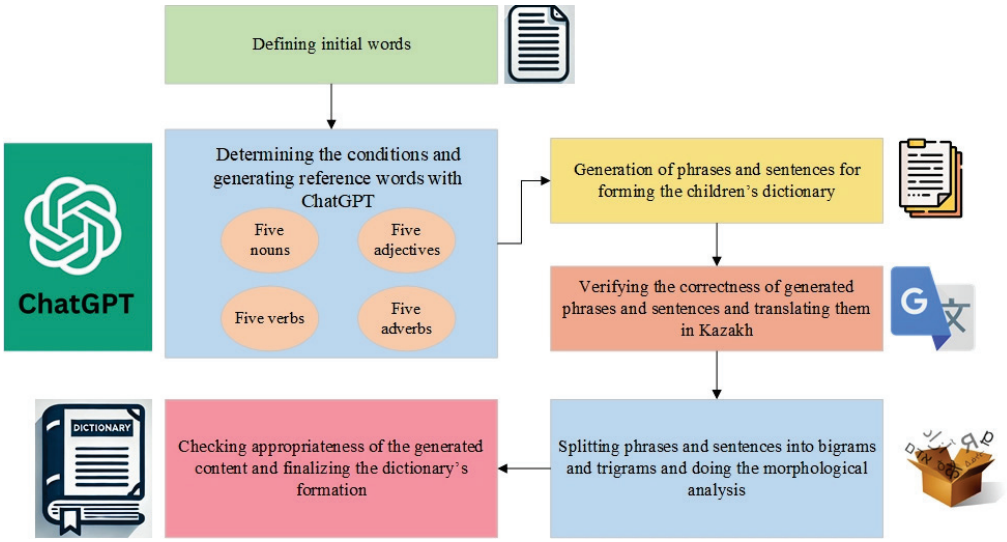


Figure 2. The structure of the proposed methodology.

Although the phrases and sentences were initially checked for usefulness and adequacy, it was essentially important to prepare them for building the structured content of the dictionary and subsequent audio and visualization module. Therefore, additional evaluation metrics, such as cosine similarity, Euclidean distance, and Manhattan distance, were utilized to check the importance of phrases and the order in which children are going to train and educate themselves with the designed module. The description of the phrases and sentence generation is described in Section 5.

The formed dictionary of phrases and sentences is then extended with the audio and visualization module that consists of a didactic game developed as the primary tool for delivering the dictionary's content. The visualization component is crafted with child-friendly graphics and animation of phrases and sentences from the formed dictionary to capture children's attention and facilitate learning. The designed cognitive load ensures that visual elements support rather than overwhelm the learning process. The audio component is synchronized with visual elements and provides a seamless learning experience. The development and specifications of the audio and visualization module is shown in Section 6.

The importance of eye contact and other communication cues is also evident in studies of how and when young children learn from video and other forms of digital media, which was considered in [48].

The study found that communicative pedagogical contexts promote the development of conceptual thinking in children aged 4 and 5, helping them to analyze and classify information more consciously and purposefully [49]. Interaction with an adult who actively draws the child's attention to specific properties of objects helps children better understand and use this information in various situations.

All the important stages of the children's dictionary formation are shown in Sections 3.1–3.3.

3.1. The Generation of Reference Words for the Children's Corpus Using ChatGPT

In this study, an important task was the generation of reference words. The generation is realized with the use of ChatGPT, which was developed by OpenAI and is based on the GPT (Generative Pre-trained Transformer) architecture, trained on vast amounts of textual data. It is considered one of the most advanced models in the field of natural language processing nowadays. Since Kazakh remains a low-resource language [50], AI models, including ChatGPT, still make significant errors when generating sentences and texts in this language. Therefore, using English as the source language is preferable, as virtually all language models work excellently with it. This approach helps increase the accuracy of the initially generated data, minimizing possible errors and inaccuracies.

First, in the generation of reference words, a list of initial English words was formed based on three main themes that help children better understand their surroundings: animals, gardens, and toys.

This list of words was presented as (1):

$$W = \{w_1, w_2, \ldots, w_n\}, \tag{1}$$

where W is a set of words relevant to children aged 3 to 5.

The complete list of words for three thematic categories is presented below:

1. Animals: Hippopotamus, Eagle, Wolf, Goose, Giraffe, Hare, Goat, Cow, Cat, Swan, Lion, Fox, Bear, Mouse, Sheep, Deer, Rooster, Python, Parrot, Elephant, Dog, Tiger.
2. Garden: Banana, Carrot, Cherry, Apple, Potato, Flower, Tree, Pear, Watermelon, Cucumber, Berry, Tomato, Garden, Fruit, Apricot, Pepper, Lemon, Beet, Vegetable, Pumpkin.
3. Toys: Dolls, Cars, Toy trains, Toy helicopters, Stuffed animals, Robots, Lego, Puzzles, Building blocks, Ninja turtles.

After forming the list of words according to the given categories, a list of semantically related reference words was generated for each word in the list. The ChatGPT models have an extensive vocabulary and can use various synonyms and semantically related words, increasing the output data's variability. This is especially useful where a high degree of diversity is needed but may not always be suitable for tasks requiring high accuracy. Therefore, to better control the reference words, categories of parts of speech are specified: in this case, five words are needed for the parts of speech: nouns, verbs, adjectives, and adverbs. The formulation of reference word generation is presented in the following way.

For each word $w_i \in W$, a set of semantically related words was generated with the use of ChatGPT (2):

$$R(w_i) = ChatGPT(\{r_{i1}, r_{i2}, \ldots, r_{im}\}), \tag{2}$$

where $R(w_i)$ is the set of semantically related words for w_i received by the use of ChatGPT.

The obtained variety of reference words is explained by the complexity of ChatGPT models that are trained on large volumes of data and have an extensive database of words and their associations in phrases and sentences. These models deeply analyze the contexts in which given words are used and compare them with the contexts of other words. Words that appear in similar contexts are assigned higher weight values, while words related to the original word in specific and limited scenarios receive lower weight values. For example, the following words were generated by ChatGPT as related words for "banana": ['fruit—0.95', 'plantain—0.92', 'yellow—0.89', 'tropical—0.85', 'peel—0.83', 'potassium—0.80', 'smoothie—0.78', 'ripe—0.75', 'sweet—0.72', 'smooth—0.70', 'curved—0.68', 'nutritious—0.65', 'healthy—0.63', 'delicious—0.60', 'banana—0.58', 'quickly—0.55', 'easily—0.52', 'softly—0.50', 'carefully—0.48', 'happily—0.45', 'blend—0.43', 'slice—0.40', 'peel—0.38', 'munch—0.35', 'devour—0.33', 'beat—0.30', 'mash—0.28', 'mix—0.25', 'boil—0.23', 'fry—0.20', 'bake—0.18', 'grill—0.15', 'spread—0.13', 'melt—0.10', 'dry—0.08', 'crumble—0.05', 'crush—0.03', 'blend—0.02', 'slurp—0.01', 'puree—0.005'].

The structure of reference word generation is shown in Figure 3.

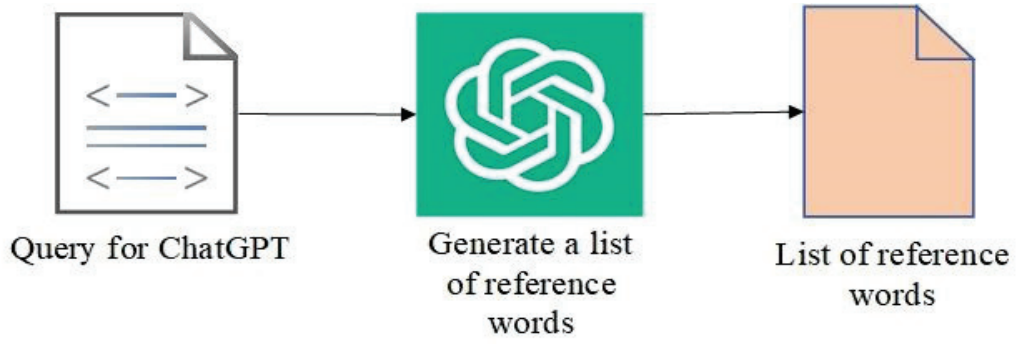


Figure 3. The structure of reference word generation.

3.2. The Generation of Phrases and Sentences for the Children's Dictionary

After forming the list of words for four parts of speech, the next stage involves generating phrases and sentences. It is expressed as (3):

$$S_i = ChatGPT(w_i, R(w_i)), \tag{3}$$

where S_i is the set of phrases and sentences generated for the word w_i and its related words $R(w_i)$.

A set of specific conditions was defined for the words in each category for better generation results, and the list of obtained words was included in the ChatGPT query. The conditions are specific to each category. Important aspects of phrases and sentences involving animals include their size, color, diet, and habitat. For garden plants, the focus is on size, color, taste, and place of growth. For toys, characteristics such as size, color, structure, and other features are essential.

For example, the following phrases and sentences were generated by ChatGPT for Banana: [yellow bananas are sweet and ripe; tropical plant grows easily in sunshine; delicious fruit with potassium and vitamins; smoothie blend with sliced banana fruit; healthy snack for happy munching time; softly peel and carefully slice banana; nutritious and delicious, easily devoured; quick snack of a yellow banana; curved fruit grows in tropical regions; easily mashed for yummy banana bread; smoothie blend of ripe banana and berries; happy kids love to slurp smoothies; healthy snack option for quickly munching; yellow bananas easily blend into smoothies; ripe fruit slices spread on toast; carefully peel bananas for baking treats; delicious and nutritious banana fries; softly mash for a healthy treat; slice, blend, and slurp banana smoothie; nutritious snack quickly devoured by kids; potassium-rich fruit for a healthy snack; easily crushed for yummy banana ice cream; ripe banana slices easily spread on toast; delicious banana smoothly blends in smoothies; healthy snack quickly devoured by children].

After the generation of phrases and sentences, it was also required to translate them from English to Kazakh. In this way, the Google Translate system was implemented (4):

$$T(S_i) = GoogleTranslate(S_i) \tag{4}$$

The whole structure of phrases and sentence generation and their translation to Kazakh is presented in Figure 4.

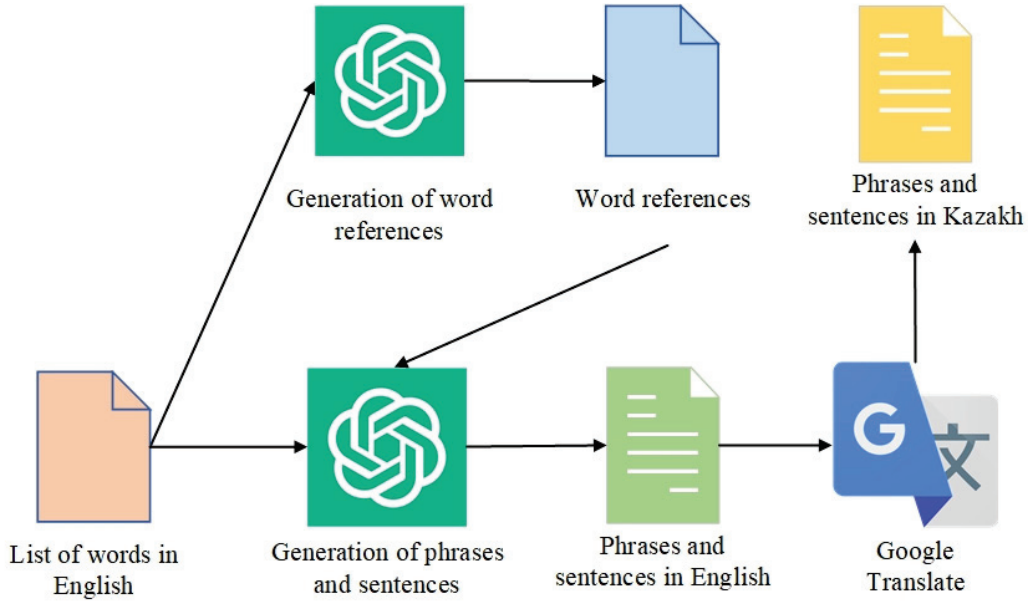


Figure 4. The system's architecture for generating phrases and sentences and translating them into Kazakh.

The translated phrases and sentences are then analyzed and filtered using morphological analysis, as described in Section 3.3.

3.3. The Morphological Analysis of Generated Phrases and Sentences for the Children's Dictionary

In the previous stages, the reference words, phrases, and sentences were generated. The key feature of those steps is that text generation for the dictionary requires a special approach and careful consideration of the target audience's characteristics to create quality content that will be interesting and accessible to young children. Generation systems often fail to produce texts that are easily understood by children aged 3 to 5 years. When generating content, it is important to consider the age-specific characteristics of children to create texts that are accessible and understandable. To ensure texts are comprehensible to young children, it is also necessary to use simple language and visual imagery. The generation stage was conducted at the sentence level, as ChatGPT struggled to generate simple phrases for children aged 3 to 5 years. It did not account for the peculiarities of children's perception and produced complex texts and phrases that children could not understand. The generated phrases were not suitable for the vocabulary of children aged 3 to 5 as they had the complexity level of adult language proficiency.

The filtering of the thematic dictionary was carried out taking into account the following parameters:

- The structure of children's speech;
- Complexity;
- Problematic sounds of Kazakh speech in children (for example, р/л[r/l], ш/с [sh/s], ө/ү [o/u], etc.).

The thematic dictionary will be presented in the following form:

$$\begin{gathered} \mathrm{D} = \left\{ u_i, f_{ij}, S_{jk} \right\} \\ \left\{ u_i \bigcup G \right\} \in f_{ij} \}, \\ \left\{ f_{ij} \bigcup G \right\} \in S_{jk} \}, \\ G := < w, R_{pos} >, \\ i = 1, n \; j = 1,5 \quad k = 1,15 \end{gathered} \tag{5}$$

where u_i are words, f_{ij} are phrases (bigrams and trigrams) for i word, s_{jk} are sentences in the Kazakh language, w is all words in Kazakh, R_{pos} is a set of rules for part of speech sequences in Kazakh. G is a set consisting of letters w and over, which is a set of rules, and R_{pos} is carried out, which is described in Table 3.

Table 3. Patterns of bigrams and trigrams for children aged 3–5 based on the proposed simple sentence structure of the Kazakh language.

Pattern	Example	Translated Version for English-Speaking Readers	
word1<n> word2<n>	Банан ағашы	[Banana tree]	n—noun v—verb adj—adjective cnjcoo—conjunction adv—adverb num—numerical prn—pronoun
word1<n> word2<v>	Банан өседі	[Banana grows]	
word1<adv> word2<v>	Ағашта өседі	[Grows on a tree]	
word1<adj> word2<n>	Тәтті банан	[Sweet banana]	
word1<num> word2<n>	Екі банан	[Two bananas]	
word1<prn> word2<n>	Оның бананы	[His banana]	
word1<prn> word2<v>	Ол алды	[He took]	
word1<n> word2<n> word3<v>	Банан ағашы өседі	[The banana tree grows]	
word1<adj> word2<n> word3<n>	Тәтті банан ағашы	[Sweet banana tree]	
word1<n> word2<v> word3<v>	Банан қарайып жатыр	[The banana is turning black]	
Word <n> word <cnjcoo> word3<n>	Банан мен алма	[Banana and apple]	
Word <adj> word <cnjcoo> word3<adj>	Тәтті және жұмсақ	[Sweet and soft]	

After the sentence generation stage, the generated sentences were translated into Kazakh using Google Translate.

Examples of generated sentences for "мысық (cat)":

Мысықт*а*р б*а*лық пен т*а*уық сияқты ет жейді. *[Cats eat meat such as fish and chicken.]*

Мысықт*а*р әдетте кішкент*а*й және жұмс*а*қ. *[Cats are usually small and soft.]*

Көптеген мысықт*а*р жылы үйлерде тұр*а*ды *[Many cats live in warm houses.]*

The received sentences were further divided into bigrams and trigrams. An example of splitting a translated generated sentence in Kazakh into bigrams:

- Example of a bigram: *(*Мысықт*а*р б*а*лық *[cats fish]) (*б*а*лық пен *[fish and]) (*пен т*а*уық *[and chicken]) (*т*а*уық сияқты *[such as chicken]) (*сияқты ет *[such as meat]) (*ет жейді *[eat meat])*.
- Example of a trigram: *(*Мысықт*а*р б*а*лық пен *[cats fish and]) (*б*а*лық пен т*а*уық *[fish and chicken]) (*пен т*а*уық сияқты *[and such as chicken]) (*т*а*уық сияқты ет *[meat such as chicken) (*сияқты ет жейді *[eat meat such as])*.

Next, morphological analysis was implemented for each bigram and trigram using the Apertium system. Apertium-Kaz is a morphological transducer and disambiguator for Kazakh [51]. The morphology of the Kazakh language and its morphological analyzers are described in works [52–61]. The Kazakh language is a complex language with rich morphology.

Example of morphological tagging for phrases [https://github.com/apertium/apertium-kaz, accessed on 28 September 2024]:

^Мысықт*а*р/мысық<*n*><*pl*><*nom*>$ ^б*а*лық/б*а*лық<*n*><*nom*>$

^пен/мен<*cnjcoo*>$ ^т*а*уық/т*а*уық<*n*><*nom*>$

^сияқты/сияқты<*post*>$ ^ет/ет<*n*><*nom*>$

^жейді/же<*v*><*tv*><*aor*><*p3*><*sg*>$^./.<*sent*>$

[^Cats/cat<n><pl>$ ^fish/fish<n><sg>$

^and/and<cnjcoo>$ ^chicken/chicken<n><sg>$

^such as/such as<post>$ ^meat/meat<n><sg>$

^eat/eat<v><vblex><inf>>$^./.<sent>$]

In the next stage, patterns of frequently used combinations of parts of speech in the Kazakh language for children aged 3 to 5 were developed. The patterns are presented in

the table below. In the Kazakh language, there are nine parts of speech: noun, adjective, numeral, pronoun, verb, adverb, postposition, onomatopoeia, and interjection.

The simple sentence structure in the Kazakh language is the following [55]: "prn v", "prn adv v", "prn n v", "prn adj n v", "prn n n v", "prn n adj n v", "prn n adv adv v", "prn n adv v", "adv n v", "prn n v adv n v", "prn v prn n adj, adj n v", "prn adv adj, adj n v", "prn adv adj n v". This approach is based on a complete set of suffixes of the Kazakh language. The creation of the complete set was developed by Tukeyev and described in [62].

In the proposed work [63], simple sentence construction is not suitable for children aged 3 to 5 as they do not use phrases longer than three words in a sentence. Table 3 presents patterns for children aged 3 to 5 based on the proposed simple sentence structure of the Kazakh language.

Filtered bigrams and trigrams based on the Kazakh language part-of-speech combination templates with analyses are presented below (excerpt):

- б*а*лық<*n*> пен<*cnjcoo*> т*а*уық<*n*> *[fish<n> and<cnjcoo> chicken<n>]*
- ет<*n*> жейді<*v*> *[meat<n> eat*і<*v*>*]*

The resulting list of bigrams and trigrams (excerpt) for children based on the example sentences:

- Б*а*лық пен т*а*уық *[fish and chicken]*
- Ет жейді *[eat meat]*

The filtered list has been added to the phrase level in the dictionary. The thematic dictionary for children in the final result is shown in Table 4 (excerpt):

Table 4. Thematic dictionary for children (part).

Words in Kazakh	Phrases in Kazakh (for Children 3–4 Years Old)	Translated Version for English-Speaking Readers	Sentences in Kazakh (for Children 3–4 Years Old)	Translated Version for English-Speaking Readers
Мысық [Cat]	Балық пен тауық Ет жейді кішкентай және жұмсақ үй жануары сары мысық ұйықтағанды жақсы көреді ...	[Fish and chicken Eats meat small and soft pet yellow cat likes to sleep] ...	Мысықтар балық пен тауық сияқты ет жейді. Мысықтар әдетте кішкентай және жұмсақ. Көптеген мысықтар жылы үйлерде тұрады. Мысықтар далада ұйықтағанды жақсы көреді. Бізде сары мысық бар. ...	[Cats eat meat such as fish and chicken. Cats are usually small and soft. Many cats live in warm houses. Cats like to sleep outside. We have a yellow cat.] ...
Банан [Banana]	сары банан тәтті банан тропикалық жеміс тәтті және піскен . . .	yellow banana sweet banana tropical fruit sweet and ripe ...	Сары банан тәтті және піскен Тропикалық өсімдік күн сәулесінде оңай өседі Калий мен дәрумендері бар дәмді жеміс Иілген жемістер тропиктік аймақтарда өседі Бақытты балалар коктейльді жегенді жақсы көреді. ...	The yellow banana is sweet and ripe. The tropical plant grows easily in the sunlight. A delicious fruit with potassium and vitamins. Curved fruits grow in tropical regions. Happy children like to eat smoothies. ...
Қуыршақ [Doll]	қызғылт қуыршақ қуыршағым әдемі барби қуыршақ . . .	pink doll my doll is beautiful Barbie doll ...	әдемі қызғылт қуыршақ ойыншығын әперді. менің қуыршағым әдемі көрінеді. ойын уақытына арналған көне барби қуыршақ үйі киіну және таңдану үшін сүйкімді қуыршақ ...	bought a beautiful pink doll toy. my doll looks beautiful. a vintage Barbie dollhouse for playtime. a cute doll for dressing up and admiring. ...

4. Practical Results—Evaluation of the Content of the Developed Dictionary

The results of the words and phrases obtained for the thematic dictionary of the Kazakh language are presented below in Table 5. Next, it is necessary to evaluate the results obtained for the proximity of phrases to the original word using metrics such as cosine similarity, Euclidean distance, and Manhattan distance. These metrics are used in various areas, like machine learning, data analytics, pattern recognition, etc. Euclidean distance evaluates the distance between two texts represented in a vector form. Cosine similarity measures the angle between vectors, indicating how similar the two pieces of text are based on the cosine of the angle between them. A value closer to 1 implies a higher similarity. On the other hand, Manhattan distance calculates the straight-line distance between two vectors in space, indicating how far apart the two pieces of text are. Lower values represent closer proximity. However, before calculating the values of these metrics, it is necessary to convert the textual representation of the phrases into a vector representation. Several methods are applied, including Bag of Words, TF-IDF Metric, and Word Embedding Methods like Word2Vec, FastText, and BERT. These methods have proven effective in various NLP tasks. For the calculation of word vectorization, the BERT model was applied [64]. It converts words and phrases into vectors of 768 values, which allows the application of the corresponding metrics to them. For example, there are words and phrases vectorized by the BERT model.

$$\begin{gathered} word_embedding = [-0.12224, -0.42816, -0.70929, 0.32610, -0.20067, \\ -0.75138, -0.48973, \ldots, 0.1356] \\ phrase_embedding = [-0.46193, -0.41801, 0.68839, -0.84216, 0.52520, \\ -0.61788, 0.10261, -0.12406, \ldots, 0.29184] \end{gathered}$$

Then, the metrics that work well with vectorized data are applied to these vectors. They have the following features:

- Cosine similarity is a measure that calculates the angle between vectors in space, determining their similarity based on the cosine of the angle between them. This measure is calculated using Formula (6):

$$\cos\text{ine_similarity}(\mathrm{A}, \mathrm{B}) = \frac{A \times B}{\parallel A \parallel \times \parallel B \parallel}, \tag{6}$$

where $A \times B$ is the dot product of vectors A and B; $\parallel A \parallel$ is the norm of vector A, calculated $\sqrt{\sum_{as\ i=1}^{n} A_i^2}$, and $\parallel B \parallel$ is the norm of vector B, calculated as $\sqrt{\sum_{i=1}^{n} B_i^2}$.

- Euclidean distance represents the direct distance between two points in space, defined as the square root of the sum of the squares of the differences in the coordinates of the points. It is calculated using Formula (7):

$$\mathrm{d}(\mathrm{A}, \mathrm{B}) = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}, \tag{7}$$

where A and B are points in n-dimensional space.

Manhattan distance is calculated as the sum of the absolute differences between the coordinates of the points along each axis. It is computed using Formula (8):

$$d_{Manhattan}(\mathrm{A}, \mathrm{B}) = \sum_{i=1}^{n} |A_i - B_i|, \tag{8}$$

These metrics can be useful when creating a dictionary for evaluating and comparing texts or other data based on their vector representation. They are shown in Table 5.

Table 5. The results of evaluating the filtered thematic dictionary of the Kazakh language (excerpt).

Words	Cosine Similarity	Translated Version for English-Speaking Readers	Euclidean Similarity	Translated Version for English-Speaking Readers	Manhattan Distance	Translated Version for English-Speaking Readers
Бегемот [Hippopotamus]	бегемот баяу: 0.7387	[hippopotamus slow: 0.7387]	жасыл өсімдіктерді жейді: 11.5877	[eats green plants: 11.5877]	жасыл өсімдіктерді жейді: 255.5542	[eats green plants: 255.5542]
	күшті бегемот: 0.7100	[strong hippopotamus: 0.7100]	күшті бегемот: 7.4231	[strong hippopotamus: 7.4231	күшті бегемот: 164.1093	[strong hippopotamus: 164.1093]
	жасыл өсімдіктерді жейді: 0.3927	[eat green plants: 0.3927]	бегемот баяу: 6.8238	[hippopotamus] slow: 6.8238]	бегемот баяу: 151.3433	[hippopotamus slow: 151.3433]
Қоян [rabbit]	ақ қоян: 0.8353	[white rabbit: 0.8353]	тез жүгіреді: 11.0343	[runs fast: 11.0343]	тез жүгіреді: 241.4868	[runs fast: 241.4868]
	қоян секіреді: 0.6306	[rabbit jumps: 0.6306]	қоян секіреді: 8.7781	[rabbit jumps: 8.7781]	қоян секіреді: 194.1711	[rabbit jumps: 194.1711]
	тез жүгіреді: 0.4350	[runs fast: 0.4350]	ақ қоян: 5.8949	[white rabbit: 5.8949]	ақ қоян: 130.0948	[white rabbit: 130.0948]
Мысық [cat]	сары мысық: 0.8190	[yellow cat: 0.8190]	үй жануары: 9.9984	[pet: 9.9984]	үй жануары: 220.0396	[pet: 220.0396]
	ұйықтағанды жақсы көреді: 0.6115	[loves to sleep: 0.6115]	ұйықтағанды жақсы көреді: 8.9627	[loves to sleep: 8.9627]	ұйықтағанды жақсы көреді: 196.4853	[loves to sleep: 196.4853]
	үй жануары: 0.4968	[pet: 0.4968]	сары мысық: 6.4826	[yellow cat: 6.4826]	сары мысық: 142.3616	[yellow cat: 142.3616]
Банан [banana]	сары банан: 0.7332	[yellow banana: 0.7332]	тропикалық жеміс: 10.4304	[tropical fruit: 10.4304]	тропикалық жеміс: 229.1729	[tropical fruit: 229.1729]
	жұмсақ банан: 0.6780	[soft banana: 0.6780]	жұмсақ банан: 8.7045	[soft banana: 8.7045]	жұмсақ банан: 191.9145	[soft banana: 191.9145]
	тропикалық жеміс: 0.5293	[tropical fruit: 0.5293]	сары банан: 7.9889	[yellow banana: 7.9889	сары банан: 173.0328	[yellow banana: 173.0328]
Сәбіз [carrots]	сары сәбіз: 0.8279	[yellow carrot: 0.8279]	қоян жейді: 8.5807	[rabbit eats: 8.5807]	қоян жейді: 188.2283	[rabbit eats: 188.2283]
	сәбіз көкөніс: 0.7868	[carrot is a vegetable: 0.7868]	сәбіз көкөніс: 6.4967	[carrot is a vegetable: 6.4967]	сәбіз көкөніс: 143.0402	[carrot is a vegetable: 143.0402]
	қоян жейді: 0.6035	[rabbit eats: 0.6306]	сары сәбіз: 5.7763	[yellow carrot: 5.7763]	сары сәбіз: 123.5438	[yellow carrot: 123.5438]

The applied three metrics allow the generated phrases to be in the appropriate order for the following visualization and audio module. It is especially useful for children to pay attention to the most significant aspects of the word category. Here, every metric could be used to order word phrases, but it is better to give preference to the one where the difference between values is larger.

In addition, the average cosine similarity [65], Euclidean similarity [66], and Manhattan distance [67] metrics were calculated for every topic used, highlighting the most significant of them in the conducted experimental research. It demonstrates which topic out of the three is preferred to be shown first to the children. The average proximity score for the received thematic dictionary is presented in Table 6.

Table 6. Average proximity score by category in the thematic dictionary.

Categories	Cosine Similarity	Euclidean Similarity	Manhattan Distance
garden	0.6724	8.328	182.941
animals	0.578	9.204	202.1627
toys	0.59	9.1	200.4

As we can see from the table, the closer the cosine similarity value is to 1, the more similar the vectors are and, consequently, the phrases or texts they represent. Thus, in phrase creation, a cosine similarity closer to 1 indicates that the phrases have a higher degree of similarity with each other. When creating a thematic dictionary, the values of Euclidean similarity indicate that some phrases or keywords will have a greater influence in forming the thematic dictionary.

Table 7 provides a compilation of curated phrases suitable for children aged 3 to 5 years. The table was generated through a systematic and multi-step process involving the application of AI tools and linguistic analysis. An English word list was initially formed and categorized into specific themes, including animals, the garden, and toys. Using ChatGPT, semantically related words were generated for each entry in this list. Subsequently, phrases and sentences were constructed utilizing these semantically related words.

Table 7. Filtered phrases for children aged 3–5 years (fragment).

Category	Words, Phrases, and Sentences (Kazakh)	Translation (English)
animals	‘Қоян’: [‘қоян секіреді’, ‘тез жүгіреді’, ‘ақ қоян’], ‘Ешкі’: [‘кішкентай қозы’, ‘ешкінің сүті’, ‘мүйізді ешкі’], ‘Мысық’: [‘сары мысық’, ‘үй жануары’, ‘ұйықтағанды жақсы көреді’],	‘Rabbit’: [‘rabbit jumps’, ‘runs fast’, ‘white rabbit’], ‘Goat’: [‘small kid’, ‘goat’s milk’, ‘horned goat’], ‘Cat’: [‘yellow cat’, ‘pet’, ‘likes to sleep’],
garden	‘Банан’: [‘жұмсақ банан’, ‘сары банан’, ‘тропикалық жеміс’], ‘Сәбіз’: [‘сары сәбіз’, ‘сәбіз көкөніс’, ‘қоян жейді’], ‘Алма’: [‘жасыл алма’, ‘қызыл алма’, ‘алма ағашы’],	‘Banana’: [‘soft banana’, ‘yellow banana’, ‘tropical fruit’], ‘Carrot’: [‘yellow carrot’, ‘carrot vegetable’, ‘rabbit eats’], ‘Apple’: [‘green apple’, ‘red apple’, ‘apple tree’],
toys	Қуыршақ: [‘әдемі қуыршақ’, ‘тірі қуыршақ’, ‘барби қуыршақ үйі’, ‘миниатюралық манекен’, ‘сүйкімді қуыршақ’, ‘әдемі жасалған қуыршақ’, ‘қуыршақ үйі’] Лего: [түрлі-түсті пластик, ойыншық кірпіштері, лего конструкциясы, пластикалық бөлшек, ойнап жобалайды, ойнап құрастырады, түрлі-түсті модульдік лего, құрылыс жинағы]	Doll: [‘beautiful doll’, ‘living doll’, ‘Barbie dollhouse’, ‘miniature mannequin’, ‘cute doll’, ‘well-made doll’, ‘dollhouse’] Lego: [‘colorful plastic’, ‘toy bricks’, ‘Lego construction’, ‘plastic piece’, ‘plays and designs’, ‘plays and builds’, ‘colorful modular Lego’, ‘construction kit’]

The generated sentences were translated from English to Kazakh to accommodate the target language requirements. These sentences were divided into bigrams and trigrams following translation to facilitate detailed linguistic analysis. Morphological analysis was performed using the Apertium system, ensuring correct usage and suitability for children. Finally, a filtering process based on part-of-speech (POS) tag templates was employed to refine the bigrams and trigrams. This resulted in a comprehensive and appropriate list of phrases for the intended age group. This rigorous approach ensured the educational tool’s accuracy, relevance, and effectiveness for young learners. The general collected data for the thematic dictionary are shown in Table 8.

Table 8. The results of the collected data for children.

Categories	Number of Unigrams	Number of Bigrams	Number of Trigrams	Number of Generated Long Sentences	Number of Generated Short Sentences
garden	40	183	37	703	573
animals	35	159	33	806	895
toys	30	108	19	596	341

These key stages helped create a list of words and phrases in the Kazakh language that can be used for teaching children. Moreover, the use of morphological analysis made

the list more accurate and relevant to educational needs. Overall, this method allows for the creation of educational materials in the Kazakh language based on semantically related words and phrases in English, which can enhance the learning of the Kazakh language for children, taking their age into account.

In this research, speech therapists are drawn on their knowledge of language development and bilingualism to create content that aligns with children's cognitive and linguistic abilities. By taking into account the unique challenges and opportunities presented by learning the Kazakh language, speech therapists actively consider the contextual and age factors that shape children's language learning experiences and speech for Kazakh-speaking children to develop their communication skills.

Speech therapist teachers checked the quality and compliance of the content of the developed dictionary in three categories: corresponds (1 point), does not correspond (0 points), and partially corresponds (0.5 point). Table 9 shows children's average scores for agreement and inconsistency of words, phrases, and sentences. Based on the results of the speech therapist's assessment of the correspondence of the dictionary's content and the sentences generated for children from 3 to 5 years old, it is concluded that the generated phrases and sentences are suitable for the target age group.

Table 9. The results of a speech therapist's assessment of the dictionary content for compliance with age groups (3–5 years).

Categories	Average Percent of Correctness for Words, %	Average Percent of Bigrams and Trigrams, %	Average Percent of Correctness for Sentences, %
garden	100	96	87
animals	100	98	91
toys	92	87	83

Initially, creating wordlists and phrases in English before translating into Kazakh can lead to inaccuracies and cultural inconsistencies.

To address the issue of English-based dependency when creating wordlists and phrases in Kazakh, the following steps are taken:

1. Research the context: Before creating a wordlist, it is important to understand the context in which the words will be used. This will help avoid misunderstandings and improve the accuracy of the translation.
2. Work with native speakers: Collaborating with native Kazakh speakers or experienced translators will help you better understand the cultural and linguistic nuances, which will help avoid inaccuracies.
3. Testing and feedback: Once the list is created, it is important to test it with the target audience and collect feedback. This will help identify potential errors and improve the quality of the content.

Analysis of Dataset Completeness

To determine the completeness of the data used in the dataset consisting of words used in children's speech, it is possible to use text statistics analysis.

Text statistics visualization is a simple but very informative method.

They include:

- Word frequency analysis;
- Sentence length analysis;
- Average word length analysis;
- Others.

They really help to explore the fundamental characteristics of text data.

To do this, follow these steps:

(1) Determine the number of characters in each sentence;

(2) Determine the number of words in a sentence;
(3) Determine the average word length.

As shown in Table 8, for each category of words, there are more bigrams than trigrams. However, in terms of the number of generated words, there are more long sentences. Since the corpus is for children, using long sentences can lead to difficulty in understanding the text and speech.

Looking at the category of animals, it is possible to determine the frequency of use of words in sentences. The frequency distribution of word usage is depicted in Figure 5, while Figure 6 illustrates the distribution of the top 10 most frequent bigrams.

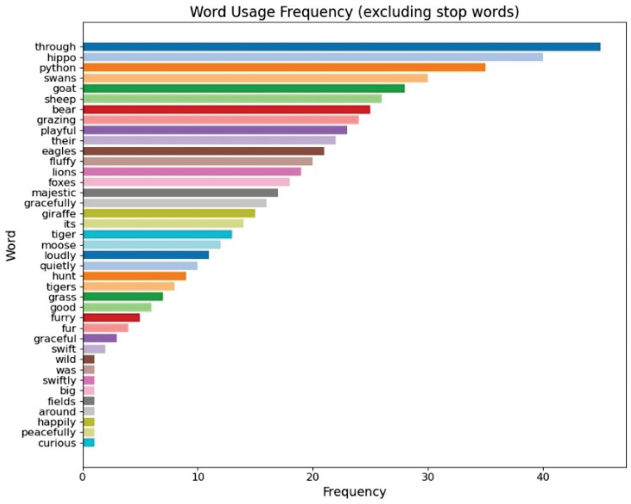


Figure 5. Word usage frequency distribution.

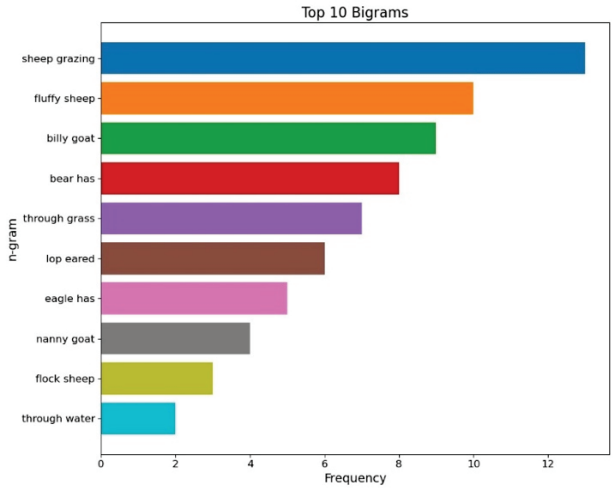


Figure 6. Top 10 bigram word frequency distribution.

Assessing the quality of the dataset is crucial to determining the overall effectiveness of the system. When assessing completeness, it is necessary to consider the needs of users in order to fully measure the quality of big data [68]. The quality assessment is demonstrated in Table 10.

Table 10. The results of a speech therapist's assessment of the dictionary content for compliance with age groups (3–5 years).

Categories	Completeness Rate, %	Missing Data Percentage, %	Weighted Completeness, %
garden	61.6	38.4	62.2
animals	74.6	25.4	75.3
toys	53.9	46.1	54.4

Completeness Rate is the percentage of non-null values in a dataset (5). Missing Data Percentage is the percentage of missing or null values in a dataset (6). Weighted Completeness is assigning weights to different data elements or words and phrases based on their importance (7) [69].

$$\text{Completeness Rate} = \left(\frac{m}{n}\right) \times 100\% \quad (9)$$

$$\text{Missing Data} = \left(\frac{v}{n}\right) \times 100\% \quad (10)$$

$$\text{Weighted Completeness} = \sum_{i=0}^{k} (\text{Completeness Rate}_i \cdot w_i) \quad (11)$$

where m is of non-null values, n is the total number of values, k is the missing values number, w is the word's weight in the dataset.

Assessing the completeness of the collected data shows that despite significant efforts to collect information, there are limitations in the accuracy and completeness of the data.

In this study, factors such as the applicability of the collected data to children were taken into account. A speech therapist checked the data for applicability to children, so the missing data does not refer to the part where words, phrases, and sentences in the Kazakh language were not selected.

There are problems with confirming the reliability of the data with the help of a speech therapist. Some categories of the dictionary may be incomplete (the category "toys"), which may lead to omissions or distortions in the analysis results. Some categories of the dictionary may be fully complete and detailed in our case with animals.

Thus, despite the existing limitations, our approach to the process of collecting and analyzing children's speech data gave holistic results and can be considered quite complete.

5. The Development of the Audio and Visualization Module for Delivering the Dictionary's Content

The main tool for developing preschool children's vocabulary is the didactic game. Didactic games are a widely used method of vocabulary work; these are educational games that can enrich children's vocabulary.

Visualization and voiceover of data can significantly influence how children perceive information. When data are presented in the form of pictures, graphs, or animations, children can better understand and remember the information, as visual materials can be more appealing and comprehensible to them.

Therefore, using visualization and voiceover can significantly improve information perception in children and contribute to their successful learning. Implementing audio and visual support in the developing system for learning new words can improve children's learning process and understanding. To simplify the task's implementation, the following module was developed using AI tools:

- Generation of images according to the algorithm of the developed children's thematic dictionary;
- Implementation of text to speech.

Study of available TTS for the Kazakh language: currently available options are KazakhTTS by ISSAI [70], Narakeet [71], Azure AI (Microsoft) [72].

An expert evaluation was conducted based on the following parameters: sound quality, generation speed, cost, ability to deploy locally, and file format. To assess sound quality, we will calculate the SNR (Signal-to-Noise Ratio).

Signal-to-Noise Ratio (SNR) is a measure used to evaluate the quality of an audio signal by comparing the power of the signal to the power of background noise or interference. SNR is measured in decibels (dB) and is typically defined as the ratio of signal power to noise power.

$$\text{SNR} = \frac{P_{signal}}{P_{noise}} = \left(\frac{U_{signal}}{U_{noise}}\right)^2, \quad (12)$$

All audio evaluation results are shown in Table 11.

Table 11. Evaluation results are based on a 45-character audio recording in the Kazakh language.

Characteristic	ISSAI KazakhTTS	Narakeet KazathTTS	Azure AI KazakhTTS
SNR, dB	22.8	21.7	25.6
Generation speed, sec	1	10	2
Free of charge	Yes	No	No
The ability to install locally	Yes	No	No
File format	wav, mp3	mp3, base64	m4a
File Size	160 kB	442 kB	740 kB

ISSAI KazakhTTS, according to the research evaluations, is the best option as it is free and can be installed on local computers. Its file size is the smallest, and the file format is convenient for application use. Although the SNR is slightly lower than that of Azure AI, it can be improved through subsequent recording processing. Since the generation speed is very low, additional processing will not significantly affect the program's performance.

To implement audio support based on our evaluation and research [73], it was decided to use KazakhTTS [70], developed by the ISSAI team in 2021. In 2022, ISSAI released the second version, increasing the size of the audio data from 90 to 271 h, which we decided to use. KazakhTTS offers five voices: Iseke, Rayya, Asel, Duman, and Gulzhanat. Using the standard phrase "Менің атым Қожа болады. Ал сіздің атыңыз кім?" ["My name is Kozha. What is your name?"], expert testing of the available voices was conducted with the requirements of moderate playback speed, soft sound, and no errors in repeated generations.

Test results:

1. Iseke: The voice has a guttural and rough sound, not suitable for our task.
2. Rayya: Playback is very fast, likely due to the audio data; this issue occurs with any text, making it unsuitable for our task.
3. Asel: There is a slight echo, and errors occur with several text-to-speech attempts. The voice is soft but not suitable for our task.
4. Duman: The sound is muted, the voice timbre is appropriate, and the playback speed is moderate, making it suitable for our task.
5. Gulzhanat: The playback speed is moderate, no errors occur with multiple text-to-speech attempts, and the voice is soft, making it suitable for our task.

For convenience, the results of this testing are published on Yandex Disk [https://disk.yandex.ru/d/JUD_QStMnREMnw, accessed on 28 September 2024]. In the final version, it was decided to use the voices of Duman and Gulzhanat.

The next step is to run KazakhTTS on a server on our local computer for a more detailed study of its capabilities. The computer has the following specifications and installed components:

- GPU: GeForce GTX 1070;
- CUDA: CUDA Toolkit 11.8;
- Torch: PyTorch Stable (2.3.0).

To optimize the program's performance, a database was created to store the recording name, file path, recording text, and usage count. Before generating a new recording, the database is checked to see if it already exists. If a matching recording is found, it is used, saving resources by avoiding the generation of a new recording. Each time a recording is used from the database, its use count is incremented by one. This was implemented to save database resources; if a recording is created but not reused within a certain period, it is deleted, which helps conserve space in the database. If the recording is needed again in the future, it will be regenerated.

After conducting experiments on short, medium, and long texts, it was observed that there were sound distortions and errors with short texts of seven characters.

For the comparative analysis of available image generation models based on text descriptions, currently there are three popular image generation models: DALL-E 3 [74], Midjourney [75], and Stable Diffusion [76].

The work [77] discusses validating the accuracy and appropriateness of illustrations produced by models such as DALL-E 3 for educational purposes. To achieve this, a comprehensive analysis of the technical foundations of the DALL-E 3 and MidJourney models is carried out, highlighting their neural network architecture and iterative improvement processes. The study reveals significant progress in generating high-quality photorealistic images from text queries, confirming the reliability and accuracy of the generated images. DALL-E 3's unique capabilities to control the image generation process through query modification allow users, including educators and librarians, unprecedented control over the creative process. This ensures that the illustrations produced are accurate and relevant to the specific educational context. The study also addresses the ethical implications of using AI-generated images in education, given potential issues such as cost and accessibility. A comprehensive approach to evaluating model performance ensures that it meets educational standards and also takes into account ethical and accessibility issues. Validation is based on detailed technical analysis, iterative improvement, query management capabilities, and ethical considerations. Thus, the article makes a compelling case for the applicability of the DALL-E 3 and MidJourney models in education and librarianship. After researching various tools, a comparative analysis was carried out. The results of the analysis are presented in Table 12.

Table 12. Comparison of available models.

Characteristic	DALL-E 3	Midjourney	Stable Diffusion
Image quality	High	Artistic	High
Generation rate	Fast	Depends on the load	Fast (locally)
Price	On request	From USD 10 to USD 600 per month	For Free
Flexibility of configuration	Limited	High	Very high
Kazakh language support	Limited; the main language is English	No, English	No, English

DALL-E 3: Developed by OpenAI, DALL-E 3 is the latest version of their model for generating images from textual descriptions. The model provides high-quality images and can generate photorealistic details.

Midjourney: This independent model focuses on creating highly artistic and stylized images. Users can control the process using specialized commands to adjust styles.

Stable Diffusion: An open model developed by Stability AI that allows users to generate images locally. It offers greater flexibility in modifying and using the model.

Figure 7 shows the generated images based on the children's thematic dictionary in Kazakh using the Midjourney program.



Figure 7. Generated drawings from the thematic dictionary.

Figure 8 shows an example of the implementation of the thematic dictionary in a mobile application. For learning purposes, the word itself is presented along with an image of the word and its pronunciation. Additionally, in the "Description of Word" section, an explanation of the word is provided depending on the child's age, using the developed approach. The educational textual data are accompanied by audio in the Kazakh language. When the page loads, the server is accessed to request the WAV file for the pronunciation of the word and its description. These recordings can be played by pressing the corresponding buttons shown in Figure 8.

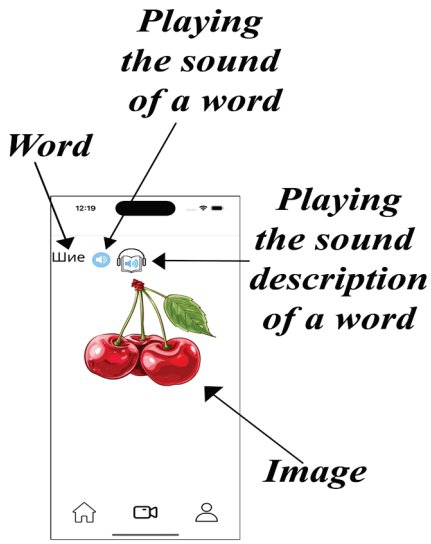


Figure 8. An example of the implementation of a thematic dictionary in a mobile application.

6. Discussion

Vocabulary is a major component of improving children's literacy as it is the basic, smallest, and most important component of language learning [78]. The dictionary was found to be recognized as an excellent medium that is critical to the development of students' literacy skills. This relates to the statements of [79], who believe that the dictionary can be called one of the most common and easy-to-use media because it is the central repository of information for language learners. The dictionary provides extensive knowledge of vocabulary, grammar, meaning, pronunciation, phrases, and parts of speech. Moreover, the results of previous research showed that vocabulary development is necessary to improve students' literacy and learning skills in language [80–82].

Developing applications and systems based on dictionaries for speech development and studying low-resource languages using modern AI technologies is very relevant. Considering that very few such means are aimed at children, we can assume that the research has further development and relevance. Existing open-source tools such as Anki [83], Duolingo [84], Rosetta Stone [85], and Babbel [86] offer polished user interfaces

and rich multimedia content. However, these tools are typically designed for older students and adults and cannot meet preschool-age children's linguistic and cognitive needs.

Additionally, the high cost associated with proprietary instruments limits their availability for widespread use in educational settings. Open-source resources for the Kazakh language are especially limited and often require additional configuration and implementation.

The developed application is based on a children's educational dictionary for a low-resource language using AI tools; it is very convenient and effective for teaching preschool children. The SUS method is often used to evaluate the usability of products and services, and it has been applied to determine the effectiveness, complexity, ease of learning, and user satisfaction of the developed application.

The SUS evaluation showed an average score of 71.3, indicating that users found the application to be user-friendly, with good visualization and voiceover of the material for children in the Kazakh language. The app's average usability score was 71.3 out of 100, indicating that users were generally satisfied with the usability level.

Based on the results of the SUS evaluation, the following measures were taken to improve the usability of the application:

- Improved navigation by adding a navigation menu and shortcut buttons to the application's main functions;
- Improving the interface by increasing the size and contrast of elements and optimizing the layout of information on the screen.

These changes were realized into the final implementation of the application, resulting in improved usability and increased user satisfaction. Figure 9 presents the outcomes of the SUS method assessment for the developed application.

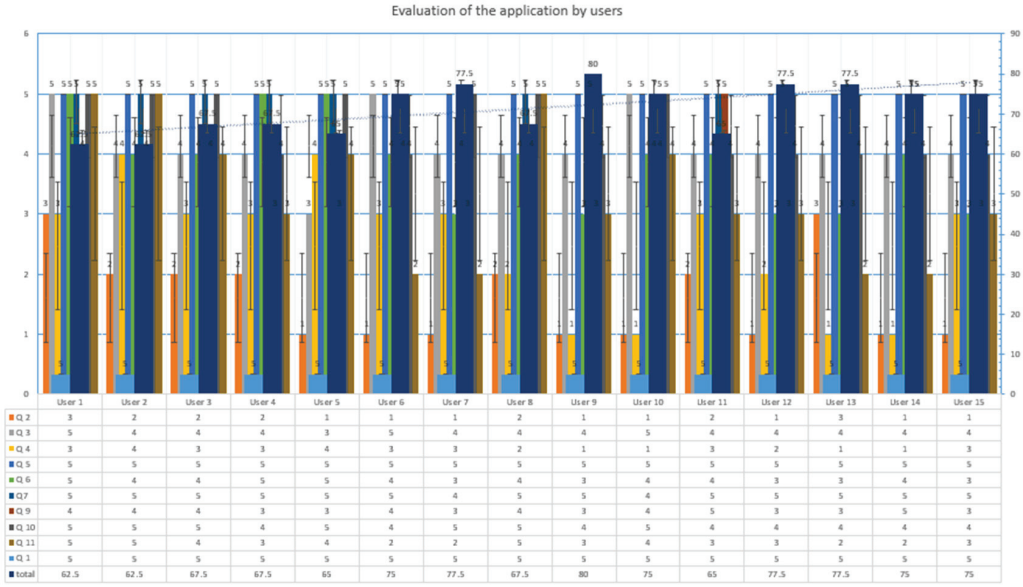


Figure 9. The result of evaluating the application of the SUS method.

The Kazakh language has few resources, so applications developed for teaching the Kazakh language are in great demand among parents and their children. The SUS method was also used to analyze and compare previously developed applications. Among the analyzed applications, it is possible to note "Qonzhyqapp" [87], "Literacy: The Talking Alphabet" [88], and "Self-speech therapist!" [89], mobile applications developed in Kazakhstan.

They contribute to the preservation and development of the mother tongue by offering children interesting and interactive ways of learning. However, these applications offer a simple way to learn the Kazakh language. That is, speaking is taught on the basis of dictionaries compiled on several topics based on animation, sound, and various exercises.

The “Qonzhyqapp”, “Literacy: The Talking Alphabet”, and “Self-speech therapist!” applications do not use AI tools. They contain a dictionary of 800–1200 words intended for children aged 2–5 years. “Qonzhyqapp” is designed for Android and iOS systems. Their main services are activated for a fee.

AI tools in the proposed application are more convenient and effective than the compared applications. Their evaluation is shown in Figure 10. Although there are many educational, language, and speech teaching applications, it is clearly necessary to create a new application that targets the specific needs of Kazakh-speaking children.

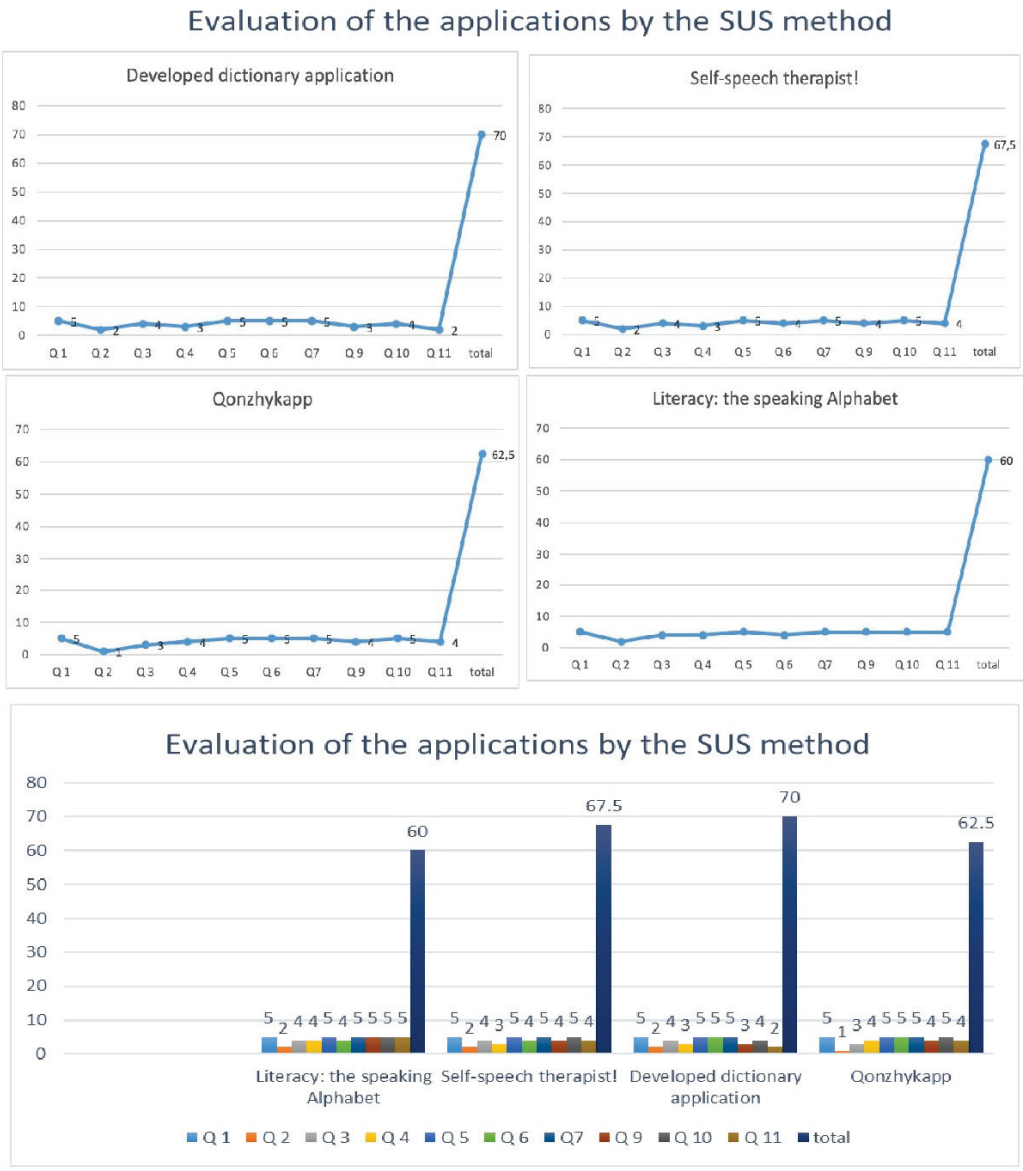


Figure 10. A questionnaire for evaluating the application of the SUS method.

The diagram compares the SUS (System Usability Scale) evaluation of four applications: “Qonzhyqapp”, “Сауат ашу: сөйлейтін әліппе” (“Literacy: the speaking Alphabet”), “Өз-өзіне Логопед!” (“Self-speech therapist!”), and developed a dictionary application. The developed dictionary application has the highest SUS evaluation of 70, indicating it is the most user-friendly and effective among the compared applications. This result showcases the superior usability and quality of our dictionary, emphasizing its effectiveness in supporting children’s education in a low-resource language through AI tools. The higher

SUS evaluation demonstrates our application's better design, ease of use, and overall user satisfaction compared to the other applications evaluated.

The developed dictionary has a child-friendly interface with colorful illustrations and easy navigation, making it very accessible and fun for young children. The integration of AI-generated images and text-to-speech (TTS) technology ensures content is visually appealing and accessible to children of varying reading levels. Feedback from teachers and children during testing in a preschool showed that the application was easy to use and very fun, which confirms its ease of use in an educational environment. It can also be noted that the dictionary was created on the basis of a complete explanatory dictionary [90] and a thematic dictionary developed specifically for the development of children's speech [91].

The fully developed audio and visualization application is available on the GitHub platform [92]. The rest of the information about the formed children's dictionary is shown in Appendix A.

7. Limitations and Future Directions

The limitations of this study cover the problems of confirming the reliability of the data with the help of a speech therapist. Some dictionary categories may be incomplete (the category "toys"), which may lead to omissions or distortions in the analysis results. Some dictionary categories may be fully complete and detailed, as in our case with animals.

The research covered in this work will be continued in future works. The developed system will be adapted to the children's level of academic performance and mastery of the material. It will offer them the next stage of learning and specific exercises and tools for speech development for each person individually. After fully testing the developed approach and populating the system with data, a mobile application on marketplace platforms is planned for launch. It will be part of a social project to support early childhood education development in remote regions and promote learning of low-resource languages like Kazakh.

In addition, in future works, the children's dictionary will be extended with new themes, phrases, and sentences related to them. Adding new images and audio recordings within previous and new themes will also increase the functionality of the visualization and voiceover module. Applying this approach to other Turkic languages with similar linguistic properties and dictionary data is also possible.

Within the framework of this research, the following is planned:

1. The plans to expand the thematic coverage of the dictionary will include:
    - Research interests and needs: Conducting surveys among children, parents, and teachers to identify the most relevant topics for expansion.
    - Regular content updating: Creating a mechanism for regularly updating and adding new topics depending on cultural and educational changes in preschool programs and society.
2. The plans to implement additional multimedia functions and interactive elements will include:
    - Animated explanations and video lessons: Implementing animated content simplifies complex concepts and creates engaging learning.
    - Interactive games and quizzes: Enriching the learning process through game mechanics that motivate children to participate actively.

8. Conclusions

In conclusion, it is important to note that studying children's speech development and creating a thematic dictionary for low-resource languages is significant in using IT and AI tools in modern society, where speech disorders are increasingly observed. This article presents a comprehensive approach to developing a thematic education dictionary for children aged 3 to 5 using AI technology. The approach involves creating linguistic resources through a hybrid method, including GPT-4 technologies (for English) and machine

translation, followed by post-processing based on morphological and semantic analysis in Kazakh. The study involved several key stages, from forming lists of words and sentences in English to filtering bigrams and trigrams using POS tag patterns to adapt the list for children in Kazakh. Practical research on the developed approach focused on three themes: gardens, animals, and toys. After thorough processing, the dictionary consists of 654 words, 1809 phrases, and 2105 sentences.

The experimental results were evaluated using the cosine similarity, Euclidean distance, and Manhattan distance metrics. The average cosine similarity value of 0.62 indicates high similarity between phrases. The average Euclidean similarity value of 8.9 suggests that some phrases or keywords have a greater influence in forming the thematic dictionary. The evaluation results demonstrate the approach's success in forming and presenting material for children's understanding. Using models like DALL-E 3, Midjourney, and Stable Diffusion allowed for creating visual materials that complement the dictionary content, aiding in children's more effective learning of new words. Additionally, the open-source TTS technology for the Kazakh language provided voiceovers for the dictionary content, adjusted for speed and tone to be easily perceived by children. The developed approach was tested in the educational process of a specialized preschool institution. The results show a 12.6% improvement in material comprehension among children and ease of use for staff (teachers and speech therapists). Thus, the proposed method enriches children's vocabulary and enhances language skills development within the modern educational environment. Using AI tools in educational processes for preschool children can make learning more engaging, accessible, and effective. Plans include implementing personalized learning by creating individual study plans for each child, considering their knowledge level, capabilities, and learning pace, as well as adding multimedia features such as video files, animation, and interactive pictures to the educational dictionary. The experimental results demonstrate the high efficiency of the proposed approach and the possibility of its widespread use for educational purposes.

Author Contributions: Conceptualization, D.R.; data curation, A.T., Z.D. and R.A.; formal analysis, A.K. and V.K.; methodology, D.R. and A.K.; project administration, D.R.; resources, A.T. and R.A.; software, V.K.; visualization, Z.D.; writing—original draft, D.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the grant project «Research and Development of Learning Methods and Electronic Resources for Children with Speech Impairment Based on AI Technology» (grant number IRN AP 19577833) of the Ministry of Science and Higher Education of the Republic of Kazakhstan.

Data Availability Statement: This study analyzed publicly available datasets. The results obtained and datasets can be found here: https://github.com/NLP-KazNU/AP19577833-Electronic-resources-for-children-with-speech-disorders-based-on-AI-technology, accessed on 28 September 2024 [85].

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Filtered phrases for children aged 3–5 years (total).

animals	‘Бегемот’: [‘бегемот баяу’, ‘күшті бегемот’, ‘жасыл өсімдіктерді жейді’], ‘Бүркіт’: [‘жылдам аңшы’, ‘күшті тырнақтары бар’, ‘биікке ұшады’], ‘Қасқыр’: [‘қасқыр ырылдайды’, ‘қасқыр жабайы’, ‘жабайы жыртқыш’], ‘Қаз’: [‘балапандары сап түзейді’, ‘дауыстап қоныс аударады’, ‘суда жүзеді’], ‘Жираф’: [‘жирафтар табыны’, ‘мойны ұзын’, ‘дақтары бар жираф’], ‘Қоян’: [‘қоян секіреді’, ‘тез жүгіреді’, ‘ақ қоян’], ‘Ешкі’: [‘кішкентай қозы’, ‘ешкінің сүті’, ‘мүйізді ешкі’], ‘Сиыр’: [‘сиыр бұзауы’, ‘ірі қара сиыр’, ‘сиырдың сүті’], ‘мысық’: [‘сары мысық’, ‘үй жануары’, ‘ұйықтағанды жақсы көреді’], ‘Аққу’: [‘ақ аққу’, ‘аққу көлде’, ‘суда сырғиды’], ‘арыстан’: [‘аң патшасы’, ‘джунглидің патшасы’, ‘жыртқыш аң’], ‘Түлкі’: [‘қу түлкі’, ‘жыртқыш түлкі’, ‘тауық аулайды’], ‘Аю’: [‘гризли аю’, ‘ақ аю’, ‘қоңыр аю’], ‘Тышқан’: [‘кішкентай тышқан’, ‘тез жүгіреді’, ‘кеміруші жануар’], ‘Қой’: [‘қой отары’, ‘шөп жейді’, ‘жайлауда жайылады’]	‘Hippopotamus’: [‘slow hippopotamus’, ‘strong hippopotamus’, ‘eats green plants’], ‘Eagle’: [‘fast hunter’, ‘has strong talons’, ‘flies high’], ‘Wolf’: [‘wolf growls’, ‘wild wolf’, ‘wild predator’], ‘Goose’: [‘goslings march in line’, ‘migrates loudly’, ‘swims in water’], ‘Giraffe’: [‘giraffe herd’, ‘long neck’, ‘spotted giraffe’], ‘Rabbit’: [‘rabbit jumps’, ‘runs fast’, ‘white rabbit’], ‘Goat’: [‘small kid’, ‘goat’s milk’, ‘horned goat’], ‘Cow’: [‘cow calf’, ‘large cow’, ‘cow’s milk’], ‘Cat’: [‘yellow cat’, ‘pet’, ‘likes to sleep’], ‘Swan’: [‘white swan’, ‘swan on the lake’, ‘glides on water’], ‘Lion’: [‘king of the animals’, ‘king of the jungle’, ‘predator’], ‘Fox’: [‘cunning fox’, ‘predatory fox’, ‘hunts chickens’], ‘Bear’: [‘grizzly bear’, ‘polar bear’, ‘brown bear’], ‘Mouse’: [‘small mouse’, ‘runs fast’, ‘rodent’], ‘Sheep’: [‘sheep flock’, ‘eats grass’, ‘grazes in the pasture’]
garden	‘Банан’: [‘жұмсақ банан’, ‘сары банан’, ‘тропикалық жеміс’], ‘Сәбіз’: [‘сары сәбіз’, ‘сәбіз көкөніс’, ‘қоян жейді’], ‘Шие’: [‘шие ағашы’, ‘шырынды шие’, ‘қызы шие’], ‘Алма’: [‘жасыл алма’, ‘қызыл алма’, ‘алма ағашы’], ‘Ағаш’: [‘емен ағашы’, ‘жасыл жапырақ’, ‘қайын ағаш’], ‘Алмұрт’: [‘сары алмұрт’, ‘жасыл алмұрт’, ‘піскен алмұрт’], ‘Қарбыз’: [‘дөңгелек қарбыз’, ‘шырынды қарбыз’], ‘Қияр’: [‘қытырлақ қияр’, ‘жасыл қияр’, ‘маринадталған қияр’], ‘Жидек’: [‘шырынды жидек’, ‘қызыл құлпынвй’, ‘қыщқыл жидек’], ‘Қызанақ’: [‘қызыл қызанақ’, ‘дөңгелек көкөніс’, ‘қызанақ пен қияр’], ‘Бақша’: [‘шырынды жемістер’, ‘жасыл бақ’, ‘гүлдерге толы’], ‘Жеміс’: [‘тропикалық жеміс’, ‘цитрус жемістері’], ‘Өрік’: [‘қара өрік’, ‘өрік джемі’, ‘тәтті өрік’], ‘Бұрыш’: [‘жасыл бұрыш’, ‘қызыл бұрыш’, ‘қара бұрыш’], ‘Лимон’: [‘цитрус жемісі’, ‘лимон шырыны’, ‘қышқыл лимон’], ‘Қызылша’: [‘күлгін қызылша’, ‘піскен қызылша’, ‘қант қызылшасы’], ‘Көкөніс’: [‘бақша көкөністері’, ‘дәмді көкөністер’, ‘жасыл өсімдік’], ‘Асқабақ’: [‘асқабақ пирогы’, ‘қызғылт сары асқабақ’, ‘асқабақ тұқымы’]	‘Banana’: [‘soft banana’, ‘yellow banana’, ‘tropical fruit’], ‘Carrot’: [‘yellow carrot’, ‘carrot vegetable’, ‘rabbit eats’], ‘Cherry’: [‘cherry tree’, ‘juicy cherry’, ‘red cherry’], ‘Apple’: [‘green apple’, ‘red apple’, ‘apple tree’], ‘Tree’: [‘oak tree’, ‘green leaf’, ‘birch tree’], ‘Pear’: [‘yellow pear’, ‘green pear’, ‘ripe pear’], ‘Watermelon’: [‘round watermelon’, ‘juicy watermelon’], ‘Cucumber’: [‘crispy cucumber’, ‘green cucumber’, ‘pickled cucumber’], ‘Berry’: [‘juicy berry’, ‘red strawberry’, ‘sour berry’], ‘Tomato’: [‘red tomato’, ‘round vegetable’, ‘tomato and cucumber’], ‘Garden’: [‘juicy fruits’, ‘green garden’, ‘full of flowers’], ‘Fruit’: [‘tropical fruit’, ‘citrus fruits’], ‘Apricot’: [‘black apricot’, ‘apricot jam’, ‘sweet apricot’], ‘Pepper’: [‘green pepper’, ‘red pepper’, ‘black pepper’], ‘Lemon’: [‘citrus fruit’, ‘lemon juice’, ‘sour lemon’], ‘Beetroot’: [‘purple beetroot’, ‘ripe beetroot’, ‘sugar beet’], ‘Vegetable’: [‘garden vegetables’, ‘tasty vegetables’, ‘green plant’], ‘Pumpkin’: [‘pumpkin pie’, ‘orange pumpkin’, ‘pumpkin seeds’]

Table A1. *Cont.*

toys	Құыршақ: [‘әдемі құыршақ’, ‘тірі құыршақ’, ‘барби құыршақ үйі’, ‘миниатюралық манекен’, ‘сүйкімді құыршақ’, ‘әдемі жасалған құыршақ’, ‘құыршақ үйі’] Көлік: [‘ашық қызыл’, ‘ойыншық машина’, ‘жарыс вагоны’, ‘кішкентай көк көлік’, ‘ойыншық автомобиль’, ‘жолда жарысады’, ‘сары жүк көлік’, ‘шағын мотоцикл’, ‘пластмассадан жасалған арба’, ‘ойыншық гараж’, ‘миниатюралық лимузин’, ‘жасыл джип’, ‘фургондарды ойнау’, ‘әдемі ойыншық машина’, ‘жылдам ойыншық автомобиль’, ‘ойыншық автомобиль’] Пойыз: ‘түрлі-түсті ойыншық пойыз’, ‘теміржол жолы’, ‘локомотив қозғалтқышы’, ‘түрлі-түсті теміржол’, ‘оқу ойыншық пойыз’ тікұшақ: [‘Кішкентай пластик ойыншық’, ‘тікұшақ ұшады’, ‘түрлі-түсті миниатюралық ойыншық’, ‘жылдам ұшады’, ‘ашық ауада ұшады’, ‘тікұшақтарды басқаруды жақсы көреді’, ‘ойнағанды жақсы көреді’] ойыншық жануарлар: [құшақтауға арналған, жұмсақ түкті серіктер, сүйкімді тұлыптар, ерке аю, икемді күшіктер, үлпілдек қоян ойыншық, сүйкімді мысық, сүйкімді құыршақ] Роботтар: [‘пластиктен жасалған’, ‘кішкентай қызыл робот’, ‘робот ойыншық’, ‘электронды автомат’, ‘көк роботты машина’, ‘кішкентай роботтар’, ‘роботты ойыншықтар’, ‘көңілді робот ойыншық’, ‘қызыл автомат’] Лего: [түрлі-түсті пластик, ойыншық кірпіштері, лего конструкциясы, пластикалық бөлшек, ойнап жобалайды, ойнап құрастырады, түрлі-түсті модульдік лего, құрылыс жинағы] Ниндзя тасбақалары: [‘ниндзя тасбақа’, ‘зұлым адамдармен күреседі’, ‘ниндзя қаруымен шайқасады’, ‘мутант тасбақа’, ‘батыл ниндзя жауынгер’]	Doll: [‘beautiful doll’, ‘living doll’, ‘Barbie dollhouse’, ‘miniature mannequin’, ‘cute doll’, ‘well-made doll’, ‘dollhouse’] Vehicle: [‘bright red’, ‘toy car’, ‘racing wagon’, ‘small blue car’, ‘toy automobile’, ‘races on the road’, ‘yellow truck’, ‘small motorcycle’, ‘plastic cart’, ‘toy garage’, ‘miniature limousine’, ‘green jeep’, ‘playing with vans’, ‘beautiful toy car’, ‘fast toy automobile’, ‘toy automobile’] Train: [‘colorful toy train’, ‘railroad track’, ‘locomotive engine’, ‘colorful railway’, ‘educational toy train’] Helicopter: [‘small plastic toy’, ‘helicopter flies’, ‘colorful miniature toy’, ‘flies fast’, ‘flies outdoors’, ‘loves controlling helicopters’, ‘loves playing’] Toy animals: [‘for hugging’, ‘soft furry companions’, ‘cute plush toys’, ‘playful bear’, ‘flexible puppies’, ‘fluffy bunny toy’, ‘cute cat’, ‘cute doll’] Robots: [‘made of plastic’, ‘small red robot’, ‘robot toy’, ‘electronic automaton’, ‘blue robot machine’, ‘small robots’, ‘robot toys’, ‘fun robot toy’, ‘red automaton’] Lego: [‘colorful plastic’, ‘toy bricks’, ‘Lego construction’, ‘plastic piece’, ‘plays and designs’, ‘plays and builds’, ‘colorful modular Lego’, ‘construction kit’] Ninja Turtles: [‘ninja turtle’, ‘fights evil people’, ‘battles with ninja weapons’, ‘mutant turtle’, ‘brave ninja warrior’]

Questions	points, 1-5 (choose only 1 from 1 to 5 points in each row)				
Was it easy for your child to figure out how to use the basic functions of the app?	5	4	3	2	1
How comfortable was your child using the feedback or support feature in the app?	5	4	3	2	1
How easy was it for your child to find and launch the app?	5	4	3	2	1
How will your child appreciate the personalization and customization options of the app?	5	4	3	2	1
Was the app interesting and engaging for your	5	4	3	2	1
How clear were the instructions and tips in the app for your child?	5	4	3	2	1
How will your child appreciate the design and layout of the application?	5	4	3	2	1
How quickly did the app load and run on your child's device?	5	4	3	2	1
How convenient was it to use the application menu?	5	4	3	2	1
How easy was it for your child to find the information they needed in the app?	5	4	3	2	1

Figure A1. An example of a questionnaire for evaluating the application of the SUS method.

Table A2. Abbreviations based on the document content.

Abbreviation	Full Term
AI	Artificial Intelligence
TTS	Text to Speech
IRN	International Research Number
AP	Advanced Placement
NLP	Natural Language Processing
Kazakh TTS	Kazakh Text to Speech
SNR	Signal-to-Noise Ratio
ISSAI	Institute of Information Systems and Artificial Intelligence
POS	Part of Speech
BERT	Bidirectional Encoder Representations from Transformers
SUS	System Usability Scale
DALL-E	A deep learning model developed by OpenAI to generate images from textual descriptions
ChatGPT	Chat Generative Pre-trained Transformer
MATEC	International Conference on Materials and Technology

Table A3. Evaluation Questionnaire.

Column Name	Drop-Down List Options
User ID	Predefined list of user IDs
Age Group	Dropdown with age ranges (e.g., 3–4, 4–5, etc.)
Rating (1–5)	Dropdown with numerical ratings from 1 to 5
Comments	Text field for user comments
Date of Evaluation	Date picker

References

1. Brodin, J.; Hollerer, L.; Renblad, K.; Stancheva-Popkostadinova, V. Preschool teachers' understanding of quality in preschool: A comparative study in three European countries. *Early Child Dev. Care* **2015**, *185*, 968–981. [CrossRef]
2. Ziolkowski, R.; Goldstein, H. Effects of an Embedded Phonological Awareness Intervention During Repeated Book Reading on Preschool Children with Language Delays. *J. Early Interv.* **2008**, *31*, 67–90. [CrossRef]
3. Grabmeier, J. Preschoolers' Reading Skills Benefit from One Modest Change by Teachers. *Educ. Dig. Essent. Read. Condens. Quick Rev.* **2012**, *78*, 63–64.
4. Lennox, S. Interactive Read-Alouds—An Avenue for Enhancing Children's Language for Thinking and Understanding: A Review of Recent Research. *Early Child. Educ. J.* **2013**, *41*, 381–389. [CrossRef]
5. Wilson, P. Why Invest in the Pre-School Years? In *Thinking Ahead. Why We Need to Improve Children's Mental Health and Wellbeing*; Jolley, R., Ed.; Faculty of Public Health (FPH): London, UK, 2011; pp. 1–9.
6. Hargrave, A.C.; Sénéchal, M. A Book Reading Intervention with Preschool Children Who Have Limited Vocabularies: The Benefits of Regular Reading and Dialogic Reading. *Early Child. Res. Q.* **2000**, *15*, 75–90. [CrossRef]
7. Skolverket. *Läroplan för Förskolan Lpfö 98*, Ny rev. ed.; Skolverket: Stockholm, Sweden, 2010.
8. Ministry of Education and Science of the Republic of Kazakhstan, National Scientific and Practical Center of Correctional Pedagogy, Laboratory of Special Education for Children with Special Educational Needs. Analysis of the Study of the Problem of Forming the Native Language System during the Speech Ontogenesis of Children of Early and Preschool Age (Analytical reference). 2019, p. 24. Available online: https://special-edu.kz/analitspr/%D0%90%D0%BD%D0%B0%D0%BB.%D1%81%D0%BF%D1%80%D0%B0%D0%B2%D0%BA%D0%B0%20%D0%BA%D0%B0%D0%B7%D0%BB%D0%B0%D0%B1.pdf (accessed on 28 March 2024).
9. Verburg, M. The Making of My First Van Dale: A Preschool Dictionary. In Proceedings of the Euralex Conference, Torino, Italy, 6–9 September 2006; pp. 357–362.
10. Nkomo, D. An African User-Perspective on English Children's and School Dictionaries. *Int. J. Lexicogr.* **2016**, *29*, 31–54. [CrossRef]
11. Bekmanova, G.; Sharipbay, A.; Kabdylova, D.; Amangeldy, N.; Sairanbekova, A.; Omarbekova, A.; Zulkhazhav, A. Development of an electronic dictionary of terminology for school textbooks. *J. Electr. Syst.* **2024**, *20–24s*, 904–911. [CrossRef]
12. Zhao, D.; Xu, H. Improving the Compilation of English–Chinese Children's Dictionaries: A Children's Cognitive Perspective. *Lexikos* **2022**, *32*, 49–65. [CrossRef]
13. Liu, X.; Lyu, J.; Zheng, D. For a Better Dictionary: Revisiting Ecolexicography as a New Paradigm. *Lexikos* **2021**, *3*, 182–207. [CrossRef]
14. Li, Q.; Russell, M. An Analysis of the Causes of Increased Error Rates in Children's Speech Recognition. In Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002), Denver, CO, USA, 16–20 September 2002; pp. 366–369.

15. Callaghan, T.; Reich, S. Tools for evaluating educational apps for young children: A systematic review of the literature. *Interact. Technol. Smart Educ.* **2018**, *18*, 18–49.
16. Hirsh-Pasek, K.; Zosh, J.M.; Golinkoff, R.M.; Gray, J.H.; Robb, M.B.; Kaufman, J. Putting Education in "Educational" Apps: Lessons from the Science of Learning. *Psychol. Sci. Public Interest* **2015**, *16*, 3–34. [CrossRef] [PubMed]
17. Kolak, J.; Norgate, S.H.; Monaghan, P.; Taylor, G. Developing Evaluation Tools for Assessing the Educational Potential of Apps for Preschool Children in the UK. *J. Child. Media* **2021**, *15*, 410–430. [CrossRef]
18. Rideout, V.; Robb, M.B. The Common Sense Census: Media Use by Kids Age Zero to Eight, 2020: Common Sense Media. Common Sense Media: Ratings, Reviews, and Advice, 17 November 2020. Available online: https://www.commonsensemedia.org (accessed on 28 April 2024).
19. Kervin, L. Powerful and Playful Literacy Learning with Digital Technologies. *Aust. J. Lang. Lit.* **2016**, *39*, 64–73. [CrossRef]
20. Arnold, A. A Randomized Controlled Trial of an Educational App to Improve Preschoolers' Emergent Literacy Skills. *J. Child. Media* **2021**, *15*, 457–475. [CrossRef]
21. Dore, R.A. Education in the App Store: Using a Mobile Game to Support U.S. Preschoolers' Vocabulary Learning. *J. Child. Media* **2019**, *13*, 452–471. [CrossRef]
22. Kirkorian, H.L.; Choi, K.; Pempek, T.A. Toddlers' Word Learning from Contingent and Non-Contingent Video on Touchscreens. *Child Dev.* **2016**, *87*, 405–413. [CrossRef]
23. Russo-Johnson, C. All Tapped Out: Touchscreen Interactivity and Young Children's Word Learning. *Front. Psychol.* **2017**, *8*, 578. [CrossRef]
24. Parmawati, L.; Ratminingsih, N.M.; Budasi, I.G. The Development of a Multilingual Thematic Digital Dictionary for Elementary School Students. *E-Link J.* **2022**, *9*, 60–75. [CrossRef]
25. Cheng, J.; Wang, H.; Lu, S. Context-Aware Learning for Sentence-Level Sentiment Analysis with Posterior Regularization. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Xi'an, China, 25–30 July 2020; pp. 555–564.
26. Why Does a Child's Tongue Come Out Late? *Speech Therapist Tips for Parents.* Available online: https://informburo.kz/kaz/bala-tl-nege-kesh-shyady-logopedt-ata-analara-keester.html (accessed on 5 April 2024).
27. Speech Retardation in a Child. Available online: https://daryn.online/article/2919# (accessed on 9 April 2024).
28. Mamyrbayev, O.; Alimhan, K.; Oralbekova, D.; Bekarystankyzy, A.; Zhumazhanov, B. Identifying the Influence of Transfer Learning Method in Developing an End-to-End Automatic Speech Recognition System with a Low Data Level. *East.-Eur. J. Enterp. Technol.* **2022**, *1*, 84–92. [CrossRef]
29. Mussakhojayeva, S.; Dauletbek, K.; Yeshpanov, R.; Varol, H.A. Multilingual Speech Recognition for Turkic Languages. *Information* **2023**, *14*, 74. [CrossRef]
30. Rakhimova, D.; Matanov, N.; Rzagaziyeva, A. Development of a Dictionary for Preschool Children with Weak Speech Skills Based on the Word2Vec Method. In Proceedings of the ICCCI, Budapest, Hungary, 27–29 September 2023; pp. 191–202.
31. Łukasik, M. Corpus Linguistics and Generative AI Tools in Term Extraction: A Case of Kashubian—A Low-Resource Language. *Appl. Linguist. Pap.* **2024**, *39*, 34–45. Available online: http://alp.uw.edu.pl/wp-content/uploads/sites/315/2024/02/5-ALP_Lukasik-Marek.pdf (accessed on 8 April 2024). [CrossRef]
32. Ponnusamy, R. Tamil Text Generation using ChatGPT-3 Models. In Proceedings of the International Conference on Tamil Computing, Mangaluru, India, 13–14 October 2023; p. 30.
33. Oranga, J. Benefits of Artificial Intelligence (ChatGPT) in Education and Learning: Is Chat GPT Helpful? *Int. Rev. Pract. Innov. Technol. Green Energy (IRPITAGE)* **2023**, *3*, 46–50.
34. Pisarev, I.A. Methods for Automated Thesauri Development of Thematic Corpora in the Learning Process Support Systems. In Proceedings of the IEEE Conference on Soft Computing and Measurements, St. Petersburg, Russia, 25–27 May 2016; pp. 39–40.
35. Yelibayeva, G.; Sharipbay, A.; Bekmanova, G.; Omarbekova, A. Ontology-Based Extraction of Kazakh Language Word Combinations in Natural Language Processing. In Proceedings of the International Conference on Data Science, E-learning and Information Systems 2021 (DATA'21), Petra, Jordan, 5–7 April 2021; pp. 60–75.
36. Piaget, J. *The Origins of Intelligence in Children*; International Universities Press: New York, NY, USA, 1952.
37. Vygotsky, L.S. *Mind in Society: The Development of Higher Psychological Processes*; Harvard University Press: Cambridge, MA, USA, 1978.
38. Verhallen, M.J.; Bus, A.G. Multimedia stories for kindergarten children: The effects of modality and narration style on comprehension and vocabulary acquisition. *J. Educ. Psychol.* **2010**, *102*, 652–660.
39. Pane, J.F.; Griffin, B.A.; McCaffrey, D.F.; Karam, R. *Effectiveness of Personalized Learning: Findings from a Study of Four Charter Schools*; RAND Corporation: Santa Monica, CA, USA, 2017.
40. Wood, S.G.; Moxley, J.H.; Tighe, E.L.; Wagner, R.K. Does use of text-to-speech and related read-aloud tools improve reading comprehension for students with reading disabilities? A meta-analysis. *J. Learn. Disabil.* **2005**, *41*, 75–84. [CrossRef]
41. Sunder, P. The Effectiveness of Audio-Visual Aids in Teaching-Learning Process. *Int. J. Creat. Res. Thoughts* **2018**, *6*, 1509–1515.
42. Sundqvist, A.; Barr, R.; Heimann, M.; Birberg-Thornberg, U.; Koch, F.-S. A longitudinal study of the relationship between children's exposure to screen media and vocabulary development. *Acta Paediatr.* **2023**, *113*, 517–522. [CrossRef]
43. Online Nation Report 2019 (OFCOM). Available online: https://www.gov.uk/find-digital-market-research/online-nation-report-2019-ofcom (accessed on 10 September 2024).

44. Ronniger, P.; Rennecke, L.; Petermann, F.; Melzer, J. Differenzierungsfähigkeit des Sprachstandserhebungstest für Kinder im Alter zwischen 3 und 5 Jahren. *Monatsschr Kinderheilkd* **2022**, *170*, 422–429. [CrossRef]
45. Jablonski, S. Inhibitory control and literacy development among 3- to 5-year-old children. *L1-Educ. Stud. Lang. Lit.* **2014**, *14*, 1–25. [CrossRef]
46. Schebesta, A.; Kunter, G. Constituent durations in English NNN compounds: A case of strategic speaker behavior? *J. Phon.* **2022**, *94*, 101164. [CrossRef]
47. Vania, C.; Lopez, A. From Characters to Words to in Between: Do We Capture Morphology? In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, BC, Canada, 30 July–4 August 2017; Association for Computational Linguistics: Vancouver, BC, Canada, 2017; pp. 2016–2027.
48. Roseberry, S.; Hirsh-Pasek, K.; Golinkoff, R.M. Skype me! Socially contingent interactions help toddlers learn language. *Child Dev.* **2014**, *85*, 956–970. [CrossRef]
49. Allen, L.; Kelly, B.B. *Committee on the Science of Children Birth to Age 8: Deepening and Broadening the Foundation for Success; Board on Children, Youth, and Families*; Institute of Medicine; National Research Council. Transforming the Workforce for Children Birth Through Age 8: A Unifying Foundation; National Academies Press (US): Washington, DC, USA, 2015. [PubMed]
50. Septiyana, L.; Aneka, A.; Mandasari, B.; Ramadhan, F.S.S. Development of A Trilingual E-Dictionary for Early Childhood; Indonesia-English-Lampung. *Tapis: J. Penelit. Ilm.* **2023**, *7*, 160–172. Available online: https://e-journal.ejournal.metrouniv.ac.id/index.php/tapis/article/view/810 (accessed on 10 July 2024). [CrossRef]
51. Apertium. Apertium-Kaz. Available online: https://github.com/apertium/apertium-kaz (accessed on 15 March 2024).
52. Kessikbayeva, G.; Cicekli, I. Rule Based Morphological Analyzer of Kazakh Language. *Linguist. Lit. Stud.* **2014**, *4*, 46–54.
53. Makhambetov, O.; Makazhanov, A.; Sabyrgaliyev, I.; Yessenbayev, Z. Data-Driven Morphological Analysis and Disambiguation for Kazakh. In *Computational Linguistics and Intelligent Text Processing*; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 1–13.
54. Makhambetov, O.; Makazhanov, A.; Yessenbayev, Z.; Sabyrgaliyev, I. Towards a Data-driven Morphological Analysis of Kazakh Language. In Proceedings of the 2nd International Conference on Turkic Languages Processing, Rouen, France, 3–5 September 2014; pp. 69–74.
55. Bekmanova, G.; Sharipbay, A.; Altenbek, G.; Adalı, E.; Zhetkenbay, L.; Kamanur, U.; Zulkhazhav, A. A Uniform Morphological Analyzer for the Kazakh and Turkish Languages. In Proceedings of the Sixth International Conference on Analysis of Images, Social Networks, and Texts (AIST 2017), Moscow, Russia, 27–29 July 2017; pp. 1–11.
56. Assylbekov, Z.; Washington, J.; Tyers, F.; Nurkas, A.; Sundetova, A.; Karibayeva, A.; Abduali, B.; Amirova, D. A Free/Open-Source Hybrid Morphological Disambiguation Tool for Kazakh. In Proceedings of the TurCLing 2016, Turkey, Konya, 3–9 April 2016; pp. 18–26.
57. Rakhimova, D.; Karibayeva, A.; Turarbek, A. The Task of Post-Editing Machine Translation for the Low-Resource Language. *Appl. Sci.* **2024**, *14*, 486. [CrossRef]
58. Bekbulatov, E.; Kartbayev, A. A Study of Certain Morphological Structures of Kazakh and Their Impact on the Machine Translation Quality. In Proceedings of the 8th International Conference on Application of Information and Communication Technologies (AICT), Astana, Kazakhstan, 15–17 October 2014; pp. 1–5.
59. Kalimoldayev, M.N.; Musabayev, R.R.; Mamyrbayev, O.J. The Algorithm of Morphological Analysis of the Kazakh Words. In Proceedings of the 10th International Conference, Hyogo, Japan, 27 December 2012; pp. 121–124.
60. Toleu, A.; Tolegen, G.; Makazhanov, A. Character-Based Deep Learning Models for Token and Sentence Segmentation. In *Proceedings of the 5th International Conference on Computer Processing of Turkic Languages (TurkLang 2017)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 67–71.
61. Tukeyev, U. Automaton Models of the Morphology Analysis and the Completeness of the Endings of the Kazakh Language. In Proceedings of the International Conference "Turkic Languages Processing" TURKLANG-2015, Kazan, Russia, 17–19 September 2015; pp. 91–100.
62. Tukeyev, U.; Sundetova, A.; Abduali, B.; Akhmadiyeva, Z.; Zhanbussunov, N. Inferring of the Morphological Chunk Transfer Rules on the Base of Complete Set of Kazakh Endings. In Proceedings of the Computational Collective Intelligence, ICCCI 2016, Lecture Notes in Computer Science, Halkidiki, Greece, 28–30 September 2016; Volume 9876, pp. 563–574.
63. Tukeyev, U.; Karibayeva, A.; Abduali, B. Neural Machine Translation System for the Kazakh Language Based on Synthetic Corpora. In Proceedings of the MATEC Web of Conferences, Sibiu, Romania, 5–7 June 2019; pp. 1–6.
64. Jothi, S.A.; Pandeeswari, N. Enriched Glove of Word Embedding and BERT of Contextual Embedding in Sarcasm Detection with AdaGrad Optimization Technique. In Proceedings of the 2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bengaluru, India, 3–5 March 2023.
65. McElroy, E.; Wood, T.; Bond, R.; Mulvenna, M.; Shevlin, M.; Ploubidis, G.B.; Hoffmann, M.S.; Moltrecht, B. Using natural language processing to facilitate the harmonisation of mental health questionnaires: A validation study using real-world data. *BMC Psychiatry* **2024**, *24*, 530. [CrossRef]
66. Abrar Jahin, M.; Talapatra, S. A Natural Language Processing-Based Classification and Mode-Based Ranking of Musculoskeletal Disorder Risk Factors. *Decis. Anal. J.* **2024**, *11*, 100464. [CrossRef]
67. Shirude, S.B.; Kolhe, S.R. Agent-Based Architecture for Developing Recommender System in Libraries. In *Knowledge Computing and its Applications*; Margret Anouncia, S., Wiil, U., Eds.; Springer: Berlin/Heidelberg, Germany, 2018; pp. 157–181.

68. Li, Y.; Song, H.; Xu, Y. Studies on Data Quality Evaluation Index System for Internet Plus Government Services in Big Data Era. *Information* **2020**, *11*, 123.
69. Youdi, G.; Guangzhen, L.; Yunzhi, X.; Rui, L.; Lingzhong, M. A survey on dataset quality in machine learning. *Inf. Softw. Technol.* **2023**, *162*, 107268.
70. Institute of Smart Systems and Artificial Intelligence. KazakhTTS. Available online: https://issai.nu.edu.kz/ru/tts2-rus/ (accessed on 10 July 2024).
71. Narakeet. Kazakh Text-to-Speech. Available online: https://www.narakeet.com/languages/kazakh-text-to-speech/ (accessed on 10 July 2024).
72. Microsoft Azure. Speech Service. Available online: https://learn.microsoft.com/en-us/azure/ai-services/speech-service/ (accessed on 10 July 2024).
73. Balabekova, T.; Kairatuly, B.; Tukeyev, U. Kazakh-Uzbek Speech Cascade Machine Translation on Complete Set of Endings. *Adv. Comput. Collect. Intell.* **2023**, *1864*, 430–442.
74. OpenAI. DALL-E 3. Available online: https://openai.com/dall-e-3/ (accessed on 10 July 2024).
75. Midjourney. Home. Available online: https://www.midjourney.com/home (accessed on 10 July 2024).
76. Stable Diffusion. Available online: https://stablediffusionweb.com/ (accessed on 10 July 2024).
77. Adetayo, A.J. Reimagining Learning through AI Art: The Promise of DALL-E and MidJourney for Education and Libraries. Library Hi Tech News. 2024. Available online: https://www.emerald.com/insight/content/doi/10.1108/LHTN-01-2024-0005/full/html (accessed on 10 July 2024).
78. Suniyasih, N.W.; Adnyani, L.D.S.; Marhaeni, A.A.I.N.; Suastra, I.M.; Putra, N.P.W. The Development of a Multilingual Thematic Picture Dictionary in Supporting the Literacy of Fifth Grade Elementary School Students. *J. Pendidik. Bhs. Ingg. Undiksha* **2020**, *8*, 29–39.
79. Rohmatillah, R. A Study on Students' Difficulties in Learning Vocabulary. *Engl. Educ. J. Tadris Bah. Ing.* **2014**, *6*, 75–93.
80. Adnyani, N.K.P.; Ratminingsih, N.M.; Budasi, I.G. A Multilingual Thematic Picture Dictionary for Assisting Sixth Grade Elementary School Students' Literacy. *J. Pendidik. Bhs. Ingg. Undiksha* **2021**, *9*, 1–8. [CrossRef]
81. Widhiastari, M.R.; Ratminingsih, N.M.; Budasi, I.G. The Effect of Digital Multilingual Thematic Dictionary Toward the Fifth Grade Students' Literacy Skill. *Linguist. Engl. Educ. Art (LEEA) J.* **2023**, *6*, 315–325. [CrossRef]
82. Wulandari, P.D.M.; Tantra, D.K.; Santosa, M.H. Teacher's Preparation of Lesson Plan and Constraints in English as a Foreign Language During the Pandemic Outbreak in the Junior High School. *J. Pendidik. Bhs. Ingg. Undiksha* **2021**, *9*, 43–50. [CrossRef]
83. Anki. Wikipedia. Available online: https://ru.wikipedia.org/wiki/Anki (accessed on 10 July 2024).
84. Duolingo. Available online: https://www.duolingo.com/ (accessed on 10 July 2024).
85. Rosetta Stone. Learn Turkish. Available online: https://www.rosettastone.com/buy/learn-turkish/ (accessed on 10 July 2024).
86. Babbel. Available online: https://uk.babbel.com/ (accessed on 10 July 2024).
87. Qonzhyq App. Mobile Application. Available online: https://play.google.com/store/apps/details?id=com.mariomobile&hl=en_US (accessed on 10 July 2024).
88. Сауат ашу: сөйлейтін әліппе. Mobile Application. Available online: https://play.google.com/store/apps/details?id=kz.qazdevelopment.sauatashu&hl=en_US (accessed on 10 July 2024).
89. Өз-өзіне Логопед! Mobile Application. Available online: https://play.google.com/store/apps/details?id=kz.logoped (accessed on 10 July 2024).
90. Uali, N.; Kurmanbayuly, S.; Malbakov, M.; Shoibekov, R.; Kuderinova, K.; Zhubayeva, O.; Zhanabekova, A.; Fazylzhanova, A.; Isayeva, Z.; Uderbayev, A. *Unique Large Explanatory Dictionary of the Kazakh Language*; Dauir Publishing House: Almaty, Kazakhstan, 2013; 1488p.
91. Yesenzholova, G.Z. *Speech Development in Early Childhood*; Methodological Recommendations for Parents; Ulagat Publishing House: Almaty, Kazakhstan, 2014; 88p.
92. NLP-KazNU. AP19577833-Application-for-Children-with-Speech-Disorders-Based-on-AI-Technology. GitHub. Available online: https://github.com/NLP-KazNU/AP19577833-Application-for-children-with-speech-disorders-based-on-AI-technology (accessed on 10 July 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

computers

MDPI

Article

A Study of a Drawing Exactness Assessment Method Using Localized Normalized Cross-Correlations in a Portrait Drawing Learning Assistant System

Yue Zhang [1], Zitong Kong [1], Nobuo Funabiki [1,*] and Chen-Chien Hsu [2]

[1] Graduate School of Environmental, Life, Natural Sciences and Technology, Okayama University, Okayama 700-8530, Japan; pr1u5yrh@s.okayama-u.ac.jp (Y.Z.); kou_shitou@s.okayama-u.ac.jp (Z.K.)
[2] Department of Electrical Engineering, National Taiwan Normal University, Taipei 10610, Taiwan; jhsu@ntnu.edu.tw
* Correspondence: funabiki@okayama-u.ac.jp; Tel.: +81-070-4319-6541

Abstract: Nowadays, portrait drawing has gained significance in cultivating painting skills and human sentiments. In practice, novices often struggle with this art form without proper guidance from professionals, since they lack understanding of the proportions and structures of facial features. To solve this limitation, we have developed a Portrait Drawing Learning Assistant System (*PDLAS*) to assist novices in learning portrait drawing. The *PDLAS* provides auxiliary lines as references for facial features that are extracted by applying OpenPose and OpenCV libraries to a face photo image of the target. A learner can draw a portrait on an iPad using drawing software where the auxiliary lines appear on a different layer to the portrait. However, in the current implementation, the *PDLAS* does not offer a function to assess the exactness of the drawing result for feedback to the learner. In this paper, we present a drawing exactness assessment method using a *Localized Normalized Cross-Correlation* (*NCC*) algorithm in the *PDLAS*. *NCC* gives a similarity score between the original face photo and drawing result images by calculating the correlation of the brightness distributions. For precise feedback, the method calculates the *NCC* for each face component by extracting the bounding box. In addition, in this paper, we improve the auxiliary lines for the nose. For evaluations, we asked students at Okayama University, Japan, to draw portraits using the *PDLAS*, and applied the proposed method to their drawing results, where the application results validated the effectiveness by suggesting improvements in drawing components. The system usability was also confirmed through a questionnaire with a *SUS score*. The main finding of this research is that the implementation of the *NCC* algorithm within the *PDLAS* significantly enhances the accuracy of novice portrait drawings by providing detailed feedback on specific facial features, proving the system's efficacy in art education and training.

Keywords: portrait drawing; auxiliary lines; OpenPose; OpenCV; normalized cross-correlation (NCC); exactness assessment

1. Introduction

Nowadays, portrait drawing has gained significance in cultivating painting skills and personal sentiments [1]. A portrait is commonly perceived as representing a human being's features, which is an important theme in figurative art. Unfortunately, novices in portrait drawing might find it difficult to define the proportions of facial features at the beginning of a sketch. On the other hand, with great improvements in digital technologies, portrait drawing has become possible on electronic devices such as tablets and note PCs.

With this motivation, recently, we developed a Portrait Drawing Learning Assistant System (*PDLAS*) to assist beginners to learn drawing portraits by themselves [2,3]. It provides auxiliary lines to help and guide users in drawing portraits. For this purpose, we

proposed an auxiliary line generation algorithm to extract the facial features and contours of face parts from a given face photo, using OpenPose [4] and OpenCV [5].

To evaluate the effectiveness of the *PDLAS*, we asked students at Okayama University, Japan, to draw the portraits from given face photos using the *PDLAS*. However, no method for evaluating the drawing accuracy of the result is implemented in the *PDLAS*.

In this paper, we present a drawing accuracy evaluation method that calculates the *Normalized Cross-Correlation* (*NCC*) for each face component to quantify the similarity between a user's drawing and its original face image. By localizing the range of the calculated *NCC* to each face part such as an eye, a nose, a mouth, and an eyebrow, the evaluation result can be improved. To extract a face part, the rectangular bounding box around the part is found first in our method. Then, we use the key points of facial features provided in OpenPose to define the area of the evaluation. The accuracy feedback in the proposed method not only enhances the interactivity and efficiency of portrait drawing learning but also offers valuable suggestions for improving drawing through precise similarity evaluations.

The preliminary application results of the drawing accuracy evaluation method found that the *NCC* of the nose is always low for any portrait result. After we analyze the reason, we improve the auxiliary line generation algorithm for the nose in this paper.

For evaluations, we asked 13 students in Okayama University, Japan, to draw portraits using the *PDLAS* and applied the proposed method to their output results. Then, from the obtained *NCC* scores, we analyzed issues in the portrait results and provided suggestions for improvements in drawing components as user feedback. Thus, the effectiveness of the proposed method was confirmed.

For the overall project, our research objectives are divided into two levels. For the whole-project level of the Portrait Drawing Learning Assistance System (*PDLAS*), the research objective is to present a system for improving the learning experiences of novices in portrait drawing. For the level of the contribution of the drawing accuracy evaluation method in this paper, the research objective is to present a method to provide feedback for improving specific parts in the drawn face. By improving them one by one, novices can eventually enhance learning experiences in portrait drawing. This research will help more people to learn drawing skills by themselves, especially in areas with fewer educational resources. It will contribute to the equalization of education, giving people in different regions and professions the opportunities to develop their artistic talents.

The rest of this paper is organized as follows:

Section 2 discusses related works in the literature.

Section 3 reviews the Portrait Drawing Learning Assistant System.

Section 4 discusses the proposed method for the drawing exactness evaluation method.

Section 5 discusses the application to existing drawing results.

Section 6 discusses the improvement of nose auxiliary line generation.

Section 7 discusses the evaluation results after improving nose auxiliary line generation.

Finally, Section 8 concludes this paper with our future works.

2. Related Works in the Literature

In this section, we discuss related works to this paper in the literature.

2.1. Portrait Drawing

In [6], Yi R. et al. proposed a novel asymmetric cycle mapping that enforces the reconstruction information to be visible (by a truncation loss) and only embedded in selective facial regions (by a relaxed forward cycle-consistency loss).

In [7], Yi R. et al. proposed a novel method to automatically transform face photos to portrait drawings using unpaired training data with two new features; i.e., their method can (1) learn to generate high quality portrait drawings in multiple styles using a single network and (2) generate portrait drawings in a "new style" unseen in the training data.

In [8], Takagi S. et al. proposed a novel image-to-pencil translation method that could not only generate high-quality pencil sketches but also offer the drawing process. Existing pencil sketch algorithms are based on texture rendering rather than the direct imitation of strokes, making them unable to show the drawing process but only a final result.

In [9], Looi L. et al. proposed a method to estimate drawing guidelines when given an image of a front-facing face. The portrait drawing process is complex and often requires an artist to break down a face into basic guidelines as a foundation for the artwork. They utilized a modified Histogram of Oriented Gradients and a Linear Support Vector Machine to identify the region of interest of a face in the image. An implementation of the ensemble of randomized regression trees was used to regress the location of key facial landmarks. Guideline estimation was achieved by using specific points from the facial landmark output in combination with information about the average proportions of a human head.

2.2. Drawing Assistant System

In [10], Takagi S. et al. proposed a learning support system for beginner's pencil drawing, which is the basis of pictures. The system receives a motif data set and a user's sketch image and returns advice to the user. The processing is composed of four functions: feature extraction of motifs, feature extraction of sketches, error identification, and generation and presentation of advice. They developed and experimented with a prototype system limited to treating a basic motif and principal advice.

In [11], Huang Z. et al. proposed dualFace, a portrait drawing interface to assist users with different levels of drawing skills to complete recognizable and authentic face sketches. Inspired by traditional artist workflows for portrait drawing, dualFace gives two stages of drawing assistance to provide global and local visual guidance. The former helps users draw contour lines for portraits (i.e., geometric structure), and the latter helps users draw details of facial parts, which conform to the user-drawn contour lines.

In [12], Iarussi E. et al. presented an interactive drawing tool that provides automated guidance over model photographs to help people practice traditional drawing-by-observation techniques. The drawing literature describes a number of techniques to support this task and help people gain consciousness of the shapes in a scene and their relationships. They compiled these techniques and derived a set of construction lines that they automatically extract from a model photograph.

In [13], Gao F. et al. presented a drawing robot which can automatically transfer a facial picture to a vivid portrait, and then draw it on paper in an average of two minutes. At the heart of their system is a novel deep learning-based portrait synthesis algorithm. Innovatively, they employed a self-consistency loss, which makes the algorithm capable of generating continuous and smooth brush-strokes. In addition, they proposed a componential sparsity constraint to reduce the number of brush-strokes over insignificant areas.

In [14], Xie J. et al. presented PortraitSketch, an interactive drawing system that helps novices create pleasing, recognizable face sketches without requiring prior artistic training. As the user traces over a source portrait photograph, PortraitSketch automatically adjusts the geometry and stroke parameters (thickness, opacity, etc.) to improve the aesthetic quality of the sketch. They presented algorithms for adjusting both outlines and shading strokes based on important features of the underlying source image.

In [15], Li S. et al. proposed AgeFace, an interactive drawing interface that assists users in creating facial features with age-specific features based on user input strokes. They evaluated the usability of AgeFace by a user experience experiment and a comparison experiment with baseline approaches. The results verified that AgeFace could achieve better performance in usability and better support in the creative process than baseline systems.

In summary, while the techniques and methods described above have greatly improved the image drawing process, they have focused primarily on improving the final output or providing tools to assist in drawing. They do not directly address and instruct novices in the learning process. They do not provide step-by-step instructions or basic training to help novices improve their drawing skills in a systematic way. This gap in the

research background provides an opportunity for further research and development of the system designed to help novices learn the fundamentals of portrait drawing. Such a system could provide tailored instruction, feedback, and educational resources.

3. Review of Portrait Drawing Learning Assistant System

In this section, we review the Portrait Drawing Learning Assistant System (PDLAS) [2].

3.1. System Overview

A *PDLAS* has been designed and implemented to assist novices in drawing portraits using electronic devices such as a tablet or a personal computer with a digital pen, and a drawing software. In our preliminary implementation [3], we adopted an iPad [16] and Apple Pencil with the drawing software Procreate (version 5.3.10).

To guide drawing portrait by novices, an image of the auxiliary lines of the face is generated by an auxiliary line generation algorithm, and is inserted at a different layer from the one for drawing the portrait in Procreate. This algorithm extracts the facial features and contours of face components such as the eyes, eyebrows, nose, and mouth from a given face photo, using OpenPose and OpenCV.

OpenPose is used to find the coordinates of the keypoints that represent important locations to determine the structure of a human body. The auxiliary lines for the eyes, the mouth, the lower face contour, the nose, and the head top are extracted from them. The OpenCV library is used to extract the auxiliary lines for the hair, the eyebrows, and the eyeglasses.

3.2. OpenPose

OpenPose is a popular open-source software to furnish the coordinates for 70 facial keypoints, illustrated in Figure 1. These keypoints denote the positions of the facial features.

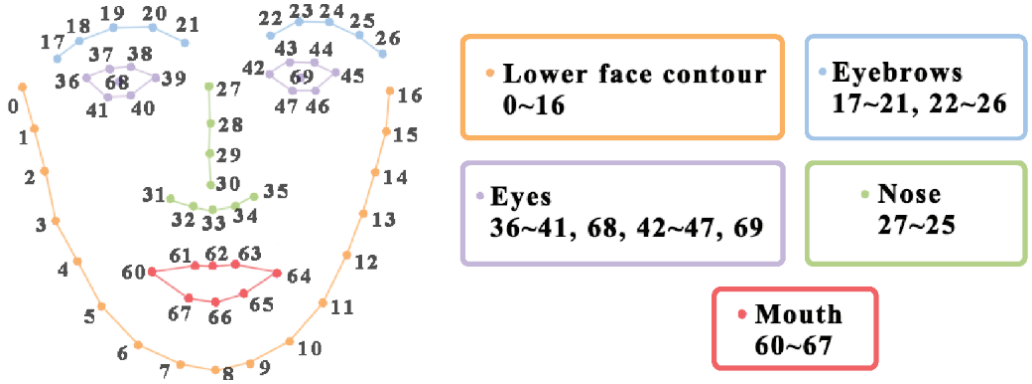


Figure 1. Seventy keypoints for facial features in OpenPose.

3.3. Auxiliary Lines

The auxiliary lines serve to assist a user drawing a portrait by easily grasping the overall structure of the face and the position of each facial feature, which include the following:

- Triangle auxiliary lines widely used in drawing methods [17];
- Outlines of the eyes, the shape of the mouth, and the lower face contour;
- Three circle auxiliary lines for the nose;
- Outlines of the hair, the eyebrows, and the eyeglasses.

The first three items are extracted from the corresponding keypoints by OpenPose, and the last item by OpenCV library functions. Figure 2 illustrates auxiliary lines generated by OpenPose and by OpenCV, which will be combined into one image for the complete auxiliary lines. Figure 3 illustrates the complete auxiliary lines.

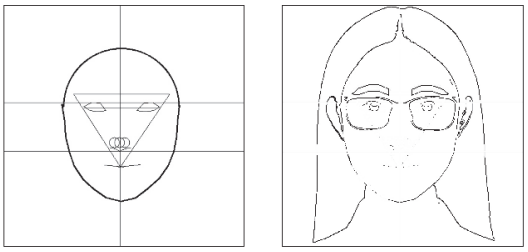


Figure 2. Auxiliary lines by OpenPose and OpenCV.

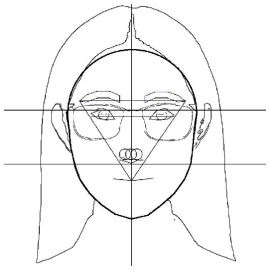


Figure 3. Complete auxiliary lines example.

3.4. Auxiliary Line Generation by OpenPose

First, we review the procedure of generating the auxiliary lines using keypoints from OpenPose in the auxiliary line generation algorithm. Table 1 shows the keypoint indices that are used to generate the corresponding auxiliary lines.

Table 1. OpenPose keypoints for auxiliary lines.

Auxiliary Lines	Corresponding Keypoints
center line	average of x-coordinates of 27th~30th
eye position line	y-coordinates of 36th, 39th, 42nd, 45th, and x-coordinates of 36th, 45th
the level line above and below the ear	average of y-coordinates of 37th, 38th, 43rd, 44th, and x-coordinates of 3rd, 13th
a large inverted triangle for positioning	endpoints: 17th, 26th, and average of 60th, 64th
eyes	36th~41st and 42nd~47th
mouth	60th~67th
lower face contour	0th~16th
three circles for nose	radius: $\frac{d}{4}$ (where d is the distance between 31st and 35th all circles' center x-coordinates: same as center line; center circle's center y-coordinate: $\frac{31st+35th}{2}$; other two circles' center y-coordinate: one radius away on either side
head top (upper half-circle)	center: 27th, radius: distance between 27th and 16th

3.5. Auxiliary Line Generation by OpenCV

Next, we review the procedure of generating the remaining auxiliary lines using functions in the OpenCV library. Actually, the functions for grayscale conversion, Gaussian blur, thresholded, color inversion, mean filtering, edge expansion, and the Canny algorithm are used to process the image for edge detections.

In this procedure, keypoint coordinates of the 17th~26th, 37th, 38th, 40th, 41th, 43th, 44th, 46th, and 47th keypoints from OpenPose are used to locate the positions of the eyebrows. In addition, they are used to eliminate unnecessary lines.

For the eyeglass in the image, first, we extract a specific area from the grayscale image as the Region of Interest (*ROI*) for the eyeglass [18]. Then, we apply mean filtering to blur the extracted eyeglass region to smooth the image. After blurring, we dilate the image to highlight the frame of the eyeglass. Finally, we use the Canny algorithm to detect and identify the edge of the eyeglass. Figure 4 shows an example of generating auxiliary lines for the eyeglass.

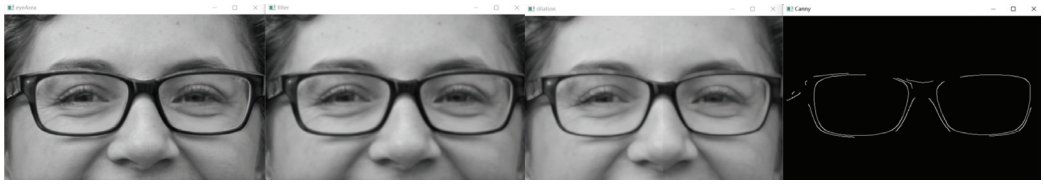


Figure 4. Auxiliary line generation example for eyeglass.

4. Proposal of Drawing Exactness Evaluation Method

In this section, we propose the drawing accuracy evaluation method using *Normalized Cross-Correlation* (*NCC*) [19] for the *PDLAS*.

4.1. Normalized Cross-Correlation (NCC)

The proposed method adopts *Normalized Cross-Correlation (NCC)* to evaluate the similarity of the user drawing result with the given face photo. *NCC* can measure the similarity between two signals. It is particularly well suited for image processing [20–24]. Correlation is widely used as an effective similarity measure in matching tasks, since it is the simplest but most effective method for similarity measurement [25]. Technically, *NCC* determines the matching point between the template and the image by searching the location of the maximum value in the image matrices [26,27]. *NCC* can provide a robust framework for evaluating likeness on a pixel-by-pixel basis by primarily reflecting the accuracy in replicating the shapes and positions of the facial features.

The *NCC* calculation formula is given as follows [28]:

$$NCC = \frac{\sum_{x,y}[T(x,y) \cdot I(x+x', y+y')]}{\sqrt{\sum_{x,y} T(x,y)^2 \cdot \sum_{x,y} I(x+x', y+y')^2}}$$

4.1.1. Template Matching

In the *NCC* formula, the term $(T(x,y))$ represents the pixel value of the template image at the coordinates $((x,y))$. $(I(x+x', y+y'))$ is the pixel value of the input image. $((x', y'))$ is the offset relative to the template image. The core idea of this template matching is to find the parts of an input image that will match a template image [29]. The template image T slides across the input image I at all the possible positions $((x', y'))$ while calculating the similarity score at each point [30].

4.1.2. Numerator Calculation

The numerator in the *NCC* formula represents the direct correlation between the template and a specific area of the input image. It is calculated by summing the product of the pixel values at the corresponding positions in the template and the input image, considering the relative offset. This sum reflects the degree to which the two sets of pixel values co-vary. A high sum suggests a strong positive correlation, indicating a potential match.

4.1.3. Denominator Calculation

The denominator in the *NCC* formula normalizes the values, considering the brightness range of each image [31]. It serves to normalize this correlation, accounting for the brightness and contrast between the template images and input images. It does so by considering the standard deviations of pixel values in both the template and the examined region of the input image. This normalization ensures that the *NCC* score is not biased by differences in image lighting or exposure, providing a more accurate measure of similarity. The *NCC* score ranges from -1 to 1, where 1, -1, and 0 indicate complete similarity, complete dissimilarity, and no correlation, respectively.

4.2. NCC *Score for Face Component*

For accurate and detailed feedback to a user, the *NCC* score is calculated for a specific face component such as an eye, a nose, an eyebrow, and a mouth in the proposed method. For this purpose, the bounding box surrounding each component with a rectangular frame is first extracted using 70 keypoints by OpenPose [32]. Then, the *NCC* score is calculated for the image inside the bounding box.

The following procedure describes the outline of the proposed *NCC* score calculation for each face component:

1. *Select Region*: we select the region in the face image that corresponds to the component to be evaluated.
2. *Extract Feature Vector*: we extract the pixel values of the region to form the feature vector. It may be necessary to pre-process the data to reduce the effects of noises and variations.
3. *Normalize Feature Vector*: we normalize the feature vector so that it has a zero mean and unit variance. This can reduce the effect of lighting or exposure differences between images.
4. *Calculate Dot Product*: we calculate the dot product of the two normalized feature vectors, where the result is a measure of their similarity.
5. *Normalize Dot Product*: we divide the dot product result by the product of the lengths of the feature vectors for the normalized correlation score to obtain the final *NCC* value.

4.3. Bounding Box for Face Component

The bounding box for each face component is extracted using the coordinates of the keypoints by OpenPose that are related to it. Since a bounding box is rectangular, the leftmost x-coordinate, the rightmost x-coordinate, the upmost y-coordinate, and the downmost y-coordinate should be extracted.

Table 2 shows the keypoints used to extract each face component's bounding box.

Table 2. *Keypoints* for bounding box.

Component	Leftmost x-Coordinate	Rightmost x-Coordinate	Upmost y-Coordinate	Downmost y-Coordinate
left eye	36th	39th	17th	28th
right eye	42nd	45th	26th	28th
left eyebrow	17th	21st	Highest of 28th, 29th, 30th	Lowest of 17th, 22nd
right eyebrow	22nd	26th	Highest of 23rd, 24th, 25th	Lowest of 22nd, 26th
mouth	48th	54th	33rd	5th
nose	39th	42nd	29th	33rd

5. Application to Existing Drawing Results

In this section, we apply the proposed method to the drawing results of the *PDLAS* in [33] that were obtained previously and discuss the results.

5.1. NCC *Score Results*

Table 3 shows the calculated *NCC* scores by the proposed method for both eyes, the nose, the mouth, and the eyebrows in the drawing results of seven students at Okayama University, Japan.

Table 3. NCC score results of different facial features.

User	Left Eye	Right Eye	Mouth	Nose	Left Eyebrow	Right Eyebrow	Average
User 1	0.56	0.74	0.59	0	0.14	0.33	0.39
User 2	0.48	0.55	0.25	0.12	0.39	0.42	0.37
User 3	0.14	0.16	0.21	0.08	0.06	0.03	0.11
User 4	0.36	0.40	0.23	−0.02	0.06	0.32	0.23
User 5	0.40	0.38	0.40	0.25	0.17	0.47	0.35
User 6	0.64	0.57	0.37	0.17	0.14	0.19	0.35
User 7	0.04	0.03	−0.09	0.05	0.10	−0.02	0.02
average	0.37	0.40	0.28	0.09	0.15	0.25	0.26

The table indicates the following:

(1) Right eye gives the highest score while nose gives the lowest.
(2) User 1 gives the highest score while User 7 gives the lowest.

We will analyze these results in the following subsections.

5.2. Analysis of Application Result of Top-Score User

First, we discuss the application result of the portrait drawing result of User 1, which gives the highest average *NCC* score. Figure 5 illustrates the face photo and the drawing result.

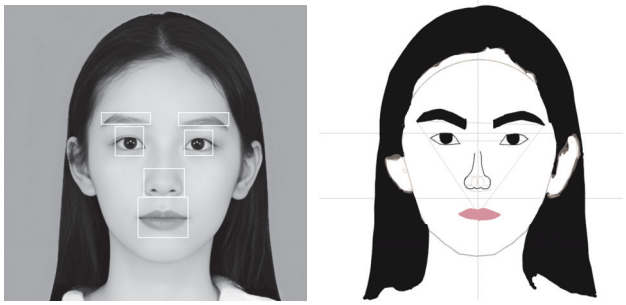


Figure 5. Drawing result of *User 1*. (Reproduced with permission from Yu H.)

The highest score is achieved for the right eye, suggesting that the user's depiction of the right eye is more similar to the reference image compared to the left eye. This user might have captured the shape and position of the right eye better. The score for the mouth is 0.59, indicating a moderate level of similarity with the reference image, but the shape or position of the mouth does not perfectly match the original image.

Notably, the *NCC* score for the nose is zero. Observing the user's rendition of the nose, we found that the auxiliary lines for the nose do not match the size and the proportion of the original nose image, due to the complexity of rendering the nose accurately among all the facial features.

5.3. Analysis of Application Result of Bottom-Score User

Next, we discuss the application result of the portrait drawing result of *User 7*, which gives the lowest average *NCC* score. Figure 6 illustrates the face photo and the drawing result.

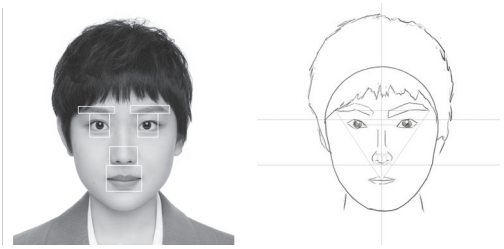


Figure 6. Drawing result of *User 7*. (Reproduced with permission from Qi H.)

Analyzing the results for User 7, this user has achieved low scores across all features. Notably, negative scores were obtained for the eyes and the right eyebrow. The left eye and the right eye both show a low positive correlation. From the drawing result, we observed that the user's drawing linework for the eyes is rather stiff, and the user has not made good use of the auxiliary lines provided by the system, resulting in the overall low score. Overall, we found the following issues in the user's drawing:

(i) The user's lines when drawing facial features are stiff.
(ii) There is a notable difference in the NCC scores between the left and right sides.

NCC Score Result

Analyzing the results of the second user, we notice that the *NCC* scores for the mouth and nose are the lowest. The score for the left eye is close to 0.5, which indicates a moderate level of similarity to the left eye in the reference image.

Similarly, the *NCC* score for the right eye is slightly higher than that for the left eye, suggesting that this user may have been more accurate or meticulous in depicting the right eye. The user's representation of the nose shows a significant difference in proportion compared to the original image, which can also be seen in the picture. This may be due to the user's lack of skill in using shadows.

We found the following issues in the user's drawing:

1. The mastery of the proportions of the facial features, especially the width of the nostrils, needs improvement.
2. Similarly, the overall drawing level of the right side of the face is higher than that of the left side.

6. Improvement of Nose Auxiliary Line Generation

In this section, we present an improvement of the auxiliary line generation algorithm for nose. The *NCC* score is very low compared with the score for other face components.

6.1. Limitation of Current Auxiliary Lines

As shown in Figure 7, the current auxiliary lines for nose cover only the bottom boundary of the nose. The other boundary is not covered due to the three-dimensional nature of the nose, where the boundary may not be clear and may be changed depending on the viewpoint [34]. Then, in the drawing result, the size of the nose at the upper part becomes different from the original one in the photo image.

6.2. Auxiliary Lines by Bezier Curve

To solve this problem, in this section, we introduce the *Bezier curve* [35] for the auxiliary lines to represent the other boundary of the nose. The *Bezier curve* is a parametric curve used in computer graphics and related fields. It needs contour points considering global shape information, with the curve passing through the first and last control points [36]. In this paper, the following quadratic curve is adopted where t represents a parameter, and P_0 and P_1 are the first and last control points, respectively.

$$B(t) = (1-t)^2 P_0 + 2t(1-t)P_1 + t^2 P_2 \quad (1)$$

For P_0 and P_1 of the *Bezier curve* for the left nose boundary, keypoint 21 at the left eyebrow and keypoint 31 at the left nose end are used. For P_0 and P_1 for the right nose boundary, keypoint 22 at the right eyebrow and keypoint 35 at the right nose end are used.

Figure 8 shows the improved auxiliary lines by the proposed method.

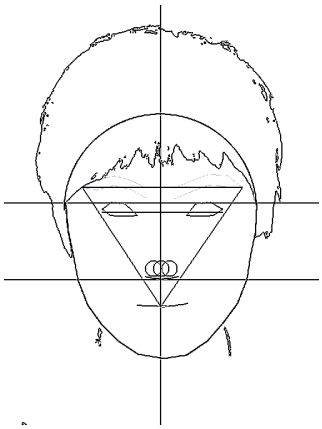


Figure 7. Auxiliary lines before improvement.

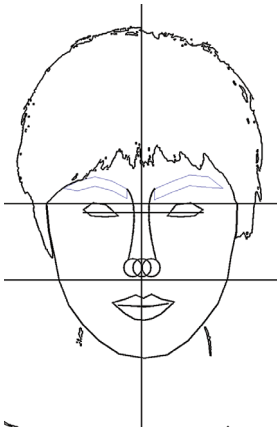


Figure 8. Improved auxiliary lines.

7. Evaluations After Improvement

In this section, we evaluate the improved auxiliary lines for the *nose* and the *drawing accuracy evaluation method* using *Normalized Cross-Correlation* (*NCC*).

7.1. Comparison of Nose NCC Score

First, to evaluate the improved auxiliary lines for the nose, we generated them and asked the same students to draw only the nose on the *PDLAS* using the same face photos. Table 4 compares the *NCC* scores before and after the improvement of nose auxiliary lines. It shows that the *NCC* score is improved by 222.22% on average.

Table 4. Comparison of *NCC* score for nose.

User	Before Improvement	After Improvement
User 1	0	0.24
User 2	0.12	0.29
User 3	0.08	0.31
User 4	−0.02	0.24
User 5	0.25	0.35
User 6	0.17	0.30
User 7	0.05	0.29
average	0.09	0.29

7.2. Application to Drawing Results for Diverse Face Photos

Second, to evaluate the *drawing accuracy evaluation method* under diverse face photos, we collected them from various countries with a range of ages and skin colors. Then, we asked new persons to draw the portraits using the *PDLAS*. Afterward, we calculated the *NCC* scores of the drawing results. Table 5 shows the results. We found that users' scores with the improved auxiliary lines were generally higher. Also, the average scores for both eyes are the same, showing no significant asymmetry problem.

Table 5. NCC scores for diverse face photos.

User	Left Eye	Right Eye	Mouth	Nose	Left Eyebrow	Right Eyebrow	Average
User 1	0.41	0.36	0.21	0.23	0.18	0.22	0.29
User 2	0.34	0.49	0.27	0.19	0.32	0.26	0.31
User 3	0.27	0.28	0.32	0.22	0.02	0.08	0.20
User 4	0.28	0.33	0.37	0.20	0.39	0.23	0.3
User 5	0.37	0.32	0.18	0.23	0.29	0.22	0.27
User 6	0.33	0.26	0.41	0.37	0.22	0.24	0.31
User 7	0.64	0.60	0.27	0.22	0.35	0.25	0.39
User 8	0.43	0.50	0.27	0.14	0.31	0.29	0.32
User 9	0.67	0.67	0.11	0.15	0.43	0.47	0.42
User 10	0.50	0.48	0.65	0.23	0.04	0.17	0.35
User 11	0.68	0.65	0.31	0.16	0.16	0.10	0.34
User 12	0.57	0.54	0.52	0.25	0.32	0.41	0.44
average	0.46	0.46	0.32	0.22	0.25	0.25	0.33

Unlike Tables 3 and 5 shows the results of the first use of the improved auxiliary line by the new user group. This helps to understand the initial adaptation of the new auxiliary line to new users. Despite the fact that the user groups are different, the overall mean scores of the new users in Table 5 increased, indicating the general effectiveness of the improved auxiliary line. The NCC scores for different facial features also varied slightly due to the different drawing abilities of each user. This also suggests that the PDLAS is able to detect different users' drawing abilities in different parts of the face through the scores, and also informs subsequent evaluations.

7.3. Evaluation of System Usability Scale

To evaluate the usability of the *PDLAS* and the drawing accuracy evaluation method, we asked the users to respond to the following 10 questions using a five-point scale: 1 (strongly disagree), 2 (disagree), 3 (neutral), 4 (agree), and 5 (strongly agree).

1. I think the auxiliary lines are easy to see.
2. I found the auxiliary lines do not match the image.
3. I think the auxiliary lines are useful for drawing.
4. I think this system is difficult to use.
5. It was enjoyable to use this system.
6. I feel it takes a long time to draw with this system.
7. I want to use the system to draw other faces.
8. I think the NCC scoring feedback is NOT helpful for improving drawing skills.
9. I am satisfied with the accuracy of the NCC scores for different features.
10. I think the size and position of each rectangle accurately reflect the area of each feature.

Then, the *System Usability Scale (SUS) score* is calculated from the answer results. The *SUS score* is an efficient and inexpensive tool for assessing the usability of a system, especially for small samples. It evaluates the system usability based on validity, frequency of use, and satisfaction. The *SUS final score* for each user is obtained through the following procedure:

1. Obtain the score by subtracting 1 from the answer for questions 1, 3, 5, 7, and 9.
2. Obtain the score by subtracting the answer from 5 for questions 2, 4, 6, 8, and 10.
3. Obtain the total score by summing all the scores.
4. Obtain the *SUS final score* by multiplying total score by 2.5.

Table 6 shows the questionnaire and *SUS score* results. The highest score is 97.5, the lowest score is 62.5, and the average score is 84. Therefore, the proposed method has a good usability [37].

Table 6. Questionnaire and SUS score results.

User	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Total Score	Final SUS Score
User1	5	2	4	2	5	1	4	2	4	1	34	85
User2	5	1	5	2	4	1	4	1	4	1	36	90
User3	5	1	5	1	5	2	5	1	5	1	39	97.5
User4	5	2	4	1	4	1	5	2	4	1	35	87.5
User5	4	3	4	2	4	1	3	3	4	1	26	65
User6	5	1	5	1	4	1	4	1	5	1	38	95
User7	4	1	5	1	5	1	5	2	4	1	37	92.5
User8	4	1	5	1	4	1	4	1	4	1	36	90
User9	5	1	5	2	5	1	5	2	5	2	37	92.5
User10	4	2	5	2	3	1	4	1	4	2	32	80
User11	5	2	4	3	4	3	5	2	3	2	25	62.5
User12	4	2	4	2	3	1	4	2	4	1	31	77.5
User13	5	2	4	2	4	1	3	2	4	2	31	77.5
average	4.6	1.6	4.5	1.7	4.2	1.2	4.2	1.7	4.2	1.3	33.6	84

7.4. Limitations of Proposed Method

Here, we discuss limitations of the proposed drawing accuracy evaluation method for the *PDLAS* that we experienced through experiments and evaluations.

7.4.1. Sensitive to Local Change

The adopted *NCC* algorithm relies on the local similarity between two images. This means that even small displacements, scaling, or rotations of a component in an image can result in a large drop in the *NCC* score. In particular, this can easily happen for face components with a small size such as the eyes and nose.

7.4.2. Insufficient Shading Evaluation

The *NCC* algorithm is sensitive to changes in the image brightness and contrast [38]. In particular, for a component with a three-dimensional structure such as the nose, *NCC* cannot evaluate the shading of it correctly.

7.4.3. Single Evaluation Dimension

The *NCC* algorithm focuses on structural similarities between two images, and is weak at evaluating other artistic dimensions such as technical accuracy, creativity, and emotional expressions. Adopting AI, evaluations from professional artists should be considered for other dimensions of the artwork, which will be outlined in future works.

8. Conclusions

This paper presented a drawing exactness assessment method using a *Localized Normalized Cross-Correlation* (*NCC*) algorithm in a Portrait Drawing Learning Assistant System (*PDLAS*). The *PDLAS* has been developed as a self-study tool to make art education more accessible and interesting for drawing beginners. It simplifies the process of learning portrait drawing and integrates technologies into traditional education.

The proposed algorithm gives the similarity score with the correlation between the original face photo and drawing result images for each face component. In addition, the auxiliary lines for the nose were improved.

For evaluations, the proposed algorithm was applied to the drawing results of students in Okayama University, Japan, using the *PDLAS*, where the results validated the effectiveness of suggesting improvements in drawing components. The system usability was confirmed through the questionnaire with the *SUS score*.

However, our study of the PDLAS still has some limitations. Firstly, the drawing results that are filled in color have higher scores than those that use lines only. Coloring or not has some effect on the results. Secondly, evaluations of other features of the face have not yet been realized, such as glasses and accessories. Thirdly, the system has been used by a limited number of young students from Asia. It should be used by a wider population including different ethnicities, cultures, and ages. Moreover, in the system, no user feedback is available currently. Specific suggestions on how to improve the drawing should be provided. In the future, we will continue the PDLAS study to solve these limitations.

Author Contributions: Y.Z. wrote the code according to the research objectives, and was mainly responsible for writing part of the auxiliary line algorithm section and the evaluation section, and collecting user results. N.F. led the research from the beginning and developed the research objectives and the system functionality. Z.K. wrote the code according to the research objectives, and was mainly responsible for writing the auxiliary line algorithm section. C.-C.H. refined the paper writing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The authors would like to assure readers that all participants involved in this research were informed, and all photographs used were supplied voluntarily by the participants themselves. The authors have ensured that all ethical considerations, particularly regarding participant consent and data protection, have been rigorously adhered to.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: No new data were created or analyzed during this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Blake, W.; Lawn, J. *Portrait Drawing*; Watson-Guptill Publications Inc.: New York, NY, USA, 1981; pp. 4–5.
2. Kong, Z.; Zhang, Y.; Funabiki, N.; Huo, Y.; Kuribayashi, M.; Harahap, D.P. A Proposal of Auxiliary Line Generation Algorithm for Portrait Drawing Learning Assistant System Using OpenPose and OpenCV. In Proceedings of the 2023 IEEE 12th Global Conference on Consumer Electronics (GCCE), Nara, Japan, 10–13 October 2023.
3. Zhang, Y.; Kong, Z.; Huo, Y.; Funabiki, N.; Harahap, D.P. A Preliminary Evaluation of Portrait Drawing Learning Assistant System. In Proceedings of the 2023 IEEE 12th Global Conference on Consumer Electronics (GCCE), Nara, Japan, 10–13 October 2023.
4. CMU Perceptual Computing Lab. Available online: https://cmu-perceptual-computing-lab.github.io/openpose/web/html/doc/index.html (accessed on 30 June 2024).
5. OpenCV. Available online: https://opencv.org/ (accessed on 30 June 2024).
6. Yi, R.; Liu, Y.; Lai, Y.; Rosin Paul, L. Unpaired Portrait Drawing Generation via Asymmetric Cycle Mapping. In Proceedings of the IEEE CVPR, Seattle, WA, USA, 13–19 June 2020.
7. Yi, R.; Liu, Y.-J.; Lai, Y.-K.; Rosin, P.L. Quality Metric Guided Portrait Line Drawing Generation From Unpaired Training Data. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 905–918. [CrossRef] [PubMed]
8. Tong, Z.; Chen, X.; Ni, B.; Wang, X. Sketch Generation with Drawing Process Guided by Vector Flow and Grayscale. *arXiv* **2020**, arXiv:2012.09004.
9. Looi, L.; Green, R. Estimating Drawing Guidelines for Portrait Drawing. In Proceedings of the IVCNZ, Tauranga, New Zealand, 9–10 December 2021.
10. Takagi, S.; Matsuda, N.; Soga, M.; Taki, H.; Shima, T.; Yoshimoto, F. An educational tool for basic techniques in beginner's pencil drawing. In Proceedings of the PCGI, Tokyo, Japan, 9–11 July 2003.
11. Huang, Z.; Peng, Y.; Hibino, T.; Zhao, C.; Xie, H. DualFace: Two-stage drawing guidance for freehand portrait sketching. *Comp. Visual Media* **2022**, *8*, 63–77. [CrossRef]

12. Iarussi, E.; Bousseau, A.; Tsandilas, T. The drawing assistant: Automated drawing guidance and feedback from photographs. In Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology, St. Andrews Scotland, UK, 8–11 October 2013.
13. Gao, F.; Zhu, J.; Yu, Z.; Li, P.; Wang, T. Making Robots Draw A Vivid Portrait In Two Minutes. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Las Vegas, NV, USA, 24 October 2020–24 January 2021.
14. Xie, J.; Hertzmann, A.; Li, W.; Winnemöller, H. PortraitSketch: Face sketching assistance for novices. In Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology, Honolulu, HI, USA, 5–8 October 2014.
15. Li, S.; Xie, H.; Yang, X.; Chang, C.-M.; Miyata, K. A Drawing Support System for Sketching Aging Anime Faces. In Proceedings of the International Conference on Cyberworlds (CW), Kanazawa, Japan, 27–29 September 2022.
16. Hebsi, M.A.A.; Akre, V.L. A framework for integrating iPAD as a pedagogical device in higher education. In Proceedings of the ICEDSA, Ras Al Khaimah, United Arab Emirates, 6–8 December 2016.
17. Hiromasa, U. *Hiromasa's Drawing Course: How to Draw Faces*; Kosaido Pub.: Tokyo, Japan, 2014.
18. OpenCV: Open Source Computer Vision. Available online: https://docs.opencv.org/4.x/d3/df2/tutorial_py_basic_ops.html (accessed on 30 June 2024).
19. Gonzalez, R.C.; Woods, R.E.; Eddins, S.L. *Digital Image Processing with Matlab*; Gatesmark Publishing: Knoxville, TN, USA, 2020; pp. 311–315.
20. Sarvaiya, J.; Patnaik, S.; Bombaywala, S. Image Registration by Template Matching Using Normalized Cross-Correlation. In Proceedings of the International Conference on Advances in Computing, Control, and Telecommunication Technologies, Bangalore, India, 28–29 December 2009.
21. Pele, O.; Werman, M. Robust Real-Time Pattern Matching Using Bayesian Sequential Hypothesis Testing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1427–1443. [CrossRef] [PubMed]
22. Li, G. Stereo matching using normalized cross-correlation in LogRGB space. In Proceedings of the International Conference on Computer Vision in Remote Sensing, Xiamen, China, 16–18 December 2012.
23. Dawoud, N.N.; Samir, B.; Janier, J. N-mean kernel filter and normalized correlation for face localization. In Proceedings of the IEEE 7th International Colloquium on Signal Processing and its Applications, Penang, Malaysia, 4–6 March 2011.
24. Chandran, S.; Mogiloju, S. Algorithm for Face Matching Using Normalized Cross-Correlation. *Int. J. Eng. Adv. Technol.* **2013**, *2*, 2249–8958.
25. Feng, Z.; Qingming, H.; Wen, G. Image matching by normalized cross-correlation. In Proceedings of the ICASSP, Toulouse, France, 14–19 May 2006.
26. Zhang, B.; Yang, H.; Yin, Z. A region-based normalized cross correlation algorithm for the vision-based positioning of elongated IC chips. *IEEE Trans. Semicond. Manuf.* **2015**, *28*, 345–352. [CrossRef]
27. Hisham, M.B.; Yaakob, S.N.; Raof, R.A.A.; Nazren, A.B.A.; Wafi, N.M. Template matching using sum of squared difference and normalized cross correlation. In Proceedings of the SCOReD, Kuala Lumpur, Malaysia, 13–14 December 2015.
28. Ban, K.D.; Lee, J.; Hwangm, D.H.; Chung, Y.K. Face image registration methods using normalized cross correlation. In Proceedings of the 2008 International Conference on Control, Automation and Systems, Seoul, Republic of Korea, 14–17 October 2008.
29. Fouda, Y.; Ragab, K. An efficient implementation of normalized cross-correlation image matching based on pyramid. In Proceedings of the iCAST, 2013 & UMEDIA, Aizu-Wakamatsu, Japan, 2–4 November 2013.
30. Cui, Z.; Qi, W.; Liu, Y. A Fast Image Template Matching Algorithm Based on Normalized Cross Correlation. *J. Phys. Conf. Ser.* **2020**, *1693*, 012163. [CrossRef]
31. Baloch, S.; Ghazali, F.; Mir, N.; Farrukh, F.; Ismail, H.; Qadri, M.Y. Hardware realization of locally normalized cross correlation algorithm. In Proceedings of the ICSAI, Shanghai, China, 15–17 November 2014.
32. Gao, Q.; Chen, H.; Yu, R.; Yang, J.; Duan, X. A robot portraits pencil sketching algorithm based on face component and texture segmentation. In Proceedings of the ICIT, Melbourne, VIC, Australia, 13–15 February 2019.
33. Zhang, Y.; Kong, Z.; Funabiki, N.; Puspitaningayu, P. A Drawing Accuracy Evaluation Method with Localized Normalized Cross-Correlation for Portrait Drawing Learning Assistant System. In Proceedings of the ITET, Tottori, Japan, 10–12 May 2024.
34. Civardi, G. *Drawing Portraits: Faces and Figures*, 1st ed.; Search Press Limited: Tunbridge Wells, Kent, UK, 2002; p. 20.
35. Hearn, D.; Baker, P.M. *Computer Graphics, C Version*, 2nd ed.; Prentice Hall: Upper Saddle River, NJ, USA, 1997; pp. 27–331.
36. Lee, Y.-H.; Lee, Y.; Park, J.-H.; Han, W.; Kim, Y. Emotional Recognition System Using Curve Fitting with Bezier Cubics. In Proceedings of the 2013 International Conference on Information Science and Applications (ICISA), Jeju Island, Republic of Korea, 18–22 June 2013; pp. 1–2.
37. Sharfina, Z.; Santoso, H.B. An Indonesian adaptation of the System Usability Scale (SUS). In Proceedings of the ICACSIS, Malang, Indonesia, 15–16 October 2016.
38. Abdulsada, N.T.; Ali, S. Automatic face recognition using normalized cross correlation (NCC) function with variable template size. *AIP Conf.* **2022**, *2437*, 020116.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

computers

MDPI

Article

A Platform for Integrating Internet of Things, Machine Learning, and Big Data Practicum in Electrical Engineering Curricula

Nandana Jayachandran [1], Atef Abdrabou [1,*], Naod Yamane [1] and Anwer Al-Dulaimi [2]

[1] Department of Electrical and Communication Engineering, College of Engineering, UAE University, Al-Ain P.O. Box 15551, United Arab Emirates; 700046308@uaeu.ac.ae (N.J.); 201850217@uaeu.ac.ae (N.Y.)
[2] College of Technical Innovations, Zayed University, Abu Dhabi P.O. Box 144534, United Arab Emirates; anwer.al-dulaimi@zu.ac.ae
* Correspondence: atef.abdrabou@uaeu.ac.ae; Tel.: +971-3-713-5149

Abstract: The integration of the Internet of Things (IoT), big data, and machine learning (ML) has pioneered a transformation across several fields. Equipping electrical engineering students to remain abreast of the dynamic technological landscape is vital. This underscores the necessity for an educational tool that can be integrated into electrical engineering curricula to offer a practical way of learning the concepts and the integration of IoT, big data, and ML. Thus, this paper offers the IoT-Edu-ML-Stream open-source platform, a graphical user interface (GUI)-based emulation software tool to help electrical engineering students design and emulate IoT-based use cases with big data analytics. The tool supports the emulation or the actual connectivity of a large number of IoT devices. The emulated devices can generate realistic correlated IoT data and stream it via the message queuing telemetry transport (MQTT) protocol to a big data platform. The tool allows students to design ML models with different algorithms for their chosen use cases and train them for decision-making based on the streamed data. Moreover, the paper proposes learning outcomes to be targeted when integrating the tool into an electrical engineering curriculum. The tool is evaluated using a comprehensive survey. The survey results show that the students gained significant knowledge about IoT concepts after using the tool, even though many of them already had prior knowledge of IoT. The results also indicate that the tool noticeably improved the students' practical skills in designing real-world use cases and helped them understand fundamental machine learning analytics with an intuitive user interface.

Keywords: IoT; AI; machine learning; big data; GUI; tool; education; engineering; curricula

1. Introduction

The integration of the Internet of Things (IoT), artificial intelligence (AI), and big data analytics has revolutionized various industries, and the field of electrical engineering is no exception. Teaching electrical engineering students about the benefits of integrating these technologies is crucial for preparing them for the rapidly evolving technological landscape.

The IoT [1] has enabled the interconnection of devices and systems, allowing for seamless communication and data exchange. By incorporating IoT concepts into the curriculum, electrical engineering students can learn how to design and develop smart systems that collect and analyze data from various sources. This knowledge is essential as IoT applications continue to expand in areas such as smart grids, smart homes, and industrial automation.

AI, on the other hand, empowers machines to perform tasks that typically require human intelligence. Teaching AI to electrical engineering students equips them with the skills to develop intelligent systems capable of making autonomous decisions, optimizing energy consumption, and improving overall system efficiency. AI algorithms can analyze large datasets generated by IoT devices, enabling predictive maintenance, fault detection, and energy optimization in electrical systems.

Big data analytics complements the integration of IoT and AI by providing the tools and techniques to extract valuable insights from vast amounts of data. Electrical engineering students can learn how to process, analyze, and visualize big data to identify patterns, trends, and anomalies. This knowledge can be applied to optimize power distribution, improve energy efficiency, and enhance system reliability.

By integrating IoT, AI, and big data analytics into the curriculum, electrical engineering students can gain a comprehensive understanding of the potential benefits and challenges associated with these technologies. They also can develop the skills necessary to design and implement intelligent systems, leverage data-driven decision-making, and contribute to the development of sustainable and efficient electrical systems.

Many research works in the literature, such as [2,3], primarily concentrate on utilizing big data and AI to assist instructors in their instructional roles, whereas others focus on the individual introduction of one of these technologies in specialized courses (e.g., [1,4]). However, these approaches do not intend to provide the students with hands-on experience regarding the practical integration of IoT, big data, and ML in real-world applications, hindering their exploration of the interplay between these technological concepts.

This paper bridges this gap by developing an integrated platform that combines IoT, big data, and ML concepts. This platform is an open-source emulation software tool called IoT-Edu-ML-Stream. https://github.com/nandanajayachandran/IoT-Edu-ML-Stream-Tool (accessed on 11 August 2024) that can be incorporated as a part of a course or a course component of a laboratory style into an electrical engineering curriculum, enabling students to gain a deep learning experience of these technologies. The tool allows the students to create scenarios or use cases using emulated or real IoT devices that generate correlated realistic IoT sensor data in real time for various types of sensors and publish this data via the message queuing telemetry transport (MQTT) protocol [5] to a big data streaming platform, such as Apache Kafka [6]. Based on the generated data, the tool also helps the students build ML models of their choice out of different available techniques. In addition, the tool offers students near real-life experience in training, testing, and parameter-tuning their ML models.

The paper offers two main contributions. First, it introduces the design of the aforementioned emulation software tool with a graphical user interface. The tool provides practical, hands-on experience for electrical engineering students in creating IoT networks with real-time realistic IoT data generation, streaming, monitoring, and intelligent big data analytics using ML. Second, the paper provides proffered course learning outcomes that can be covered when the proposed tool is integrated into electrical engineering courses or course components. Furthermore, it introduces the reflection of these course learning outcomes on commonly used engineering program learning outcomes to help assess students' performance using an outcome-based learning approach.

The rest of the paper is organized as follows. Section 2 provides an overview of the most relevant works in the literature. Section 3 introduces the study content and significance. Section 4 details the research methodology, including research questions, the proposed tool design approach, its features, user interface, integration into electrical engineering curricula, and evaluation procedure. Section 5 presents the tool evaluation results and discussion. Finally, the paper is concluded in Section 6, highlighting future research directions.

2. Related Works

Artificial intelligence (AI) is a rapidly growing field with applications in various domains, including STEM education, big data analytics, materials informatics, and machine learning [1,4,7,8].

Several studies emphasize the integration of AI and big data into the realm of education. Khan et al. [2] propose the use of adaptive e-learning systems powered by AI and big data in higher educational institutions to assist educators and students in the teaching process and provide insights into the educational implications of these technologies in

higher education [2]. In [3], the authors discuss the design of an innovative teaching model based on AI technology, incorporating artificial intelligence hardware and software such as big data and cloud computing. The model covers pre-class, in-class, and after-class teaching, providing insights into how AI can be integrated into the teaching process [3].

Moreover, some studies describe the importance of exploring the use of learning analytics, computational analysis techniques, and AI in higher education. In [9], Montuori et al. highlight the value of learning analytics in tracking student progress, analyzing university data, and designing assessments. In [10] explores the role of human learning and learning analytics in the age of artificial intelligence, emphasizing the importance of understanding human learning processes and leveraging learning analytics to enhance educational outcomes.

Other approaches target applying AI and big data analytics in domain-specific education. For instance, in [11], the author discusses the use of big data-driven systems in advanced mathematics education. The author highlights how these systems can efficiently analyze data information, design teaching strategies, and improve teaching quality. Barenkamp et al. [12] discuss the applications of AI in classical software engineering, highlighting the automation of routine jobs, structured analysis of big data pools, and systematic evaluation using neural networks. Wei in [13] discusses the integration of information infrastructure and big data analytics in an online learning framework for rural education.

Some research works focus on enhancing employment opportunities for students via big data analytics courses. Xu et al. [14] explore the impact of digital education and the role of big data analytics courses in enhancing the skills and employability of engineering students. Tsai et al. [15] discuss precision education using statistical learning and deep learning, focusing on the impact of digital education and the role of big data analytics courses in enhancing students' skills and employability.

On the other hand, educators are recognizing the importance of incorporating AI concepts into their curricula to prepare students for the future workforce [1,4,7,8]. One approach to teaching AI is through interactive, hands-on methodologies that promote learning by doing. Xie-Li et al. [1] propose a methodology that uses open-source electronic platforms, such as Arduino, Snap Circuits, Raspberry Pi, and Circuit Playground, to teach AI concepts. Gil et al. [16] describe another way to teach AI using a workflow paradigm to capture and reuse end-to-end analytic methods, enabling students to learn about software reuse, parallel distributed programming, and high-end computing. Ayotte et al. [7] compare two different methods of teaching machine learning concepts to undergraduate students in electrical engineering. They found that a workshop approach may be more effective for student learning than a side-by-side teaching approach. Filsinger [8] discusses the design of an AI course tailored to the needs of electrical and computer engineering technology students, offering a survey of various theoretical techniques presented in a mathematically accessible way. Mannodi-Kanakkithodi et al. [4] emphasizes the importance of educating materials scientists in data science and AI. They suggest providing researchers with workshops to learn and apply AI/ML tools to materials' data.

A few research papers address the introduction of IoT into electrical engineering curricula. A remote Arduino lab designed for practicing microcontroller and IoT programming is introduced in [17] for electrical and computer engineering students. Another hardware-based platform, Espressif, is proposed in [18] to develop IoT devices, enabling the students to build their own projects. In [19], Al Taai et al. cover the use of IoT in education, its role in addressing educational challenges, and its significant applications for students and teachers. It also highlights IoT's contributions to sustainable development, environmental protection, and various application areas. Some studies have introduced IoT traffic generators [20] and analyzed those data using ML without focusing on the education sector.

To harness the potential of the Internet of Things (IoT) and blockchain in education, several research studies have emphasized innovative applications and benefits. One

notable study by Alammary et al. [21] conducted a systematic review of blockchain-based applications in education, highlighting the growing interest in leveraging blockchain technology to enhance educational processes. The study provides insights into the various ways blockchain can be utilized in educational settings to improve security, transparency, and efficiency. Blockchain technology provides a secure and transparent platform for storing and sharing academic records, thereby enhancing the integrity and accessibility of educational resources, as well as streamlining administrative processes [22]. Through the use of blockchain, educators and learners can track learning outcomes, address educational challenges, and verify the authenticity of educational documents [23,24]. Furthermore, blockchain can support decentralized education systems, fostering collaboration among educational institutions and driving advancements in education quality [25].

Moreover, IoT technologies can monitor students' physical fitness data, improve supply chain traceability, and enhance logistics management within educational settings [26,27]. IoT serves as a valuable educational tool for learning programming, addressing educational issues, and offering essential applications for both students and teachers [19]. Additionally, IoT can help overcome barriers to education, such as geographical constraints and economic disparities, providing new opportunities for educational access and inclusivity [28].

Some other research works address the combination of blockchain and IoT in education. For instance, blockchain technology can enhance data security, traceability, and interoperability in IoT applications, ensuring the integrity of educational data and systems [29]. By merging blockchain and IoT, educational stakeholders can explore decentralized learning models, enhance documentation processes, and improve the overall efficiency of educational operations [30]. The adoption of blockchain and IoT in education tackles current challenges and sets the stage for future innovations and progress in the education sector [31].

In summary, the aforementioned research works mainly address two research directions. The first is the usage of IoT, big data analytics, and AI in education to facilitate the teaching process, enhance learning outcomes, and improve the employability chances of students. Additionally, this research direction explores the use of IoT technologies to monitor and enhance educational environments, as well as the application of blockchain technology to secure and optimize educational processes. The second direction is teaching AI concepts in different domains, including interactive, hands-on methodologies, intelligent workflow systems, materials informatics workshops, and tailored AI courses for specific disciplines. However, these works do not tackle preparing students with hands-on experience in the practical integration and interplay of IoT, big data, and ML in real-world applications.

Different from the reviewed research literature, the proposed tool offers a hands-on approach that enhances students' practical experience by immersing them in real-world IoT, big data, and AI environments. By working with actual hardware and software components, students gain valuable skills in data generation, network configuration, big data handling, and machine learning integration. This practical exposure prepares them for the job market, equipping them with the knowledge and confidence needed to tackle real-world challenges in the rapidly evolving fields of IoT, big data, and machine learning. Through this tool, the students will experience the transition from theoretical understanding to practical proficiency, making them highly competitive in their future careers.

To the best of our knowledge, no other educational tool has been proposed with target learning outcomes in the literature for electrical engineering students, with seamless integration of IoT, big data, and ML technologies. This integration is achieved via a GUI platform capable of real-time data generation using either real hardware or software emulation, allowing students to configure IoT devices, handle large volumes of data, and apply machine learning techniques through a user-friendly interface.

3. Study Content and Significance

3.1. Internet of Things (IoT)

The Internet of Things (IoT) is a vast network that connects everyday objects and devices, granting them the ability to observe, communicate, and collaborate effortlessly. As

IoT continues to evolve, it holds the potential to transform various industries, enhancing connectivity, intelligence, and efficiency while enriching our daily experiences. At its core, IoT relies on devices and objects equipped with embedded sensors that establish connections with one another via the Internet [32]. These devices share data with a central managing platform, which, in turn, provides insights through specialized applications tailored to specific needs. Thus, IoT can stream and automate numerous daily tasks across industries for many objectives, such as reducing costs, increasing productivity, enhancing safety, improving customer experiences, and generating new revenue streams, to name a few. In essence, IoT can empower engineers in various engineering disciplines to make more informed decisions.

As more and more devices become connected to the internet, it is becoming increasingly important for electrical engineering students to understand the principles of IoT. This knowledge can be applied to a wide range of promising industries, including, but not limited to, e-healthcare, smart transportation, and smart cities, which pave the way for students to pursue their careers after graduation in fields related to these industries. By studying IoT, students can learn about the various components that make up an IoT system, including industry-standard protocols and communication mechanisms used to connect these components, such as the MQTT protocol.

3.2. MQTT

MQTT is a lightweight and robust open messaging protocol designed for efficient IoT/machine-to-machine communication using a publish/subscribe technique [5]. It is tailored for IoT, offering secure bidirectional communication, real-time message delivery, and scalability to connect millions of devices via publish/subscribe operations even when not simultaneously connected.

When the MQTT client seeks to establish a connection with a central node called the MQTT broker, it initiates a TCP (Transmission Control Protocol) connection to the broker's designated IP address. Once connected to the broker, the MQTT client can send messages to specific topics, which serve as data transmission channels or destinations. The message can encompass a wide array of data types. MQTT imposes no limitations on the payload format, making it adaptable for diverse IoT applications. The MQTT client designates the topic to which it intends to send the message. Other MQTT clients, functioning as subscribers, can subscribe to particular topics to receive messages sent to those topics. This subscription enables them to receive messages published on a topic a client has subscribed to. Upon receiving the message, the broker forwards it to all subscribers interested in that topic. The broker ensures the delivery of that message to the subscribing client. The connection establishment diagram is shown below in Figure 1.

MQTT has developed a reputation as an efficient protocol that empowers IoT devices. Its minimal overhead lowers bandwidth usage, computing, and memory load on resource-constrained communicating nodes, such as IoT devices. MQTT achieves efficient data transmission in large-scale IoT networks through its lightweight nature, which minimizes the communication network traffic [5]. MQTT is currently a common choice for applications such as remote monitoring, patient health tracking, and danger alerts. Its efficient and scalable nature, combined with the publish/subscribe model and support for real-time communication, makes MQTT popular in the industry and, hence, a good choice for the proposed platform to help students gain valuable practical experience.

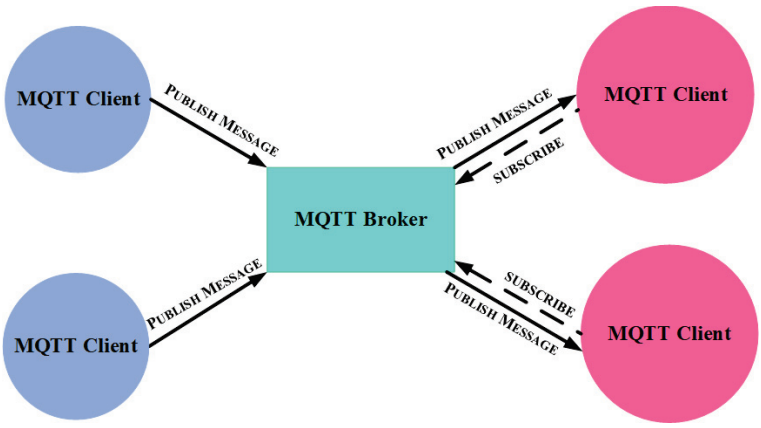


Figure 1. MQTT connection establishment.

3.3. Big Data and Big Data Platforms

The widespread adoption of technologies like IoT, social media platforms, and the significant increase in digitization across various sectors has led to the generation of massive amounts of data, commonly referred to as "big data". Big data encompasses vast and complex datasets that exceed the capacity of traditional data processing tools in terms of capturing, storing, managing, and analyzing [16]. It is characterized by three primary attributes: volume, velocity, and variety. Volume pertains to the sheer scale of available data, velocity refers to the speed at which data accumulates, and variety encompasses the diverse data sources. Additionally, two more attributes are commonly associated with big data: veracity (consistency and reliability of data) and value (usefulness of insights extracted from the data).

Big data platforms are specifically designed to handle large, dynamically generated, complex datasets in real time. These tools, developed by data management vendors, play a pivotal role in enhancing the scalability, accessibility, performance, and security of organizations relying on big data analytics. In today's landscape, a wide range of big data tools is available, spanning data stores, development platforms, integration tools for reporting, and advanced analytics. Prominent examples include Hadoop, Apache Spark, Apache Kafka, Cassandra, and Rapid Miner [14].

Notably, Apache Kafka, a renowned open-source platform, has emerged as a favored choice within the realm of big data, enabling real-time data streaming and processing at scale. It is a powerful and extensible open-source event streaming platform designed to facilitate the creation of real-time data pipelines and streaming applications [6]. It utilizes a publish–subscribe messaging system that excels at rapid and robust data processing. One of Kafka's standout features lies in its complete decoupling of data producers and data consumers, offering unmatched scalability. Kafka's architecture revolves around the concept of topics, which act as data containers accommodating multiple producers and consumers/subscribers, facilitating the repeated reading and processing of data/events. Importantly, Kafka's performance remains consistent regardless of data volume, making it an ideal choice for long-term data retention.

3.4. Machine Learning

Machine learning is a field that involves the development and understanding of models that have the ability to learn from data. ML algorithms enable models to approximate any function when provided with sufficient data in the right format. These algorithms fine tune themselves based on the data they are trained on. Thus, it is a sort of artificial intelligence that allows computer systems to learn and develop without being explicitly programmed. It entails using algorithms and statistical models to analyze data, detect trends, and make predictions or judgments based on the data without human intervention. ML popularity is fueled by the advances in chip computation, the abundance of data we

can now collect, and the continuous increase in computational power required for machine learning algorithms to run efficiently.

Teaching ML to electrical engineering students is important due to the current increasing demand for intelligent systems and the need to analyze vast amounts of data collected in real time from IoT devices. This makes machine learning a critical component of many modern technologies, including self-driving cars, voice assistants, predictive maintenance systems, and many others. Therefore, teaching ML to these students is essential for preparing them for the future of technology and ensuring that they have the skills needed to succeed in a rapidly changing job market.

3.5. Integration of IoT, a Big Data Platform and ML

The integration of a big data platform with IoT and machine learning not only enhances data processing capabilities but also facilitates real-time actions based on intelligent insights. Figure 2 shows a simple integration of the domains. With a big data platform, e.g., Kafka, providing a robust data transport mechanism, organizations can efficiently manage the influx of data from diverse IoT sources, ensuring data integrity and timely delivery. Machine learning models integrated into this pipeline can analyze data for patterns, trends, and anomalies in real time. For example, in a multidisciplinary application of agriculture and electrical engineering, wireless sensors in farms can transmit environmental data to Kafka, where machine learning models can provide insights into optimal irrigation schedules or pest control strategies. This synergy creates a closed-loop system in which insights generated by machine learning algorithms can trigger immediate actions, such as adjusting settings on connected devices and automating decision-making processes. The integration of Kafka, IoT, and machine learning opens doors to a wide array of applications that drive efficiency, cost savings, and innovation across industries.

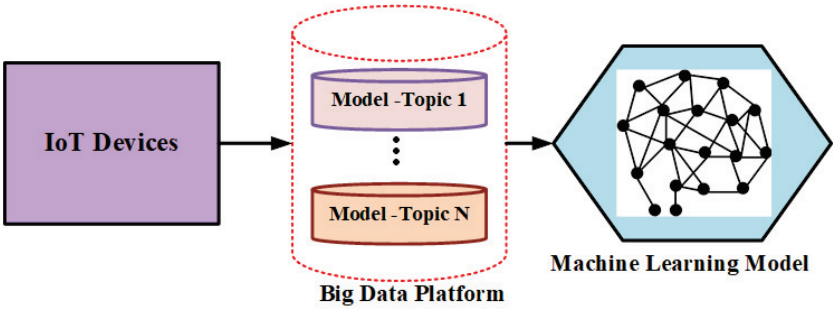


Figure 2. Integration of IoT, big data platform, and ML.

Students who gain hands-on training in these technologies will be at the forefront of innovation, leading the charge in industries like e-healthcare, smart agriculture, smart transportation, and smart cities. Practical knowledge of these technologies is crucial, as it equips students with the skills required to work in the real world. Training allows students to understand the complexities of these technologies, including the protocols and communication mechanisms used to connect various components. Furthermore, practical training helps students develop analytical and problem-solving skills. By working with real data sets, students learn to analyze and interpret data, identify patterns and trends, and derive insights that can be used to drive decision-making processes. This knowledge prepares graduates to make informed decisions, ultimately driving innovation and efficiency across many different industries.

4. Research Methodology

This section introduces the research questions, the tool design, and the integration of the tool into the curriculum.

4.1. Research Questions

This research mainly tackles the following research question: How can we design an educational tool that provides electrical engineering students with hands-on experience in the integration of IoT with big data analytics using basic ML techniques? Another related question is: How can such a tool be integrated into typical electrical engineering curricula?

4.2. Tool Design Approach and Features

Our main design objective is to create a GUI-based tool that integrates IoT, big data, and machine learning (ML) to provide hands-on experience for electrical engineering students in these fields. Consequently, the tool design should equip the students with hands-on experience in emulating IoT devices using the MQTT protocol, introduce them to real-time big data platforms, and encourage them to explore and analyze different ML techniques in a user-friendly manner. Therefore, the tool shall have an iterative structure that advocates correcting errors and developing proficiency through the educational journey.

To achieve the tool design objective, we followed a GUI paradigm that enables the students to interact with the tool via a step-by-step option-based approach. This approach guides them first to set up either real hardware or synthesized (virtual) IoT devices by selecting the types of devices, adding sensors to them, and configuring them as MQTT clients. A packet capturing tool (Pyshark [33]) is used to obtain the IoT data generated in real time either from hardware or data synthesis. Afterward, the data are streamed to a big data platform (Kafka). The streamed data are subsequently utilized to conduct machine learning analysis based on the machine learning algorithm chosen by the student user from a list of available algorithm options.

Furthermore, the tool design exemplifies scalability, effectively handling substantial volumes of IoT data. The tool is exclusively developed using Python, the industry standard for AI development. It can be further customized and extended to meet specific use cases and requirements, offering flexibility to adapt its functionality as needed. Figure 3 highlights the design features of the tool.

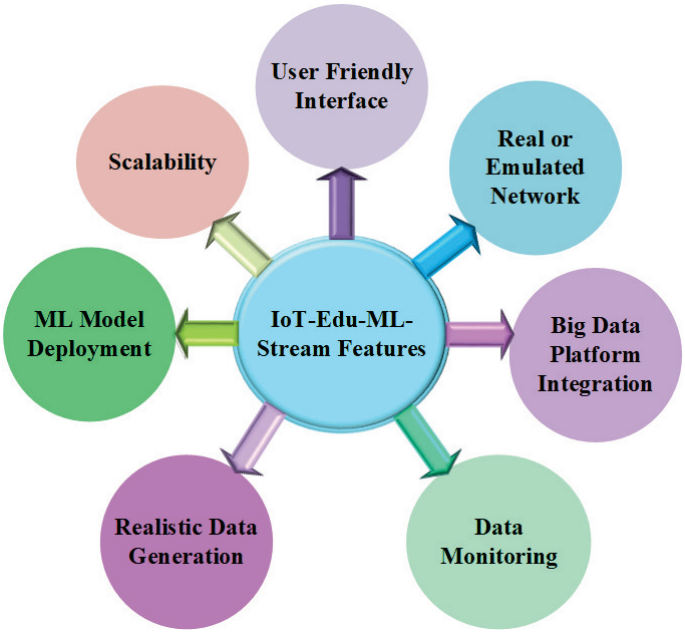


Figure 3. IoT-Edu-ML-Stream features.

Since the tool usage covers the interplay between IoT, big data, and machine learning, the tool design structure is modular, as illustrated in Figure 4. Initially, an IoT data generation module generates IoT data, which is subsequently captured, processed, and sent to the big data module that streams data in real time to the big data platform (Kafka). The

real-time streaming data are then used for ML data analytics through the chosen machine learning module. The design and functionality of each module are detailed in the sequel.

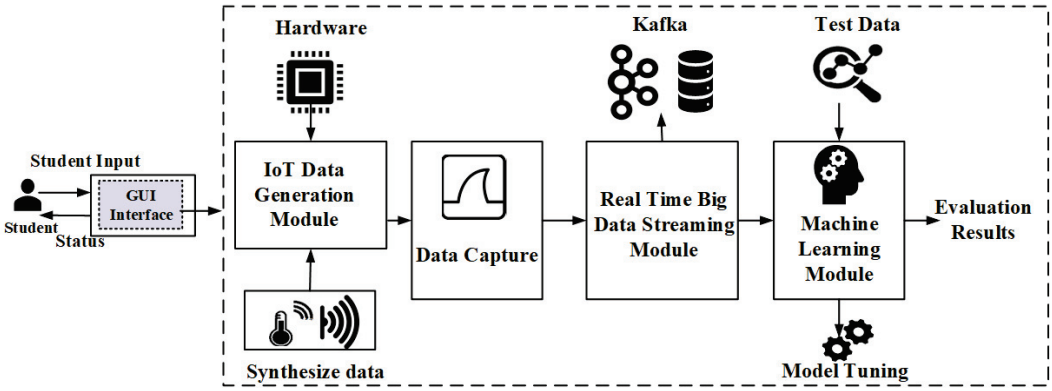


Figure 4. Design approach.

4.2.1. IoT Data Generation Module

This module within the tool emulates IoT devices or interacts with real ones. The output of this module is MQTT network packets that contain realistic and correlated sensor readings. Initially, students set up essential parameters in this module, enabling the creation of IoT data from one of two options, namely, real hardware devices or emulated (synthesized) sensors of different types (e.g., temperature, humidity, air pressure, etc.) with realistic and correlated readings based on widely available datasets. In this module, student users can define the IoT network devices of their designed use cases by adding devices and attaching sensors to them. Currently, the tool supports only the MQTT protocol for transferring the sensed data.

Clients and MQTT Broker

The IoT devices take on the role of MQTT clients. These clients are responsible for sending and receiving data within the IoT network. The key component in MQTT communication is the MQTT broker, which serves as an intermediary, ensuring seamless communication between MQTT clients. It plays a pivotal role in routing messages from publishers to subscribers, facilitating efficient data exchange.

IoT Network Configuration

Students can create use cases, add and configure devices, and specify the number of devices and sensors per device. These configurations collectively form the IoT network configuration.

Data Generation Profile

Students can set up the data generation profile for periodic or random data generation and define periods or time intervals for synthesized data generation. This configuration effectively emulates real-time sensor data. For generating data from real hardware, the hardware sensor nodes should be configured to follow the same profiles.

Data Generation

For the hardware option of the tool, a physical setup can be used to generate real-time IoT data, an example of which is shown in the tool evaluation Section 4.4. On the other hand, for the synthesizing option, we developed an IoT data generator that utilizes pre-trained machine learning models. These pre-trained models are capable of generating correlated sensor readings. We have included temperature, humidity, motion, and light sensors in the current data generation module. Moreover, the synthesizer is customizable to generate other sensor readings whenever required. It uses a Random Forest-based pre-trained model to generate correlated data and the MQTT protocol for data publishing. The pre-trained ML model is developed using a combination of real-time open-source datasets [34] and hardware-synthesized data. It operates by loading the pre-trained model and retrieving configuration settings provided by the students, which include the MQTT

topics, IP addresses for the clients and broker, and the data publishing frequency. Based on the network configurations, the pre-trained model predicts correlated sensor readings and publishes them to the specified MQTT topics. This approach allows students to emulate IoT data generation scenarios or use real hardware to generate them, providing a comprehensive learning experience.

4.2.2. Real-Time Streaming Module

The real-time streaming module is responsible for capturing the IoT data generated by the previous module and streaming it to a Kafka-based big data platform. The module operates in the following stages:

Data Capture

In this stage, PyShark is used as a Python module designed to serve as a bridge between Tshark and Python, enabling parsing network packets in Python by utilizing Wireshark dissectors [33].

Big Data Platform Integration

Captured data are seamlessly transferred to a real-time streaming big data platform (Kafka) in this stage. The Kafka processes involve establishing a cluster, creating meaningful topics to categorize and classify data logically, and then aligning with the IoT data generation module by naming topics after device names and sensor types. Integration with Kafka is accomplished through well-defined producer and consumer configurations [6].

The Kafka producer is set up with settings such as idempotence, ensuring that messages are delivered exactly once without duplicates. The compression type is set to "lz4", reducing data size to increase transmission efficiency. Other settings include batch size, defining the amount of data sent in one go to optimize throughput, retries to handle message resend attempts in case of failures, and delivery timeouts to ensure messages are sent on time.

For the consumer, configurations include settings to read data from Kafka topics and perform pre-processing steps, such as removing redundancy and converting byte data into a readable format before forwarding it to the appropriate Kafka topics. Additional settings involve isolation levels, offset management, and auto-commit configurations. Isolation levels ensure data integrity by specifying how transactions are handled. Offset management tracks the reading progress within each topic, enabling the consumer to resume from the correct position. Auto-commit settings streamline read position updates by automatically recording the read progress. These setups ensure the correct formatting and readiness for further use. In essence, this integration involves the creation of model topics within Kafka to manage and organize incoming data efficiently. This real-time streaming process should be continuous and uninterrupted.

4.2.3. Machine Learning Module

The last module involves machine learning analysis of the real-time streamed data, where students can either load a dataset already created or generate the dataset in real time. The ML module is capable of handling both regression and classification tasks. The main steps involved are described below.

Data Preprocessing and Dataset Splitting:

After loading the dataset, the student has the option to perform data preprocessing. This step involves cleaning and transforming the data to ensure it is suitable for machine learning. Preprocessing includes tasks such as handling missing values and outliers, scaling features, and encoding categorical variables if necessary.

Since students can generate datasets in real time, they have the option to specify the number of data records they want. There is no need for explicit programming from the students. Once the data are preprocessed, the student is prompted to split the dataset into training and testing subsets. This step is crucial for assessing the ML model's performance. Typically, a portion of the data is allocated for training, while the remaining portion is

reserved for testing. This separation allows the model to learn patterns from the training data and evaluate its generalization on unseen data.

Regression and Classification:

The current machine learning analysis available in the IoT-Edu-ML-Stream tool encompasses both regression and classification tasks, depending on the nature of the use case. For regression, the goal is to predict continuous output variables based on input features, such as predicting temperature from sensor readings. The tool uses typical evaluation metrics for regression, such as mean squared error (MSE) and R-squared, which assess prediction accuracy. The tool can be extended for classification tasks to include models that assign data points to predefined categories. Metrics such as accuracy, precision, recall, and F1 score can be used to evaluate classification performance, helping users understand the model's effectiveness in categorizing data.

Model Fitting and Optimization

Once the machine learning model is chosen and configured, it is fitted to the training data. Model fitting involves adjusting the model's parameters to find the best-fitting model for the given data. The iterative training process explores various combinations of parameters and configurations to optimize model performance.

The tool currently contains four model options available: a linear model (linear regression or logistic regression, depending on the task), a tree-based model (Random Forest), an ensemble method (Gradient Boosting), and a neural network model. This setup is customizable, allowing the future addition of any number of supervised algorithms. The trained machine learning models are then saved for future predictions within the same tool session. These trained models can make predictions based on the input data obtained from the model topic, which could include real-time IoT sensor readings. Such predictions are crucial for identifying patterns and correlations among various inputs, such as temperature and humidity. Students can gain insights into how different factors influence the output variables by analyzing the model's predictions, helping them understand and better tune the IoT data generation module parameters.

4.2.4. Operation Procedure

The flowchart in Figure 5 summarizes the flow through the tool operation procedure. As the figure shows, when initiating the IoT-Edu-ML-Stream tool, students will have two choices: start a new project or load an existing one. Next, they can either generate data from hardware devices or synthesize it. In either case, they should configure the IoT network by adding devices and selecting the required sensors. Once the devices and the MQTT broker are configured, they will create big data topics (according to the devices and sensors they added). This step will also be reached if they choose to load an existing project at the beginning. After that, they start the IoT data generation. Following data generation, the generated IoT data are captured using a data capturing module employing PyShark. Students can view the big data topics and their contents as needed.

Subsequently, the ML module comes into play. Students are supposed to begin by configuring the ML model, starting with the option to select the desired time limit for loading the data being streamed. Since the data are streamed in real time, setting a longer time limit for data generation guarantees the availability of a sufficient amount of data in the corresponding Kafka topics. After loading the dataset, student users can specify the ML model inputs from the big data platform topics. If necessary, data preprocessing can be performed at this stage. Since the streamed data are in real time, there might be missing values in the dataset, which should be treated before fitting the model. Here, the students save the dataset and are provided with options to view it. If they find the need for preprocessing, they can do it; otherwise, they can skip this step. According to the ML model selected, students will have to adjust the parameters for the model in the ML tuning step. For each selected model, the parameters to be tuned will be different.

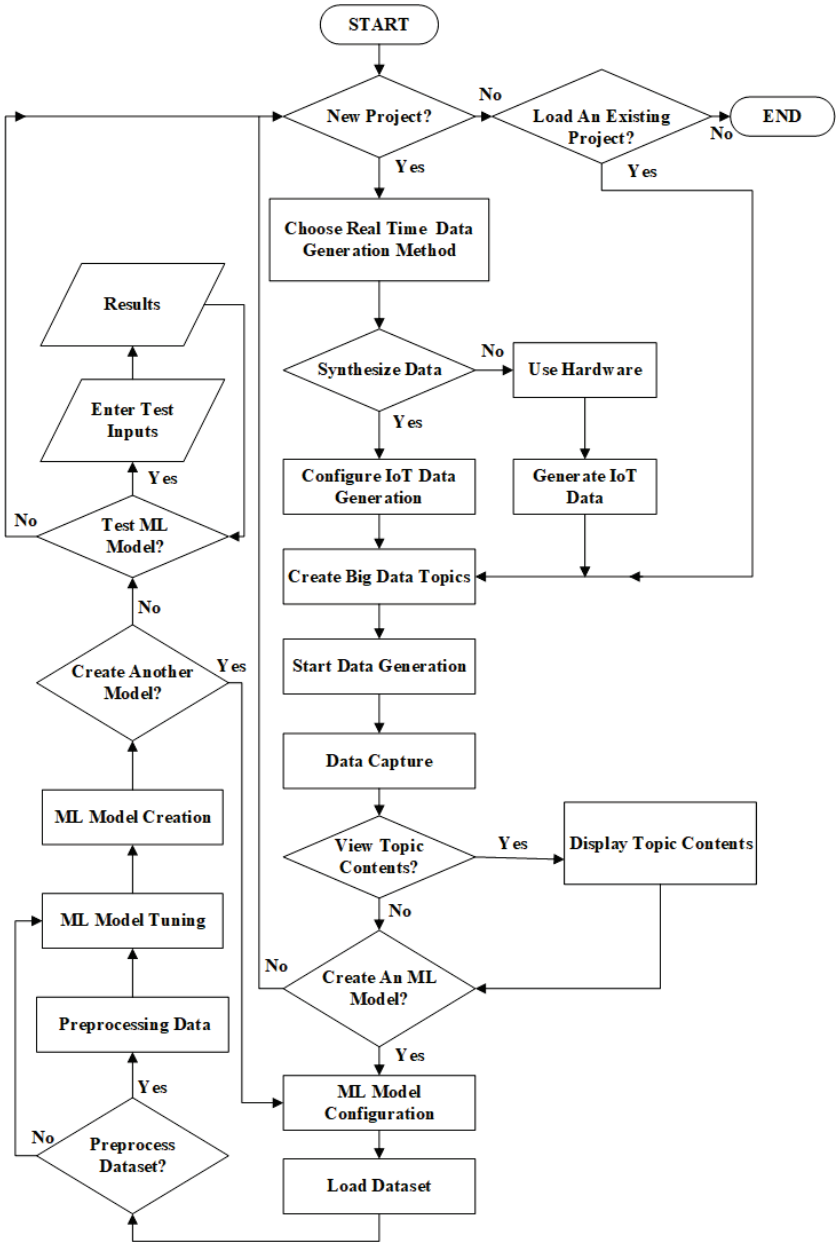


Figure 5. IoT-Edu-ML-Stream flowchart.

Once an ML model is created, students are prompted to decide whether to create another one. If they choose to do so, they will be directed to the ML model creation step again. Otherwise, they can proceed to test the ML model by adding user inputs and viewing the results. If they choose not to test the model, they will be redirected to the new project creation menu, where they have the option to exit the tool.

4.2.5. The Tool User Interface

The IoT-Edu-ML-Stream GUI tool offers a user-friendly interface for setting up IoT data generation, real-time streaming, and ML analysis of the streamed data. The screens are designed with help text added to guide students through the concepts and ensure smooth navigation. The process begins with the students launching the GUI of the tool and selecting whether to generate real-time data from hardware or synthesize data in real time through emulation, as shown in Figure 6. This initial choice determines the data source for their IoT network configuration.

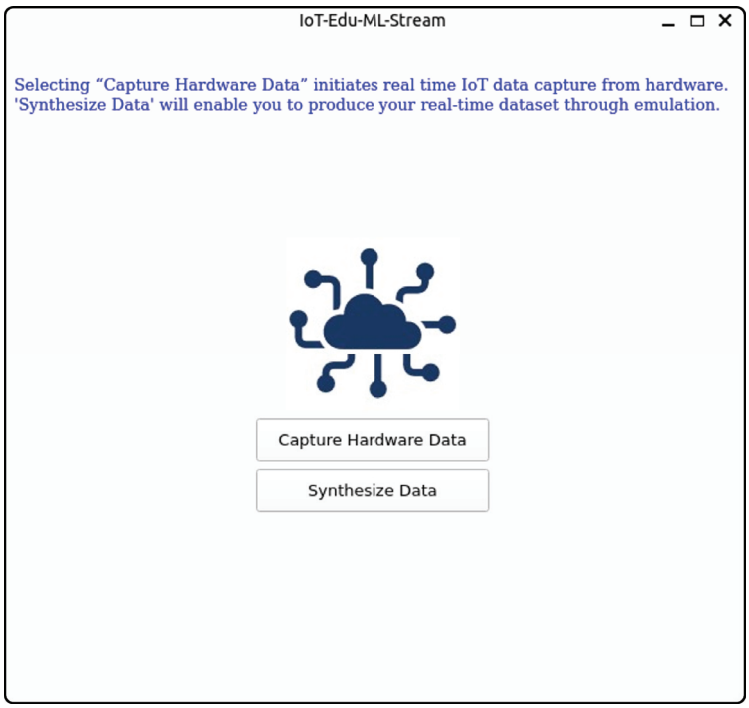


Figure 6. Screen to select data generation method.

After selecting the data generation method, students create their IoT network by selecting devices and adding sensors, as revealed in the screenshot of Figure 7.

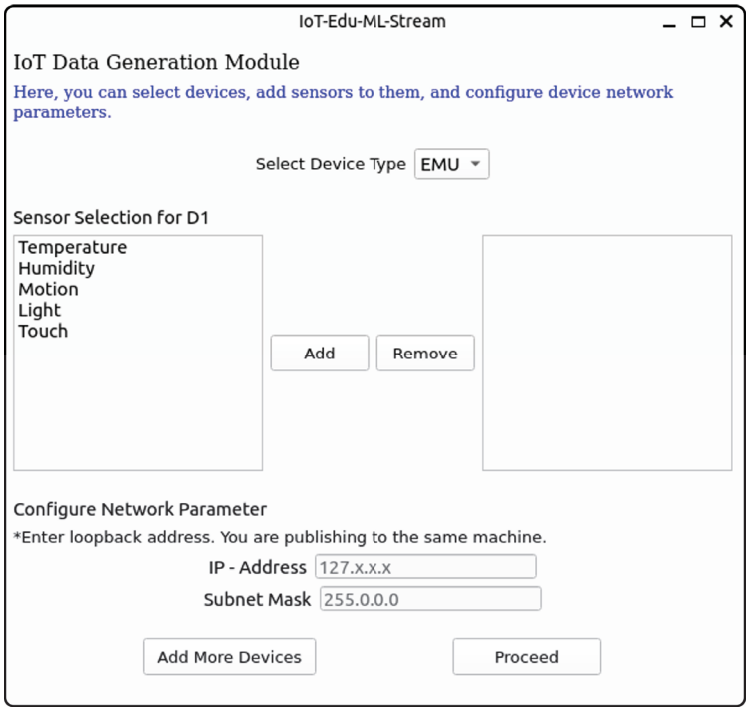


Figure 7. Screen to create IoT network.

Students will then configure the network, which consists of the IoT devices (acting as MQTT clients) and the MQTT broker, by entering the IP addresses and network masks of all the network entities. As shown in Figure 8, loopback addresses are used for both the broker and the client as they communicate within the same computer in the synthesized data scenario. The screen in Figure 9 provides a summary of the selected devices

and IoT network configuration. Clicking "Proceed" initiates data emulation, enabling communication between the MQTT broker and the defined clients.

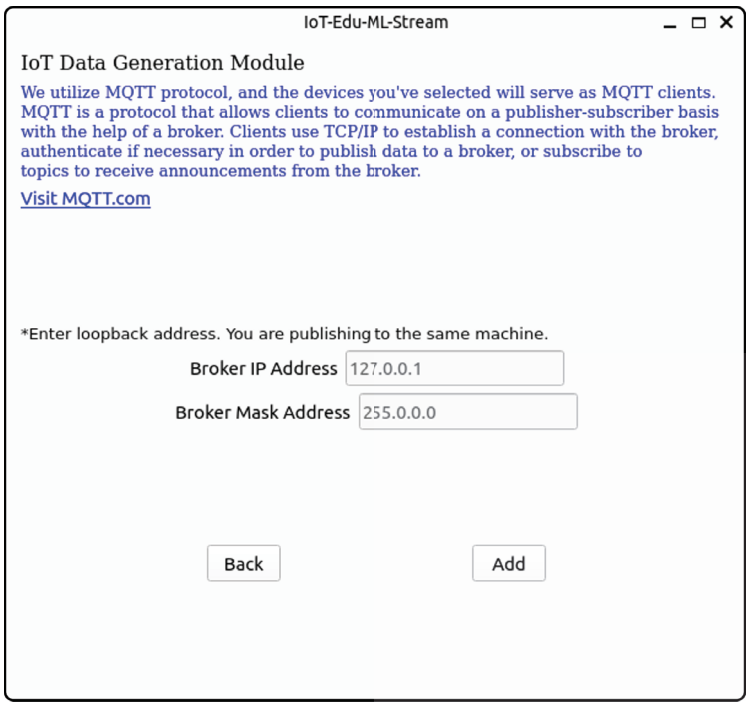


Figure 8. Screen for IoT network configuration.

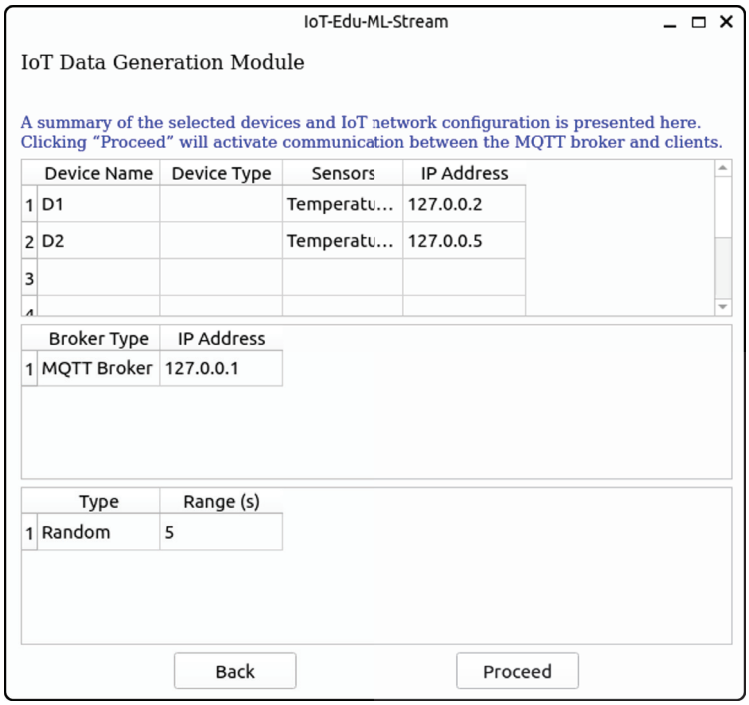


Figure 9. Screen showing network configuration summary.

Data streamed from the IoT network are stored in the big data platform (Kafka) under specific topics. The screenshot in Figure 10 shows the topic names to be created and includes an option to start the data streaming by creating the big data topics. The nomenclature used here is as follows: for "Device 1" with the sensor "Temperature", the topic name will be written as "D1-Temperature".

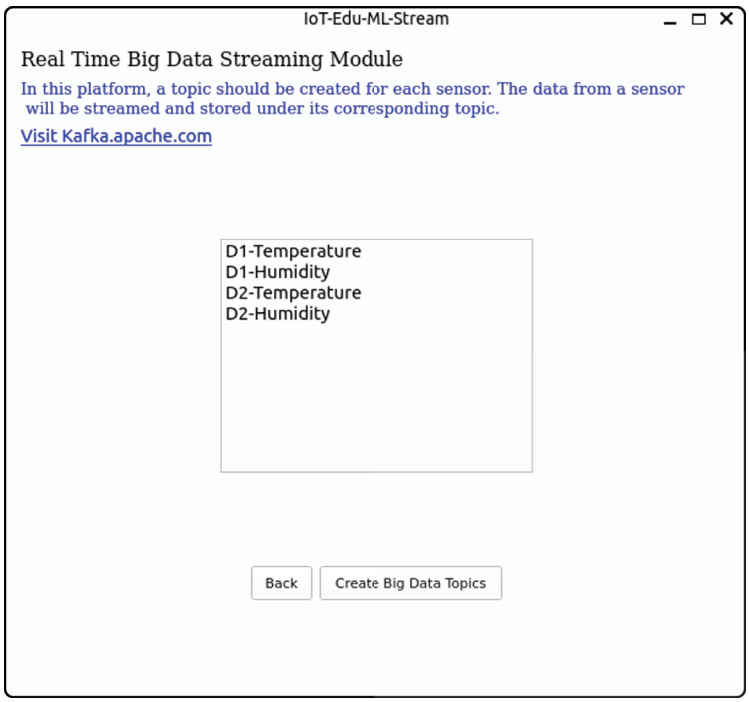


Figure 10. Screen to create big data topics.

Students begin building their machine learning models by defining the input data. They create datasets for ML tasks from the data stored in the big data platform. The process of selecting the input features for the model is shown in Figure 11.

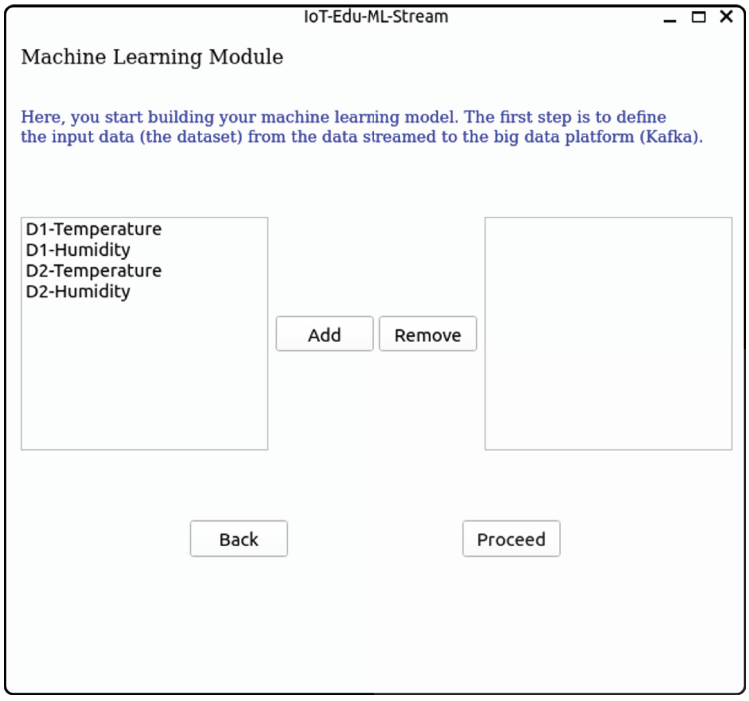


Figure 11. Screen for choosing ML input data.

Student users are then allowed to save the prepared dataset in a comma-separated value (CSV) format, as shown in the exhibit of Figure 12 by specifying the number of data records needed in the dataset file.

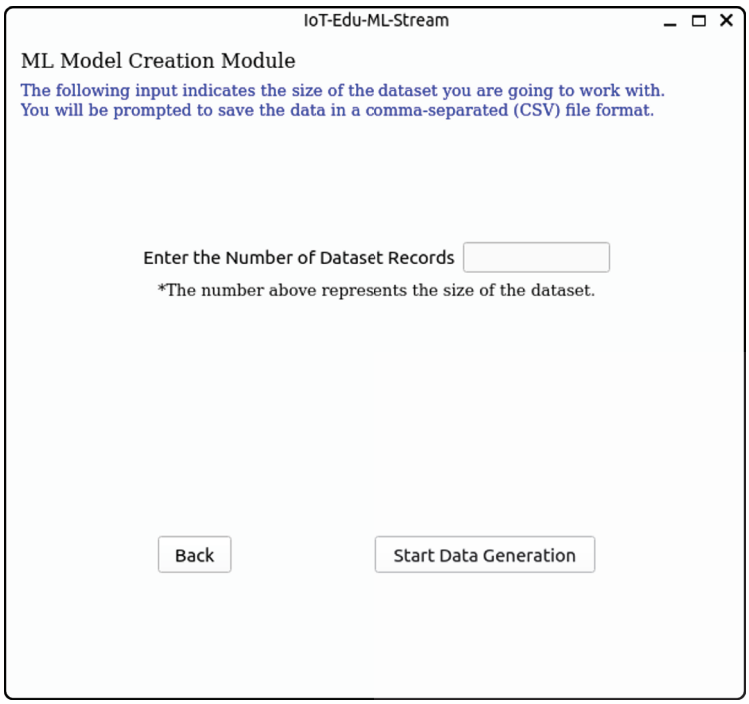


Figure 12. Option to save the dataset in CSV format.

This CSV file can then be uploaded for ML tasks. After uploading the CSV file, students can select a suitable machine learning algorithm to build their model, as revealed in Figure 13. Each algorithm is accompanied by a brief description to aid in selection, guiding students in choosing the appropriate technique for their task.

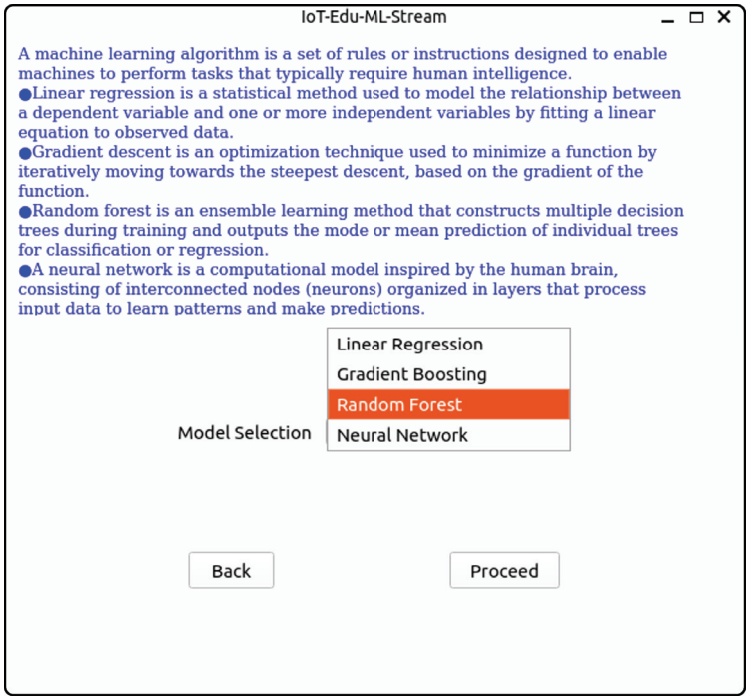


Figure 13. Screen to choose ML algorithm.

The screenshot in Figure 14 shows how the students can configure the parameters of their chosen machine learning algorithm. For instance, if the students select the linear regression (LR) model, they can tune parameters such as whether to fit the intercept and normalize the data. If they choose the neural network (NN) model, they will adjust

parameters such as the number of layers, neurons per layer, and the optimization algorithm (e.g., Adam, SGD, or RMSprop). In this instance, they are tuning the parameters of the Random Forest model, where they can set the number of estimators and the maximum depth of the trees. Once the parameters are set, clicking "Train" initiates the model training process, allowing the model to learn from the training data based on the specified configurations.

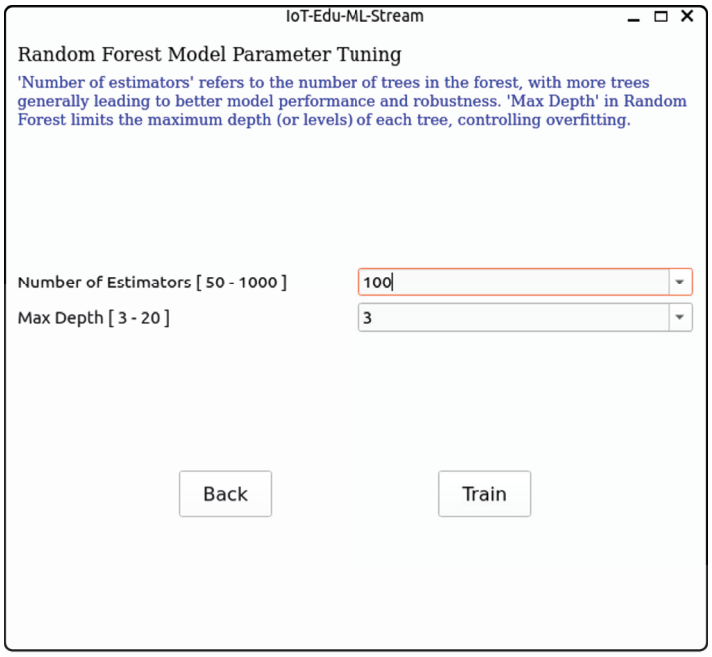


Figure 14. Configuration of the parameters.

The tool later displays the evaluation metrics for the trained model, as shown in Figure 15. After this step, students can choose to test the created model, train the model with a different algorithm, or use a different dataset for training the model. This screen allows them to evaluate and refine their machine learning models.

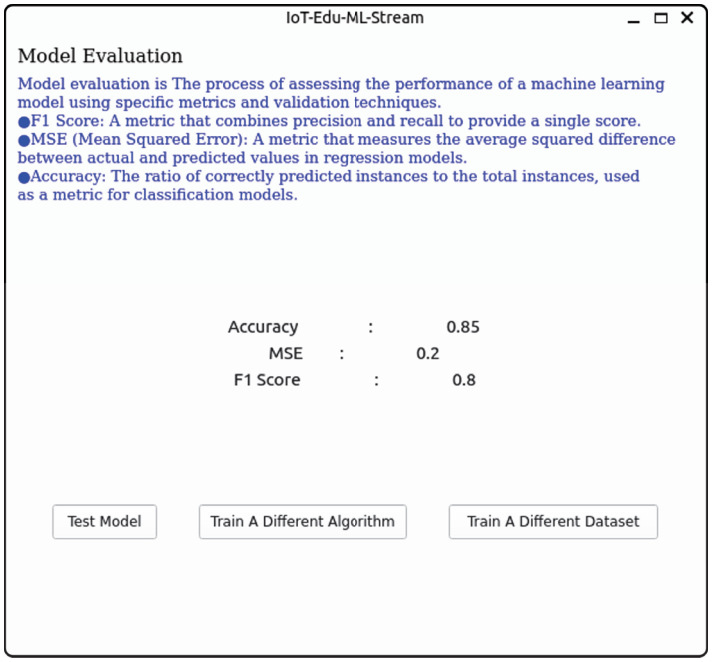


Figure 15. Model metrics and options.

4.3. Tool Integration in Electrical Engineering Curriculum

The IoT-Edu-ML-Stream tool can benefit electrical engineering students if introduced to support a general course about IoT, ML, and big data. It can also be introduced as a component in a senior-level specialized course in the electrical engineering curriculum addressing modern application use cases that require IoT sensing and communication with intelligent big data analytics, such as courses about industrial automation and smart power grids.

This is because a specialized course or course component for electrical engineering students that integrates ML, IoT, and big data should provide an in-depth understanding of the fundamental concepts of AI and machine learning. Students are supposed to gain knowledge and skills related to various machine learning techniques, such as supervised learning. This understanding is essential for designing and implementing the use cases addressed by the course subject. Moreover, students shall also learn about the basic principles of IoT and how it relates to AI. This requires covering topics related to data acquisition, sensor/device networking, and data processing, which are essential for building intelligent systems interacting with the physical world. Furthermore, the students need to learn how to manage and analyze large amounts of data, which is essential for making decisions based on data-driven insights.

By using the proposed tool, students are supposed to gain the skills and knowledge to design and implement ML, IoT, and big data use cases. They will be able to analyze data using various ML techniques and evaluate the performance of these techniques.

In the following, we introduce proposed course learning outcomes for integrating the tool into an electric engineering curriculum and outline the benefits of this integration and some limitations.

4.3.1. Proposed Learning Outcomes for Curriculum Integration

Following an outcome-based learning approach, some suggested course learning outcomes (CLOs) that can be targeted while using the IoT-Edu-ML-Stream tool in a specialized course or course components are introduced in the following.

After using the tool and by the end of studying the required material, students should be able to:

- CLO1: Outline the fundamental concepts of AI and machine learning.
- CLO2: Identify the basic principles of IoT and how they relate to AI, including data acquisition, processing, and analytics.
- CLO3: Recognize big data platforms and related big data technologies.
- CLO4: Design an IoT-based application use case involving intelligent decision-making via data analytics.
- CLO5: Compare the performance of different ML techniques.

Since many engineering programs all over the world are accredited by the Engineering Accreditation Committee of the Accreditation Board for Engineering and Technology (ABET), we introduce here some student outcomes or program learning outcomes (PLOs), following the ABET guidelines (with the same numbering of the outcomes) [35]. These PLOs (seven) have been proposed by ABET as general student learning outcomes that can be targeted by any engineering program. These PLOs define the specific knowledge, skills, and behaviors the students are expected to possess. These outcomes are indicative of the progression students make throughout the program, reflecting the acquisition of essential skills, knowledge, and behaviors. These PLOs can be addressed by including the proposed tool in an electrical engineering curriculum. Table 1 outlines the mapping between the suggested CLOs and four ABET engineering PLOs. This mapping shows some of the PLOs (defined by the ABET number and description in the first two columns) that can be covered by the proposed CLOs and the most relevant CLOs (last column) to each. This implies that the attainment of relevant CLOs refers to the attainment of their corresponding PLO. The table shows that the suggested CLOs can map to more than half of the program outcomes

set by ABET guidelines (seven) [35], indicating the proposed tool's effective contribution to electrical engineering programs.

Table 1. The ABET program learning outcomes [35] addressed by the usage of IoT-Edu-ML-Stream mapped to the suggested CLOs.

PLOs	Description	CLO Mapping
First	"*An ability to identify, formulate, and solve complex engineering problems by applying principles of engineering, science, and mathematics*"	CLO-2,4,5
Second	"*An ability to apply engineering design to produce solutions that meet specified needs with consideration of public health, safety, and welfare, as well as global, cultural, social, environmental, and economic factors*"	CLO-4,5
Sixth	"*An ability to develop and conduct appropriate experimentation, analyze and interpret data, and use engineering judgment to draw conclusions*"	CLO-4,5
Seventh	"*An ability to acquire and apply new knowledge as needed, using appropriate learning strategies*"	CLO-1,2,3

4.3.2. Integration Benefits

Integrating the IoT-Edu-ML-Stream platform into the curriculum provides several key benefits. Students gain a comprehensive understanding of the combined effects of IoT, big data, and machine learning. They learn essential IoT concepts, including network configuration, client–broker communication, and IP address setup, through both emulation and hardware options. The tool also exposes students to real-time big data platforms, including data streaming and topic creation, offering hands-on experience with real-time data processing. In the ML module, students generate datasets, learn the importance of data in machine learning, and practice splitting datasets into input features and target variables. They engage in training models, understand the differences between regression and classification, and explore preprocessing and parameter tuning. The tool guides students through evaluating and testing models, providing a complete ML process experience. Additionally, the tool's user interface offers instructions, helpful hints, and links to further resources, enhancing student engagement and encouraging deeper learning.

4.3.3. Limitations

The current version of the tool is designed to handle only the MQTT protocol, which may limit its conformance with other IoT communication protocols. The tool setup is a one-time process but requires expertise in the Linux operating system to ensure the needed software packages are properly installed. The hardware devices, if required to generate data, should be suitably pre-configured to ensure data communication is working as it should. Additionally, the machine learning module within the tool is currently limited to handling only supervised learning techniques, restricting the scope of ML applications that can be explored using the platform.

4.4. Tool Evaluation

To assess the effectiveness of the IoT-Edu-ML-Stream platform, we conducted an evaluation survey among senior undergraduate electrical engineering students, formatted as a five-point Likert questionnaire. The survey consists of six questions categorized into three sections. In the first section of the survey, respondents rated their prior knowledge related to IoT, ML, and big data. The questions in this section are meant to analyze the participant's understanding of IoT with big data and machine learning concepts after using the tool to evaluate the knowledge gained compared to prior knowledge. The second section focuses on the effectiveness of designing application use cases with the tool. In the third section, we evaluated the feedback on how well the tool introduced real-world scenarios. By organizing the survey this way, we aimed to cover all aspects comprehensively. The survey questions are listed as follows:

- Q1—Rate your prior knowledge in the integrated domains of IoT (Internet of Things), big data, and ML (Machine Learning).

- Q2—Rate your knowledge gained about IoT concepts after using the tool.
- Q3—Rate to what extent you grasped the basic concepts of machine learning using the tool.
- Q4—Rate the level of difficulty in designing use cases using the tool.
- Q5—Rate how much the tool enhanced your practical skills for real-world applications.
- Q6—Rate how user-friendly the tool's interface was.

4.4.1. Participants and Data Collection

The survey involved the participation of 29 senior undergraduate students with different background levels in IoT and ML. The students were exposed to the proposed platform during the activity mentioned in Section 4.4.2. After they finished the activity, the provided survey responses were collected from the survey participants, who were informed about the purpose of the survey and were assured of the confidentiality of their responses. The survey was distributed using the secure and user-friendly platform SurveyMonkey, ensuring that participants could easily provide their responses anonymously. Each participant was granted access to the proposed platform for a specific period.

4.4.2. Tool Evaluation Activity

During this period, the participating students actively interacted with the tool's intuitive graphical user interface (GUI) to gain practical, real-world experience in IoT, big data analytics, and machine learning (ML). They were asked to design a practical use case for controlling the temperature adjustment of a room air conditioner (AC) without human intervention using ML. Figure 16 effectively illustrates the hardware setup and software packages required to implement the use case efficiently. The central component is a Linux box running Ubuntu 20.04, which operates all the essential software packages, including Python 3.10.12, Kafka 3.6.0, Pyshark 0.6, and the MQTT protocol implemented with Paho-MQTT (Mosquitto version 2.0.11). A desktop display screen was connected to the Linux box to display the IoT-Edu-ML-Stream interface. The Linux box is connected to a WiFi router, which serves as an access point for the entire system. Device 1 is a Raspberry Pi connected to a DHT-11 Humidity sensor, whereas Device 2 is an ESP-32 connected to a DS18B20 digital temperature sensor. Both devices are connected wirelessly to the router for network communication.

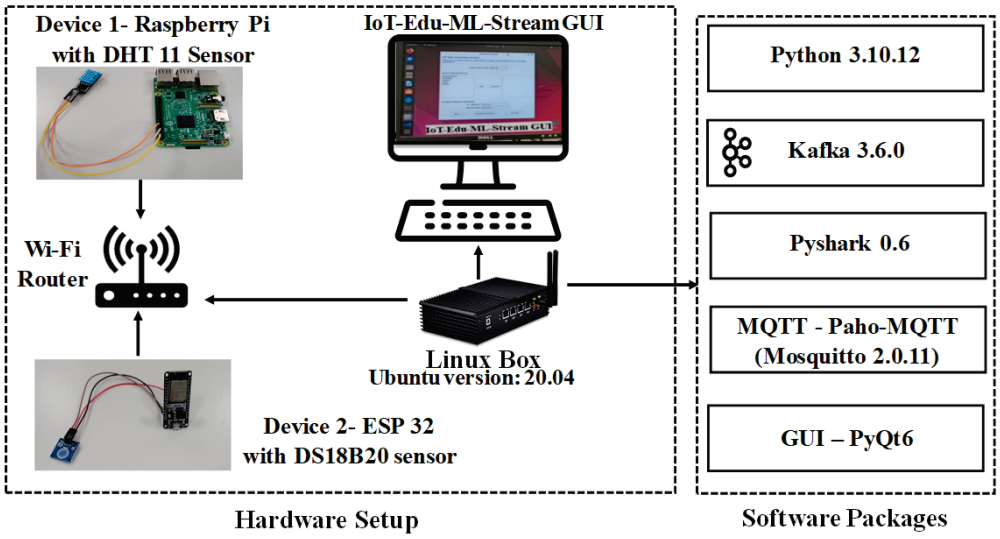


Figure 16. Block diagram outlining the required hardware and software setup for the case study.

Students were provided with the pre-configured hardware devices shown in Figure 16 to understand the generation of IoT data through hardware. They generated data by setting all the necessary configurations with the tool, thus enabling the streaming of IoT

data from humidity and temperature sensors indoors and outdoors. Figure 16 showcases the indoor hardware setup. The students were also asked to generate more data using the data synthesizer option to have a dataset of a sufficient number of data records. They emulated inside and outside devices equipped with humidity and temperature sensors, configured them using the GUI tool, synthesized real-time data, and streamed it through the big data platform.

In the ML phase, the students entered the number of data points they wanted in their dataset, captured real-time data, and saved it in a dataset. Students then labeled and uploaded the dataset and performed preprocessing. They selected the ML model they wanted to train, including linear regression, Gradient Boosting, Random Forest, and neural network models. After selecting the model, the students tuned the parameters for the chosen model and followed the explanations on the screen to proceed with training the ML model. Since they carried out the classification, they evaluated the trained model by calculating metrics such as accuracy and F1 score. The students built all the models, compared the accuracy of each, and increased the data records to see the improvement in accuracy. With the best model, they made decisions on the intelligent automated adjustment of the room AC. This hands-on activity allowed them to experience the entire process, from data generation to the practical application of ML algorithms for decision-making.

The students were asked to answer the survey questions after finishing the activity. Note that Q1, Q2, Q3, and Q5 are used to measure the student's perception of prior and gained knowledge and practical skills after completing the activity. Suppose this activity is a part of some course in an electrical engineering curriculum. In that case, the survey results can also be used as a measure of the student's perception regarding the attainment of the proposed CLOs using Table 2 (i.e., the question that receives a high average rating indicates its related CLO(s) are attained from the student's perspective). Attaining the CLOs is an essential measure of the success of integrating the proposed platform into an electrical engineering curriculum.

Table 2. Knowledge and skills survey questions mapped to the suggested CLOs.

Survey Question	Related CLO(s)
Q1	CLO-1
Q2	CLO-2
Q3	CLO-2, CLO-5
Q5	CLO-3, CLO-4

It is worth noting that the proposed platform can be applied to any scenario involving IoT sensor data. Beyond temperature control, it can be used in various use cases. For instance, in computer network security applications, students could configure IoT devices or generate emulated data for different applications. These devices can stream real-time data by properly configuring the IoT networks and training the ML models to detect unusual patterns in the reported data in order to characterize data integrity security threats. In another example, within the context of electric grids, the tool can synthesize data obtained from various sensors installed on vital equipment, such as transformers, or receive this data live via a configured IoT network from a real hardware version of the equipment installed in a university laboratory setup. These data can then be processed to build ML models to predict and identify faults in this transformer, facilitating predictive maintenance of electric grids. These examples highlight how the tool offers practical, hands-on experience across various real-world scenarios of significant importance to electrical engineering students' future careers.

5. Tool Evaluation Results and Discussion

This section presents the survey results and provides a discussion of the findings.

5.1. Tool Evaluation Results

The analysis of the survey results is presented in the bar charts below. The survey started by addressing students' prior knowledge related to IoT, big data fundamentals, and ML. The results indicated a diverse range of responses, as shown in Figure 17. The mean rating of the responses is 3.62, and the standard deviation (SD) is 1.25, indicating the variability in students' prior knowledge. About 62% of the participating students had a reasonably good understanding (rated 5 and 4). Another substantial portion perceived their knowledge to be average, while a smaller portion of about 10% indicated limited knowledge.

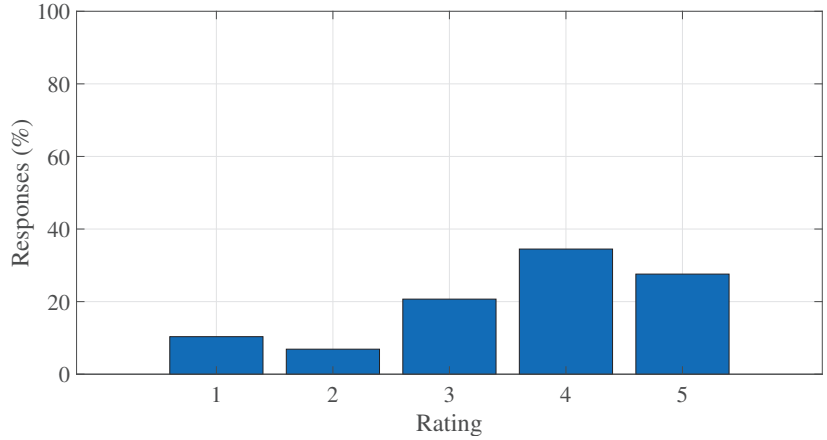


Figure 17. Q1 survey response.

Regarding developing their IoT knowledge after using the tool, Figure 18 shows that approximately 65.5% of students reported significant improvement (rated 5) in the fundamentals of these topics. In comparison, around 35% indicated very good knowledge gain. No student gave a rating below 4. The high mean rating of 4.66 and low standard deviation of 0.425 highlight the tool's effectiveness in enhancing IoT knowledge.

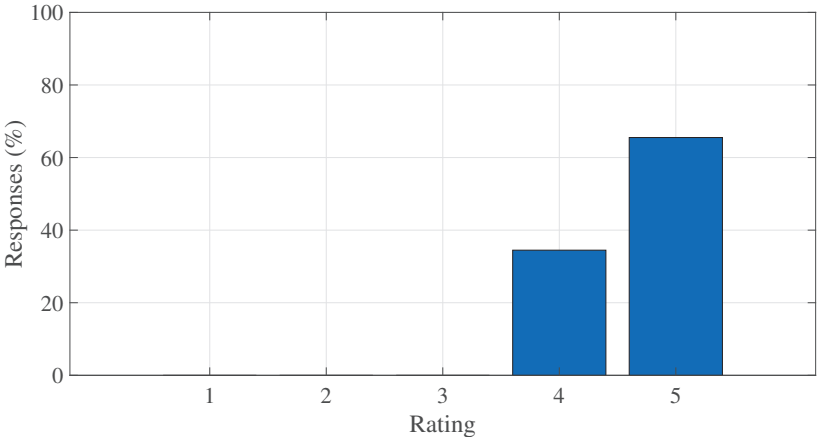


Figure 18. Q2 survey response.

When asked the participants how they grasped ML concepts conveyed by the tool, about 93.1% of students understood the concepts clearly (rated 5 and 4), whereas 7% found it moderately clear (rated 3), as shown in Figure 19. This translates to a mean rating of 4.59 and a standard deviation of 0.588, indicating that, despite the good grasping of the ML concepts, there were some minor variations in students' responses.

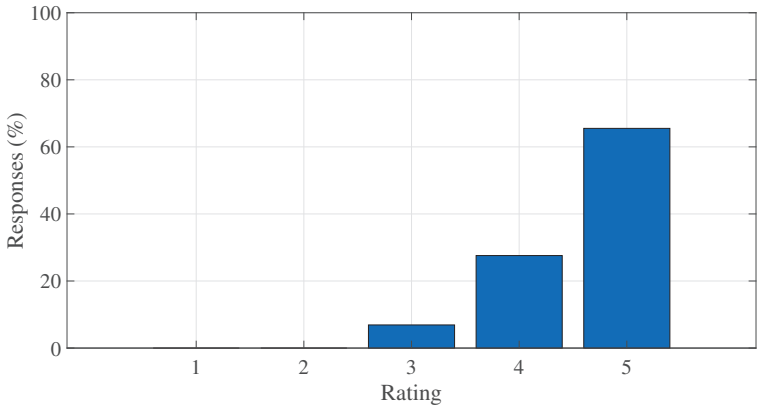


Figure 19. Q3 survey response.

The survey results indicate that the majority of students found it very easy or easy (rated 5 or 4) to design the use cases with the tool, as presented in Figure 20. Only very few of them found it neutral. This gives a mean of 4.76 and a standard deviation of 0.489, suggesting that designing use cases with the tool is manageable and easy.

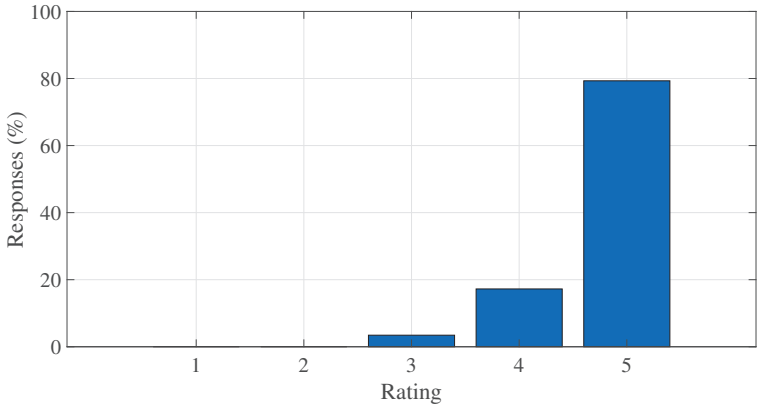


Figure 20. Q4 survey response.

Figure 21 provides insights into the tool's impact on participants' practical skills development for real-world applications. Approximately 62% expressed very high levels of satisfaction, with 34.5% rated 4, and only very few students rated 3. The statistical measures also indicate significant strides toward practical proficiency, with a mean of 4.59 and a standard deviation of 0.526.

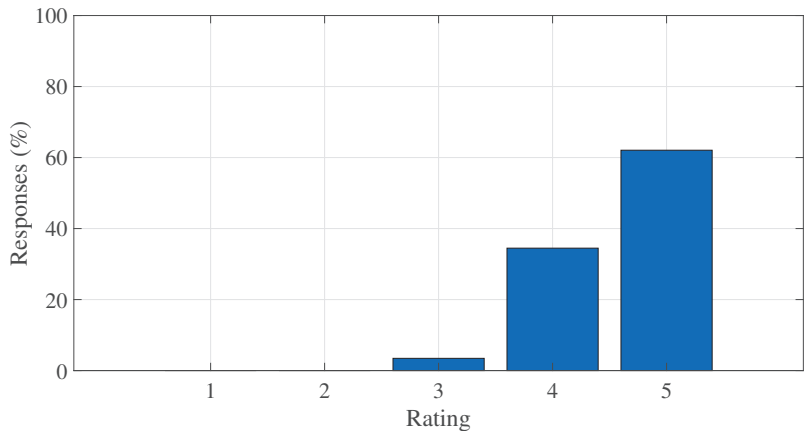


Figure 21. Q5 survey response.

A key aspect of the survey was to ask the students about the tool's user interface. All the students found it either extremely friendly or very friendly (rated 5 and 4), resulting in a mean of 4.62 and a standard deviation of 0.492. The survey results are shown in Figure 22.

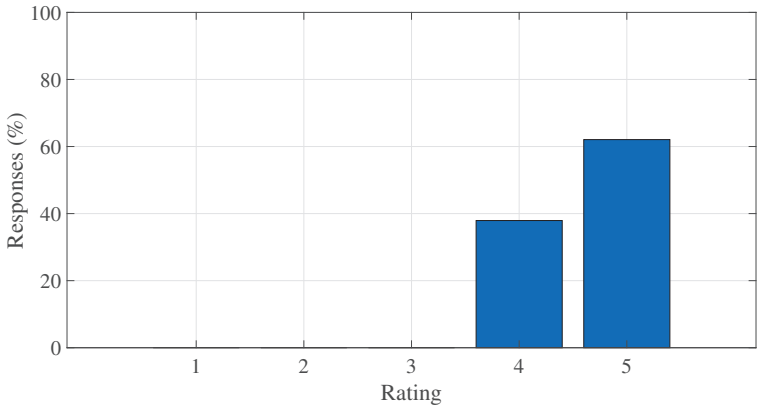


Figure 22. Q6 survey response.

5.2. Discussion

The survey findings underscore the positive impact of utilizing the IoT-Edu-ML-Stream platform to instruct electrical engineering students on the fundamentals of integrating IoT, big data, and ML.

The tool aids in grasping IoT concepts, as indicated by the feedback about the knowledge gained, as reported in Q2 results with an average rating of (4.66 ± 0.155). Here, we used a 95% confidence interval to provide a meaningful analysis of the survey results of each question. This result indicates that the proposed tool enriched the students' knowledge of IoT concepts. This is despite the 62% expressed that they have good knowledge of IoT and fundamentals of ML, as indicated by an average rating of (3.62 ± 0.45) for Q1 results. This enhancement likely stems from the tool's amalgamation of hardware devices and IoT network configurations within its user interface.

However, there are areas for improvement highlighted by the responses to Q3, with an average rating of (4.59 ± 0.214). The current version of the tool's limited support for ML algorithms may impede its effectiveness in comprehensively teaching ML principles. Additionally, the necessity for more time for participants to explore various use cases and scenarios suggests that further refinement and expansion of the tool's capabilities may be warranted to optimize its effectiveness.

Q4 serves the purpose of evaluating the tool user interface and the overall experience of designing a complete use case that requires IoT communication with big data analytics. The response to Q4 is reflected in an average rating of (4.76 ± 0.178), which affirms the impact of the hands-on approach the tool follows in conveying the principles of designing a complete application use case from configuring the IoT-based network to the intelligent analytics of the collected data.

Furthermore, the majority of respondents highly rated the tool's efficacy in illustrating the practical interplay between IoT, big data, and ML, suggesting that it bolsters and enhances their existing knowledge in these domains, as indicated by an average rating of (4.59 ± 0.1915) for Q5 results. This observation emphasizes the significance of hands-on, experiential learning in reinforcing theoretical concepts, particularly in intricate and interdisciplinary fields like IoT and ML. Again, the need to spend more time with the tool in working out more use cases affected the survey response for this question, similar to Q3.

The friendliness of the tool user interface is measured by the response to Q6. The result (4.62 ± 0.179) reflects its intuitiveness and usability. This is crucial to reducing the students' dependency on the instructor, giving them self-confidence in learning the IoT, big data, and ML integration fundamentals.

In essence, the survey outcomes affirm that interactive and immersive learning environments, such as those facilitated by the proposed emulation platform, are indispensable for effectively conveying complex concepts in high-tech fields like IoT, big data, and ML. The survey results also imply that the proposed tool bridges the gap between academic theory and practical application by designing use cases addressing real-world scenarios. It furnishes students with invaluable skills and insights that qualify them for the practical job market. This is reflected in the high average rating of the survey questions, indicating the attainment of the proposed CLOs.

6. Conclusions and Future Directions

Although some AI-oriented courses are available in some curricula, they are for specific majors and mainly target AI techniques without addressing how these techniques can be integrated and applied in practice in essential disciplines such as electrical engineering. This makes electrical engineering students lack the knowledge and practice of IoT, machine learning (ML), and big data fundamentals and their relevance to the other courses they study in the mainstream curriculum.

Thus, this paper offers an emulation-based software platform for integrating IoT, ML, and big data into electrical engineering curricula. The proposed platform is a GUI-based tool called IoT-Edu-ML-Stream. By employing a friendly user interface, it allows students to design a wide range of use cases for data analytics by configuring an MQTT-based IoT network of sensor devices of various types, generating IoT data with a realistic pattern from these sensors, streaming it to a real-time big data platform, and finally performing intelligent ML-based analysis and decision-making.

Moreover, the paper introduces an example set of learning outcomes that can be addressed when using the tool in any course in an electrical engineering curriculum and their mapping to typical electrical engineering program outcomes. A survey was conducted among student participants to evaluate the tool. The results were analyzed, showing overall positive feedback regarding the tool usage and the knowledge gained.

In the future, the tool can be expanded to include other machine learning techniques (e.g., unsupervised learning), thereby increasing its applicability. Also, it can be modified to adopt other protocols for IoT communication. With these advancements, the IoT-Edu-ML-Stream tool should become an even more helpful teaching tool for IoT, big data, and ML education and experimentation.

Author Contributions: N.J. contributed to designing and developing the proposed tool. She also contributed to paper writing. N.Y. contributed to the development of the tool user interface. A.A. contributed to the architectural design of the proposed tool, the proposed lab tasks, and the course learning outcomes. He also participated in writing the paper. A.A.-D. contributed to fund acquisition and paper writing review. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the UAE University UPAR grant number 12N107.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, D.X.; Mendez, E.A. Artificial Intelligence in STEM Education: Interactive Hands-on Environment using Open Source Electronic Platforms. *Rev. Tecnol. En Marcha* **2023**, *36*, 45–52. [CrossRef]
2. Khan, M.A.; Khojah, M.; Vivek. Artificial intelligence and big data: The advent of new pedagogy in the adaptive e-learning system in the higher educational institutions of Saudi Arabia. *Educ. Res. Int.* **2022**, *2022*, 1263555. [CrossRef]
3. Liu, Y.; Chen, L.; Yao, Z. The application of artificial intelligence assistant to deep learning in teachers' teaching and students' learning processes. *Front. Psychol.* **2022**, *13*, 9175. [CrossRef] [PubMed]
4. Mannodi-Kanakkithodi, A.; McDannald, A.; Sun, S.; Desai, S.; Brown, K.A.; Kusne, A.G. A framework for materials informatics education through workshops. *MRS Bull.* **2023**, *48*, 560–569. [CrossRef]

5. Kantawong, K.; Chaichumpa, S.; Pravesjit, S.; Yaibuates, M. A Lightweight Framework for Retrieve IP Device Status Based on MQTT Protocol. In Proceedings of the 2022 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON), Chiang Rai, Thailand, 26–28 January 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 46–49. [CrossRef]
6. Foundation, A.S. Appache Kafka. Available online: https://kafka.apache.org/ (accessed on 4 December 2023).
7. Sahu, C.; Ayotte, B.; Banavar, M.K. Integrating machine learning concepts into undergraduate classes. In Proceedings of the 2021 IEEE Frontiers in Education Conference (FIE), Lincoln, NE, USA, 13–16 October 2021; pp. 1–5. [CrossRef]
8. Filsinger, M. Designing an Artificial Intelligence Course for Electrical and Computer Engineering Technology Students. In Proceedings of the 2006 Annual Conference and Exposition, Saint Louis, MO, USA, 4–7 June 2006; pp. 11–417. [CrossRef]
9. Montuori, L.; Alcazar-Ortega, M.; Vargas-Salgado, C.; Alfonso-Solar, D. Learning Analytics as Data-driven Decision Making in Higher Education: A Case Study. In Proceedings of the INNODOCT 2022. International Conference on Innovation, Documentation and Education, Valencia, Spain, 2–7 November 2022; pp. 83–89. [CrossRef]
10. Starcic, A.I. Human learning and learning analytics in the age of artificial intelligence. *Br. J. Educ. Technol.* **2019**, *50*, 2974–2976. [CrossRef]
11. Tan, J. Information analysis of advanced mathematics education-adaptive algorithm based on big data. *Math. Probl. Eng.* **2022**, *2022*, 7796681. [CrossRef]
12. Barenkamp, M.; Rebstadt, J.; Thomas, O. Applications of AI in classical software engineering. *AI Perspect.* **2020**, *2*, 1. [CrossRef]
13. Wei, N. Decreasing land use and increasing information infrastructure: Big data analytics driven integrated online learning framework in rural education. *Front. Environ. Sci.* **2022**, *10*, 1025646. [CrossRef]
14. Xu, L.; Zhang, J.; Ding, Y.; Sun, G.; Zhang, W.; Philbin, S.P.; Guo, B.H. Assessing the impact of digital education and the role of the big data analytics course to enhance the skills and employability of engineering students. *Front. Psychol.* **2022**, *13*, 974574. [CrossRef]
15. Tsai, S.C.; Chen, C.H.; Shiao, Y.T.; Ciou, J.S.; Wu, T.N. Precision education with statistical learning and deep learning: A case study in Taiwan. *Int. J. Educ. Technol. High. Educ.* **2020**, *17*, 12. [CrossRef]
16. Gil, Y. Teaching big data analytics skills with intelligent workflow systems. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, Arizona, 12–17 February 2016; Volume 30. [CrossRef]
17. Panagiotakis, S.; Karampidis, K.; Garefalakis, M.; Tsironi-Lamari, A.; Rallis, I.; Kamarianakis, Z.; Papadourakis, G. Remote Arduino Labs for Teaching Microcontrollers and Internet of Things Programming. In Proceedings of the 31st Annual Conference of the European Association for Education in Electrical and Information Engineering (EAEEIE), Coimbra, Portugal, 29 June–1 July 2022; pp. 1–6. [CrossRef]
18. Hercog, D.; Lerher, T.; Truntič, M.; Težak, O. Design and Implementation of ESP32-Based IoT Devices. *Sensors* **2023**, *23*, 6739. [CrossRef]
19. Hassan Al-Taai, S.H.; Abbas Kanber, N.H.; Mohammed al Dulaimi, W.A. The Importance of Using the Internet of Things in Education. *Int. J. Emerg. Technol. Learn. (Ijet)* **2023**, *18*, 19–39. [CrossRef]
20. Ghazanfar, S.; Hussain, F.; Rehman, A.U.; Fayyaz, U.U.; Shahzad, F.; Shah, G.A. IoT-Flock: An Open-source Framework for IoT Traffic Generation. In Proceedings of the 2020 International Conference on Emerging Trends in Smart Technologies (ICETST), Karachi, Pakistan, 26–27 March 2020; pp. 1–6. [CrossRef]
21. Alammary, S.; Alhazmi, S.; Almasri, M.; Gillani, S. Blockchain-based Applications in Education: A Systematic Review. *Appl. Sci.* **2019**, *9*, 2400. [CrossRef]
22. Karale, A.; Khanuja, H. Implementation of Blockchain Technology in Education System. *Int. J. Recent Technol. Eng.* **2019**, *8*, 2823–2828. [CrossRef]
23. Shmatko, O.; Borova, T.; Yevseiev, S.; Milov, O. Tokenization of Educational Assets Based on Blockchain Technologies. *Sci. Pedagog. Educ.* **2021**, *3*, 4–10. [CrossRef]
24. Han, L. Exploring the Mode of Inter-University General Education Consortium Based on Blockchain Technology. *J. Educ. Educ. Res.* **2023**, *6*, 113–115. [CrossRef]
25. Uspabayeva, A.; Sattarova, A.; Mirza, N.; Kubeeva, M.; Abdualievich, Z.K.; Rysbayeva, G. Evaluation of High School Students' New Trends in Education: Internet of Things. *Int. J. Emerg. Technol. Learn. (Ijet)* **2022**, *17*, 159–175. [CrossRef]
26. Sang, Y.; Wang, L. Physical Fitness Data Monitoring of College Students Based on the Internet of Things and Blockchain. *Front. Public Health* **2022**, *10*, 940951. [CrossRef]
27. Yujie, H.; Qiuxia, H. Innovative Mode of Logistics Management of "Internet of Things + Blockchain"-Integrated E-Commerce Platform. *Comput. Intell. Neurosci.* **2022**, *2022*, 7766228. [CrossRef]
28. Shahin, Y.M. Technological Acceptance of the Internet of Things (IOT) in Egyptian Schools. *Int. J. Instr. Technol. Educ. Stud.* **2020**, *1*, 6–10. [CrossRef]
29. Jeong, Y.S. Secure IIoT Information Reinforcement Model Based on IIoT Information Platform Using Blockchain. *Sensors* **2022**, *22*, 4645. [CrossRef] [PubMed]
30. Kyun, S.; JaeKyung, Y.; Jang, J.Y. A Decentralized Approach to Education Powered by Blockchain Technology. *Asia-Pac. J. Converg. Res. Interchange* **2021**, *7*, 131–141. [CrossRef]
31. Wang, Z.; Wu, W.; Zeng, C.; Yao, J.L.; Yang, Y.; Xu, H. Graph Neural Networks Enhanced Smart Contract Vulnerability Detection of Educational Blockchain. *arXiv* **2023**, arXiv:2303.04477.

32. Miorandi, D.; Sicari, S.; De Pellegrini, F.; Chlamtac, I. Internet of things: Vision, applications and research challenges. *Ad Hoc Netw.* **2012**, *10*, 1497–1516. [CrossRef]
33. KimiNewt. PyShark. Available online: https://pypi.org/project/pyshark/ (accessed on 4 December 2023).
34. Sartori, I.; Walnum, H.T.; Skeie, K.S.; Georges, L.; Knudsen, M.D.; Bacher, P.; Candanedo, J.; Sigounis, A.M.; Prakash, A.K.; Pritoni, M.; et al. Sub-hourly measurement datasets from 6 real buildings: Energy use and indoor climate. *Data Brief* **2023**, *48*, 13. [CrossRef]
35. ABET. Criteria for Accrediting Engineering Programs, 2019–2020. Available online: https://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-engineering-programs-2019-2020/ (accessed on 4 December 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

computers

MDPI

Article

Student Teachers' Perceptions of a Game-Based Exam in the Genial.ly App

Elina Gravelsina and Linda Daniela *

Faculty of Education Sciences and Psychology, University of Latvia, LV-1586 Riga, Latvia; elina.gravelsina@lu.lv
* Correspondence: linda.daniela@lu.lv

Abstract: This research examines student teachers' perceptions of a game-based exam conducted in the Genial.ly app in the study course "Legal Aspects of the Pedagogical Process". This study aims to find out the pros and cons of game-based exams and understand which digital solutions can enable the development and analysis of digital game data. At the beginning of the course, students were introduced to the research and asked to provide feedback throughout the course on what they saw as the most important aspects of each class and insights on how they envisioned the game-based exam could proceed. The game-based exam was built using the digital platform Genial.ly after its update, which provided the possibility to include open-ended questions and collect data for analyses. It was designed with a narrative in which a new teacher comes to a school and is asked for help in different situations. After reading a description of each situation, the students answered questions about how they would resolve them based on Latvia's regulations. After the exam, students wrote feedback indicating that the game-based exam helped them visualize the situations presented, resulting in lower stress levels compared to a traditional exam. This research was structured based on design-based principles and the data were analyzed from the perspective of how educators can use freely available solutions to develop game-based exams to test students' knowledge gained during a course. The results show that Genial.ly can be used as an examination tool, as indicated by positive student teachers' responses. However, this research has limitations as it was conducted with only one test group due to administrative reasons. Future research could address this by including multiple groups within the same course as well as testing game-based exams in other subject courses for comparison.

Keywords: game-based exam; game-based learning; feedback; Genial.ly; student teachers

1. Introduction

In the evolving and ever-changing landscape of education, game-based learning has emerged as a transformative approach, offering a dynamic pedagogical alternative that not only engages students but also fosters deeper understanding and critical thinking skills. In educational research and practice, various terms describe the use of game elements to enhance learning. These terms include "educational games", "game-based learning", "serious games", "gamification", and "play-based learning". While they might sound similar, each has its specific focus and application [1–3]. In this article, the term "game-based learning" will be used. This refers to a method where educators use tools like board games, escape games, or digital games designed with a clear educational strategy and specific learning goals in mind [4,5]. It is important to note that this is not just about making learning fun. The design process is goal-oriented, aiming for a balance between an educational challenge and enjoyment [6], all while the educator guides and monitors the process [6,7]. Incorporating game-based learning also aligns with modern pedagogical trends that emphasize active learning and student-centered approaches [8,9].

Beyond the delivery of knowledge, game-based learning's potential also lies in its capacity for assessment. When designed with precision, games can evaluate students'

understanding, track their progress, and pinpoint areas needing attention [10,11]. Ifenthaler et al. [12] suggest that adding assessment features to game-based learning is still a developing concept, mainly because it takes much longer to design such tools. Platforms like Kahoot, Quizlet, and Nearpod have harnessed this dual role, transforming traditional assessments into interactive experiences [13–15]. Although these types of assessment are common—as seen from a scoping review of 200 articles (details provided further in the research)—only one article [16] offered insight into using game-based learning for final course summative assessments. The lack of research highlights the need for further studies in this field. To identify research where final course assessments in higher education are specifically designed as games, authors suggest using the term "game-based exam".

Within this framework, the "Legal Aspects of the Pedagogical Process" course embarked on an innovative journey, leveraging Genial.ly as a tool for a summative game-based exam to assess knowledge gained during the course. The objective of this research is to find out how educators can develop game-based exams incorporating principles of serious games analytics using available digital solutions and to ensure that knowledge gained during the course is tested. Genial.ly, a platform renowned for its interactive content creation capabilities [17], was chosen not just for its visual appeal but also for its recent advancements in facilitating assessment opportunities in the Genial.ly "Master" plan. This case study tests available solutions for educators who lack specific technological competencies and where no specific game-based programs are available to test students' knowledge. This research is important at all levels of education, where innovative assessment methods can significantly impact student learning and engagement. In higher education, regular exams are a critical part of each course, and exploring alternative assessment methods like game-based exams can provide valuable insights into improving the examination process. Moreover, this study focuses on student teachers who will eventually transfer their knowledge and practices to their own students. By engaging with game-based exams, these future educators can adopt and possibly continue such innovative practices in their teaching careers. This multiplier effect could lead to broader implementation of game-based assessments, enhancing learning experiences across various educational settings. Design-based research conducts a comprehensive analysis of this initiative, and comments on the advantages and potential challenges of using Genial.ly for game-based exams.

2. State of the Art

This section discusses how game-based learning enhances the learning process and the role of games in assessment. In the many articles defining game-based learning, the explanation that game-based learning is opposite to traditional techniques, methods, or conventional instructions stood out [5,8,18]. In this sense, "traditional" refers to the teacher as the primary source of knowledge and lessons most often being delivered through lectures. Game-based learning, on the other hand, represents a shift toward a more student-centered approach. Instead of passively absorbing information, students are actively engaged in the learning process. Coleman and Money [8] state that student-centered learning is a meaningful keyword for game-based learning. It can be observed that student-centered learning is not only an outcome of game-based learning but also the common keyword that links it with active learning strategies.

Games, by their very nature, are interactive and require players to make decisions, solve problems, and think critically. The purpose of using game-based learning comes down to the educator and how they support reaching set goals. Mostly, it is used to deliver knowledge and skills to the learners [5,8,17]; however, when the game's design is carefully planned, it can do more than just teach. For instance, students' understanding can also be assessed when proper mechanics are integrated within the game [10–12,19]. While students "play", the game can track their progress or the game outcome highlights what they have mastered and points out areas the educator might need to help them with.

This dual role of teaching and assessing makes game-based learning a valuable tool in modern education. Some games and online platforms that students and teachers prefer to

use like Kahoot, Quizlet, Quizziz, and Nearpod [13–15,20] have specially designed quiz games that transform traditional quizzes and lecture formats into dynamic sessions where students are actively involved, motivated, and challenged. Although these platforms are highly evaluated by the previously mentioned sources, there are still unidentified aspects, like inclusion, stress levels, the real knowledge revealed, etc. In "The Views of the Teacher Candidates on the Use of Kahoot as a Gaming Tool", Uzunboylu and colleagues [21] examined Kahoot and identified some negative aspects, such as the application's dependency on the internet and the issue of leaving user account names to be chosen by students, which could lead to potential problems including unethical usernames. However, they provide generally positive views of the platform.

While platforms like Kahoot offer assessment capabilities, they may not match the interactive potential found in Genial.ly. Several articles have highlighted the positive learning experiences facilitated by Genial.ly's digital tools [17,22–24]. However, it is worth noting that very few of these, if any, delve into the negative aspects; consider potential risks associated with their use; or analyze their potential in using them for course exams.

There are many aspects to consider when creating a game-based learning experience, including educational content, game design, technical issues, and resource management. One of the most common problems is that the game mechanics and design are not well-aligned with the educational objectives [25–27]. Another often overlooked issue is the pressure on teachers who create these materials. This process combines all the aforementioned aspects and can lead to teacher burnout [28]. When designing game-based exams with the prior notion to test them as a potential tool for teachers, these aspects need to be taken into consideration.

Having the prior notion of using this digital platform as a tool for examination, it was important to understand the possibilities of designing it for summative assessment. Digital game-based learning goes beyond traditional assessment methods to create immersive learning experiences that actively engage students. Specifically, game-based learning is designed to assist learners in addressing and overcoming potential obstacles. These challenges may include feelings of anxiety, diminished motivation, comprehension difficulties, and reduced engagement [10,29]. Furthermore, game-based learning aims to enhance the level of interaction, thereby offering substantial educational advantages and fostering a more conducive learning environment.

While many articles provide insight into using games-based learning for formative assessment, there is little research conducted to examine this field for summative assessments that are meant for university course exams. We conducted a scoping review in the Web of Science database using the keywords "game-based examination" or "exam". This search included articles, book chapters, and proceeding papers, applying filters for the English language and research fields such as computer sciences, education, and social sciences. The articles were gathered from 2020 until the summer of 2024. The 5189 search results were sorted by relevance, and the first 100 articles were examined to gather insights and context for this study. Ninety-seven articles used the word "examination" to evaluate methods or, in other contexts, to integrate game-based learning elements into the instructional phase of the course rather than be used as a synonym of summative assessment that happens at the end of a course or program. This is mainly performed to prepare students for an exam or to use an exam to see if game-based learning works. Only in three articles the final "exam" or "examination" was made into a game or used gamified solutions. One of these was conducted in a real-life escape room [30] and another did not use game-based learning but rather separate game elements such as a point system [31] so they are not relevant to this study. The only relevant article from the first hundred on game-based learning exams found on Web of Science was "Digital game-based examination for sensor placement in context of an Industry 4.0 lecture using the Unity 3D engine—a case study" by Julian Koch, Martin Gomse, and Thorsten Schüppstuhl [16]. With the same filters, similar findings were identified using the keywords "game-based assessment" and "summative assessment". One article on a real-life escape room was included [32] but otherwise, from two-hundred

and ninety-three articles, only the first hundred that were the most relevant were examined, and no other inquiries relevant to this study were found.

The research by J. Koch, M. Gomse, and T. Schüppstuhl [16] highlighted the practical implementation of using a game-based examination with the complex Unity 3D engine. Although it is contrary to this research idea—that course teachers with no prior knowledge of coding can make their own game—this article provides an example relevant to the research topic. Three main conclusions that can be used from this article are as follows: First, the knowledge that a digital game-based examination can be an effective tool for assessing complex problem-solving skills and practical knowledge in a more realistic, engaging manner. Second, with its interactive and immersive design, the game allowed students to demonstrate their understanding in a more applied context compared to traditional exams. The gamified environment was found to be more stimulating and enjoyable, as observed by students showing high levels of engagement and motivation while participating in the game-based examination.

Overall, the findings suggest that while game-based learning is frequently used to enhance the instructional process, and the words "examination" and "assessment" are commonly used to refer to the process of evaluating a method or its elements and for formative assessment, they are not synonymous to a summative assessment conducted at the end of a course or program. Formative assessments relative to summative ones, and especially course exams, measure short learning periods. Exams have to have an overview of the whole course. To differentiate between these two types of game-based assessments, in this research, the authors would like to propose the term "game-based exam" to refer to the process where the final course examination or summative assessment is made into an educational game.

The efficacy of technological tools in assessing educational outcomes remains a topic of debate. In game-based learning experiences, there are three predominant methods to evaluate learning data: analyzing in-game scores, utilizing external assessments, and incorporating internal game assessments [12]. A recent advancement in this field is the introduction of "learning analytics", also known as "serious games analytics" [33,34]. This approach adapts the game environment in response to the player's performance. It not only considers player-specific data such as demographics and prior knowledge but also examines in-game behaviors, including duration of play, encountered challenges, and achieved objectives that can be obtained from the game and used to improve personalized learning [34,35].

Considering these factors, this study seeks not only to explore the technical aspects of employing Genial.ly as a tool for game-based exams but also to grasp the viewpoints of both students and educators on this matter.

3. Research Design

Building on the foundational understanding of game-based learning and its transformative potential in education, an opportunity arose to change the traditional learning process. Given the multitude of benefits and firsthand experience with game-based learning, the chance to adjust it to a university course arose. This section aims to lay out the process and results obtained while designing and implementing a game-based learning course exam in the Genial.ly app.

The approach of using a game-based exam to test students' knowledge gained through lectures is new for higher education in Latvia. Another novelty is that the game was developed by educators without specific competencies in coding, and the research data provides insight into how freely available tools can be incorporated into the study process to test knowledge gained, as most previous research is conducted on ready-made games or simple quizzes that cannot be used for the final exams. A design-based research strategy was employed to collect and evaluate the data on the effectiveness of game-based exams in testing the results of students' learning during the course. It is an empirical research methodology that focuses on the iterative design, testing, and refinement of interventions

in real-world settings. It bridges the gap between theoretical research and practice by emphasizing collaboration between researchers and practitioners [36]. For this article, employing design-based research is particularly advantageous as it allows for systematic study and improvement of a game-based exam for the course "Legal Aspects of the Pedagogical Process". By integrating a design-based research strategy, this research not only captures the effectiveness of the intervention but also contributes to the broader theoretical understanding, ensures that the game-based exam is both pedagogically sound and practically relevant for knowledge testing, and ensures it is accessible for educators without specific competencies in coding who wish to develop game-based solutions for their teaching process. In a similar piece of research, Sidonie Pors [37] also used design-based research and suggested three design stages: pre-implementation, implementation, and post-implementation. For a better overview of this design-based research, please see Figure 1.

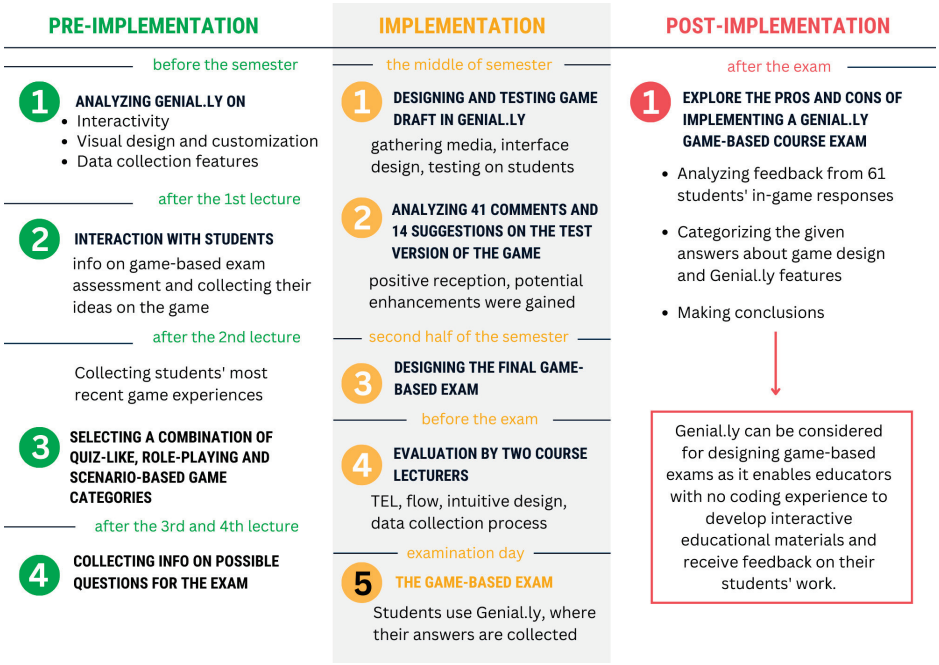


Figure 1. Overview of the design-based research method with 3 main phases and connected parts.

For this research, the digital game-based knowledge-testing experience was designed in Genial.ly, a digital platform and tool that allows users to create interactive and animated content, such as presentations, infographics, reports, quizzes, and more. Customization options enable users to design visually engaging content without the need for coding. Positively reviewed by other authors [17], and successfully implemented in other projects such as by the University of Latvia's Faculty of Education Sciences and Psychology, Genial.ly was chosen for its capability to deliver a rich visual experience and facilitate knowledge testing for students. While Genial.ly is an innovative tool for creating interactive content, it needs more comprehensive capabilities for serious game analytics. However, it does provide some basic data on user interactions, such as click-through rates, time spent in the game, and overall engagement metrics.

Consequently, an innovative approach was initiated to develop an exam in the form of a game for the "Legal Aspects of the Pedagogical Process" course. This initiative aimed to assess students' comprehension of the subject matter and offer an engaging, interactive, and immersive examination experience. This approach provides a firsthand experience of situations that potential teachers may encounter. "Legal Aspects of the Pedagogical Process" is a course that student teachers take in the second semester of their four-year degree. It provides an overview of the Latvian and European regulations concerning compulsory school students, their parents, and teachers. The course was conducted online in real time,

allowing immediate interaction and engagement. Additionally, sessions were recorded, providing students with the opportunity to revisit and review content at their convenience. Each lecture was supplemented with discussions and mid-lecture interactive questions in the platform Mentimeter, and students were prompted to share their reflections after every session.

On a couple of occasions during the course, students were asked to offer insights into how they envisioned the presented information being transformed into a game. In the middle of the semester, students were presented with a first glimpse of the game-based learning experience, and adaptations were made based on their feedback. There were 89 students on the course but only 65 took the final test, and 61 of them filled out the evaluation questionnaire. The participants were full-time student teachers enrolled in a four-year program. The course is offered in the second semester of the first year, and all group members are student teachers specializing in various subjects. By the end of their studies they receive two qualifications, which they choose at the beginning of their studies. In the academic year, the number of active participants fluctuates due to the possibility that existing students may take academic leave, drop out, switch to part-time programs, or transfer to other programs. This course, "Legal Aspects of the Pedagogical Process", is a mandatory course for all student teachers, regardless of their specialization.

The exam was conducted in person in computer labs, supervised by the course professor. Organized as a Genial.ly game, students logged in with their credentials to participate in an online quiz and role-play scenarios designed to assess their knowledge of various real-life situations related to the course material. The results of this exam contributed 30% toward the students' final grade for course completion.

The pre-implementation phase included analyzing Genial.ly's technical aspects, conducting surveys inquiring insight into how students envisioned the game-based learning exam, and determining their thoughts on the course material. In the implementation phase, students were asked for their opinions through a single open-ended question regarding the initial draft of the Genial.ly game-based learning experience to make necessary corrections. At the end of the semester, the final version of the game-based exam was designed and tested, including by the course professor (the second author of this paper). This game-based exam constituted the final part of the implementation phase. In the post-implementation phase, after the exam was administered, the students' results and feedback were analyzed.

The final exam included the following:

- quiz-based questions where students had to choose one or several correct answers;
- situations that they had to analyze and provide the solution for based on legislative frameworks;
- open-ended questions where students had to provide their opinion.

To better understand the game content, which is directly related to the course "Legal Aspects of the Pedagogical Process", five main topics that were included in the exam have been outlined:

1. Disciplinary procedures and regulations: this includes handling physical altercations that a teacher may observe, school actions on student violence, and restrictions on teaching;
2. Assessment and evaluation: questions related to summative, diagnostic, and formative assessments as well as support during state examinations and what teachers' obligations are;
3. Legal and ethical frameworks: covers normative acts, ethical values in education, and legal reporting obligations for teachers experiencing violence or abuse;
4. Teacher qualifications and responsibilities: focusing on the qualifications required for teaching specific subjects, conditions for exclusion from pedagogical work, and responsibilities towards parents and guardians;
5. Student welfare and rights: addresses the entitlement to free school meals, data privacy regulations, social support, etc.

Although there are many topics that students need to cover, the primary objective of the course is for student teachers to know where to find relevant information and which strategies they can use to find potential solutions to problems. It was important to evaluate that students demonstrated the ability to solve problems according to legislative requirements that are in force in Latvia. They were allowed to search for documents online during the exam but the professor suggested keeping track of the time allocated for the exam. For example, situations like two boys fighting in the hallway—which the student teachers had to address—were based on actual experiences gathered from in-service teachers working in schools.

The data for the research was collected in several phases, as follows:

1. Evaluating the Genial.ly platform based on the possibility of developing the game-based exam and collecting data

The platform's flexibility was evaluated from a technical standpoint by considering several key factors. First, interactivity, such as the game mechanic capabilities, including the ability to incorporate multimedia content and branching scenarios (essential for testing purposes). Second, visual design and customization, to provide a user-friendly interface. Third, data collection features, focusing on its capacity to gather user interactions, responses, and performance metrics within the game environment for assessment possibilities;

2. Collecting students' opinions about possible features of the game for the final examination after each class

Students who attended lectures answered two or three open-ended questions after each class regarding a possible game-based exam and the lecture's content (what they found important and what a new teacher would need to know in a school environment);

3. Collecting students' opinions about developed scenarios (similar to those used in the final exam)

Students who attended lectures in the middle of the semester were given a game similar to the one developed for the final exam and instructed: test the game's introduction and provide feedback regarding its potential. This resulted in 41 positive comments and 14 suggestions from the students on how they envisioned the game;

4. Evaluating the game for final corrections

The technology-enhanced learning (TEL) and course content specialist, responsible for course development and examination topic selection, was asked to assess the game based on content relevance, potential responses, and TEL principles, emphasizing the significance of information flow within digital environments for effective learning outcomes;

5. Collecting students' opinions on a game developed for knowledge testing at the end of the final exam

The last question in the exam was an open-ended question where students were asked: please share some brief thoughts on this style of examination.

One of the primary limitations of this study is that it was conducted with only one test group due to administrative reasons. Every spring semester there is only one group in this course and no data are saved from previous students' assessments. This restriction limits the generalizability of the findings. Future research could address this limitation by including multiple groups within the same course and by testing game-based exams in other subject courses for comparison. Such studies would provide a broader understanding of the effectiveness and applicability of game-based exams administered through the platform (for this study it was Genial.ly) as an examination tool across different educational contexts.

All procedures performed in studies that involve human participants should be conducted in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Since the results of specific students' data were not analyzed for this research, but rather the process of designing a game-based exam, the University of Latvia guidelines indicated that approval from the ethics committee was not required. Everything was coordinated at the beginning of the semester, and the students helped to organize this process by voluntarily providing feedback and were given the possibility of completing a regular exam if necessary.

4. Discussion and Results

In this section, the outcomes and results of the design-based research will be analyzed following the logic used during research implementing the three design stages. Each stage provides insights into the development and assessment of the game-based learning exam.

4.1. Pre-Implementation

In the pre-implementation phase, Genial.ly underwent a comprehensive assessment across three key categories to ensure its compatibility with previously set objectives of facilitating an interactive course examination. The first key category, interactivity, was of the highest value for the design. Within this context, it was established that content creators could integrate a wide array of multimedia elements, including images, videos, audio, animations, and transitions. Furthermore, it is possible to incorporate external resources using embedding links. Users can actively engage with both internal and external links to reveal hidden or extended content. Moreover, the platform allows users to move objects, providing them with the flexibility to organize and structure information in a customized manner.

The second key category is visual design and customization. Although the platform includes customizable templates, it also empowers users to design visually compelling materials with a user-friendly interface and a wide range of design tools, shapes, and text formatting options. Users can adapt the platform to their specific pedagogical needs. This flexibility extends to personalized branding, design choices, and content creation, making it a versatile tool for a variety of educational contexts.

Originally the game was intended to serve as a visual supplement for students. In this scenario, the learning data would have been evaluated by utilizing external assessments (student worksheets); however, with the platform's recent feature enhancements—including the addition of interactive questions and the capability to insert open-ended queries—an opportunity to fully integrate the course examination within the game itself emerged. Consequently, it was possible to utilize the game's features to internally assess students' work and obtain an overview of their performance. Taking this into account, assessment is the third key category that was essential to build this experience. Educators can seamlessly integrate quizzes, surveys, and interactive evaluations within the platform's content, allowing real-time feedback and analysis (Figure 2).

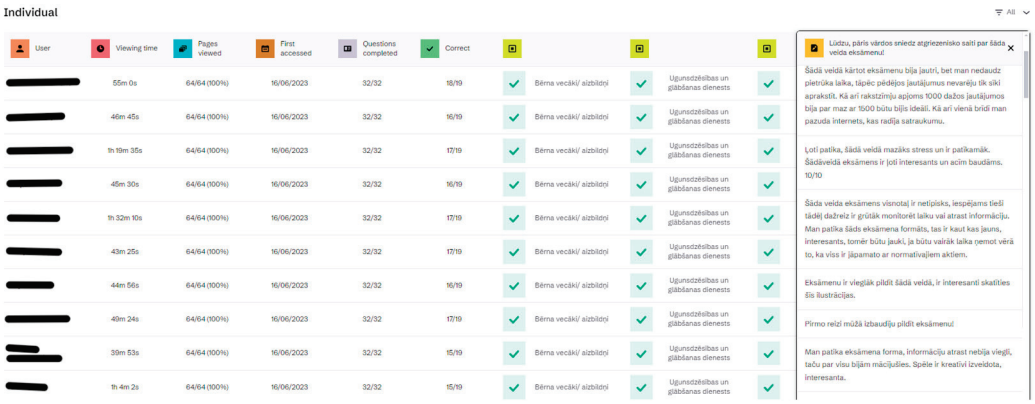


Figure 2. Feedback and analysis from Genial.ly's "individual activity" view that includes answers to both test-based questions and open-ended questions where students provided their opinions.

During the first lecture, students were introduced to the concept of a game-based summative assessment being implemented at the end of the semester. Following this lecture, students were invited to provide suggestions regarding the type of game that could enhance their understanding of the subject matter, and their opinions were collected using Google Forms. The question asked after the lecture was as stated: following today's presentation and lecture, in your opinion, what type of game would help to reflect this topic? The student survey revealed that of the 49 students who participated in the lecture, 35% envisioned this game as a simulation, scenario role-play, or branched narrative. Some reflected their desire to engage in scenarios that would immerse them in various roles, including that of a teacher or student. They tried to identify and address challenges within these scenarios, offering solutions to improve the situation. Some students even envisioned resolving conflicts with students and their family members, class teachers, or social pedagogues, fostering a sense of empathy and encouraging everyone to advocate for their beliefs. These insights highlight their interest in dynamic and participatory learning experiences, focusing on understanding and problem-solving within diverse roles and situations. Ten students (20% of respondents) perceived the final exam as resembling a quiz-type game. They drew parallels with popular platforms like Kahoot, where questions are closely connected to their respective answers. One student elaborated on a real-life quiz game format that included a game board and random questions. Other students mentioned card games, matching games, and ball games as exam analogies.

While a range of game-based assessment methods was suggested, the initial two examples, i.e., comparing final exams to quiz-like games and scenario-based games, appear to be more coherent and relatable when considering the student teachers' perspective. In 13 responses, it was evident that a few students may not have had a complete grasp of implementing game-based learning. A few students mentioned active learning methods that lacked gamified elements, such as straightforward discussions or activities like drawing a brainstorming cloud. These comments indicate varying levels of understanding and familiarity with the concept of game-based learning. Since this activity aimed to collect their ideas, not evaluate their understanding, this aspect will not be discussed further.

After the second lecture, 45 students who attended the class answered the question, what games have you played lately in the last few months, naming 59 different games. "Uno" emerged as the most popular choice, with twelve respondents indicating they had played it recently, closely followed by "Monopoly", named by ten students; "Alias", a word-guessing game, garnered the interest of seven participants, while five individuals reported playing various tabletop games. While there were ten mentions of mobile or video games, most students indicated a recent experience with board games. These results confirm that students mostly engage in tangible face-to-face interactions facilitated by board games, which presents a challenge for educators who wish to use digital game-based learning. Consequently, a decision was made to develop a game draft that combined quiz-like, role-play, and scenario-based elements. This game was provided in the middle of the semester to introduce students to the concepts planned for the final exam.

Two additional questions were asked after each lecture on the lesson's content as follows: What do you think was the most important thing for new teachers to know today? Give one specific example and please formulate one question about today's topic that you find interesting (must be different from the answer to the previous question). The first question encourages reflection on the significance of specific knowledge for aspiring teachers, potentially highlighting key concepts or challenges they perceive as crucial in this class. Meanwhile, the second question prompts students to actively engage with the topic by formulating an inquiry, showcasing their curiosity and understanding of the material. The questions aimed to gather perspectives and engage students with the course content and were regarded as possible questions for the exam.

4.2. Implementation

During the implementation phase, the first game draft was introduced to students and their feedback was gathered to assess their opinions on the game's effectiveness and to identify areas for improvement. Questions from this draft about the course content were not included in the final test. This phase aimed to gauge student responses and suggestions for refining the game to better align with their learning needs and preferences and to introduce students with the principles of game-based knowledge testing planned for the final exam. The students were asked: test the game's introduction and provide feedback regarding its potential. A total of 35 students attended class that day and in response, they provided 41 positive comments and 14 suggestions. For the data collection, Google Forms was used. Student teachers generally responded positively, praising the game itself or its visual and audio elements. The game's intuitiveness and the fact that it can be used on mobile devices were also well-received. Student teachers provided positive feedback, highlighting their enthusiasm for addressing crisis situations and their eagerness to find solutions when role-playing as the new teachers. For some, it was a novelty. The recommendations included making the visuals more engaging and colorful in case they were meant to capture children's attention. As this was not the case and the game would be used for adults, the previous comments suggested the visualizations with the photography collage were engaging and intuitive enough. Some feedback suggested improving the audio quality for a more pleasant experience. Others suggested addressing language errors, providing additional explanations in some cases, and considering a name change. The feedback collected provided valuable insights into the initial reception of the game-based learning tool and potential enhancements, and based on these comments, the game was improved to develop the game-based exam.

Before administering the game-based exam to students, it was evaluated by two course lecturers, one of whom was an expert in educational game development and the other, an expert in TEL and course content. The expert in game development evaluated the game from the perspective of how students managed to navigate through the game and how the data were collected in the system to evaluate students' performance. The expert in TEL and course content, who developed the course and chose the topics for examination, evaluated the game from the perspective of content and possible answers and from the perspective of principles of TEL, where the flow of information in the digital space is an important aspect of learning success. After this evaluation, all the comments and suggestions were implemented to ensure that the game could be used for the students' final exam.

The final part of the implementation phase was to conduct the game-based exam. Bearing in mind that the main points of focus of the research were on Genial.ly as a potential platform for exams, testing the development of game-based exams by educators with no specific competencies in coding, and the pros and cons of the use of games for final exams to test students' knowledge and opinions rather than analyzing exam results—which cannot be compared with other groups' results because no information was collected on students' learning habits, previous knowledge, cognitive capabilities, etc.—therefore, it is not possible to compare different groups of students at this stage. Instead, this paper discusses a case study where researchers (educators with no specific competencies in coding) tried to develop an interactive game-based final exam and test the possibility of replacing a traditional exam with a game-based exam to assess knowledge gained during the course. After the exam, the results of the examination process were evaluated from the perspective of how students managed the process and how they could give reasonable answers. This evaluation was made by one of the experts participating in this research, who had previously administered final exams for students on the same course and had previous experience with knowledge testing. The time allotted for the final exam was 1.5 h, and students were asked questions about all the topics learned during the course. The number of questions in the game-based exam was about the same as it was in previous years with other groups of students. It was concluded that there were no differences compared to the way students had previously answered the exam questions, meaning that a game-based

exam in Genial.ly did not restrict them from retrieving their knowledge; this means that the game-based exam fits for the purpose of knowledge evaluation but changes the way the test is administered. The reason for changing the form of administering the final exam was based on the idea that a game-based exam with storytelling elements would help them retrieve their knowledge and reduce their stress levels. At this stage, it can be concluded that such a testing strategy has the potential to reduce stress and anxiety, as also mentioned in previous research [10,29], but this should be tested in further research steps.

4.3. Post-Implementation

The primary objective of this case study was to explore the pros and cons of implementing a Genial.ly game-based course exam in the "Legal Aspects of the Pedagogical Process" course from a student teacher's perspective. By analyzing these diverse perspectives, this study aims to provide a comprehensive understanding of the efficacy and areas of improvement for game-based examinations in an academic setting. While these insights can offer a glimpse into user behavior, they do not delve deep into the complex learning outcomes and detailed player performance metrics that serious game analytics typically encompass.

At the end of the course, 65 students took the final exam and 61 left feedback about the game-based exam. Acknowledging that game-based learning might be a novel experience for some, it was reassuring that most students (60 of who gave feedback) found the experience positive. One student who gave feedback did not reflect on the game experience, only on the questions and time limit. This reaffirmed the necessity of the pre-implementation and implementation stages to coax the students into thinking about the upcoming experience. It may be said that in this game-based exam, students had a unique and engaging experience that differed from traditional written exams.

In the exam, in response to the last question (Please share some brief thoughts on this style of examination), student teachers expressed a range of opinions about this format, providing an extensive overview. As mentioned before, 60 students found the exam fun and interactive. In one student's feedback, it was said, "For the first time in my life, I enjoyed taking an exam". Most appreciated (15/60 responses) was the possibility of immersing themselves in various scenarios. Only one student out of sixty expressed dissatisfaction with the dialogues between the questions, saying they were not necessary. Most students appreciated the opportunity to role-play, and it was mentioned that the interactive nature of the exam allowed them to explore real-life situations and apply their knowledge. They liked the fact that the exam felt like a conversation rather than a test. One student said, "Interactivity and real-life situations allowed me to better understand the role of a teacher". This may also be due to the visualization element of the game (Figure 3). Visualizations given in this material are from the game developed for this particular reason, and were presented in Latvian but a translation of the text is given below for context. Eleven students mentioned that the visualizations helped them better understand and engage with the content. As one student stated, "It was visually pleasing to answer the questions".

Interestingly, game-based learning reduces stress, as found by Dziob [10], and this was supported by ten students who mentioned that the game-based format reduced their stress compared to traditional exams. They felt that it was a more relaxed and enjoyable way to demonstrate their knowledge. One student remarked, "It felt less formal, so the stress around the exam was much less". Although almost all students (60 students) found this exam format to be different and more interesting than traditional written exams, there is still room for improvement. Only one student mentioned that a traditional exam would be preferable, and this was mentioned in the context of a lack of time, which most students found challenging. In all, twelve students mentioned that there was not enough time to write everything they wanted, and four students mentioned that 1000 characters were insufficient to elaborate their answers. The game did not have specific elements that reacted to time, e.g., "You have 2 minutes to answer the following question". The time for the exam was set by the lead professor, who tested the game before the examination. It was deliberately set to ensure the students focused on the main information and did not provide

unnecessary details. While the exam was constrained to a brief 1.5-hour timeframe, its primary objective was to assess the students' ability to articulate and demonstrate their knowledge effectively.



Figure 3. The game visualizes a conversation in the hallway.

Another aspect that is always a risk factor when using digital solutions is technical problems. Six students mentioned that the classroom's internet could have been better, and they had concerns about not being able to save their answers due to these issues. However, if the connection is poor, they can refresh the page, log in again using their username, and return to the point where they left off. From a design aspect, one student mentioned in his feedback that the progress percentage in the corner of the screen was a useful asset during the exam and that it would be good to have it all the time. For now, the progress indicator is shown after each topic is played out in the game, but this is something that can be added to every slide (Figure 4).

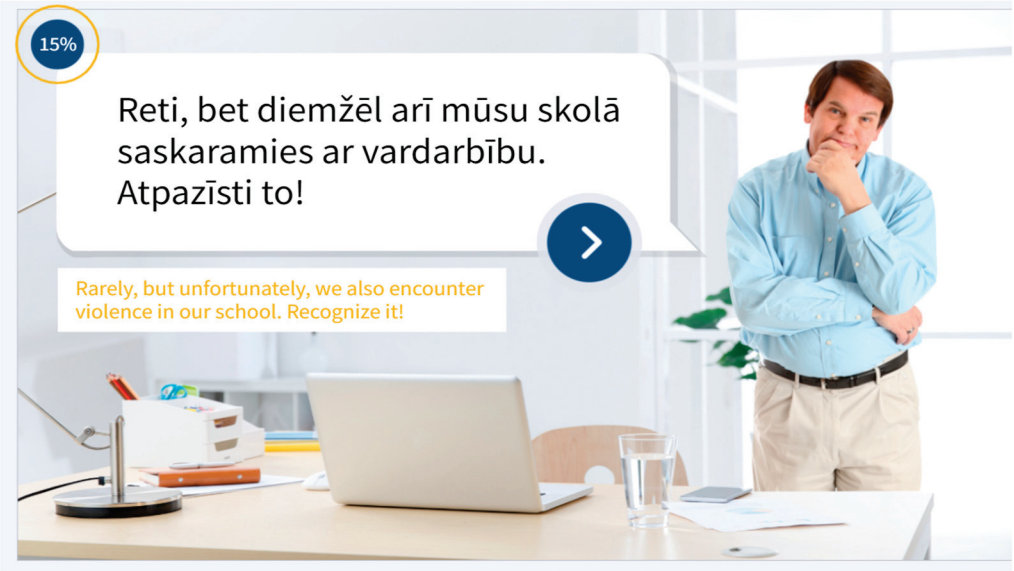


Figure 4. The game progress indicator, displayed in the upper left corner.

In four cases, students mentioned the lack of space to write their answers to open-ended questions. While it does not take much effort to change the basic visual elements, the designer is restricted by some Genial.ly features, such as the limit of 1000 characters that users can write within their open answers or the possibility of changing the words on the buttons and write-in windows into foreign languages. For example, while questions can be written in Latvian, to submit their answer, students need to press the "Send" button, which cannot be translated from English to any other language (Figure 5).

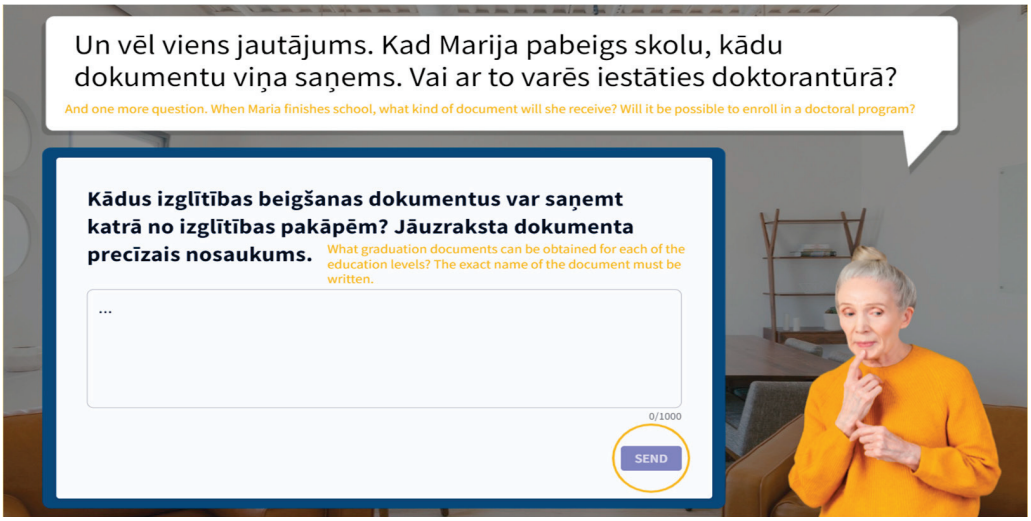


Figure 5. Inability to customize the words of the interactive question interface.

Although all the students had a good level of English, it would be better to have the possibility to create all the content in the same language. In the context of adult education, this does not pose a significant concern; however, when contemplating its application in youth education, it could potentially become problematic.

When comparing the main insights from the article "Digital game-based examination for sensor placement in context of an Industry 4.0 lecture using the Unity 3D engine—a case study", [16] showed similarities to our study. First, game-based exams can assess problem-solving skills and practical knowledge in a realistic and engaging manner, and although the study course does not provide the student teachers with hands-on experience, from their answers it was clear that it was easier for them to envision the situation and try to find a solution as they could in real life. This is mainly due to the second point—the interactive and immersive design of the game—that allows students to demonstrate their understanding in a more applied context compared to traditional exams. As some students also mentioned they enjoyed the provided game, it can be agreed that a gamified environment is more stimulating and enjoyable, leading to higher student engagement and motivation during game-based examinations. Nevertheless, these two studies have their differences in their complexity but show similarities in the conclusions reached.

To summarize, Table 1 provides a clear overview of the pros and cons of using Genial.ly as a resource for making exams from the student teachers' and designers' perspectives. It highlights the benefits and drawbacks of both the game-based format and specific features of the Genially platform, allowing educators to make informed decisions about its implementation.

Observing these points can help identify areas for improvement. The concept of a game-based exam generally has more pros than cons; however, the Genial.ly app, as a platform for this purpose, shows an equal number of pros and cons based on this research. Since many of these cons are technical issues that may be resolved in future updates, Genial.ly can still be considered a viable tool for designing game-based exams as the interface and structure are intuitive and easily understandable and allow educators with

no competencies in coding to develop interactive learning materials and exams to test the knowledge gained during learning.

Table 1. Pros and cons of using genially for game-based exams.

Category	Pros	Cons
Game design	Fun, interactive	Lack of time
	Immersive scenarios to explore real-life situations and apply knowledge	
	Game-based format was more relaxed and enjoyable	Different game types not suitable for everyone
	Visualizations helped better understand and engage with the content	
	Intuitive design and progress indicator	
Genially features	Flexibility in content creation and design (embedding links, different media)	Concerns about poor internet connection and potential inability to save answers
	Addition of "interactive questions" allowed for internal assessment within the game	Limited ability to change language of interface elements (e.g., 'Send' button).
	Provides features to create intuitive and mobile-compatible designs	1000-character limit for open-ended answers
	Suitable for various educational contexts	Assigning students their profiles

The broader implications of our findings for educational technology and assessment practices are significant. Platforms like Genial.ly not only facilitate the integration of gamification elements into assessments but also allow for comprehensive data collection and analysis. This dual capability enhances the ability of educators to track and measure student performance more effectively. The Authors hope that this article will inspire more educators to provide their courses with different game-based exams as this research field is still little explored.

An unexpected discovery from our study was that the game-based assessment tended to induce less stress among a portion of students compared to traditional assessments. By reducing anxiety, game-based exams can foster a more conducive learning environment, promoting better engagement and deeper learning; however, this needs to be more accurately tested and researched.

However, this research has limitations, as it was conducted with only one test group due to administrative reasons. This single-group approach restricts the generalizability of the findings and may not fully capture the diverse experiences and outcomes that could emerge from a broader participant base.

Future research could address this limitation by including multiple groups within the same course. This would allow for a more robust comparison between different sets of students, providing a clearer picture of how game-based exams impact various learning styles and academic performance. Additionally, testing game-based exams in other subject courses could offer valuable insights into the versatility and effectiveness of this assessment method across different disciplines. By expanding the scope of the research to include multiple test groups and a variety of subjects, future studies can more comprehensively evaluate the potential benefits and challenges of implementing game-based exams in higher education. This broader approach would help in identifying best practices and refining the design of game-based assessments to enhance their applicability and impact in diverse educational settings.

Overall, these findings suggest that incorporating game-based assessments in educational practices can revolutionize the way we approach student evaluation, making it more interactive, engaging, and less stressful. This shift has the potential to improve learning outcomes and provide more meaningful insights into student performance.

5. Conclusions

In the pre-implementation phase, the Genial.ly platform was evaluated for its interactivity, visual design, and customization to meet the objectives of an interactive course examination. The platform offered flexibility in content creation and design, making it suitable for various educational contexts. The addition of 'interactive questions' allowed for internal assessment within the game itself. In the implementation phase, the first game draft was introduced to students, and their feedback indicated a positive attitude toward such an assessment strategy. Students appreciated the game's visual elements, intuitiveness, and mobile compatibility. Recommendations included enhancing the visual engagement and audio quality. In the post-implementation phase, this study found that students generally had a positive experience with the game-based exam:

- The research showed that the final exam can be organized in the form of an interactive game to help students retrieve the knowledge learned during the course;
- A total of 60 out of 61 students who gave feedback expressed the opinion that the game-based exam was a positive experience;
- Overall, 25% of students (N15) enjoyed the game's interactivity and their immersion in various scenarios;
- Eleven students noted that the visualizations in the game helped them engage with the content;
- Ten students mentioned that they believe that such an exam reduces their stress levels;
- While most students preferred this format over traditional written exams, there is room for improvement, particularly in addressing technical issues and customizing certain elements.

Based on student feedback, we recommend the following for educators interested in implementing similar assessments:

- Combine question types: use a variety of question types to assess different skills and knowledge areas, not only testing answers but also open-ended questions for their thoughts;
- Choose platforms that have data collection possibilities, to gain deeper insights into student performance;
- Ensure the design and visual elements are engaging and clear. Information design and aesthetics play a crucial role in maintaining student interest;
- Design questions and tasks that reflect real-life situations to make the assessments more relevant and practical;
- Craft questions and instructions that are unambiguous to avoid confusion and ensure that students understand what is required of them. If possible, test them on a similar audience.

From the designer's view, they can consider resources and help from other specialists to avoid burnout in designing game-based experiences as this is one of the main risks in designing game-based experiences.

The significance of this study lies in using a game-based exam to enhance the teaching of legal aspects of pedagogical processes. This research project showed that educators with no competencies in coding can develop interactive games for knowledge testing, which replace traditional knowledge testing formats. It has highlighted the benefits of interactivity, reduced stress, and visual engagement in game-based learning but research should be continued to analyze the role of game-based exams to reduce test anxiety. This study also acknowledges the limitations related to technical issues and customization. The Genial.ly app can be used as a platform for creating interactive game-based learning exams with diverse forms of knowledge testing (quizzes with right and wrong answers, scenarios where students should provide the solution for a problem, and open-ended questions where students should provide their opinions).

Author Contributions: Conceptualization, E.G. and L.D.; methodology, E.G. and L.D.; software, E.G.; validation, L.D. and E.G.; formal analysis, L.D.; investigation, E.G.; resources, E.G.; writing—original draft preparation, E.G.; writing—review and editing, E.G. and L.D.; visualization, E.G.; supervision, L.D.; project administration, L.D.; funding acquisition, L.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data supporting the reported results can be found at the following: https://drive.google.com/drive/folders/10C1SkLR1HXXPUXyCQHM8D98TsI4vg9Go?usp=sharing (accessed on 11 August 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Deterding, S.; Dixon, D.; Khaled, R.; Lennart, N. From game design elements to gamefulness: Defining gamification. In Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, MindTrek, Tampere, Finland, 28–30 September 2011; Association for Computing Machinery: New York, NY, USA, 2011; pp. 9–15.
2. Gao, F.; Li, L.; Sun, Y. A systematic review of mobile game-based learning in STEM education. *Educ. Technol. Res. Dev.* **2020**, *68*, 1791–1827. [CrossRef]
3. Yildirim, I. The effects of gamification-based teaching practices on student achievement and students' attitudes toward lessons. *Internet High. Educ.* **2017**, *33*, 86–92. [CrossRef]
4. Grāvelsiņa, E.; Daniela, L. Designing an online escape room as an educational tool. In *Smart Pedagogy of Game-Based Learning*; Daniela, L., Ed.; Springer: Cham, Switzerland, 2020; pp. 119–133.
5. Karakoç, B.; Eryılmaz, K.; Turan Özpolat, E.; Yıldırım, İ. The effect of game-based learning on student achievement: A meta-analysis study. *Technol. Knowl. Learn.* **2022**, *27*, 207–222. [CrossRef]
6. Bylieva, D.; Sastre, M. Classification of educational games according to their complexity and the player's skills. *Eur. Proc. Soc. Behav. Sci.* **2018**, *51*, 438–446. [CrossRef]
7. Pho, A.; Dinscore, A. Game-Based Learning. Tips and Trends. 2015. Available online: https://acrl.ala.org/IS/wp-content/uploads/2014/05/spring2015.pdf (accessed on 11 August 2024).
8. Coleman, T.E.; Money, A.G. Student-centred digital game-based learning: A conceptual framework and survey of the state of the art. *High. Educ.* **2020**, *79*, 415–457. [CrossRef]
9. Hartt, M.; Hosseini, H.; Mostafapour, M. Game on: Exploring the effectiveness of game-based learning. *Plan. Pract. Res.* **2020**, *35*, 589–604. [CrossRef]
10. Dziob, D. Board game in physics classes—A proposal for a new method of student assessment. *Res. Sci. Educ.* **2020**, *50*, 845–862. [CrossRef]
11. Hung, H.-T.; Yang, J.C.; Hwang, G.-J.; Chu, H.-C.; Wang, C.-C. A scoping review of research on digital game-based language learning. *Comput. Educ.* **2018**, *126*, 89–104. [CrossRef]
12. Ifenthaler, D.; Eseryel, D.; Ge, X. (Eds.) Assessment for game-based learning. In *Assessment in Game-Based Learning: Foundations, Innovations, and Perspectives*; Springer: New York, NY, USA, 2012; pp. 3–10.
13. Plump, C.M.; LaRosa, J. Using Kahoot! in the classroom to create engagement and active learning: A game-based technology solution for eLearning novices. *Manag. Teach. Rev.* **2017**, *2*, 151–158. [CrossRef]
14. Portela, F. A new approach to perform individual assessments at higher education using gamification systems. In Proceedings of the 4th International Computer Programming Education Conference (ICPEC 2023) Proceedings, Vila do Conde, Portugal, 26–28 June 2023; de Queirós, R.A.P., Teixeira Pinto, M.P., Eds.; Dagstuhl Publishing: Wadern, Germany, 2023; pp. 8:1–8:12.
15. Shakhmalova, I.; Zotova, N. Techniques for increasing educational motivation and the need to assess students' knowledge: The effectiveness of educational digital games in learning English grammatical material. *J. Psycholinguist. Res.* **2023**, *52*, 1875–1895. [CrossRef] [PubMed]
16. Koch, J.; Gomse, M.; Schüppstuhl, T. Digital game-based examination for sensor placement in context of an Industry 4.0 lecture using the Unity 3D engine—A case study. *Procedia Manuf.* **2021**, *55*, 563–570. [CrossRef]
17. Sanz-Prieto, M.; de Pablo González, G. *Gamify gamifying: Learning with breakouts. Smart Pedagogy of Game-Based Learning: Advances in Game-Based Learning*; Daniela, L., Ed.; Springer: Cham, Switzerland, 2021; pp. 103–118.
18. Wang, M.; Zheng, X. Using game-based learning to support learning science: A study with middle school students. *Asia-Pac. Educ. Res.* **2021**, *30*, 167–176. [CrossRef]
19. Ke, F. Designing and integrating purposeful learning in game play: A systematic review. *Educ. Technol. Res. Dev.* **2016**, *64*, 219–244. [CrossRef]
20. Kurniawan, R.D.K.; Avrianto, R.P.; Indrajit, R.E.; Dazki, E. Selection the best quiz applications as learning performance evaluation media using the analytical hierarchical process method. *J. Tek. Inform. (JUTIF)* **2022**, *3*, 1389–1396. [CrossRef]
21. Uzunboylu, H.; Galimova, E.G.; Kurbanov, R.A.; Belyalova, A.M.; Deberdeeva, N.A.; Timofeeva, M. The views of the teacher candidates on the use of Kahoot as a gaming tool. *Int. J. Emerg. Technol. Learn.* **2020**, *15*, 158–168. [CrossRef]

22. Cabrera-Solano, P. Game-based learning in higher education: The pedagogical effect of genially games in English as a foreign language instruction. *Int. J. Educ. Methodol.* **2022**, *8*, 719–729. [CrossRef]
23. Carceller, M.d.C.; Serna-García, M.; López-Fernández, E.; Pellín Carcelén, A.; Flacco, N. Genially as an interactive tool in the integrated curriculum in basic biomedical sciences. In Proceedings of the INTED2023: 17th International Technology, Education and Development Conference Proceedings, Valencia, Spain, 6–8 March 2023; pp. 3992–3997.
24. De Souza, R.T.M.P.; Kasseboehmer, A.C. The thalidomide mystery: A digital escape room using Genially and WhatsApp for high school students. *J. Chem. Educ.* **2022**, *99*, 1132–1139. [CrossRef]
25. Sarva, E.; Gravelsina, E.; Daniela, L. Digital solutions for gamification and game-based learning from the perspective of educators. In Proceedings of the 18th International Technology, Education and Development Conference, Valencia, Spain, 4–6 March 2024; pp. 109–118. [CrossRef]
26. Yasin, A.; Fatima, R.; Wen, L.; Zheng, J.; Niazi, M. What goes wrong during phishing education? A probe into a game-based assessment with unfavorable results. *Entertain. Comput.* **2024**, *2024*, 100815. [CrossRef]
27. Tay, J.; Goh, Y.M.; Safiena, S.; Bound, H. Designing digital game-based learning for professional upskilling: A systematic literature review. *Comput. Educ.* **2022**, *184*, 104518. [CrossRef]
28. Sobocinski, M. I gamified my courses and I hate that. . .. *World J. Sci. Technol. Sustain. Dev.* **2017**, *14*, 135–142. [CrossRef]
29. Xu, Z.; Chen, Z.; Eutsler, L.; Geng, Z.; Kogut, A. A scoping review of digital game-based technology on English language learning. *Educ. Technol. Res. Dev.* **2020**, *68*, 877–904. [CrossRef]
30. Akatsu, H.; Shiima, Y.; Gomi, H.; Hegab, A.E.; Kobayashi, G.; Naka, T.; Ogino, M. Teaching "medical interview and physical examination" from the very beginning of medical school and using "escape rooms" during the final assessment: Achievements and educational impact in Japan. *BMC Med. Educ.* **2022**, *22*, 67. [CrossRef] [PubMed]
31. Frank-Bolton, P.; Simha, R. The Reverse Exam: A Gamified Exam Structure to Motivate Studying and Reduce Anxiety. In Proceedings of the 51st ACM Technical Symposium on Computer Science Education, Portland, OR, USA, 11–14 March 2020.
32. Martínez-Cal, J.; Sandoval-Hernández, I.; Ropero-Padilla, C.; Rodriguez-Arrastia, M.; González-Sánchez, M.; Molina-Torres, G. An escape room game-based innovation for the assessment of physiotherapy students: A qualitative study. *Stud. Educ. Eval.* **2024**, *81*, 101331. [CrossRef]
33. Loh, C.S.; Sheng, Y.; Ifenthaler, D. (Eds.) Serious games analytics: Theoretical framework. In *Serious Games Analytics: Methodologies for Performance Measurement, Assessment, and Improvement*; Springer: New York, NY, USA, 2015; pp. 3–29.
34. Niemelä, M.; Kärkkäinen, T.; Äyrämö, S.; Ronimus, M.; Richardson, U.; Lyytinen, H. Game learning analytics for understanding reading skills in transparent writing system. *Br. J. Educ. Technol.* **2020**, *51*, 2376–2390. [CrossRef]
35. Kim, Y.J.; Ifenthaler, D. (Eds.) Game-based assessment: The past ten years and moving forward. In *Game-Based Assessment Revisited: Advances in Game-Based Learning*; Springer: Cham, Switzerland, 2019; pp. 3–11.
36. Baumgartner, E.; Bell, P.; Brophy, S.; Hoadley, C.; His, S.; Joseph, D.; Orrill, C.; Puntambekar, S.; Sandoval, W.; Tabak, I. Design-based research: An emerging paradigm for educational inquiry. *Educ. Res.* **2003**, *32*, 5–8. [CrossRef]
37. Pors, S. *Conducting Design-Based Research on Web-Enhanced Learning in Cambodia*; Sage Research Methods Cases Part 1; SAGE Publications, Ltd.: Thousand Oaks, CA, USA, 2014. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

computers

MDPI

Review

Intelligent Tutoring Systems in Mathematics Education: A Systematic Literature Review Using the Substitution, Augmentation, Modification, Redefinition Model

Taekwon Son

The Department of Mathematics Education, Korea National University of Education, Chengju 28173, Republic of Korea; sontaekwon7@gmail.com

Abstract: Scholars have claimed that artificial intelligence can be used in education to transform learning. However, there is insufficient evidence on whether intelligent tutoring systems (ITSs), a representative form of artificial intelligence in education, has transformed the teaching and learning of mathematics. To fill this gap, this systematic review was conducted to examine empirical studies from 2003 to 2023 that used ITSs in mathematics education. Technology integration was coded using the substitution, augmentation, modification, redefinition (SAMR) model, which was extended to suit ITSs in a mathematics education context. How different contexts and teacher roles are intertwined with SAMR levels were examined. The results show that while ITSs in mathematics education primarily augmented existing learning, recent ITS studies have transformed students' learning experiences. ITSs were most commonly applied at the elementary school level, and most ITS studies focused on the areas of number and arithmetic, algebra, and geometry. The level of SAMR varied depending on the research purpose, and ITS studies in mathematics education were mainly conducted in a way that minimized teacher intervention. The results of this study suggest that the affordance of an ITS, the educational context, and the teacher's role should be considered simultaneously to demonstrate the transformative power of ITSs in mathematics education.

Keywords: technology integration; mathematics education; SAMR model; intelligent tutoring systems

MSC: 97U50

1. Introduction

The use of technology in mathematics education is recommended to form a problem-solving and inquiry-based educational environment in terms of constructivism and epistemological approaches [1]. To address this purpose, mathematics educators should use technology in a transformative way rather than simply assimilated into traditional instruction practices [2–5]. Artificial Intelligence (AI) technology has developed rapidly, AI in Education (AIED) can support student learning through customized learning, realize an educational design that is impossible in existing educational modes, and provide new opportunities, potential, and challenges for educational innovation [6–8]. Intelligent tutoring systems (ITSs) are a representative form of transformative application in AIED [9]. An ITS is an AI application that aims to provide immediate, personalized instruction or feedback to learners without the intervention of a human teacher [10]. An ITS has the potential to play a pivotal role in supporting and supplementing the individualized educational needs of learners in mathematics education, such as personalized learning paths [11], problem-solving steps [12], and scaffolding [13].

However, despite abundant empirical evidence that ITSs are effective in mathematics learning, several scholars have pointed out that ITSs have not reached their full perceived potential in mathematics education [7,14,15]. For example, some ITSs do not provide

constructivist opportunities for learners to construct their own knowledge but simply replace or augment traditional practices [15]. Therefore, it not only matters to integrate ITSs in mathematics learning, but it matters how instructors integrate them, because the use of ITSs may otherwise continue to reflect traditional teaching methods [16]. Scholars have also postulated the use of AIED can transform learning (e.g., [6–8]), but there is insufficient evidence to determine whether the ITSs at the heart of AIED are transforming mathematics learning or are being used to replicate past practices. A literature review that systematically synthesizes existing empirical evidence can help mathematics education stakeholders make more informed decisions. Existing literature reviews and meta-analyses have addressed ITSs in education, including the effect of ITSs [9], characteristics, application, and evaluation of ITSs [17], and the effectiveness of K–12 mathematics learning [18]. However, little research has addressed explicitly how ITSs are integrated into mathematics education and what challenges impede the integration process.

To address this gap, this study systematically analyzed 20 years of empirical studies on the use of ITSs in mathematics education using the substitution, augmentation, modification, redefinition (SAMR) model [19], which is a taxonomy-based approach for selecting, using, and evaluating technology in educational settings. This model is suitable for describing the level of technology integration in a specific educational context [1]. The present study used the SAMR model to investigate whether ITSs were used to change learning in mathematics education or to replicate traditional learning. The findings provide educators and researchers with insight into why the potential of ITSs has not yet been fully realized in mathematics education and guide future research directions.

2. Background

2.1. Technology Integration in Mathematics Education

Integrating technology into mathematics education has the potential to support student learning and fundamentally transform teaching and learning [1,20]. Cheung and Slavin [21] conducted a meta-analysis of the effects of educational technology applications on mathematics achievement in K–12 classrooms over 30 years. Cheung and Slavin found that technology had a positive effect on student achievement compared to traditional methods, with differential impacts depending on the type of educational technology used. Advances in various information, communication, and computing technologies (e.g., mobile technology; [22]), information and communication technology (ICT) [23], and augmented reality [24] have provided new opportunities for learning mathematics. However, a recent literature review of empirical studies reported that technology use has not lived up to its perceived potential to transform mathematics learning experiences [1]. Students most often experience digital technologies to enhance traditional practices in the context of mathematics education. When technology is used primarily as training and practice in mathematics learning, its impact on student achievement may be minimal or even negative [25,26].

Several scholars have recommended that it is better to use technology for tasks that change significantly using technology rather than tasks that can be performed without technology [1–3,5]. In these environments, students have the opportunity to use mathematics for productive purposes rather than following learning procedures according to predefined scenarios [5]. However, technology does not fulfill its potential on its own, instead requiring instructive, careful implementation to realize its potential and avoid negative impacts [27]. Drijvers [28] found the affordances of the technology, the educational context, and the teacher's role while using technology are important factors in integrating technology into mathematics classrooms successfully. Drijvers emphasized that three elements should be interrelated to make technology work in mathematics education; to do this, teachers should orchestrate learning by (a) synthesizing the results of technology-rich activities, (b) helping students develop ways to use tools effectively, and (c) relating experiences within the technology environment to mathematical activities [29]. Learners will not use the technology independently in ways that result in positive learning gains [30].

2.2. ITSs in Mathematics Education

An ITS is a type of educational technology designed to support one-on-one adaptive learning by providing feedback and hints [31]. Typically, an ITS features a user interface that presents students with instructional material and allows them to answer structured questions, often breaking each question into several steps to prevent student failure [17]. Since the introduction of the first ITS by Jaime R. Carbonell in 1970, ITSs have been used in mathematics education to provide personalized learning activities tailored to learners' characteristics and needs. The growing interest in ITSs within mathematics education is largely driven by the limitations of human resources and time, as it is often difficult for a single teacher to meet the educational needs of all students [32]. From an economic perspective, ITSs hold the potential to address this issue.

With recent advancements in AI, ITSs have evolved into adaptive learning systems that incorporate AI technologies. These new AI-based systems guide learners through various stages of learning by generating hints and feedback from expert knowledge databases as needed [33]. As AI continues to be integrated into educational contexts, ITSs have emerged as one of the most prominent fields where AI is applied in mathematics education [34]. The primary objective and strength of AI-driven ITSs lie in their ability to assess learners, define models, and provide appropriate feedback. These unique functional competencies can be combined with using games [35], fostering exploratory learning environments [15], and working with multiple representations [36] to form the basis of opportunities for driving curriculum changes, assessment changes, and pedagogical changes [37]. ITSs may offer a chance to reshape how mathematics is assessed, taught, and learned.

However, applying ITSs in mathematics education does not guarantee positive educational outcomes. It is well known that students do not always use technology in ways that lead to positive learning benefits [30]. Despite the unique affordances of ITSs and their potential to transform learning experiences, the innovative potential of ITSs is not always fully realized in mathematics learning [33]. Several scholars have noted that mathematics activities are often not designed to take full advantage of ITSs, with the technology primarily replacing traditional instructional practices [7]. For instance, ITSs typically diagnose learners and provide feedback while structuring the learning process into multiple steps to prevent failure. This design feature minimizes failure and adheres to standardized instructional methods, potentially overlooking the value of productive failure and alternative instructional approaches, such as collaborative learning, inquiry-based learning, and group discussions [7]. Therefore, while ITSs offer unique potential and limitations and the opportunity to fundamentally transform learning, fully realizing these innovative functions requires a deeper understanding of how ITSs can be effectively utilized in educational settings (e.g., [7,14,15]).

2.3. SAMR Model

Several perspectives describe technology integration, and one representative model that classifies technology integration in educational contexts is the SAMR model [19]. The SAMR model was developed based on the technological, pedagogical, content knowledge (TPACK) framework [38] and the replacement, amplification, and transformation (RAT) framework [39].

The SAMR model categorizes the degree of technology integration into two ranges and four levels within each range (see Figure 1). This classification reflects the view that technology integration in education falls into one exclusive category [40]. The substitution level is a situation in which technology is used to directly substitute for traditional methods without functional or conceptual changes, such as transferring the learning content to be taught through technology. In the augmentation level, technology substitutes for traditional teaching methods, but with functional improvements. It is possible to change small-scale improvement, which does not imply strong changes to teaching and learning at this level. In the modification level, technology allows for the redesign of typical lesson goals, activities, and tasks. Physical barriers in the classroom are eliminated as students' expressions of

their positions become more flexible and closer to learning that occurs in a variety of contexts. Finally, the redefinition level is the use of technology for learning in ways that could not be implemented without technology. Two levels at the bottom represent using technology to enhance learning, and two levels at the top represent using technology to transform learning.

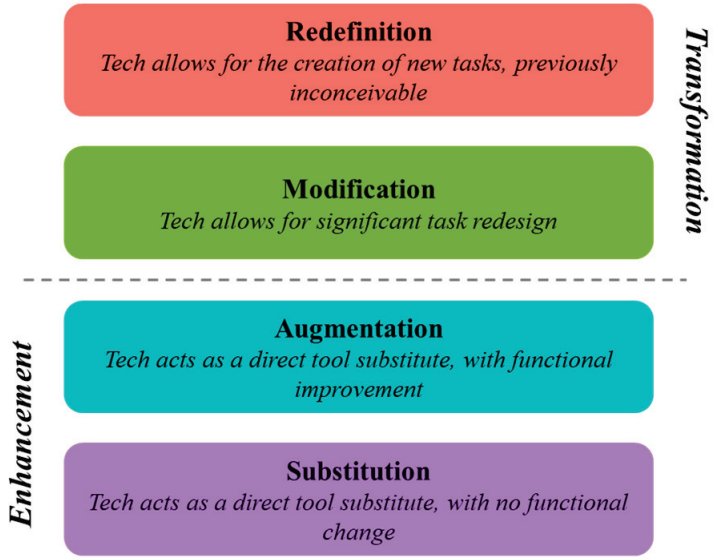


Figure 1. Hierarchy of the SAMR model (adapted from Puentedura, [19]).

The SAMR model reflects the idea that when teaching and learning mathematics, technology should not simply replicate classroom practices but provide opportunities to use mathematics for more productive purposes [1,5]. Therefore, several scholars recommended that it is better to use technology for tasks that are significantly changed by the use of technology rather than tasks that can be performed without technology [1–3,5]. In these environments, students have the opportunity to use mathematics for real productive purposes rather than following learning procedures according to predefined scenarios [5]. Based on this perspective, the SAMR model provided a classification standard for integrating technology into mathematics education, such as a literature review of technology [1], interactive whiteboard [41], and digital technology [42]. Researchers have found that the use of technology in mathematics education has been primarily at the replacement and augmentation level (e.g., [1,43]). These results suggest that technology does not fulfill its potential in mathematics education and that educational and careful implementation is needed to fulfill its potential and avoid negative impacts [27].

The SAMR model can lead to misunderstanding and confusion in explaining or understanding how to understand, interpret, and apply the levels due to subjective interpretation [40]. Accordingly, some researchers have expanded the SAMR model to fit specific research contexts, such as learning processes [44], Bloom's taxonomy [45], and teaching and learning in AIED [46]. In this study, the SAMR model was extended in the context of ITS use in mathematics education. The extended SAMR model for ITSs of mathematics education is described in the methods section.

2.4. Review of Previous Research

As the publication of ITS research papers in the field of education increases, it is necessary to organize and review related issues systematically. Table 1 outlines the existing literature review and meta-analysis studies that analyzed ITS-related research papers in the field of education and provides comprehensive results.

Table 1. Previous studies on ITSs in education.

Author	Number of Studies	Timeline	Focus of Review
Wang et al. [9]	40	2011–2022	Effectiveness on studies that applied social experiment methods
Mousavinasab et al. [17]	53	2007–2017	Characteristics, applications, and evaluation
Kulik and Fletcher [33]	50	1990–2013	Effectiveness in education
Steenbergen-Hu and Cooper [47]	35	1990–2011	Effectiveness for college students
Steenbergen-Hu and Cooper [18]	26	1997–2010	Effectiveness on K–12 students' mathematical learning
VanLehn [48]	28	1975–2010	Comparing the effectiveness of human tutoring, computer tutoring, and no tutoring

Some studies have focused on examining the effectiveness of ITSs across the education field, and the results demonstrated the value of ITS integration in education. For example, VanLehn [48] conducted a meta-analysis of 28 articles and compared the effectiveness of human tutoring, computer tutoring, and no tutoring. VanLehn reported that ITSs are nearly as effective as human tutoring in increasing learning outcomes in STEM subjects, but ITS should not be used to replace the full classroom experience. Steenbergen-Hu and Cooper [47] conducted a meta-analysis of 35 studies in higher education. They reported that ITSs had a moderate positive effect on college students' academic learning, and although less effective than human tutoring, ITSs outperformed other teaching methods, such as traditional classroom instruction, computer-assisted instruction, and homework.

Some studies have addressed the overall trend in ITSs in education. Mousavinasab et al. [17] reviewed the characteristics, application, and evaluation of ITSs in 53 articles. Wang et al. [9] synthesized 40 empirical studies that examined the effects of ITSs on teaching and learning through social experiments. Other studies have focused on examining the effectiveness of ITSs in mathematics education. Steenbergen-Hu and Cooper [18] conducted a meta-analysis of studies of 34 ITSs on K–12 mathematics learning published from 1997 to 2010, concluding that ITSs did not have a negative effect on mathematics learning and had a small positive effect. Kulik and Fletcher [33] conducted a meta-analysis of 50 articles on K–12 mathematics learning published from 1990 to 2013 and reported that ITSs had stronger effects than the typical outcomes of another private tutoring source.

As a result, existing studies have mainly focused on the effectiveness and overall trends in ITSs in the field of education, particularly the effectiveness of ITSs in the mathematics subject. The findings of these existing studies suggest that the integration of ITSs in the field of education and mathematics education is effective. However, none of the mentioned studies focused on the integration of ITSs in mathematics education. Research examining the degree of integration of ITSs in mathematics education can help researchers understand how ITSs inform the design of mathematics instruction and guide future research that can transform mathematics teaching and learning. Moreover, research focusing on ITSs in mathematics education is very limited. Most studies have addressed the use of ITSs across educational sectors; some studies focused on mathematics education and ITSs (e.g., [18,33]), but they conducted meta-analyses examining effectiveness rather than using a literature review approach. Furthermore, no ITS studies in mathematics education were systematically reviewed after 2013. With the recent rapid advancements and innovations in ITS studies, these reviews may subsequently be outdated. Because the number of ITS studies in education has increased significantly over the past decade, it is timely to conduct a comprehensive and critical review of the extent to which ITSs have been integrated into mathematics education, including the latest research.

2.5. Research Questions

The purpose of this study was to present a systematic review of ITS studies in mathematics education from 2003 to 2023 and reveal how ITSs are used in mathematics education

based on the SAMR model. The relevance of the SAMR model to contextualize the teacher's role of implementing ITSs is emphasized. Three research questions guided this study:

RQ1. What levels of the SAMR levels are ITS studies integrated into mathematics education?

RQ2. What are the trends in ITS studies in mathematics education regarding the SAMR levels across contexts (e.g., publication year, educational level, mathematics domain, research purpose)?

RQ3. What are the trends in ITS studies in mathematics education regarding the SAMR levels across teachers' roles?

3. Methods

The purpose of a systematic review is to answer specific questions based on an explicit, systematic, and replicable search strategy using inclusion and exclusion criteria that can identify studies that should be included or excluded [49]. This systematic review drew on recently published literature and included empirical studies on the use of ITSs in mathematics education between 2003 and 2023. The literature screening process followed the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) proposed by Page et al. [50].

3.1. Search Strategy

The study search was conducted using the databases of Education Resources Information Center (ERIC), Web of Science, and EBSCO Education Source, which are widely used in the field of educational technology. There are other databases, such as IEEE Xplore, Scopus, and ProQuest, but they were not included in the initial literature search because their scope overlaps with the three selected databases [51]. To obtain comprehensive data, search terms were created by referring to the related search terms used in other studies [9,18]. The Boolean search is listed in Table 2. The search was conducted on March 1, 2024, and the initial search yielded 4219 results. The search string was formulated as follows to include as many ITS-related articles in mathematics education as possible: (intelligent * OR adaptive OR customized) AND (learning OR instruction OR tutoring OR mentoring) AND (system OR software OR application) AND (math OR maths OR mathematics OR mathematics education). This text string encompasses all studies addressing ITS in mathematics education in this systematic review. Search terms were separated by commas, and the asterisk character (*) was used to capture variations of intelligent, intelligence, etc.

Table 2. Searching strings.

A	B	C	D
1. intelligent * 2. adaptive 3. customized	1. learning 2. instruction 3. tutoring 4. mentoring	1. system 2. software 3. application	1. math 2. maths 3. mathematics 4. mathematics education

Note. The asterisk (*) was used to broaden a search.

3.2. Study Selection

The obtained articles were screened according the inclusion and exclusion criteria presented in Table 3 after removing 519 duplicate articles and 1389 articles published before 2003 using EndNote reference management software. The period was set from 2003–2023, as ITSs have been very actively researched over the past 20 years and have produced many publications. Only empirical studies in which ITSs were applied in a mathematics education setting and the evaluation results were revealed were included. Articles that were not published in journals were excluded because it was unclear whether they had undergone peer review, an acceptable standard for ensuring scholarly quality.

Table 3. Inclusion and exclusion criteria.

Inclusion Criteria	Exclusion Criteria
Published 2003–2023	Published before 2003
English language	Not in English
Empirical research with evaluation results	Not empirical research (e.g., commentary) or only proposing ITS design solutions or ITSs without offering evaluation results
Article has been peer reviewed	Article has not been peer reviewed
Journal article	Not a journal article
ITS application in mathematics educational setting	No mathematics educational setting
Articles in which the full text was available.	Articles in which the full text was not available.

According to the established inclusion and exclusion criteria, the title and abstract of the paper were reviewed, and the full text was secondarily reviewed. In the first screening, two separate reviewers read the titles and abstracts of all papers and coded them into three groups: (a) included, (b) ambiguous, and (c) excluded. Both reviewers had extensive publishing experience in performing coding. Coding agreement between reviewers was found to be 92%, and papers that did not receive the same coding reached 100% agreement through discussion. An initial review of 2311 studies resulted in the removal of 2007 articles. The remaining 304 articles were retrieved in full text for Stage 2 screening and evaluated by two reviewers. In this process, 21 articles in which the full text was not freely retrievable were excluded. In addition, 225 articles were excluded because the contents of the full text referred superficially to mathematics education or focused on the development process of ITSs rather than the context of mathematics education. Ultimately, 58 articles were identified as meeting the current study goals. The literature search and review procedure based on PRISMA is presented in Figure 2.

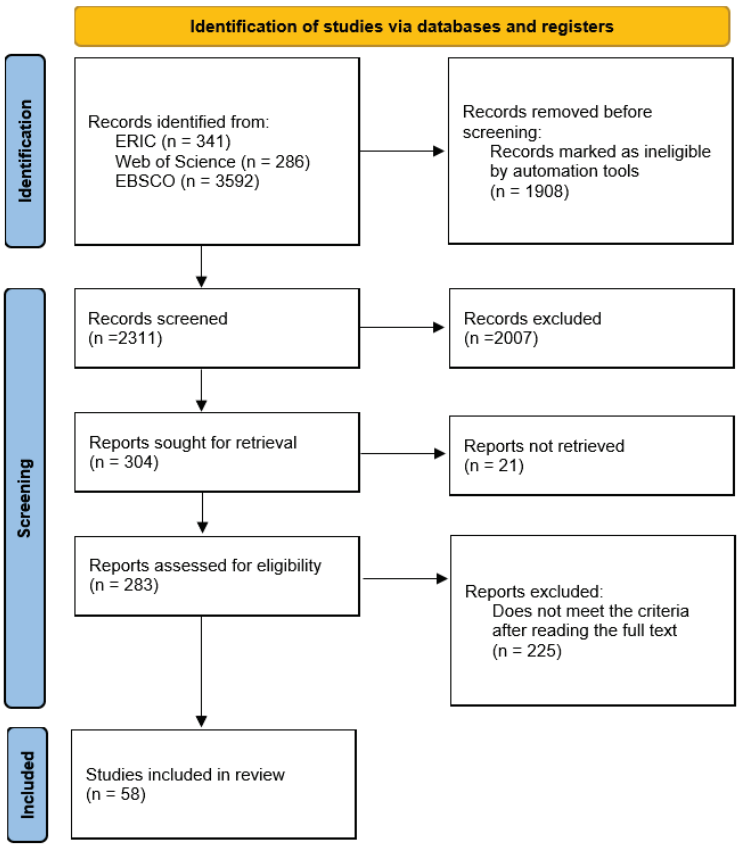


Figure 2. A diagrammatic representation of the literature search and review process based on the PRISMA recommendation statement.

3.3. Analysis

An electronic data extraction form based on a coding scheme was designed to reduce the potential for bias [52]. The 58 articles included in the literature review were coded based on the research questions and the background section: (a) contexts (i.e., publication year, educational level, mathematics domain, and research purpose), (b) SAMR levels, and (c) teachers' roles.

3.3.1. Contexts

In "contexts," data on the publication year, education level, mathematics domains, and research purpose were collected by reviewing 58 articles. Most of the contexts are self-explanatory; however, "research purpose" included five purposes for conducting research (see Table 4).

Table 4. Classification of research purpose.

Categories	Description	Example
Development of ITS	Focus on developing and validating new ITSs	Pai et al. [53]
Improvement of ITS	Focus on improving learning productivity by integrating other ITSs or adding new functions to an existing ITS.	Nye et al. [54]
Application of existing ITS	Focus on applying existing ITSs to a specific educational context or mathematics domain and confirming its effectiveness.	Huang et al. [55]
Investigation of factors	Focus on examining factors affecting learning mathematics using ITSs.	San Pedro et al. [56]
Exploring teaching methods	Focus on exploring effective teaching methods in mathematics learning using ITSs.	Cung et al. [57]

3.3.2. SAMR Levels

The present study developed the SAMR framework for the technology integration classification of ITSs in mathematics education. Table 5 shows the extended SAMR model for ITSs in mathematics education by referring to Crompton and Burke [44]. This framework was used to code the level of ITS integration in mathematics education.

Table 5. SAMR model for ITSs.

Level	Existing Definition	Mathematics Education and ITS Context
Redefinition	Tech allows for the creation of new tasks, previously inconceivable.	Teachers and students use ITSs to implement the teaching and learning of mathematics that cannot be inconceivable without ITSs.
Modification	Tech allows for significant task redesign.	Teachers and students can use ITSs to redesign the goals and tasks of teaching and learning of mathematics.
Augmentation	Tech acts as a direct tool substitute with functional improvement.	The ITS acts the same as teaching and learning of mathematics, in that teachers and students can implement it in classrooms, but with functional improvements.
Substitution	Tech acts as a direct tool substitute, with no functional change.	ITSs act the same in the teaching and learning of mathematics, in that teachers and students can implement it in classrooms.

At the substitution level, an ITS directly substitutes for traditional classes without functional or conceptual changes. For example, the ITS can be used simply to provide practice problems or content to be taught. At the augmentation level, the ITS substitutes for the teacher's role in a typical classroom, but with functional improvements. For example, researchers can use ITSs to analyze learner performance and provide appropriately tailored content. In a real classroom, it is difficult for a single teacher to respond individually to the individual needs of multiple students, so an ITS is a functional improvement over the role played by traditional teachers. At the modification level, ITSs can redesign the objectives, activities, and assignments of mathematics classes. Learners can use the ITS to design their own tasks or conduct learning in a game-based ITS. In this type of technology integration, the goals and tasks of mathematics classes are redesigned based on learner

autonomy. Finally, at the redefinition level, ITS can create new tasks that are inconceivable in traditional mathematics classes. For example, an ITS can act as a virtual student seeking learning, and a student can learn by teaching the virtual student. This type of learning is impossible to achieve without ITSs, as students with the same level of knowledge and background context cannot exist in a real-context classroom. Two levels at the bottom are using technology to enhance learning (i.e., substitution and augmentation levels), and two levels at the top are using technology to transform learning.

3.3.3. Teachers' Roles

Kessler et al. [58] found four types of roles in which teachers intervene in interactions between students and ITSs in an educational environment using Cognitive Tutor. In this study, three out of four types of teachers' roles were identified, with the exception of the role where the teacher directly interacted with the ITS on behalf of the student: (a) the ITS environment as designed, (b) teacher facilitating the ITS environment, and (c) teacher facilitating mathematics. In an ITS environment as the designed role, teachers either have no involvement in students' mathematics learning or only technically supervise it, and pedagogical interaction between students and teachers is minimal. All mathematics teaching and learning processes occur between ICTs and students. In the teacher facilitating the ITS environment role, the teacher assists the student using the ITS through the process and provides feedback upon the student's help. Although the mathematical teaching and learning process still occurs between students and the ITS, teachers intervene in this process as mediators and play a role in facilitating students' ITS learning. In the teacher facilitating mathematics role, teachers lead mathematics learning through ITSs from lesson design to lesson conducting. Teachers determine what learning content is provided to students in the ITS or connect ITS learning with meaningful mathematical activities outside of the ITS learning environment. Teachers interact with students about their specific mathematical thinking, and students can physically move away from the ITS.

3.3.4. Coding

Table 6 summarizes the coding framework in three categories (i.e., contexts, SAMR levels, and teachers' roles). Two coders coded 58 articles independently based on the coding framework and compared the results. The agreement rate for the initial coding was 92%. The two coders met online to review articles with inconsistent coding and reached 100% agreement through discussion.

Table 6. Coding framework.

Category	Code	Example
Context	Publication year	2013
	Educational level	Elementary School
	Mathematics domain	Algebra
	Research Purpose	Development of ITS
SAMR level	Substitution	Substitution
	Augmentation	
	Modification	
	Redefinition	
Teachers' roles	ITS environment as designed	ITS environment as designed
	Teacher facilitating the ITS environment	
	Teacher facilitating mathematics	

4. Results

A total of 58 studies were finally reviewed and coded based on the coding scheme. Coding results are presented in the Appendix A. This section presents overall trends in the studies reviewed and findings for each of the three research questions that guided the systematic review.

4.1. RQ1. What Level of the SAMR Levels Are ITS Studies Integrated into Mathematics Education?

All four levels of the SAMR framework were identified in the reviewed 58 studies (see Figure 3). The substitution level comprised 1.7% (n = 1) of the total article. The ITS substituted for the teacher's role without functional improvements at this level, such as presenting problems to students and providing feedback. For example, del Olmo-Muñoz et al. [14] used an ITS called HINTS (Hypergraph-based INtelligent Tutoring System) to address the gap in low-income households. Students solved practice problems provided by the system, read solutions, or solved similar additional problems. To prevent students from relying on messages from the ITS, aids on-demand was prevented. This did not seem to have any added value compared to teachers giving students worksheets with practice problems and providing additional feedback.

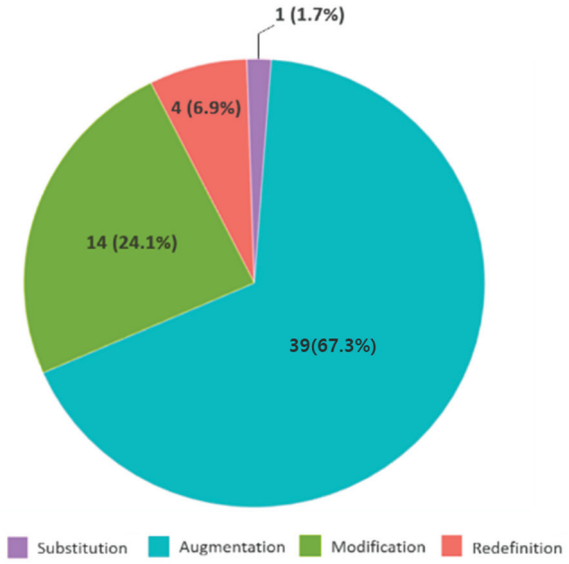


Figure 3. Frequency of studies coded in the SAMR model.

The augmentation level comprised the largest proportion of reviewed articles at 67.3% (n = 39). This result indicates that the ITS has been used to replace traditional approaches in mathematics education with some functional or conceptual improvements directly. For example, Walkington and Bernacki [59] applied Cognitive Tutor to learning algebra, which provides customized problems tailored to students' individual interests. Students learned mathematics according to the learning path provided by the ITS, which did not involve lesson redesign or self-inquiry learning.

The modification level comprised 24.1% (n = 14) of total articles. For example, Çetin et al. [35] developed ArtiBos, which allows students to perform problem-posing activities based on the information provided by the system. Students were able to select appropriate questions and problems and solve and evaluate them following Polya's problem-solving steps. Students are given autonomy to redesign course objectives and tasks rather than learning according to a rigid ITS structure at this level.

The redefinition level comprised 6.9% (n = 4) of all articles. Matsuda et al. [60] used SimStudent, which enables learning by teaching, to provide students with an artificial peer learning environment. The ITS acted as a student seeking learning, and the student learned by teaching the ITS. Implementing such learning is impossible without technology, as real-world educational contexts cannot produce students with the same level of knowledge and background.

In summary, most of the reviewed literature used ITSs for the substitution and augmentation of mathematics learning. It is desirable to use the technology for tasks that fit the two higher levels of the SAMR hierarchy [1,5]. However, only 31% of the reviewed

literature used ITSs in this way, and the redefinition level included only 6.9%. Despite the smaller number of the transformation range compared to the enhancement range, studies classified as modification and redefinition indicate ongoing efforts to promote change in mathematics learning.

4.2. RQ2. What Are the Trends of ITS Studies in Mathematics Education Regarding the SAMR Levels across Contexts?

4.2.1. Publication Year

Figure 4 illustrates the change in the number of published articles from 2003 to 2023. There was a peak in 2019 in terms of the number of published studies on ITSs in mathematics education. The ITS studies were categorized into two time periods per decade unit, that is, 2003–2012 and 2013–2023. The number of publications in the latter period (n = 44, 75.9%) was over 3 times that of the previous period. This shows that ITS research in mathematics education has grown rapidly over the past 10 years.

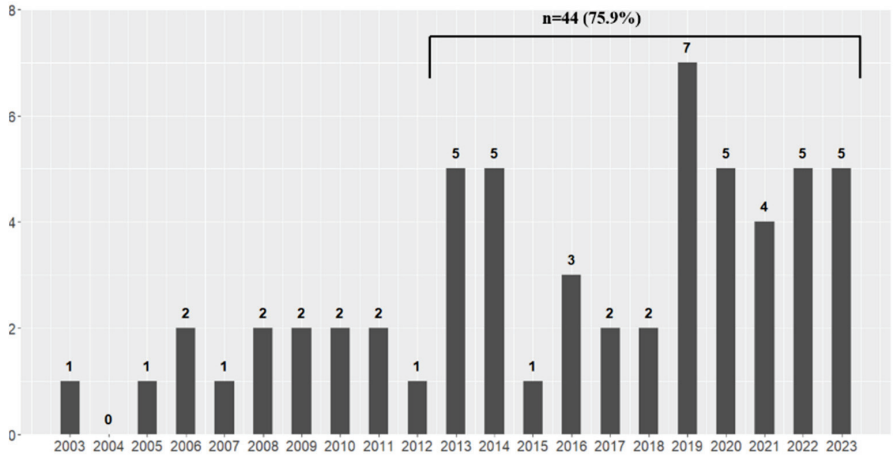


Figure 4. Trend in changes in the number of papers from 2003 to 2023.

Figure 5 shows trends in changes in SAMR levels across published studies by decade. In the first decade, ITSs were integrated into mathematics education mainly at the augmentation level (n = 12, 85.7%). There were two studies (14.3%) on the modification level, and no study on the transformation range was conducted. On the other hand, in the latter decade, several studies were conducted at the level of modification 14.3% (n = 12) and redefinition 8.5% (n = 4). These results suggest most ITS research in the transformation range in mathematics teaching and learning has been conducted in the last decade.

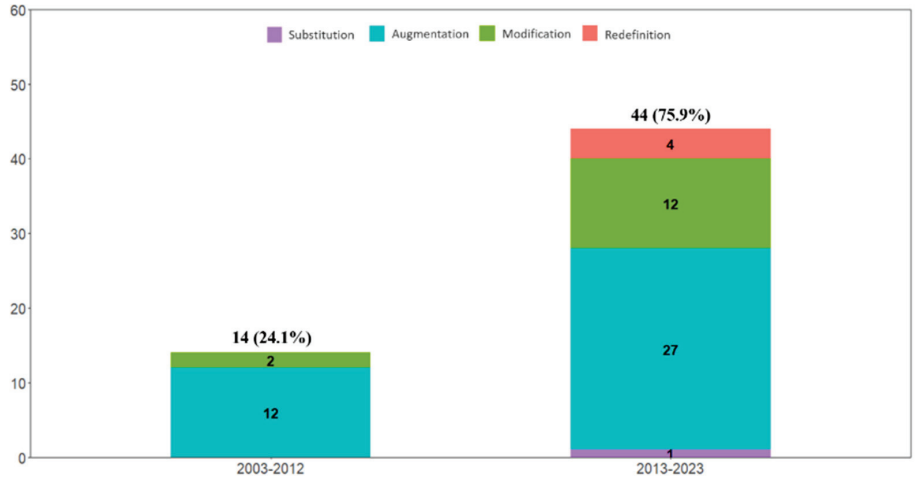


Figure 5. Trends in changes in the number of papers per decade unit.

4.2.2. Educational Levels

Figure 6 shows the distribution of SAMR levels across educational levels covered in the reviewed literature. Some studies ([61,62]) applied ITSs to multiple educational levels, so 60 educational levels were included. The most ITS research was conducted in elementary school at 41.7% (n = 25), followed by middle school at 25% (n = 15), high school at 18.3% (n = 11), colleges at 13.3% (n = 8), and teacher at 1.6% (n = 1). Most studies were conducted in the context of K–12 mathematics learning; however, there was no ITS study conducted at the Pre–K educational level, and the fewest ITS studies were conducted for teacher education.

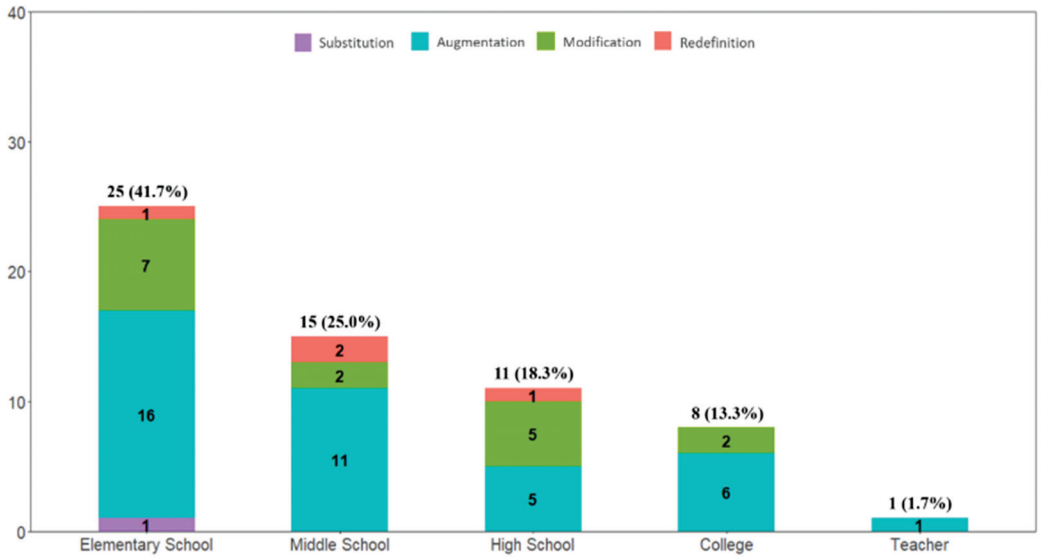


Figure 6. SAMR distribution across educational levels.

Across all educational levels, researchers used ITSs primarily at the augmentation level (elementary school 61.5%, middle school 73.3%, high school 50%, college 75%, teachers 100%). This shows that mathematics teaching and learning at educational levels are not fundamentally changed by ITSs but are mainly enhanced. However, studies at the modification level or redefinition level were also conducted at diverse educational levels. Specifically, at the high school education level, studies at the transformation range (modification 41.7%, redefinition 8.3%) were conducted similarly to the augmentation level (50%). For example, Walker et al. [63] applied adaptive collaborative learning support, which provides collaborative learning environments to promote collaborative interactions among individual students. Students became tutors and provided feedback on the problem-solving process of their tutee peers and received feedback from the system on their own reflective behavior. These findings provide evidence for multiple examples in which an ITS has been transformative for mathematics learning. These researchers did not substitute for traditional practices but rather focused on pedagogical practices that go beyond existing teaching methods.

4.2.3. Mathematics Domain

Figure 7 shows the distribution of SAMR levels across the mathematics domains. The reviewed studies covered six content domains of mathematics education. Most studies were conducted around the domains of numbers and arithmetic (38.7%), algebra (32.3%), and geometry (11.3%). Seven studies classified as 'Not specified' did not explicitly describe to which mathematical domain the ITS was applied. In some studies, the ITS applied to multiple domain contexts (e.g., [12,64]), and the total number of mathematical domains was 62. For example, Chang et al. [12] applied ITSs to multiple domains, such as the operation of fractions, the area of triangles, and unit cost.

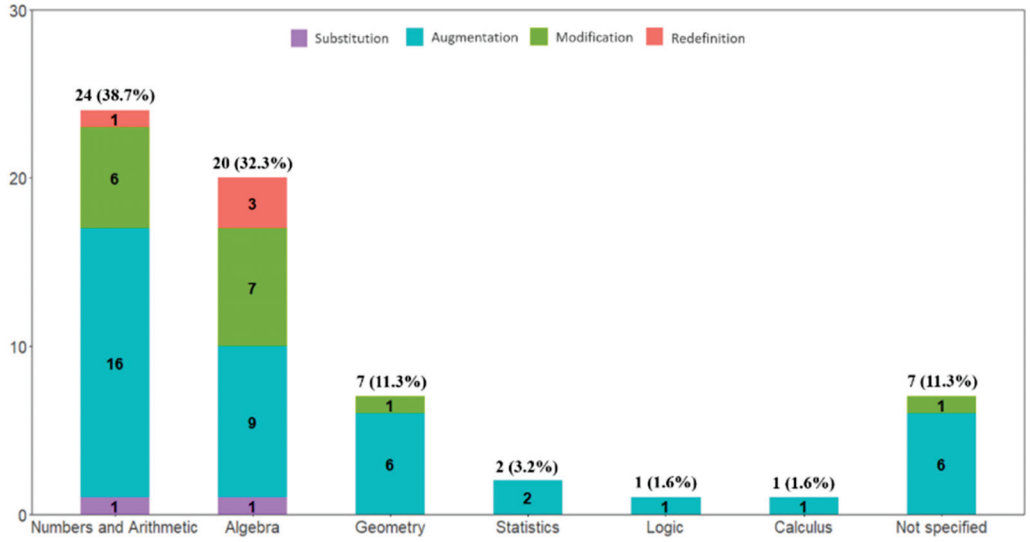


Figure 7. SAMR distribution across mathematics domains.

In numbers and arithmetic, ITSs augmented mathematics learning primarily through problems solutions, and feedback to practice basic arithmetic and fraction operations (e.g., [65–67]). This type of ITS presented students with mathematics problems step by step, taking individual differences into account and providing immediate feedback. On the other hand, some studies have transformed mathematics learning in the domain of numbers and arithmetic using ITSs. For example, Kong and Kwok [68] applied ITSs to learn fraction equivalence. Students were able to engage in exploratory activities that allowed them to formulate and test hypotheses, and the ITS provided a profitable space for interaction. In algebra, the ITS was most frequently used at the modification level (n = 7; 35%) and redefinition level (n = 3; 15%) compared to other domains. The ITS at these levels provided an opportunity to redesign modeling lessons [69], support self-regulated learning by applying an open-learner model [70], and implement artificial teachable agents that would be difficult to inconceivable without an ITS [60]. In geometry, ITSs were most often used at the augmentation (n = 6; 85.7%) level, and there was only one study that applied ITSs at the modification level. Butcher and Aleven [71] applied ITSs to promote rule-diagram mapping where students connect knowledge to diagrams. Students first made solvability decisions through a reasoning process about a given problem, specifying the geometric rules and diagram features that they would use to arrive at the correct answer. Statistics, logic, and calculus had fewer studies than other domains. Although it is difficult to draw generalizable conclusions about SAMR levels in the three domains due to the limited number of studies, all three domains have been studied at the augmentation level.

4.2.4. Research Purpose

Figure 8 shows that the reviewed literature conducted various studies with five research purposes. "Development of ITS" was the research purpose of most studies (n = 17; 28.3%). This finding means researchers have paid the most attention to developing new ITSs in mathematics education for two decades. Although there are slight differences in the number of studies conducted with each research purpose set, they are generally evenly distributed. This result indicates that research on ITSs in mathematics education has been conducted for various purposes without being biased.

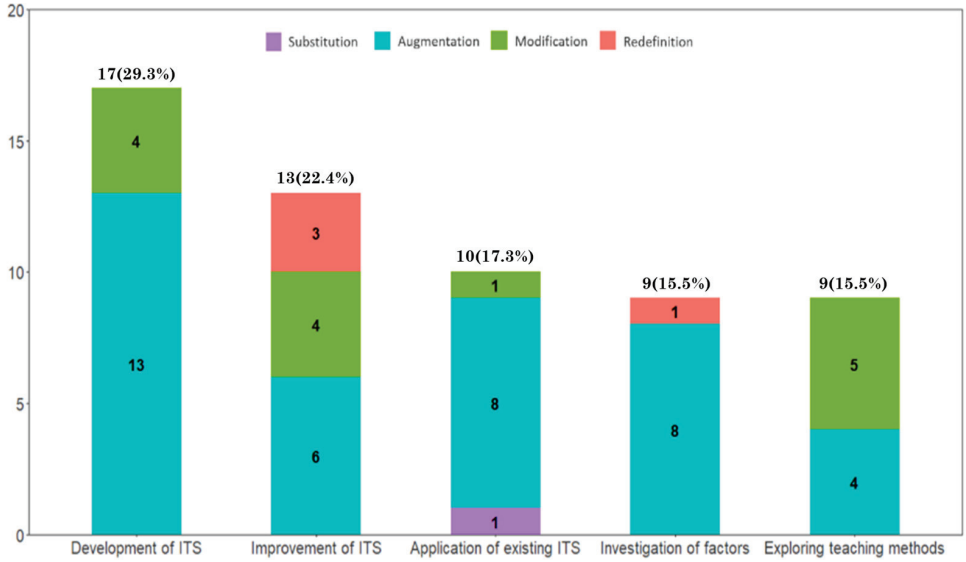


Figure 8. SAMR distribution across the research purpose.

4.3. RQ3. What Are the Trends of ITS Studies in Mathematics Education Regarding the SAMR Levels across Teachers' Roles?

Figure 9 shows the distribution of SAMR levels across teachers' roles. One ITS study for teachers was excluded from the analysis [72]. Some studies reported diverse roles for teachers [66,73]. For example, Phillips et al. [73] reported that although teachers were encouraged to consider different models for blended learning, most teachers did not intervene in ITS learning, and only some teachers integrated ITSs into mathematics learning. As a result, a total of 59 papers were analyzed. The teacher's role is predominantly one of managing or not intervening in learning between students and ITSs (72.9%). There were 16.9% (n = 10) of studies in which teachers played a role in facilitating learning between students and the ITS, and the role of teachers in leading ITS learning and connecting ITS learning with external mathematics activities accounted for the lowest percentage of all studies (n = 6; 10.2%). These results implicitly reveal the dominant perspective that ITSs can replace the role of teachers in mathematics education.

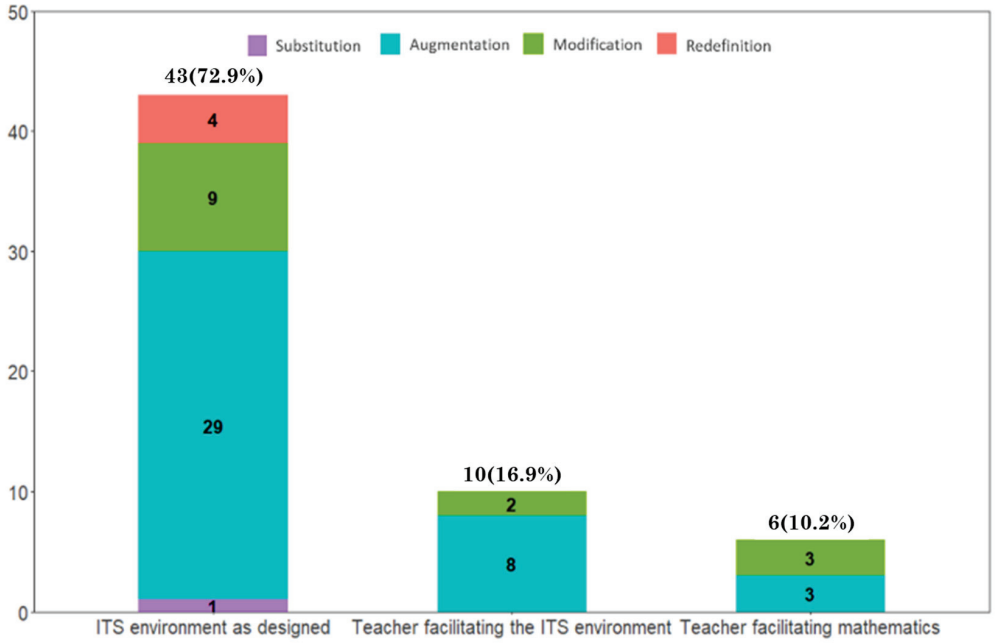


Figure 9. SAMR distribution across teachers' roles.

In the ITS environment as the designed role, studies at the augmentation level were the most dominant, with all levels of SAMR presented in this teacher's role. This finding could be interpreted to suggest that teacher intervention did not have a significant impact on the task design of ITSs. These studies mainly attempted to prove that ITSs can replace teachers by comparing the effects of ITS learning and teacher-led learning (e.g., [74–76]). Because mathematics learning takes place between the ITS and the student and the teacher's role is minimal, the effectiveness of mathematics learning is difficult to exceed the affordances of the ITS.

In the teacher facilitating the ITS environment role, most studies were conducted at the augmentation level (n = 8; 80%), and some studies were conducted at the modification level (n = 2; 20%). Teachers intervene and provide assistance with students' ITS learning, but mathematics learning occurs between the ITS and students, and it is difficult for teachers' intervention to change SAMR levels. For example, Stillson and Alsup [77] found that although teachers supported ITSs by providing additional help in understanding explanations, students either gave up participating or experienced confusion and frustration due to the difficulty of using the ITS.

In the teacher-facilitating-mathematics role, studies were conducted at 50% each at the augmentation level and 50% at the modification level. In these studies, students engaged in self-directed learning, engaging in ITS and off-ITS learning. For example, Kong and Kwok [68] designed a learning cycle for ITS learning. Students participated individually and sometimes in groups, with support from the ITS in the early stages and without ITS support in the later stages. Students were encouraged to freely explore the ITS according to their interests and mainly used the ITS as a communication medium or to conduct group discussions to consolidate the objectives of the learning activities. In these processes, the ITS was one part, not the whole of learning, and the teacher played a role in designing and leading the entire mathematics lesson.

5. Discussion

The purpose of this study was to provide an updated synthesis of the level at which ITS studies have been integrated into mathematics education. The multiple contexts and teachers' roles that determine successful technology integration in mathematics education were examined simultaneously at the SAMR levels. Although the reviewed literature does not provide a clear picture of the actual integration of ITSs in the classroom, it does provide a clear guide to research trends.

The results of the study showed that ITSs have been mainly used (67.2%) at the augmentation level in mathematics education over the past 20 years. The use of ITSs in mathematics education at lower levels of the SAMR framework supports the argument of several scholars that the externally recognized potential of technology generally lags behind in improving teaching and learning [1,5,14]. However, 31% of total studies implemented using ITS were found to be at the level of modification and redefinition. Although this number is not sufficient to drive transformation in mathematics education, recent ITS studies in mathematics education provide evidence that they are transforming experiences of students' mathematics learning. The SAMR classification results for publication year support this evidence. In recent decades, researchers have been using ITSs to redesign the mathematics task and realize teaching and learning of mathematics that would be unimaginable without ITSs (see Figure 5). Some examples in this range include game-based learning [78], exploratory environments [36], and teachable learning [60]. These approaches demonstrate that researchers recognize the affordances of ITSs in allowing learning beyond what learning activities performed without technology can achieve.

ITS research on mathematics education was conducted in the context of K–12 mathematics learning, focusing primarily on elementary schools (41.7%). This result is inconsistent with existing literature studies [9], which revealed that ITS studies in education mainly focus on secondary education. Elementary school students may have more opportunities to use ITSs because they are relatively free from the constraints of the scholastic aptitude

test. There is little empirical ITS research with teachers and preschoolers. This is consistent with Wang et al. [9], which reported that ITS research on adult education and early childhood education in the education field is limited. Teacher professional development related to technology can lead to greater confidence in the benefits of technology to support mathematics learning [79]. Early mathematics is a strong predictor of later learning, so the use of technology in mathematics is encouraged from early childhood [80,81]. Therefore, more research is needed to better understand the affordances of ITS-supported learning in mathematics education at the teacher level and preschool level.

Studies using ITSs at the educational level have mainly been conducted at the augmentation level (see Figure 6). These results show that the affordances offered by ITS learning have not been fully explored at the educational level. However, it is interesting to note that research on transformation range has primarily been conducted at the high school education level compared to other education levels. One example is research that promotes positive collaborative learning and encourages reflective processes among colleagues [63]. Walker et al. [63] used an ITS to focus on pedagogical practices that go beyond traditional classroom practices rather than replacing them. These types of learning provide opportunities for students to apply, evaluate, and analyze knowledge regardless of the educational content (Bloom's taxonomy; [45]) and need to be expanded to other educational levels.

This study found that the three domains of number and arithmetic, algebra, and geometry were the most studied domains of ITSs in mathematics education over the past 20 years. On the other hand, there has been little empirical research on reasoning, statistics, and calculus. Most studies have been conducted at the augmentation level, and this tendency may be due to the characteristics of the domain. Since most ITSs guide learning in a step-by-step manner toward a mastery goal [7], the process of calculating the operations of natural numbers and fractions and measuring the area of shapes in the step-by-step design method is intuitively appropriate for ITS development. However, the potential of technology to transform mathematics education goes beyond simply memorizing and mastering knowledge. If mathematics is characterized as a subject that simply uses rote learning methods, it will be almost impossible to redefine ITSs in mathematics education [44]. Technology integration in mathematics education should be based on realistic contexts, inquiry-based approaches, and collaborative approaches [1]. This gap between research findings and recommendations in mathematics education presents a challenging domain and potential opportunity for applying ITSs and suggests that further research is needed.

This study also found that SAMR levels differ depending on the research purposes. Studies conducted for the purpose of developing an ITS, applying an existing ITS, or exploring factors affecting ITS performance have mainly been conducted at the augmented level. On the other hand, studies aimed at improving the performance of ITSs or exploring effective teaching methods have primarily been conducted on the level of technology integration that can transform mathematics learning. These findings may stem from differences in the research purpose. Studies conducted with the purpose of applying existing ITSs to a specific educational context or exploring factors that influence ITS learning may be difficult to learn mathematics to exceed the limits of the affordances of the predesigned ITS. On the other hand, studies conducted for the purpose of improving the functionality of ITSs and exploring effective teaching methods generally start from considering the limitations of existing ITSs or teaching methods and are therefore more likely to change from ITS-led learning to learner-centered learning. Interestingly, studies conducted for the purpose of the "Development of ITS" mainly focus on enhancing mathematics learning. Considering the cost of designing and developing a new ITS, it may be more effective to improve the functionality of an existing ITS. However, these results do not mean that ITS research in mathematics education should be conducted for specific research purposes. Rather, it is important to recognize the limitations of existing ITSs and teaching methods and to make challenging attempts that consider the affordances of ITSs, educational context, and the teachers' roles [28]. These attempts can lay the foundation to help learners redesign their lesson goals and realize mathematics lessons that would be unimaginable without ITSs.

This study found that ITS studies in mathematics education primarily were conducted in a way that minimizes teacher intervention. Almost 72% of ITS studies excluded or minimized teacher involvement, reflecting an implicit belief that ITSs could replace teachers to teach students. This finding counters the dominant perspective that the success of technology use in mathematics education depends on the teacher (e.g., [28,29]). del Olmo-Muñoz et al. [14] criticized research on the application of technology in mathematics education as having a very limited vision that prioritizes the use of teacher intervention, and most studies in this study have seemed to emphasize the standalone use of ITSs. However, several scholars have reported that students face various challenges in learning ITSs with minimal teacher intervention. For example, Shih et al. [75] noted that teaching using ITSs is well designed only for specific course topics, requiring teacher intervention to flexibly adjust the lesson content and handle unexpected student responses. Matsuda et al. [82] reported that teachers' intervention is essential because students using teachable ITSs do not accurately recognize their own errors, unconsciously make inappropriate instructional decisions, and provide incorrect feedback and hints. Bartelet et al. [66] pointed out that teachers' explanations and interventions are essential for the positive effects of ITSs on students. These findings show that even though an ITS is designed to minimize the teacher's role, teacher intervention is essential to increase student engagement and minimize errors. Therefore, the present study contributes to guiding teachers on how to use ITSs to promote productive student learning and create exploratory learning environments. For example, teachers can use ITSs to promote student engagement in exploratory activities, develop learning tasks, or test student hypotheses to innovate mathematics lessons using ITSs (e.g., [68,71]). Furthermore, teachers can use the extended SAMR model as a guide when teaching mathematics using ITSs to determine whether ITS use is transforming the mathematics classroom or simply replicating existing practices and to reflect on their own teaching practices.

6. Conclusions

The present study offers a new perspective for investigating the integration of ITSs in mathematics education using the SAMR framework as a lens. This study extends existing research on the level of technology integration by including explicit definitions of mathematics education and ITSs at the SAMR level. The results of this study suggest that ITSs were mainly integrated into mathematics education at the augmentation level, but recent changes are taking place to transform mathematics learning. Given the affordances of AI for educational enrichment, the reason for using ITSs may not be limited to replicating and improving teaching methods that can be implemented in the classroom. The most powerful use of ITSs is in using it to transform learning. When an ITS implements teaching and learning that cannot be realized in traditional mathematics classes, ITSs will become an infrastructure for specific teaching and learning, and their use will become essential [83]. However, an ITS itself does not guarantee success in mathematics education. When applying ITSs to education, priority should be given to examining what is pedagogically feasible rather than pursuing what is technologically possible [20]. It is recommended that teachers and researchers provide opportunities to engage students in new teaching and learning activities that would be inconceivable without using ITSs. For ITSs to be useful in mathematics education, the affordances of ITSs, educational context, and teachers' roles in lesson design should be considered together.

7. Limitations and Future Research Agenda

This study had limitations. The keywords, databases, search periods, and inclusion criteria used in the literature search may not provide a description of all studies that have integrated ITSs into mathematics education. Despite this limitation, the findings of this review offer several practical research directions for stakeholders in mathematics education. First, it may not be desirable to expand the use of ITSs across all educational levels and content areas in mathematics education. The reviewed ITS studies minimized

teacher involvement and focused on teaching-specific content areas at the elementary level. Therefore, decisions about teacher adoption of ITSs should be made in ways that contribute to effective mathematics learning. For example, if mathematical learning is better supported by other activities, such as inquiry-based learning, discussions, or group projects, ITSs could be utilized as a tool for certain instructional activities rather than as a standalone method. Therefore, future research should explore why teachers utilize ITSs in mathematics education and how they utilize ITSs in their lessons.

Second, although an ITS is considered a technology for implementing personalized learning in mathematics education, only a limited number of ITS interventions provided fully productive and inquiry-based learning experiences. Moreover, students using ITSs continue to face persistent challenges, such as technical malfunctions and high costs. These limitations may hinder the broader adoption of ITSs in more general curriculum instruction. Thus, to generate collective evidence on the effectiveness of ITSs in mathematics education, a shift from anecdotal findings within specific topics to a broader academic context is necessary. Such a shift would enhance the generalizability of research outcomes and uncover the complex relationships between mathematical content and technological affordances.

Third, researchers should shift their focus from ITS technology itself to ITS-based pedagogy, emphasizing the effective integration of technology and educational theory. The effectiveness of an ITS, once it becomes an integral part of mathematics curricula, will largely depend on whether it can support instructional design and implementation. Therefore, future studies need to investigate the theoretical foundations of ITS interventions and their intricate connections to educational theory.

Funding: This research received no external funding.

Data Availability Statement: The data supporting this study's findings are available from the authors upon reasonable request.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A

Table A1. Summary of the major features of the 58 ITS studies.

Author	Educational Level	Domain	Teachers' Roles	SAMR Model	Research Purpose
Shih et al. [75]	Elementary	Fraction	ITS environment as designed	A	Development of ITS
Wang et al. [9]	Middle	Not specified	ITS environment as designed	A	Application of existing ITS
Çetin et al. [35]	High	Not specified	ITS environment as designed	M	Development of ITS
Spitzer and Moeller [62]	Elementary and middle	Not specified	ITS environment as designed	A	Application of existing ITS
del Olmo-Muñoz et al. [14]	Elementary	Basic arithmetic, Algebra	ITS environment as designed	S	Application of existing ITS
del Olmo et al. [67]	Elementary	Basic arithmetic	ITS environment as designed	A	Investigation of factors
Rebolledo-Mendez et al. [84]	Middle	Statistics	ITS environment as designed	A	Investigation of factors
Joaquim et al. [78]	Elementary	Algebra	Teacher facilitating the ITS environment	M	Exploring teaching methods
Mavrikis et al. [15]	Elementary	Fraction	ITS environment as designed	M	Exploring teaching methods
de MORAIS and Jaques [85]	Middle	Algebra	ITS environment as designed	A	Improvement of ITS
Bush [36]	Elementary	Fraction	ITS environment as designed	M	Application of existing ITS
Pai et al. [53]	Elementary	Basic arithmetic	ITS environment as designed	A	Development of ITS

Table A1. *Cont.*

Author	Educational Level	Domain	Teachers' Roles	SAMR Model	Research Purpose
Zhang et al. [86]	Elementary	Fraction	ITS environment as designed	M	Improvement of ITS
Glaze et al. [72]	Teacher	Geometry	-	A	Investigation of factors
Oker et al. [87]	Elementary	Basic arithmetic	ITS environment as designed	A	Exploring teaching methods
Phillips et al. [73]	High	Algebra	Teacher facilitating the ITS environment, Teacher facilitating mathematics	A	Exploring teaching methods
VanLehn et al. [69]	College	Algebra	ITS environment as designed	M	Development of ITS
Matsuda et al. [60]	Middle	Algebra	ITS environment as designed	R	Improvement of ITS
Borracci et al. [88]	College	Algebra	Teacher facilitating the ITS environment	A	Investigation of factors
Cung et al. [57]	College	Calculus	ITS environment as designed	A	Exploring teaching methods
Wu [76]	Elementary	Fraction	ITS environment as designed	A	Development of ITS
Miller and Bernacki [89]	College	Not specified	ITS environment as designed	A	Investigation of factors
Rajendran et al. [90,91]	Elementary	Basic arithmetic, Algebra	ITS environment as designed	A	Improvement of ITS
Tärning et al. [91]	Elementary	Basic arithmetic	ITS environment as designed	R	Improvement of ITS
Olsen et al. [92]	Elementary	Fraction	Teacher facilitating the ITS environment	M	Exploring teaching methods
Walkington and Bernacki [59]	High	Algebra	ITS environment as designed	A	Improvement of ITS
Nye et al. [54]	College	Algebra	ITS environment as designed	A	Improvement of ITS
Bernacki and Walkington [93]	High	Algebra	Teacher facilitating the ITS environment	A	Improvement of ITS
Wu et al. [94]	Elementary	Fraction	ITS environment as designed	A	Development of ITS
Long and Aleven [70]	Middle	Algebra	ITS environment as designed	M	Improvement of ITS
Bringula et al. [95]	High	Algebra	ITS environment as designed	R	Investigation of factors
Bartelet et al. [66]	Middle	Basic arithmetic, Ratio	ITS environment as designed, Teacher facilitating the ITS environment	A	Investigation of factors
Huang et al. [55]	Elementary	Not specified	ITS environment as designed	A	Application of existing ITS
González-Calero et al. [13]	Middle	Algebra	ITS environment as designed	A	Exploring teaching methods
Pane et al. [61]	Middle, High	Algebra	Teacher facilitating mathematics	M	Exploring teaching methods
Arnau et al. [65]	Elementary	Basic arithmetic	ITS environment as designed	A	Development of ITS
Walker et al. [63]	High	Algebra	ITS environment as designed	M	Improvement of ITS
San Pedro et al. [56]	High	Statistics	ITS environment as designed	A	Investigation of factors
Khachatryan et al. [96]	Elementary	Not specified	Teacher facilitating mathematics	A	Development of ITS
Arnau et al. [97]	College	Algebra	ITS environment as designed	M	Development of ITS
Craig et al. [74]	Middle	Not specified	ITS environment as designed	A	Application of existing ITS
Abramovich et al. [98]	Middle	Proportional reasoning	Teacher facilitating mathematics	A	Investigation of factors
Butcher and Aleven [71]	High	Geometry	Teacher facilitating mathematics	M	Improvement of ITS

Table A1. *Cont.*

Author	Educational Level	Domain	Teachers' Roles	SAMR Model	Research Purpose
Matsuda et al. [82]	Middle	Algebra	ITS environment as designed	R	Improvement of ITS
Xu et al. [99]	Middle	Geometry	ITS environment as designed	A	Development of ITS
Arroyo et al. [100]	Middle	Basic arithmetic	ITS environment as designed	A	Application of existing ITS
Roll et al. [101]	High	Geometry	ITS environment as designed	A	Improvement of ITS
Beal et al. [102]	Elementary	Basic arithmetic, Fraction	ITS environment as designed	A	Development of ITS
Maloy et al. [103]	Elementary	Basic arithmetic	Teacher facilitating mathematics	M	Exploring teaching methods
Keleş et al. [31]	College	Sequences	ITS environment as designed	A	Development of ITS
Mendicino et al. [64]	Elementary	Basic arithmetic, Algebra, Geometry	Teacher facilitating mathematics	A	Application of existing ITS
Chen [11]	Elementary	Fraction	ITS environment as designed	A	Development of ITS
Hwang et al. [104]	Middle	Basic arithmetic	Teacher facilitating mathematics	A	Application of existing ITS
Lanzilotti and Roselli [105]	Elementary	Logic	ITS environment as designed	A	Development of ITS
Chang et al. [12]	Elementary	Basic arithmetic, Fraction, Geometry	ITS environment as designed	A	Development of ITS
Aleven et al. [106]	High	Geometry	Teacher facilitating the ITS environment	A	Development of ITS
Kong and Kwok [68]	Elementary	Fraction	Teacher facilitating mathematics	M	Development of ITS
Stillson and Alsup [77]	College	Algebra	Teacher facilitating the ITS environment	A	Application of existing ITS

Note. S: substitution level, A: augmentation level, M: modification level, R: redefinition level.

References

1. Bray, A.; Tangney, B. Technology Usage in Mathematics Education Research–A Systematic Review of Recent Trends. *Comput. Educ.* **2017**, *114*, 255–273. [CrossRef]
2. Laborde, C. Integration of Technology in the Design of Geometry Tasks with Cabri-Geometry. *Int. J. Comput. Math. Learn.* **2002**, *6*, 283–317. [CrossRef]
3. Oates, G. Sustaining Integrated Technology in Undergraduate Mathematics. *Int. J. Math. Educ. Sci. Technol.* **2011**, *42*, 709–721. [CrossRef]
4. Oldknow, A. ICT Bringing Mathematics to Life and Life to Mathematics. *Electron. J. Math. Technol.* **2009**, *3*, 137–148.
5. Olive, J.; Makar, K.; Hoyos, V.; Kor, L.K.; Kosheleva, O.; Sträßer, R. Mathematical Knowledge and Practices Resulting from Access to Digital Technologies. In *Mathematics Education and Technology-Rethinking the Terrain: The 17th ICMI Study*; Hoyles, C., Lagrange, J.-B., Eds.; New ICMI Study Series; Springer: Boston, MA, USA, 2010; pp. 133–177; ISBN 978-1-4419-0146-0.
6. Baker, T. *Educ-AI-Tion Rebooted? Exploring the Future of Artificial Intelligence in Schools and Colleges*; NESTA: London, UK, 2019.
7. Holmes, W.; Bialik, M.; Fadel, C. Artificial Intelligence in Education. In *Data Ethics: Building Trust: How Digital Technologies Can Serve Humanity*; Globethics Publications: Geneva, Switzerland, 2023; pp. 621–653.
8. Luckin, R.; Holmes, W. *Intelligence Unleashed: An Argument for AI in Education*; UCL Knowledge Lab.: London, UK, 2016.
9. Wang, H.; Tlili, A.; Huang, R.; Cai, Z.; Li, M.; Cheng, Z.; Yang, D.; Li, M.; Zhu, X.; Fei, C. Examining the Applications of Intelligent Tutoring Systems in Real Educational Contexts: A Systematic Literature Review from the Social Experiment Perspective. *Educ. Inf. Technol.* **2023**, *28*, 9113–9148. [CrossRef] [PubMed]
10. Psotka, J.; Massey, L.; Mutter, S. *Intelligent Tutoring Systems: Lessons Learned*; Lawrence Erlbaum Associates: Hillsdale, NJ, USA, 1988.
11. Chen, C.-M. Intelligent Web-Based Learning System with Personalized Learning Path Guidance. *Comput. Educ.* **2008**, *51*, 787–814. [CrossRef]
12. Chang, K.-E.; Sung, Y.-T.; Lin, S.-F. Computer-Assisted Learning for Mathematical Problem Solving. *Comput. Educ.* **2006**, *46*, 140–151. [CrossRef]
13. González-Calero, J.A.; Arnau, D.; Puig, L.; Arevalillo-Herráez, M. Intensive Scaffolding in an Intelligent Tutoring System for the Learning of Algebraic Word Problem Solving. *Br. J. Educ. Technol.* **2015**, *46*, 1189–1200. [CrossRef]
14. Del Olmo-Muñoz, J.; González-Calero, J.A.; Diago, P.D.; Arnau, D.; Arevalillo-Herráez, M. Intelligent Tutoring Systems for Word Problem Solving in COVID-19 Days: Could They Have Been (Part of) the Solution? *ZDM Math. Educ.* **2023**, *55*, 35–48. [CrossRef]

15. Mavrikis, M.; Rummel, N.; Wiedmann, M.; Loibl, K.; Holmes, W. Combining Exploratory Learning with Structured Practice Educational Technologies to Foster Both Conceptual and Procedural Fractions Knowledge. *Educ. Technol. Res. Dev.* **2022**, *70*, 691–712. [CrossRef]
16. Bicalho, R.N.D.M.; Coll, C.; Engel, A.; Lopes de Oliveira, M.C.S. Integration of ICTs in Teaching Practices: Propositions to the SAMR Model. *Educ. Technol. Res. Dev.* **2023**, *71*, 563–578. [CrossRef] [PubMed]
17. Mousavinasab, E.; Zarifsanaiey, N.; Niakan Kalhori, S.R.; Rakhshan, M.; Keikha, L.; Ghazi Saeedi, M. Intelligent Tutoring Systems: A Systematic Review of Characteristics, Applications, and Evaluation Methods. *Interact. Learn. Environ.* **2021**, *29*, 142–163. [CrossRef]
18. Steenbergen-Hu, S.; Cooper, H. A Meta-Analysis of the Effectiveness of Intelligent Tutoring Systems on K–12 Students' Mathematical Learning. *J. Educ. Psychol.* **2013**, *105*, 970–987. [CrossRef]
19. Puentedura, R. Transformation, Technology, and Education; 2006. Available online: http://hippasus.com/resources/tte/ (accessed on 3 April 2024).
20. Zawacki-Richter, O.; Marín, V.I.; Bond, M.; Gouverneur, F. Systematic Review of Research on Artificial Intelligence Applications in Higher Education–Where Are the Educators? *Int. J. Educ. Technol. High. Educ.* **2019**, *16*, 39. [CrossRef]
21. Cheung, A.C.; Slavin, R.E. The Effectiveness of Educational Technology Applications for Enhancing Mathematics Achievement in K-12 Classrooms: A Meta-Analysis. *Educ. Res. Rev.* **2013**, *9*, 88–113. [CrossRef]
22. Larkin, K.; Calder, N. Mathematics Education and Mobile Technologies. *Math. Educ. Res. J.* **2016**, *28*, 1–7. [CrossRef]
23. De Witte, K.; Rogge, N. Does ICT Matter for Effectiveness and Efficiency in Mathematics Education? *Comput. Educ.* **2014**, *75*, 173–184. [CrossRef]
24. Demitriadou, E.; Stavroulia, K.E.; Lanitis, A. Comparative Evaluation of Virtual and Augmented Reality for Teaching Mathematics in Primary Education. *Educ. Inf. Technol.* **2020**, *25*, 381–401. [CrossRef]
25. Hennessy, S.; Ruthven, K.; Brindley, S. Teacher Perspectives on Integrating ICT into Subject Teaching: Commitment, Constraints, Caution, and Change. *J. Curric. Stud.* **2005**, *37*, 155–192. [CrossRef]
26. Means, B. Technology and Education Change: Focus on Student Learning. *J. Res. Technol. Educ.* **2010**, *42*, 285–307. [CrossRef]
27. Jankvist, U.T.; Geraniou, E. Digital Technologies as a Way of Making Original Sources More Accessible to Students. In Proceedings of the Eighth European Summer University on History and Epistemology in Mathematics Education ESU, Oslo, Norway, 20–24 July 2018; Barbin, E., Jankvist, U.T., Kjeldsen, T.H., Smestad, B., Tzanakis, C., Eds.; Oslo Metropolitan University: Oslo, Norway, 2019; pp. 107–130.
28. Drijvers, P. Digital Technology in Mathematics Education: Why It Works (Or Doesn't). In *Selected Regular Lectures from the 12th International Congress on Mathematical Education, Seoul, Republic of Korea, 8–15 July 2012*; Cho, S., Ed.; Springer: Berlin/Heidelberg, Germany, 2015; pp. 135–151. [CrossRef]
29. Viberg, O.; Grönlund, Å.; Andersson, A. Integrating Digital Technology in Mathematics Education: A Swedish Case Study. *Interact. Learn. Environ.* **2023**, *31*, 232–243. [CrossRef]
30. Ng, W.; Nicholas, H. A Framework for Sustainable Mobile Learning in Schools. *Br. J. Educ. Technol.* **2013**, *44*, 695–715. [CrossRef]
31. Keleş, A.; Ocak, R.; Keleş, A.; Gülcü, A. ZOSMAT: Web-Based Intelligent Tutoring System for Teaching–Learning Process. *Expert Syst. Appl.* **2009**, *36*, 1229–1239. [CrossRef]
32. Buchanan, T. The Efficacy of a World-Wide Web Mediated Formative Assessment. *J. Comput. Assist. Learn.* **2000**, *16*, 193–200. [CrossRef]
33. Kulik, J.A.; Fletcher, J.D. Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review. *Rev. Educ. Res.* **2016**, *86*, 42–78. [CrossRef]
34. Hwang, G.-J.; Tu, Y.-F. Roles and Research Trends of Artificial Intelligence in Mathematics Education: A Bibliometric Mapping Analysis and Systematic Review. *Mathematics* **2021**, *9*, 584. [CrossRef]
35. Çetin, I.; Erümit, A.K.; Nabiyev, V.; Karal, H.; Kösa, T.; Kokoç, M. The Effect of Gamified Adaptive Intelligent Tutoring System Artibos on Problem-Solving Skills. *Particip. Educ. Res.* **2023**, *10*, 344–374. [CrossRef]
36. Bush, J.B. Software-based Intervention with Digital Manipulatives to Support Student Conceptual Understandings of Fractions. *Br. J. Educ. Technol.* **2021**, *52*, 2299–2318. [CrossRef]
37. Pierce, R.; Stacey, K. Mapping Pedagogical Opportunities Provided by Mathematics Analysis Software. *Int. J. Comput. Math. Learn.* **2010**, *15*, 1–20. [CrossRef]
38. Mishra, P.; Koehler, M.J. Introducing Technological Pedagogical Content Knowledge. In Proceedings of the Annual Meeting of the American Educational Research Association, New York, NY, USA, 24–28 March 2008; Volume 1, p. 16.
39. Hughes, J.E. Teaching English with Technology: Exploring Teacher Learning and Practice. Doctoral Dissertation, Michigan State University, Ann Arbor, MI, USA, 2000.
40. Hamilton, E.; Rosenberg, J. The Substitution Augmentation Modification Redefinition (SAMR) Model: A Critical Review and Suggestions for Its Use. *TechTrends* **2016**, *60*, 433–441. [CrossRef]
41. Kohen, Z. Informed Integration of IWB Technology, Incorporated with Exposure to Varied Mathematics Problem-Solving Skills: Its Effect on Students' Real-Time Emotions. *Int. J. Math. Educ. Sci. Technol.* **2019**, *50*, 1128–1151. [CrossRef]
42. Loong, E.Y.-K.; Herbert, S. Primary School Teachers' Use of Digital Technology in Mathematics: The Complexities. *Math. Educ. Res. J.* **2018**, *30*, 475–498. [CrossRef]

43. Fabian, K.; Topping, K.J. Putting "Mobile" into Mathematics: Results of a Randomised Controlled Trial. *Contemp. Educ. Psychol.* **2019**, *59*, 101783. [CrossRef]
44. Crompton, H.; Burke, D. Mobile Learning and Pedagogical Opportunities: A Configurative Systematic Review of PreK-12 Research Using the SAMR Framework. *Comput. Educ.* **2020**, *156*, 103945. [CrossRef]
45. Nair, R.S.; Chuan, T.C. Integrating Technology That Uses Modified SAMR Model as a Pedagogical Framework in Evaluating Learning Performance of Undergraduates. *Educ. Rev.* **2021**, *5*, 373–384. [CrossRef]
46. Lameras, P.; Arnab, S. Power to the Teachers: An Exploratory Review on Artificial Intelligence in Education. *Information* **2021**, *13*, 14. [CrossRef]
47. Steenbergen-Hu, S.; Cooper, H. A Meta-Analysis of the Effectiveness of Intelligent Tutoring Systems on College Students' Academic Learning. *J. Educ. Psychol.* **2014**, *106*, 331–347. [CrossRef]
48. Van Lehn, K. The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educ. Psychol.* **2011**, *46*, 197–221. [CrossRef]
49. Gough, D.; Thomas, J.; Oliver, S. *An Introduction to Systematic Reviews*; Sage Publications: Thousand Oaks, CA, USA, 2017; pp. 1–352.
50. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *Int. J. Surg.* **2021**, *88*, 105906. [CrossRef]
51. Luo, H.; Li, G.; Feng, Q.; Yang, Y.; Zuo, M. Virtual Reality in K-12 and Higher Education: A Systematic Review of the Literature from 2000 to 2019. *J. Comput. Assist. Learn.* **2021**, *37*, 887–901. [CrossRef]
52. Kitchenham, B.; Charters, S. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*; EBSE Technical Report EBSE-2007-01; Keele University: Keele, UK; Durham University: Durham, UK, 2007.
53. Pai, K.; Kuo, B.C.; Liao, C.H.; Liu, Y.M. Application of Chinese Dialogue-Based Intelligent Tutoring System in Remedial Instruction for Mathematics Learning. *Educ. Psychol.* **2021**, *41*, 137–152. [CrossRef]
54. Nye, B.D.; Pavlik, P.I.; Windsor, A.; Olney, A.M.; Hajeer, M.; Hu, X. SKOPE-IT (Shareable Knowledge Objects as Portable Intelligent Tutors): Overlaying Natural Language Tutoring on an Adaptive Learning System for Mathematics. *Int. J. STEM Educ.* **2018**, *5*, 2. [CrossRef] [PubMed]
55. Huang, X.; Craig, S.D.; Xie, J.; Graesser, A.; Hu, X. Intelligent Tutoring Systems Work as a Math Gap Reducer in 6th Grade After-School Program. *Learn. Individ. Differ.* **2016**, *47*, 258–265. [CrossRef]
56. San Pedro, M.O.Z.; Baker, R.S.J.D.; Rodrigo, M.M.T. Carelessness and Affect in an Intelligent Tutoring System for Mathematics. *Int. J. Artif. Intell. Educ.* **2014**, *24*, 189–210. [CrossRef]
57. Cung, B.; Xu, D.; Eichhorn, S.; Warschauer, M. Getting Academically Underprepared Students Ready Through College Developmental Education: Does the Course Delivery Format Matter? *Am. J. Distance Educ.* **2019**, *33*, 178–194. [CrossRef]
58. Kessler, A.; Boston, M.; Stein, M.K. Exploring How Teachers Support Students' Mathematical Learning in Computer-Directed Learning Environments. *Inf. Learn. Sci.* **2020**, *121*, 52–78. [CrossRef]
59. Walkington, C.; Bernacki, M.L. Personalizing Algebra to Students' Individual Interests in an Intelligent Tutoring System. *Int. J. Artif. Intell. Educ.* **2018**, *29*, 58–88. [CrossRef]
60. Matsuda, N.; Weng, W.; Wall, N. The Effect of Metacognitive Scaffolding for Learning by Teaching a Teachable Agent. *Int. J. Artif. Intell. Educ.* **2020**, *30*, 1–37. [CrossRef]
61. Pane, J.F.; Griffin, B.A.; McCaffrey, D.F.; Karam, R. Effectiveness of Cognitive Tutor Algebra I at Scale. *Educ. Eval. Policy Anal.* **2014**, *36*, 127–144. [CrossRef]
62. Spitzer, M.W.H.; Moeller, K. Performance Increases in Mathematics during COVID-19 Pandemic Distance Learning in Austria: Evidence from an Intelligent Tutoring System for Mathematics. *Trends Neurosci. Educ.* **2023**, *31*, 100203. [CrossRef]
63. Walker, E.; Rummel, N.; Koedinger, K.R. Adaptive Intelligent Support to Improve Peer Tutoring in Algebra. *Int. J. Artif. Intell. Educ.* **2014**, *24*, 33–61. [CrossRef]
64. Mendicino, M.; Razzaq, L.; Heffernan, N.T. A Comparison of Traditional Homework to Computer-Supported Homework. *J. Res. Tech. Educ.* **2009**, *41*, 331–359. [CrossRef]
65. Arnau, D.; Arevalillo-Herráez, M.; González-Calero, J.A. Emulating Human Supervision in an Intelligent Tutoring System for Arithmetical Problem Solving. *IEEE Trans. Learn. Technol.* **2014**, *7*, 155–164. [CrossRef]
66. Bartelet, D.; Ghysels, J.; Groot, W.; Haelermans, C.; Maassen van den Brink, H. The Differential Effect of Basic Mathematics Skills Homework Via a Web-Based Intelligent Tutoring System across Achievement Subgroups and Mathematics Domains: A Randomized Field Experiment. *J. Educ. Psy.* **2016**, *108*, 1–20. [CrossRef]
67. Del Olmo-Muñoz, J.; González-Calero, J.A.; Diago, D.; Arnau, D.; Arevalillo-Herráez, M. Using Intra-task Flexibility on an Intelligent Tutoring System to Promote Arithmetic Problem-solving Proficiency. *J. Educ. Technol.* **2022**, *53*, 1976–1992. [CrossRef]
68. Kong, S.C.; Kwok, L.F. A Cognitive Tool for Teaching the Addition/Subtraction of Common Fractions: A Model of Affordances. *Comput. Educ.* **2005**, *45*, 245–265. [CrossRef]
69. VanLehn, K.; Banerjee, C.; Milner, F.; Wetzel, J. Teaching Algebraic Model Construction: A Tutoring System, Lessons Learned and an Evaluation. *Int. J. Artif. Intell Educ.* **2020**, *30*, 459–480. [CrossRef]
70. Long, Y.; Aleven, V. Enhancing Learning Outcomes through Self-Regulated Learning Support with an Open Learner Model. *User Model. User-Adap. Inter.* **2017**, *27*, 55–88. [CrossRef]

71. Butcher, K.R.; Aleven, V. Using Student Interactions to Foster Rule–Diagram Mapping during Problem Solving in an Intelligent Tutoring System. *J. Educ. Psychol.* **2013**, *105*, 988–1009. [CrossRef]
72. Glaze, A.; Moyer-Packenham, P.; Longhurst, M. Teachers' Conceptions of Mathematics and the Use of Intelligent Tutoring Systems, Calculators, Dynamic Geometry Software and Desmos in the Classroom. *J. Comput. Math. Sci. Teach.* **2021**, *40*, 201–227.
73. Phillips, A.; Pane, J.F.; Reumann-Moore, R.; Shenbanjo, O. Implementing an Adaptive Intelligent Tutoring System as an Instructional Supplement. *Educ. Technol. Res. Dev.* **2020**, *68*, 1409–1437. [CrossRef]
74. Craig, S.D.; Hu, X.; Graesser, A.C.; Bargagliotti, A.E.; Sterbinsky, A.; Cheney, K.R.; Okwumabua, T. The Impact of a Technology-Based Mathematics after-School Program Using ALEKS on Student's Knowledge and Behaviors. *Comput. Educ.* **2013**, *68*, 495–504. [CrossRef]
75. Shih, S.C.; Chang, C.C.; Kuo, B.C.; Huang, Y.H. Mathematics Intelligent Tutoring System for Learning Multiplication and Division of Fractions Based on Diagnostic Teaching. *Educ. Inf. Technol.* **2023**, *28*, 9189–9210. [CrossRef]
76. Wu, H.-M. Online Individualised Tutor for Improving Mathematics Learning: A Cognitive Diagnostic Model Approach. *Educ. Psychol.* **2019**, *39*, 1218–1232. [CrossRef]
77. Stillson, H.; Alsup, J. Smart ALEKS... or Not? Teaching Basic Algebra Using an Online Interactive Learning System. *Math. Comput. Educ.* **2003**, *37*, 329–340.
78. Joaquim, S.; Bittencourt, I.I.; de Amorim Silva, R.; Espinheira, P.L.; Reis, M. What to Do and What to Avoid on the Use of Gamified Intelligent Tutor System for Low-Income Students. *Educ. Inf. Technol.* **2022**, *27*, 2677–2694. [CrossRef]
79. Bennison, A.; Goos, M. Learning to Teach Mathematics with Technology: A Survey of Professional Development Needs, Experiences and Impacts. *Math. Ed. Res. J.* **2010**, *22*, 31–56. [CrossRef]
80. Cross, C.T.; Woods, T.A.; Schweingruber, H.E. *Mathematics Learning in Early Childhood: Paths toward Excellence and Equity*; National Academies Press: Washington, DC, USA, 2009; ISBN 0-309-12806-4.
81. Duncan, G.J.; Dowsett, C.J.; Claessens, A.; Magnuson, K.; Huston, A.C.; Klebanov, P.; Pagani, L.S.; Feinstein, L.; Engel, M.; Brooks-Gunn, J.; et al. School Readiness and Later Achievement. *Dev. Psychol.* **2007**, *43*, 1428–1446. [CrossRef]
82. Matsuda, N.; Yarzebinski, E.; Keiser, V.; Raizada, R.; Cohen, W.W.; Stylianides, G.J.; Koedinger, K.R. Cognitive Anatomy of Tutor Learning: Lessons Learned with SimStudent. *J. Educ. Psy.* **2013**, *105*, 1152–1163. [CrossRef]
83. Kaput, J.; Hegedus, S.; Lesh, R. Technology Becoming Infrastructural in Mathematics Education. In *Foundations for the Future in Mathematics Education*; Routledge: Abingdon-on-Thames, UK, 2007; ISBN 978-1-00-306452-7.
84. Rebolledo-Mendez, G.; Huerta-Pacheco, N.S.; Baker, R.S.; du Boulay, B. Meta-Affective Behaviour within an Intelligent Tutoring System for Mathematics. *Int. J. Artif. Intell. Educ.* **2022**, *32*, 174–195. [CrossRef]
85. De MORAIS, F.; Jaques, P. Does Handwriting Impact Learning on Math Tutoring Systems? *Infor. Educ.* **2022**, *21*, 55–90. [CrossRef]
86. Zhang, J.; Gao, M.; Holmes, W.; Mavrikis, M. Interaction Patterns in Exploratory Learning Environments for Mathematics: A Sequential Analysis of Feedback and External Representations in Chinese Schools. *Interact. Learn. Environ.* **2021**, *29*, 1211–1228. [CrossRef]
87. Oker, A.; Pecune, F.; Declercq, C. Virtual Tutor and Pupil Interaction: A Study of Empathic Feedback as Extrinsic Motivation for Learning. *Educ. Inf. Technol.* **2020**, *25*, 3643–3658. [CrossRef]
88. Borracci, G.; Gauthier, E.; Jennings, J.; Sale, K.; Muldner, K. The Effect of Assistance on Learning and Affect in an Algebra Tutor. *J. Educ. Comput. Res.* **2020**, *57*, 2032–2052. [CrossRef]
89. Miller, C.J.; Bernacki, M.L. Training Preparatory Mathematics Students to Be High Ability Self-Regulators: Comparative and Case-Study Analyses of Impact on Learning Behavior and Achievement. *High Abil. Stud.* **2019**, *30*, 167–197. [CrossRef]
90. Rajendran, R.; Iyer, S.; Murthy, S. Personalized Affective Feedback to Address Students' Frustration in ITS. *IEEE Trans. Learn. Technol.* **2019**, *12*, 87–97. [CrossRef]
91. Tärning, B.; Silvervarg, A.; Gulz, A.; Haake, M. Instructing a Teachable Agent with Low or High Self-Efficacy–Does Similarity Attract? *Int. J. Artif. Intell. Educ.* **2019**, *29*, 89–121. [CrossRef]
92. Olsen, J.K.; Rummel, N.; Aleven, V. It Is Not Either or: An Initial Investigation into Combining Collaborative and Individual Learning Using an ITS. *Intern. J. Comput. Support. Collab. Learn.* **2019**, *14*, 353–381. [CrossRef]
93. Bernacki, M.L.; Walkington, C. The Role of Situational Interest in Personalized Learning. *J. Educ. Psychol.* **2018**, *110*, 864–881. [CrossRef]
94. Wu, H.-M.; Kuo, B.-C.; Wang, S.-C. Computerized Dynamic Adaptive Tests with Immediately Individualized Feedback for Primary School Mathematics Learning. *Educ. Technol. Soc.* **2017**, *20*, 61–73.
95. Bringula, R.P.; Basa, R.S.; Dela Cruz, C.; Rodrigo, M.M.T. Effects of Prior Knowledge in Mathematics on Learner-Interface Interactions in a Learning-by-Teaching Intelligent Tutoring System. *J. Educ. Comput. Res.* **2016**, *54*, 462–482. [CrossRef]
96. Khachatryan, G.A.; Romashov, A.V.; Khachatryan, A.R.; Gaudino, S.J.; Khachatryan, J.M.; Guarian, K.R.; Yufa, N.V. Reasoning Mind Genie 2: An Intelligent Tutoring System as a Vehicle for International Transfer of Instructional Methods in Mathematics. *Int. J. Artif. Intell. Educ.* **2014**, *24*, 333–382. [CrossRef]
97. Arnau, D.; Arevalillo-Herráez, M.; Puig, L.; González-Calero, J.A. Fundamentals of the Design and the Operation of an Intelligent Tutoring System for the Learning of the Arithmetical and Algebraic Way of Solving Word Problems. *Comput. Educ.* **2013**, *63*, 119–130. [CrossRef]
98. Abramovich, S.; Schunn, C.; Higashi, R.M. Are Badges Useful in Education?: It Depends upon the Type of Badge and Expertise of Learner. *Educ. Technol. Res. Dev.* **2013**, *61*, 217–232. [CrossRef]

99. Xu, W.; Zhao, K.; Li, Y.; Yi, Z. FUDAOWANG: A Web-Based Intelligent Tutoring System Implementing Advanced Education Concepts. *Int. J. Distance Educ. Technol.* **2012**, *10*, 67–90. [CrossRef]
100. Arroyo, I.; Royer, J.M.; Woolf, B.P. Using an Intelligent Tutor and Math Fluency Training to Improve Math Performance. *Int. J. Artif. Intell. Educ.* **2011**, *21*, 135–152. [CrossRef]
101. Roll, I.; Aleven, V.; McLaren, B.M.; Koedinger, K.R. Improving Students' Help-Seeking Skills Using Metacognitive Feedback in an Intelligent Tutoring System. *Learn. Instr.* **2011**, *21*, 267–280. [CrossRef]
102. Beal, C.R.; Arroyo, I.M.; Cohen, P.R.; Woolf, B.P. Evaluation of AnimalWatch: An Intelligent Tutoring System for Arithmetic and Fractions. *J. Interact. Online Learn.* **2010**, *9*, 64–77.
103. Maloy, R.W.; Edwards, S.A.; Anderson, G. Teaching Math Problem Solving Using a Web-Based Tutoring System, Learning Games, and Students' Writing. *J. STEM Educ.* **2010**, *11*, 82–90.
104. Hwang, G.; Tseng, J.C.R.; Hwang, G. Diagnosing Student Learning Problems Based on Historical Assessment Records. *Innov. Educ. Teach. Int.* **2008**, *45*, 77–89. [CrossRef]
105. Lanzilotti, R.; Roselli, T. An Experimental Evaluation of Logiocando, an Intelligent Tutoring Hypermedia System. *Int. J. Artif. Intell. Educ.* **2007**, *17*, 41–56.
106. Aleven, V.; McLaren, B.; Roll, I.; Koedinger, K. Toward Meta-Cognitive Tutoring: A Model of Help Seeking with a Cognitive Tutor. *Int. J. Artif. Intell. Educ.* **2006**, *16*, 101–128.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

MDPI AG
Grosspeteranlage 5
4052 Basel
Switzerland
Tel.: +41 61 683 77 34

Computers Editorial Office
E-mail: computers@mdpi.com
www.mdpi.com/journal/computers



Disclaimer/Publisher's Note: The title and front matter of this reprint are at the discretion of the Guest Editor. The publisher is not responsible for their content or any associated concerns. The statements, opinions and data contained in all individual articles are solely those of the individual Editor and contributors and not of MDPI. MDPI disclaims responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.





Academic Open
Access Publishing

mdpi.com

ISBN 978-3-7258-7541-2