education sciences

Special Issue Reprint

Artificial Intelligence and Blended Learning

Challenges, Opportunities, and Future Directions

Edited by
Will W. K. Ma

mdpi.com/journal/education

MDPI

Artificial Intelligence and Blended Learning: Challenges, Opportunities, and Future Directions

Artificial Intelligence and Blended Learning: Challenges, Opportunities, and Future Directions

Guest Editor

Will W. K. Ma

MDPI

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

Guest Editor
Will W. K. Ma
Centre for Innovative
Teaching and Learning
Tung Wah College
Hong Kong
China

Editorial Office
MDPI AG
Grosspeteranlage 5
4052 Basel, Switzerland

This is a reprint of the Special Issue, published open access by the journal *Education Sciences* (ISSN 2227-7102), freely accessible at: https://www.mdpi.com/journal/education/special_issues/5R19DNNZME.

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 978-3-7258-6682-3 (Hbk)
ISBN 978-3-7258-6683-0 (PDF)
https://doi.org/10.3390/books978-3-7258-6683-0

Contents

About the Editor

Will W. K. Ma

Will W. K. Ma, PhD, has served as Professor and Director of the Centre for Innovative Teaching and Learning at Tung Wah College, Hong Kong. His research interests encompass artificial intelligence, blended learning, information systems adoption, knowledge sharing and creation, and higher education. He earned his PhD from the University of Hong Kong, his MPhil and MSc in Information Systems Management from the Hong Kong University of Science and Technology, and his BA (Hons) in Business Studies from the City University of Hong Kong.

Preface

This Special Issue began with a call for papers on the theme "Artificial Intelligence and Blended Learning: Challenges, Opportunities, and Future Directions." Its aim was to explore the intersection of artificial intelligence (AI) and blended learning, emphasizing innovative research and practical applications that enhance the effectiveness and efficiency of blended learning environments.

Blended learning, which combines traditional face-to-face instruction with online or digital components, has become increasingly prevalent in educational settings. AI, with its capabilities in data analysis, machine learning, and natural language processing, promised to transform the landscape of blended learning by personalizing instruction, enabling intelligent tutoring systems, automating assessment processes, and creating immersive learning experiences.

Researchers and practitioners from diverse disciplines were invited to contribute original research papers, case studies, and theoretical perspectives that illuminated the potential of AI to enhance blended learning. The call for papers outlined a range of topics of interest for prospective authors.

Following a year-long call for submissions in 2025, eight papers were accepted for publication in *Education Sciences*: six research articles and two systematic reviews. The Special Issue provided a platform for scholars to disseminate their work to a global audience and contributed to advancing knowledge in blended learning and AI, fostering collaboration and innovation.

We thank all authors for their interest and hard work. Their high-quality contributions collectively advanced the frontiers of AI in blended learning research and applications.

Will W. K. Ma
Guest Editor

education sciences

MDPI

Article

An Integrated Individual, Social, and Technology Model for the Sustainable Adoption of Generative AI in Blended Learning

Will W. K. Ma

The Centre for Innovative Teaching & Learning (CITL), Tung Wah College, 90A Shan Tung Street, Kowloon, Hong Kong SAR, China; willma@twc.edu.hk

Abstract

Generative AI is a promising adjunct to blended learning, offering an innovative means to enhance academic performance. Its rapid diffusion has been accompanied by criticism and uncertainty, particularly regarding ethics and the potential displacement of human labor. A review of the existing research reveals persistent gaps in understanding AI use among students. This study therefore aimed to develop an integrated model to explain generative AI adoption across two distinctive time points. Employing a survey-based design, cross-sectional data were collected at two time points from college students at a local tertiary institution in Hong Kong. PLS-SEM Model testing showed that performance expectancy was the strongest and most persistent determinant of both intention to use and actual use across both data collections. Risk propensity had no effect at the outset, but at a longer usage time point, it was significantly related to intention and use through performance expectancy. Social influence exerted a direct and significant effect initially and later demonstrated both direct and indirect significant effects on intention and use via performance expectancy. The findings identify key determinants and enhance our understanding of the complex decision-making process involved in the use of generative AI.

Keywords: generative AI adoption; risk propensity; social influence; performance expectancy; blended learning

1. Introduction

Generative AI is a promising adjunct to blended learning, offering an innovative means to enhance academic performance by clarifying complex concepts, facilitating the brainstorming of ideas and solutions to real-world problems, and delivering continuous, on-demand feedback that complements instructor support. However, this does not mean that there is no problem with its acceptance and use. For example, numerous school districts and universities have historically banned or blocked access to generative AI because of concerns over plagiarism, the impact on critical thinking skills, and the potential on inaccurate information (Johnson, 2023). In a more recent study, it was still found that institutions exert efforts to maintain academic integrity through prohibitive AI policies and detection systems to influence student behavior that leads to many students opting not to use AI at all, even for legitimate academic support (Marks, 2025). Research also revealed that students are not necessarily rushing to abuse AI where students seem to have strong views on cheating, high levels of concern about its role in education, and mixed opinions on its impact on their lives (Marks, 2025). Students think that it is risky to use AI as reports

find that the main reasons putting students off using AI are being accused of cheating and obtaining false results or 'hallucinations' (Freeman, 2025). The focus in evaluating generative AI appears to be shifting from predominantly ethical considerations to greater attention on students' risk perceptions and risk propensity. While numerous studies on ethical issues and ethical concerns with generative AI (e.g., Burriss et al., 2024; Farhi et al., 2023; Huang et al., 2025), there are rare studies on risk and, in particular,, risk propensity of the individual users. Moreover, prior studies called for a more holistic framework to capture the various perspectives in order to understand the complex decision-making process (e.g., Hemdanou et al., 2024; Nazaretsky et al., 2025; Zhao et al., 2024). Understanding there may be differences to the determinants over experience gained, investigating the sustainability of the determinants would not just be interesting, but also useful to provide a rich explanation to the acceptance decision making processes and to provide insights to devise implementation strategies at different time points (e.g., Annamalai et al., 2025).

Having discussed all of the above, understanding the use of generative AI would be both relevant and important. Therefore, this study aims to develop a more holistic framework to investigate the generative AI adoption issue. The research objective of this study is to explore the sustainable key determinants of generative AI adoption from a holistic perspective. The research questions include the following:

- RQ1: What are the different perspectives in explaining generative AI adoption?
- RQ2: What are the key determinants influencing generative AI adoption?
- RQ3: What are the relationships among these key determinants?
- RQ4: Are there any differences to the relationships among these key determinants and generative AI adoption at different usage time points?

This study is organized as follows. First, it describes the generative AI phenomena. Then, a literature review was conducted to identify key determinants of generative AI adoption in the past. Moreover, an integrated framework was developed, and a number of hypotheses were defined for testing. In the Materials and Methods section, the background, participants, data collection and data analysis are reported. In the Results section, the instrument validation and the model testing results are reported. The Discussion section explains the results and benchmarks them with prior studies. The theoretical contribution and the practical contribution are discussed, and limitations and topics for further studies are provided.

1.1. Literature Review

1.1.1. Technology Adoption

Technology adoption and acceptance have been extensively investigated over the past few decades as emerging technologies continue to enhance productivity, quality of life, and well-being. A Web of Science search for "technology adoption" returns 115,163 results, including 5714 peer-reviewed articles published between 2020 and 2024. The primary aim of this research stream is to explain the mechanisms and factors underlying technology adoption, typically examined at the individual level with intention to use or actual usage as the dependent variable (e.g., Berényi & Deutsch, 2023; Gong et al., 2025; Zhao et al., 2024).

A range of theoretical frameworks has been employed to study technology adoption, including the Technology Acceptance Model (TAM) (e.g., Geng et al., 2023; Metallo et al., 2022), the Unified Theory of Acceptance and Use of Technology (UTAUT) (e.g., Queiroz et al., 2021; Sorwar et al., 2023), and Technological Pedagogical Content Knowledge (TPACK) (e.g., Zhang et al., 2025). Extended and integrated models are also common, such as TAM combined with TPACK (e.g., Li, 2025) and UTAUT combined with TPACK (e.g., Mohammad-Salehi et al., 2021). In addition to these frameworks, studies frequently

https://doi.org/10.3390/educsci16010128

incorporate context-specific constructs to account for determinants relevant to particular domains. These contextually grounded constructs are discussed below.

1.1.2. Risk Propensity

Risk propensity is defined as an individual's current tendency to take or avoid risks (Sitkin & Weingart, 1995, p. 1575) and is often treated as a stable dispositional trait. However, Sitkin and Weingart (1995) argue that risk propensity can change over time and is thus an emergent property of the decision maker (p. 1575). They attribute this changeability in part to the influence of past experience: as individuals accumulate experience, they may become less susceptible to contextual influences and more likely to exhibit adaptive, cross-situational consistency. Empirically, risk propensity is a strong predictor of risk-taking behavior (Müller et al., 2025). Because emergent technologies typically involve uncertainty and ambiguity, constructs such as risk propensity, risk perception, and risk attitude have been widely examined across technology contexts, including online shopping (Donthu & Gilliland, 1996), peer-to-peer file sharing (Xu et al., 2005), farming technologies (Brick & Visser, 2015), and auditors' adoption of artificial intelligence (Bracci et al., 2025). At the organizational level, evidence from more than 400 small and medium-sized enterprises (SMEs) indicates that risk propensity affects firm-level technology adoption (Doe et al., 2022). In the fintech domain, risk propensity significantly influences Generation Z's peer-to-peer borrowing decisions, with individuals reporting lower risk perceptions being more vulnerable to debt (Yuswandi & Hamdani, 2025). However, findings are not uniformly consistent. Some studies report no direct or indirect effects of risk propensity on behavioral intention. For example, in research on AI-assisted programming, attitudes positively predicted intentions to use ChatGPT, but risk propensity did not affect attitudes (Batac et al., 2024). Overall, risk propensity appears to be a plausible determinant of behavioral intention to use technology, but its effects may depend on contextual factors and may be mediated or moderated by other variables, suggesting potential confounding influences.

1.1.3. Social Influence

Social influence is defined as the extent to which an individual perceives that important others believe he or she should use a new system (Venkatesh et al., 2003, p. 451). Empirical evidence consistently highlights its central role in technology adoption. For preservice teachers, social influence emerged as the most significant positive predictor of behavioral intention to use artificial intelligence in lesson planning (Acquah et al., 2024). In the context of mobile platform applications, it directly and significantly shaped users' continuance intention (Liu et al., 2023). Cross-cultural research on college students in Poland and Egypt similarly found that social influence significantly affected intentions to use generative AI. In healthcare, social influence was a main determinant of intention to use AI doctors across primary, secondary, and tertiary care settings (Uymaz et al., 2024). Within online learning, it positively and significantly predicted students' intention to use Tencent Meeting/VooV Meeting for course participation (Qin & Yu, 2024). Beyond individual adoption, social influence—together with sales technology orientation—was identified as a key driver of social selling at the individual level, supported by organizational social media strategy, tools, and content (Terho et al., 2022). To sum up, social influence appears to be a plausible determinant of behavioral intention to use technology.

1.1.4. Performance Expectance

Performance expectancy—the degree to which an individual believes that using a system will enhance job performance (Venkatesh et al., 2003, p. 447)—has been widely examined across domains, geographies, and application types (e.g., Roy, 2024; Chi et al., 2022). It frequently emerges as a key determinant of adoption. For instance, it was among

https://doi.org/10.3390/educsci16010128

the most significant predictors of intention to use a biometric mobile payment system (Liébana-Cabanillas et al., 2024). Studies of autonomous delivery robots for meals and packages similarly found that performance expectancy influenced acceptance (Kaiser et al., 2024). In healthcare, performance expectancy directly predicted passive clinician resistance to implementing health information technology (E. D. Kim et al., 2023). In the context of mobile platform apps, it had a direct and significant effect on continuance intention (Liu et al., 2023). It also significantly shaped user acceptance of smart home voice assistants (Zhong et al., 2024) and was associated with both actual participation and continued engagement on video-conferencing platforms (Alajmi & Said Ali, 2022). Beyond behavioral intention, performance expectancy was positively related to satisfaction with AI-based digital assistants (Marikyan et al., 2022). Overall, empirical evidence supports performance expectancy as a central determinant of technology acceptance.

1.1.5. Indirect and Mediating Effects

Across diverse technological contexts, these three constructs frequently operate through indirect pathways rather than exerting purely direct effects on behavioral intention. Performance expectancy often functions as a central mediator, translating upstream factors (e.g., risk propensity, social influence, effort expectancy, task–technology fit, and literacy-related capabilities) into adoption-related outcomes. In several cases, indirect pathways attenuate or nullify direct effects, underscoring the importance of modeling mediation and moderation.

Evidence suggests that risk propensity shapes intention primarily via cognitive appraisals and attitudes rather than direct paths. For entrepreneurship, risk-taking propensity increases attitude toward behavior and perceived behavioral control, which in turn predict entrepreneurial intention (Mothibi & Malebana, 2025). In consumer technologies, higher risk propensity enhances perceived usefulness of voice assistants (Sestino et al., 2024), consistent with models where usefulness/performance beliefs mediate between trait risk and adoption.

Social influence is commonly significant but often weaker than core expectancy and attitude constructs. It can exert both direct and indirect effects depending on context. For instance, it had a weak direct effect on intentions to use biometric mobile payment (Liébana-Cabanillas et al., 2024), but in AI teaching preparedness, it operated directly and indirectly via professional development (Ayanwale et al., 2024). Indirect channels frequently run through performance-related beliefs: social influence affected clinician resistance only indirectly via performance expectancy (E. D. Kim et al., 2023) and was a critical antecedent of performance expectancy that ultimately shaped intention to use autonomous vehicles (Ribeiro et al., 2022). Similarly, null direct effects in nursing students' intent to use AI-based healthcare technologies may reflect mediation through performance expectancy (Kwak et al., 2022).

Performance expectancy consistently shows both direct and mediating roles. It mediated the impact of task–technology fit on usage intention in BOPS contexts, whereas social influence was nonsignificant (S. Kim et al., 2022). In automated shuttles, performance expectancy and social influence directly predicted intention, with performance expectancy mediating effort expectancy's effect on intention (Nordhoff et al., 2021).

In sum, risk propensity, social influence, and performance expectancy collectively influence technology adoption through a network of indirect effects, with performance expectancy frequently acting as the central conduit that translates upstream determinants into behavioral intention.

https://doi.org/10.3390/educsci16010128

1.1.6. Short Run and the Long Run

Prior studies suggest that technology adoption is not static and change over time (e.g., Kolil & Achuthan, 2023). As users gain experience or receive training, the relative influence of key determinants can change. Sitkin and Weingart (1995) argue that risk propensity is an emergent, experience-sensitive property rather than a fixed trait. As individuals accrue experience, they become less susceptible to transient contextual cues and are more likely to display cross-situational consistency in risk-taking. This implies that early-stage variability in risk responses may stabilize over time. In a study of older adults' smartphone acceptance before and after training, social influence significantly impacted behavioral intention only after the training (Yang et al., 2023). This suggests that once users share a common frame of reference (from training or early use), normative cues from peers, instructors, or caregivers become more actionable and influential. In the same study of older adults, performance expectancy predicted behavioral intention before training, but not after (Yang et al., 2023). Initially, beliefs about usefulness and performance gains drive intention because users anchor decisions on anticipated benefits. After training, as performance beliefs become more concrete (or variance in these beliefs narrows), their marginal impact on intention can decline, while social influence or habit takes on a larger role.

In sum, determinants of behavioral intention are dynamic: they may evolve with experience, become more salient following exposure or training, or diminish in influence as other factors assume prominence at a longer usage time point.

1.2. Model Framework and Hypotheses Development

With the above literature review, an integrated model comprising individual, social and technology perspective was proposed to explain generative AI adoption by individual college students. The model framework was composed of key determinants of risk propensity (individual perspective), social influence (social perspective) and performance expectancy (technology perspective) to the behavioral intention to use and hence actual usage of generative AI by individual users (see Figure 1).

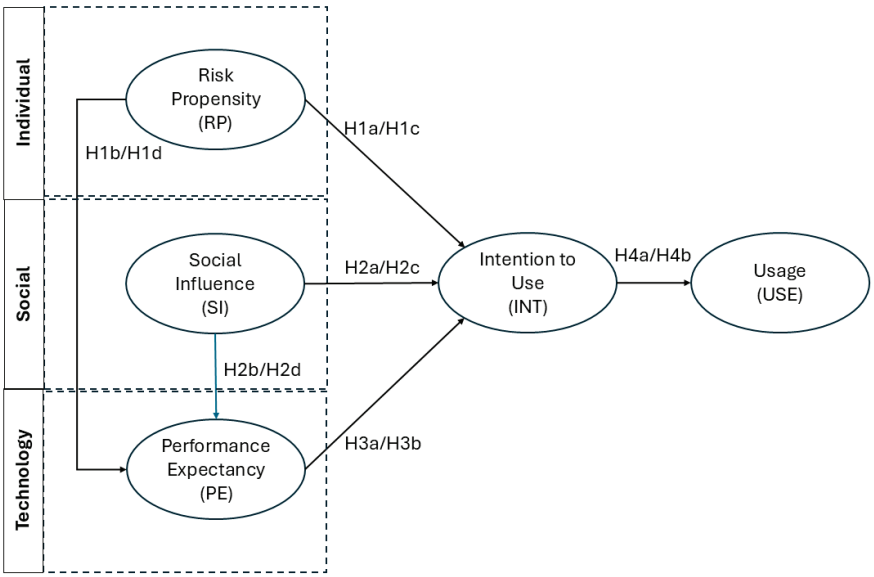


Figure 1. An integrated generative AI adoption model framework.

1.2.1. Risk Propensity

Risk propensity—defined as an individual's current tendency to take or avoid risk (Sitkin & Weingart, 1995, p. 1575)—is not fixed; it evolves with experience. The adoption of

emergent technologies typically entails uncertainty and potential loss (e.g., security and privacy threats in early online purchasing and peer-to-peer file sharing). In the context of generative AI, from the individual perspective, students may anticipate both benefits (enhanced academic performance) and salient risks (accusations of cheating, inaccurate outputs), which can deter use (Freeman, 2025; Marks, 2025). In early stages, greater risk aversion (i.e., lower risk-taking propensity) should heighten the perceived downsides and suppress perceived benefits, leading to lower performance expectancy and weaker intention to use. At a longer usage time point, however, as users accumulate uneventful experience, uncertainty typically declines, usage becomes routinized, and the familiar option is reframed as the safer status quo. Under these conditions, risk-averse individuals may prefer continued use of the now-familiar system and come to perceive higher performance gains, whereas switching to alternatives (new tools or environments) is viewed as the riskier choice. This temporal rebalancing supports distinct short-run and long-run predictions. Therefore, we test,

Hypotheses (short run)

- H1a: Greater risk aversion is associated with lower behavioral intention to use generative AI.
- H1b: Greater risk aversion is associated with lower performance expectancy regarding generative AI.

Hypotheses (long run)

- H1c: At a longer usage time point, as experience accumulates, greater risk aversion is associated with higher behavioral intention to use generative AI.
- H1d: At a longer usage time point, as experience accumulates, greater risk aversion is associated with higher performance expectancy regarding generative AI.

1.2.2. Social Influence

Social influence is the perceived expectation of important others that one should use a particular system (Venkatesh et al., 2003, p. 451). College students are embedded in social networks comprising instructors, peers, and the broader campus community. Within these networks, behavior is shaped by normative pressures and identification processes: students align with valued referents, adhere to community norms, and comply with perceived expectations to maintain belonging and avoid dissonance.

Applied to generative AI, from the social perspective, when students perceive that significant others endorse or use these tools, they are more likely to form intentions to use them.

The influence of normative pressures is also likely to have cognitive consequences. To justify conformity and maintain self-consistency, students exposed to strong social influence may actively search for, attend to, and learn about the performance benefits of generative AI, thereby elevating their performance expectancy. At a longer usage time point, as students gain experience and usage becomes routine, the direct normative pressure can attenuate; intentions may be sustained more by habit and internalized evaluations than by explicit expectations. Nevertheless, social influence can continue to exert indirect effects by shaping learning opportunities, peer support, and shared practices that reinforce perceived performance gains. Therefore, we test,

Hypotheses (short run)

- H2a: Greater social influence is associated with higher behavioral intention to use generative AI.
- H2b: Greater social influence is associated with higher performance expectancy regarding generative AI.

https://doi.org/10.3390/educsci16010128

Hypotheses (long run)

- H2c: At a longer usage time point, as experience accumulates, greater social influence is associated with higher behavioral intention to use generative AI.
- H2d: At a longer usage time point, as experience accumulates, greater social influence is associated with higher performance expectancy regarding generative AI.

1.2.3. Performance Expectancy

Performance expectancy—the belief that using a system will enhance one's task performance (Venkatesh et al., 2003, p. 447)—is central to students' technology adoption. For college students whose core activities include learning, information seeking, and solving everyday problems by applying course knowledge, generative AI can streamline key tasks. By enabling rapid information retrieval, idea generation, and interactive problem-solving, generative AI can increase efficiency relative to traditional methods (e.g., manual library searches and extensive reading). Accordingly, the more students perceive generative AI as useful for their academic tasks, the stronger their intention to use it.

Experience is likely to amplify this relationship. As students become more familiar with generative AI, they learn to formulate effective prompts, evaluate outputs, and integrate the tool into their study routines. These competencies can increase realized benefits, thereby strengthening the link between performance expectancy and intention at a longer usage time point. Therefore, we test,

Hypotheses (short run)

- H3a: Higher performance expectancy regarding generative AI is associated with higher behavioral intention to use generative AI.

Hypothesis (long run)

- H3b: At a longer usage time point, the positive association between performance expectancy and behavioral intention to use generative AI will be stronger than in the short run.

1.2.4. Intention to Use

Behavioral intention is a well-established proxy for predicting technology usage in adoption research (e.g., Gong et al., 2025; Zhao et al., 2024). Although some studies capture actual usage directly (e.g., Berényi & Deutsch, 2023), intention remains a robust antecedent of behavior: when individuals report stronger intentions to use technology, their subsequent usage tends to be higher. Applied to generative AI among college students, higher intention to use should translate into greater actual use. Moreover, this intention–behavior linkage is expected to be stable at a longer usage time point; accumulating experience may have different absolute levels of intention and usage, but not the positive association between them. Therefore, we test,

Hypotheses (short run)

- H4a: Higher behavioral intention to use generative AI is associated with higher actual usage of generative AI among individual students.

Hypothesis (long run)

- H4b: At a longer usage time point, higher behavioral intention to use generative AI remains associated with higher actual usage of generative AI among individual students.

https://doi.org/10.3390/educsci16010128

2. Materials and Methods

2.1. Background

A local tertiary institution in Hong Kong began providing free, quota-based access to ChatGPT for all staff and students on 1 September 2024 (AY2024–2025), allocating 100 credits to each student and 200 credits to each staff member. This study investigated students' usage patterns and perceptions of ChatGPT at the beginning of the initiative and again at the end of the academic year.

2.2. Subjects

The participants were students from a local tertiary institution in Hong Kong. In Stage 1, all students at the college were invited to participate via internal email. A total of 145 students completed the survey (44 male, 101 female), reflecting the institution's overall gender distribution. Participants' ages ranged from 17 to 49 years (M = 22.54). In Stage 2, one class offered in the summer semester was randomly selected. With the instructor's consent, two research assistants invited all enrolled students to participate. A total of 111 students completed the survey (23 male, 87 female, 1 not specified). This class primarily comprised nursing students enrolled in a social media communication course, a field in which female students predominated. Their ages ranged from 18 to 27 years (M = 20.36).

2.3. Measurements

Previously validated scales were adapted for this study. Risk propensity was measured with two items (Xu et al., 2005); performance expectancy (four items), social influence (four items), and intention to use (three items) were drawn from Venkatesh et al. (2003) (see Table A1 in the Appendix A). All constructs were assessed using a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree). Self-reported usage was captured with a 7-point frequency scale ranging from 1 (never/rarely) to 7 (several times per day).

2.4. Data Collection

Data collection occurred in two stages. Stage 1 took place in October 2024, approximately one month after the institution launched its ChatGPT portal on 1 September 2024. An invitation email was distributed to all students via the internal mailing list, directing them to a survey hosted on Microsoft Forms. The survey opened with an informed consent page; participants who indicated agreement proceeded with the questionnaire. Completion time averaged about 10 min, as recorded by Microsoft Forms. The survey remained open for three weeks, during which two reminder emails were sent.

Stage 2 was conducted in mid-July (approximately ten months after the portal's launch). With prior consent from the instructor, two research assistants visited a randomly selected summer-semester class. Students were provided with an informed consent form to sign and return, after which they completed the same survey administered in Stage 1. The survey took approximately 10 min to complete, and the research assistants collected the completed forms on site.

Ethical approval for the study was obtained from the College Research Ethics Committee in July–August 2024, prior to the start of the academic year.

2.5. Data Analysis

It began with a descriptive analysis to profile the sample. This included a detailed summary of participants' demographic characteristics and patterns of ChatGPT usage. Then, a descriptive analysis of all determinants, alongside the computation of means and standard deviations were presented. Building on this foundation, the measurement model

https://doi.org/10.3390/educsci16010128

underwent rigorous evaluation to establish psychometric soundness, with procedures implemented to demonstrate both reliability and validity. Following confirmation of the measurement properties, the structural model was estimated and the hypothesized relationships were systematically examined using Partial Least Squares Structural Equation Modeling (PLS-SEM) (Hair et al., 2022). This analytic approach enabled robust testing of the proposed hypotheses and provided a comprehensive evaluation of the model's explanatory power and path relationships.

3. Results

3.1. Descriptive Analysis of Respondents

System log statistics for ChatGPT usage (1 September 2024–31 August 2025). At the start of the 2024–2025 academic year (September 2024), 4457 students were enrolled. The institution provided all staff and students with access to ChatGPT via a customized online portal launched on 1 September 2024. According to the system logs, after three months of use (by 30 November 2025), there were 1223 student logins, with 865 distinct student users, and total credit consumption of 4793.05. Two students exhausted their full allocation of 100 credits, and 16 students used more than 50 credits. By the end of the academic year (31 August 2025), there were 1626 student logins and 1228 distinct student users, with total credits consumed of 10,656.11. Twelve students exhausted their 100-credit quota, and 48 students used more than 50 credits.

Survey participants. Data collection was conducted in two waves: October 2024 and July 2025. Stage 1 yielded 145 completed questionnaires, and Stage 2 yielded 111. The institutional gender distribution was 32.8% male and 67.2% female. Stage 1 closely mirrored this distribution, whereas Stage 2 was similar but showed a slight deviation from the overall population. Moreover, Stage 1 included a higher proportion of Year 1 students, whereas Stage 2 comprised more students in Years 2 and 3. This sampling issue is discussed further in the Limitations section. The table below summarizes descriptive statistics for respondent characteristics (see Table 1).

Table 1. Descriptive analysis of respondents.

	Stage 1 (October 2024) ***n* = 145**	**Stage 2 (July 2025)** ***n* = 111**
Gender	Male: 44 (30.3%); Female: 101 (60.7%)	Male: 23 (20.9%); Female = 87 (79.1%) (1 not indicated)
Year of Study	Year 1: 69 Year 2: 28 Year 3: 24 Year 4: 19 Year 5: 5	Year 1: 12 Year 2: 48 Year 3: 38 Year 4: 13 Year 5: 0

3.2. Instrument Validation

Table 2 presents descriptive statistics for all constructs—risk propensity (RP), social influence (SI), performance expectancy (PE), intention to use (INT), and usage—including means and standard deviations, Cronbach's alpha, composite reliability (CR), and average variance extracted (AVE).

Reliability denotes the degree to which a measure consistently captures the construct it is intended to assess; it reflects the proportion of true-score variance uncontaminated by measurement error (Hair et al., 2010). Multi-item scales typically yield more reliable estimates than single-item measures. Before conducting substantive analyses, scale reliability was evaluated using Cronbach's alpha, treating values in the 0.60–0.70 range as the lower bound of acceptability (Hair et al., 2010; Nunnally & Bernstein, 1994). In this study, all constructs at both Stage 1 and Stage 2 exceeded 0.70, indicating satisfactory internal

https://doi.org/10.3390/educsci16010128

consistency. Additionally, composite reliability (CR), computed from the confirmatory factor analysis, was used to assess internal consistency; all constructs exhibited CR values above 0.90, supporting strong internal consistency (Bagozzi & Yi, 1988; Hair et al., 2010) (see Table 2).

Table 2. Descriptive analysis of variables.

	Stage 1 (October 2024)				Stage 2 (July 2025)			
	M(SD)	**α**	**CR**	**AVE**	**M(SD)**	**α**	**CR**	**AVE**
RP	5.08 (1.217)	0.818	0.915	0.843	4.52 (1.298)	0.780	0.900	0.819
SI	4.77 (1.183)	0.864	0.907	0.710	4.02 (1.234)	0.891	0.924	0.753
PE	5.45 (1.194)	0.944	0.959	0.856	4.90 (1.551)	0.964	0.974	0.904
INT	5.33 (1.435)	0.965	0.977	0.934	4.75 (1.658)	0.970	0.981	0.944
Usage [1]	3.80 (1.869)				3.47 (1.925)			

[1] Single item.

With reliability established, validity was assessed. Validity refers to the extent to which a measure or set of measures accurately represents the target construct and is free from systematic or nonrandom error; it concerns how well the construct is captured by its indicators (Hair et al., 2010). Convergent and discriminant validity are two widely accepted forms of construct validity.

Convergent validity evaluates the degree to which indicators of the same construct are correlated. In this study, standardized factor loadings and average variance extracted (AVE) provided evidence of convergence: all item loadings on their intended constructs exceeded 0.80, and all AVE values were above 0.70, indicating that each construct explained more than 70% of the variance in its indicators. Taken together, these results support convergent validity (Chin, 1998; Fornell & Larcker, 1981; Hair et al., 2010) (see Table 2).

Discriminant validity assesses the extent to which conceptually related constructs are empirically distinct. Discriminant validity was evaluated using the Fornell–Larcker criterion (Fornell & Larcker, 1981; Henseler et al., 2015). For each construct, the square root of AVE exceeded its correlations with other constructs, and AVE values were greater than the squared interconstruct correlations (see Table 3), consistent with discriminant validity. To supplement these results, the cross-loadings matrix was inspected (Chin, 1998; Hair et al., 2022), which showed that each indicator loaded highest on its intended construct and substantially higher than on any other construct (see Table 4). Collectively, these findings demonstrate satisfactory discriminant validity of the scales.

Table 3. Fornell-Larcker Matrix.

	Stage 1 (October 2024)					Stage 2 (July 2025)				
	RP	**SI**	**PE**	**INT**	**Usage**	**RP**	**SI**	**PE**	**INT**	**Usage**
RP	0.918					0.905				
SI	0.278	0.842				0.541	0.868			
PE	0.256	0.705	0.925			0.572	0.635	0.951		
INT	0.251	0.618	0.728	0.966		0.5	0.473	0.7	0.972	
Usage	−0.058	0.234	0.318	0.446	1	0.244	0.348	0.508	0.495	1

Table 4. Cross-loadings Matrix.

	Stage 1 (October 2024)				Stage 2 (July 2025)			
	INT	**PE**	**RP**	**SI**	**INT**	**PE**	**RP**	**SI**
RP1	0.237	0.285	**0.942**	0.287	0.391	0.485	**0.887**	0.476
RP2	0.223	0.17	**0.894**	0.215	0.506	0.547	**0.923**	0.501
SI1	0.507	0.592	0.208	**0.857**	0.342	0.517	0.499	**0.85**
SI2	0.479	0.564	0.209	**0.832**	0.356	0.497	0.448	**0.888**
SI3	0.565	0.652	0.243	**0.867**	0.493	0.593	0.47	**0.908**
SI4	0.528	0.56	0.275	**0.813**	0.429	0.581	0.459	**0.822**
PE1	0.697	**0.924**	0.18	0.638	0.655	**0.936**	0.554	0.586
PE2	0.663	**0.934**	0.258	0.639	0.678	**0.972**	0.541	0.622
PE3	0.635	**0.932**	0.245	0.632	0.71	**0.962**	0.529	0.583
PE4	0.695	**0.91**	0.263	0.694	0.619	**0.933**	0.555	0.624
INT1	**0.973**	0.712	0.273	0.614	**0.967**	0.684	0.478	0.474
INT2	**0.964**	0.705	0.233	0.604	**0.979**	0.653	0.46	0.418
INT3	**0.962**	0.695	0.221	0.573	**0.968**	0.703	0.519	0.486

Note. Factor loadings on the target construct are shown in bold to emphasize the primary association; bold values represent the highest loading for each item.

In sum, both the reliability and validity of the scale were rigorously assessed and supported by the evidence.

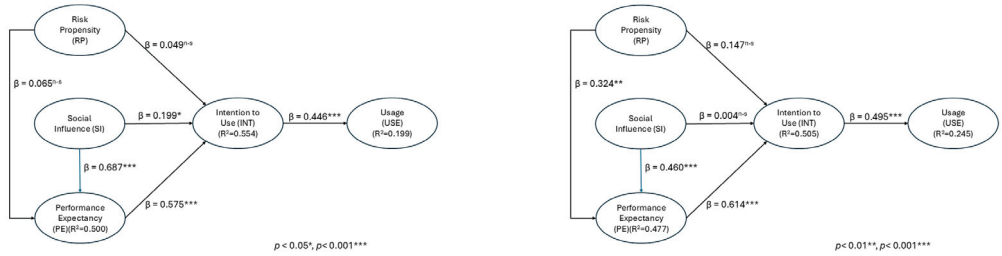
3.3. Model Testing Results

In preliminary analyses, we assessed potential control variables (gender, age, social organization participation, and parents' education level) and found no significant associations with intention to use.

Subsequently, structural model analysis and path coefficient estimation were conducted using PLS-SEM in SmartPLS 4.1.1. PLS-SEM was deemed appropriate because the study's primary objective is prediction and theory development rather than strict theory testing; the model is complex, comprising numerous constructs, indicators, structural paths, and multiple mediations, for which PLS-SEM scales effectively, including hierarchical component models; and the sample size is small to moderate, aligning with PLS-SEM's minimal distributional assumptions (Chin, 1998; Hair et al., 2022).

3.3.1. Overall Model

The model explained a moderate proportion of variance in Intention to Use at both time points (Stage 1: $R^2 = 0.554$, adjusted $R^2 = 0.552$; Stage 2: $R^2 = 0.505$, adjusted $R^2 = 0.492$), consistent with benchmark guidelines in PLS-SEM that classify R^2 values around 0.50 as moderate and around 0.75 as substantial (Hair et al., 2022; Cohen, 1988) (see Figure 2a,b; Table 5). For Usage, explanatory power increased from Stage 1 to Stage 2 ($R^2 = 0.199$ to $R^2 = 0.245$), representing an absolute gain of 0.046 (approximately 23% relative improvement). While Usage remained at the weak-to-lower-moderate boundary (with 0.25 often used as the threshold for weak-to-moderate), such levels are typical in behavioral outcomes where variance is harder to explain than intentions. These results align with field norms in technology adoption research, where R^2 values of approximately 0.40–0.60 for intention and 0.20–0.30 for behavior are commonly regarded as acceptable.



(**a**) (**b**)

Figure 2. Model testing results: (**a**) Stage 1; (**b**) Stage 2.

Table 5. Summary of hypotheses testing.

	Stage 1 (October 2024)		**Stage 2 (July 2025)**	
Hypotheses	**Coefficients**	**Support**	**Coefficients**	**Support**
H1a: RP → INT	β = 0.049, n-s	No	β = 0.147, n-s	No
H1b: RP → PE	β = 0.065, n-s	No	β = 0.324 **	Yes
H2a: SI → INT	β = 0.199 *	Yes	β = 0.004, n-s	No
H2b: SI → PE	β = 0.687 ***	Yes	β = 0.460 ***	Yes
H3: PE → INT	β = 0.575 ***	Yes	β = 0.614 ***	Yes
H4: INT → Usage	β = 0.446 ***	Yes	β = 0.495 ***	Yes
Indirect effect				
RP → PE → INT	β = 0.037, n-s		β = 0.199 **	
RP→ PE → INT → Usage	β = 0.038, n-s		β = 0.171 **	
SI → PE → INT	β = 0.395 ***		β = 0.282 **	
SI→ PE → INT → Usage	β = 0.176 ***		β = 0.140 **	

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

3.3.2. Risk Propensity

At Stage 1, the direct paths from risk propensity to intention to use (H1a) and to performance expectancy (H1b) were not significant; thus, both hypotheses were not supported. Consistent with these null direct effects, the indirect effect of risk propensity on intention to use via performance expectancy (RP -> PE -> INT) was also non-significant (β = 0.037, ns), as was the total indirect effect on usage through intention (RP -> PE -> INT -> Usage; β = 0.038, ns).

At Stage 2, the direct effect of risk propensity on intention to use remained non-significant (H1c not supported), whereas the path from risk propensity to performance expectancy became significant (H1b supported; β = 0.324, $p < 0.01$). In line with this pattern, the indirect effect of risk propensity on intention via performance expectancy was significant (RP -> PE -> INT; β = 0.199, $p < 0.01$), as was the total indirect effect on usage (RP -> PE -> INT -> Usage; β = 0.171, $p < 0.01$).

3.3.3. Social Influence

At Stage 1, both H2a (social influence → intention to use) and H2b (social influence → performance expectancy) were supported, with significant path coefficients (β = 0.199, $p < 0.05$; β = 0.687, $p < 0.001$, respectively). Mediation analysis indicates partial mediation by performance expectancy: social influence had a significant direct effect on intention (β = 0.199, $p < 0.05$) and a significant indirect effect via performance expectancy (SI → PE → INT; β = 0.395, $p < 0.001$). The total indirect effect of social influence on usage through intention was also significant (SI → PE → INT → Usage; β = 0.176, $p < 0.001$).

At Stage 2, the direct path from social influence to intention to use became non-significant (H2c not supported), whereas the path to performance expectancy remained

significant (H2d supported; β = 0.460, $p < 0.001$). In this stage, performance expectancy fully mediated the effect of social influence: the indirect effect on intention via performance expectancy was significant (SI → PE → INT; β = 0.282, $p < 0.01$), and the total indirect effect on usage through intention was likewise significant (SI → PE → INT → Usage; β = 0.140, $p < 0.01$).

3.3.4. Performance Expectancy

Across both Stage 1 and Stage 2, hypotheses H3a and H3b (performance expectancy → intention to use) were supported. The path coefficients were significant and robust at each stage: β = 0.575 ($p < 0.001$) in Stage 1 and β = 0.614 ($p < 0.001$) in Stage 2.

3.3.5. Intention to Use

Across both Stage 1 and Stage 2, hypotheses H4a and H4b (intention to use → usage) were supported. The path coefficients were significant in both stages: β = 0.446 ($p < 0.001$) in Stage 1 and β = 0.495 ($p < 0.001$) in Stage 2.

3.3.6. Independent Analysis on Risk Propensity

As a post hoc analysis, the independent effects of risk propensity on behavioral intention and usage was examined (see Figure 3a,b). At Stage 1, the direct effect of risk propensity on intention was significant (β = 0.251, $p < 0.05$). At Stage 2, this effect approximately doubled and strengthened (β = 0.503, $p < 0.001$). Correspondingly, the explained variance (R^2) in intention increased from 0.063 at Stage 1 to 0.253 at Stage 2—an over fourfold increase—indicating a substantial rise in the proportion of variance in intention accounted for by risk propensity at a longer usage point of time. This is further discussed in the limitation section.

(**a**) (**b**)

Figure 3. Independent risk propensity model testing results: (**a**) Stage 1; (**b**) Stage 2.

4. Discussion

4.1. Key Findings

Key findings are summarized below:

- Model performance: The model showed moderate explanatory power for intention to use (Stage 1: $R^2 = 0.554$, adjusted $R^2 = 0.552$; Stage 2: $R^2 = 0.505$, adjusted $R^2 = 0.492$), consistent with PLS-SEM benchmarks (Hair et al., 2022; Cohen, 1988), and weak-to-lower-moderate but improving explanatory power for usage ($R^2 = 0.199 \rightarrow 0.245$; +0.046, ~23% gain), aligning with field norms that regard ~0.40–0.60 for intention and ~0.20–0.30 for behavior as acceptable.
- Individual factor—risk propensity: Overall, risk propensity shows no influence on intention or usage at Stage 1—either directly or via performance expectancy—indicating no short-run mediation; by contrast, at Stage 2, with the direct effect on intention remaining non-significant and the indirect pathways becoming significant, risk propensity's impact on both intention and usage is fully mediated by performance expectancy.
- Social factor—social influence: The influence of social factors shifts from a combination of direct and indirect pathways at Stage 1 to a purely indirect pathway via performance expectancy at Stage 2.

- Technology factor—performance expectancy: Performance expectancy was a consistent and central determinant, exerting both direct and indirect effects on intention to use at both the initial stage and a longer usage time point.
- Behavioral linkage: Intention to use significantly predicted actual usage

4.2. Individual Factor—Risk Propensity

In the initial post-launch period (Stage 1), college students' risk propensity showed no significant association with their intention to use generative AI, a result that contradicts the hypothesized relationship yet aligns with some prior work (e.g., Batac et al., 2024). This suggests that, early on, decisions to use generative AI were not shaped by students' dispositional tendency to take or avoid risk; in other words, students did not perceive using generative AI as a form of risk-taking that would be moderated by their risk orientation. Such null findings may help explain the limited attention to risk propensity in earlier technology-adoption studies: if treated as a stable trait, risk propensity might appear irrelevant to adoption decisions.

Guided by Sitkin and Weingart's (1995) conceptualization that risk propensity can evolve with experience, its role was examined at two distinctive time points. Although the direct effect of risk propensity on intention remained non-significant at Stage 2, the path from risk propensity to performance expectancy became significant, yielding a significant indirect (mediated) effect on intention and, in turn, on usage. This pattern is consistent with studies that identify risk propensity as a predictor of technology-related intentions (e.g., Bracci et al., 2025; Brick & Visser, 2015; Donthu & Gilliland, 1996; Xu et al., 2005). A plausible explanation is that as students accumulate experience, their risk-related assessments of generative AI—and their beliefs about its academic utility—become more favorable, thereby strengthening performance expectancy and subsequently intention. This temporal dynamic may also account for mixed results in the literature, as findings likely depend on the stage of users' exposure and experience at the time of data collection.

Importantly, the absence of a direct effect on intention indicates that students do not adopt generative AI merely because they deem it "safe"; rather, adoption is driven by perceived performance benefits that develop over time. Practically, institutions seeking to increase adoption—especially among risk-averse students—should emphasize evidence-based best practices, training, and workshops that reinforce both the legitimacy and the academic usefulness of generative AI. These efforts can address salient concerns reported in recent surveys (e.g., fear of cheating accusations, misinformation, and potential harm to academic integrity and performance; Marks, 2025), thereby enhancing performance expectancy and, through it, intention to use. By shifting attention from general ethical concerns to the less-explored role of risk propensity and its evolution, this study advances understanding of student decision-making and offers actionable guidance for implementation strategies.

4.3. Social Factor—Social Influence

Prior empirical research consistently identifies social influence as a significant predictor of technology adoption (Acquah et al., 2024; Liu et al., 2023; Qin & Yu, 2024; Terho et al., 2022; Uymaz et al., 2024). In the initial phase of use (e.g., the first month), college students' intention to use generative AI is positively associated with perceived social influence, aligning with the dominant findings in the literature. Over time, the direct effect of social influence on intention diminishes and becomes fully mediated by performance expectancy; that is, social influence affects intention only through its impact on perceived usefulness. This pattern is consistent with studies documenting indirect and mediating pathways to intention via other belief constructs (Liébana-Cabanillas et al., 2024; Ayanwale et al., 2024). Therefore, higher perceived social influence increases students' performance expectancy

(i.e., perceived usefulness) of generative AI, which in turn raises their intention to use and ultimately their actual usage.

These results provide significant practical implications. Institutions and instructors can shape students' engagement with generative AI. In early stages, interventions that leverage social norms and peer endorsement may be effective. As time progresses, promotional efforts should pivot to demonstrating concrete benefits and providing guidance on how generative AI supports academic tasks, thereby strengthening performance expectancy and sustaining usage.

4.4. Technology Factor—Performance Expectancy

Performance expectancy—defined as the belief that a technology enhances task performance—has been repeatedly identified as a central determinant of technology adoption (Chi et al., 2022; Roy, 2024; Kaiser et al., 2024; Liébana-Cabanillas et al., 2024; Venkatesh et al., 2003). Consistent with this literature, the present study finds that performance expectancy robustly predicts both intention to use generative AI and hence self-reported usage across time. More importantly, the study demonstrates that performance expectancy operates as a key mediator within an integrated adoption framework: it transmits the effects of the individual trait risk propensity and the contextual factor social influence onto intention to use. In other words, risk propensity and social influence do not primarily drive intention through direct pathways; rather, they shape intention insofar as they elevate perceived usefulness, which then increases both intention and actual usage.

The mediating mechanism is straightforward. Individuals with higher risk propensity are more willing to experiment with novel tools, which provides opportunities to observe concrete performance gains (for example, faster drafting, clearer structuring, or more accurate summarization). These mastery experiences raise performance expectancy and, in turn, strengthen intention to use and subsequent usage. Thus, risk propensity motivates initial exploration, but it is the realized utility—captured by performance expectancy—that sustains adoption. Likewise, social influence—via peer endorsement, instructor modeling, or institutional norms—exposes learners to credible use cases, best practices, and vicarious evidence of benefits. Observing others achieve efficiency or quality improvements and receiving guidance that reduces uncertainty increases perceived usefulness; this elevated performance expectancy becomes the proximal driver of intention and behavior. Normative pressure alone is insufficient without clear signals of utility.

These findings suggest actionable strategies for educational institutions and instructors. Early interventions can leverage social endorsement to catalyze trial, but sustaining adoption requires systematically cultivating performance expectancy: provide discipline-specific use cases, structured practice with feedback, and transparent metrics of improvement (e.g., time saved, accuracy gains, output quality). Translating risk propensity and social influence into usefulness signals—through demonstrations, peer showcases, and curated prompt libraries—will more effectively promote ongoing use than risk assessment and normative messaging alone.

4.5. Theoretical Contribution

This study investigated an integrated model of generative AI adoption among college students. Drawing on a systematic literature review, we identified key determinants and assembled them into a testable framework. The results not only confirmed significant associations between these determinants and both intention to use and self-reported usage, but also illuminated the dynamics of the adoption process. Specifically, performance expectancy emerged as a central mediator through which both risk propensity and social influence exert their effects on intention and usage. Moreover, by analyzing two cross-

sectional datasets collected ten months apart, the study documented temporal shifts in the strength and pathways of these effects. Collectively, these findings provide empirical support that advances technology adoption theory and contributes to the broader literature on AI adoption in higher education.

4.6. Practical Contribution

The study's findings enrich our understanding of the generative AI adoption process and yield actionable implications for effective implementation. Regardless of individual differences in risk propensity or the strength of perceived social influence, performance expectancy emerges as both the principal determinant of adoption and the pivotal mediator through which these factors shape intention and use. Accordingly, implementation strategies should prioritize training that introduces best practices tailored to academic contexts, explicitly demonstrating what generative AI can do and how to use it to maximize learning benefits in a legitimate and proper way. Communication should likewise foreground safe and ethical use, offering clear guidance to ensure that AI supports—rather than undermines—learning outcomes and academic integrity, thereby alleviating students' concerns about accusations of cheating and the propagation of false or misleading information when using generative AI.

4.7. Limitations and Future Research

This study has several limitations that qualify the interpretation of its findings. First, the design relied on two cross-sectional samples rather than a longitudinal panel; participants at Stage 1 (n = 145) and Stage 2 (n = 110) were different individuals, precluding within-person analyses of change and limiting causal inference over time. Although the two samples differed in gender distribution and year-of-study composition, both were drawn from the same institutional cohort in which ChatGPT was introduced simultaneously. Future research is recommended to employ a longitudinal panel design to corroborate these findings. Second, the modest sample sizes reduce statistical power, precision, and generalizability; future work would benefit from larger samples (e.g., over 200 per wave) to yield more reliable estimates and support more complex model testing. Third, the model may omit relevant individual-level determinants—such as self-efficacy and risk perception—identified in prior research, raising concerns about omitted-variable bias. Notably, an independent analysis revealed a significant direct effect of risk propensity on intention to use ($\beta = 0.251$, $p < 0.05$), suggesting that the integrated model may exclude important confounders or untested mediating pathways. Collectively, these observations indicate that the current framework is informative but incomplete. Future studies should employ longitudinal designs, increase sample sizes, expand the construct set to include potential mediators and moderators (e.g., trust, perceived risk, risk attitude, self-efficacy, anxiety), and compare alternative model specifications to determine whether the effect of risk propensity on intention is direct, indirect, or fully mediated through other constructs.

5. Conclusions

This study examines a timely and important issue: the adoption of generative AI by college students. On one hand, generative AI represents an innovative tool that can enhance academic performance by clarifying complex concepts, enabling brainstorming of ideas and solutions to real-world problems, and providing continuous, on-demand feedback that complements instructor support. On the other hand, many academic institutions enforce restrictive use policies and deploy AI-detection tools to deter misuse, aiming to prevent plagiarism, cheating, and the spread of misinformation—measures that may inadvertently discourage students from using generative AI altogether. Drawing on the prior literature,

we developed and tested an integrated framework encompassing individual, social, and technological factors to identify the key determinants of adoption and to illuminate the complexity of students' decision-making. The analysis revealed significant relationships among risk propensity, social influence, and performance expectancy with intention to use, which in turn was associated with actual usage. These findings contribute to the literature by advancing theoretical understanding of generative AI adoption and by offering practical implications for implementation within higher education.

Funding: This research received no external funding.

Institutional Review Board Statement: The study received the ethical review and approval by the Research Ethics Committee of TUNG WAH COLLEGE (protocol code REC2024216 on 28 August 2024).

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

Data Availability Statement: Research data is available for sharing.

Acknowledgments: During the preparation of this manuscript/study, the author(s) used ChatGPT gpt-5-2025-08-07 model for the purposes of polishing the language. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
PLS-SEM	Partial Least Squares Structural Equation Modeling

Appendix A Measurement Items

Table A1. Measurement items.

Constructs (Sources)/Items	
Risk Propensity (Xu et al., 2005)	
RP1	I avoid risky things.
RP2	I would rather be safe than sorry.
Social Influence (Venkatesh et al., 2003)	
SI1	People who influence my behavior think that I should use ChatGPT.
SI2	People who are important to me think that I should use ChatGPT.
SI3	The senior management of this business has been helpful in the use of ChatGPT.
SI4	In general, the organization has supported the use of ChatGPT.
Performance Expectancy (Venkatesh et al., 2003)	
PE1	I would find ChatGPT useful in my study.
PE2	Using ChatGPT enables me to accomplish tasks more quickly.
PE3	Using ChatGPT increases my productivity.
PE4	If I use ChatGPT, I will increase my chances of getting better academic performance.
Intention to Use (Venkatesh et al., 2003)	
INT1	I intend to use ChatGPT in the next 4 weeks.
INT2	I predict I would use ChatGPT in the next 4 weeks.
INT3	I plan to use ChatGPT in the next 4 weeks.

References

Acquah, B. Y. S., Arthur, F., Salifu, I., Quayson, E., & Nortey, S. A. (2024). Preservice teachers' behavioural intention to use artificial intelligence in lesson planning: A dual-staged PLS-SEM-ANN approach. *Computers and Education: Artificial Intelligence*, *7*, 100307. [CrossRef]

Alajmi, M. A., & Said Ali, M. (2022). Video-conference platforms: Understanding the antecedents and consequences of participating in or attending virtual conferences in developing countries. *International Journal of Human–Computer Interaction*, *38*(13), 1195–1211. [CrossRef]

Annamalai, N., Bervell, B., Mireku, D. O., & Andoh, R. P. K. (2025). Artificial intelligence in higher education: Modelling students' motivation for continuous use of ChatGPT based on a modified self-determination theory. *Computers and Education: Artificial Intelligence*, *8*, 100346. [CrossRef]

Ayanwale, M. A., Ntshangase, S. D., Adelana, O. P., Afolabi, K. W., Adam, U. A., & Olatunbosun, S. O. (2024). Navigating the future: Exploring in-service teachers' preparedness for artificial intelligence integration into South African schools. *Computers and Education: Artificial Intelligence*, *7*, 100330. [CrossRef]

Bagozzi, R. P., & Yi, Y. (1988). On the evaluation of structural equation models. *Journal of the Academy of Marketing Science*, *16*(1), 74–94. [CrossRef]

Batac, C. A., Baroja, M. J., Caballero, D. J. D., Coloma, L. G., Tan, L. M., & Ebardo, R. (2024, April 26–29). *Do human beliefs and traits influence the adoption of ChatGPT among programming students?* 2024 10th International Conference on Computing and Artificial Intelligence (pp. 339–344), Bali, Indonesia.

Berényi, L., & Deutsch, N. (2023). Technology adoption among higher education students. *Vezetéstudomány/Budapest Management Review*, *54*(11), 28–39. [CrossRef]

Bracci, E., Tallaki, M., & Ebua Otia, J. (2025). Propensity factors of artificial intelligence technology adoption by public sector auditors. *Journal of Public Budgeting, Accounting & Financial Management*, *ahead-of-print*. [CrossRef]

Brick, K., & Visser, M. (2015). Risk preferences, technology adoption and insurance uptake: A framed experiment. *Journal of Economic Behavior & Organization*, *118*, 383–396. [CrossRef]

Burriss, S. K., Hutchins, N., Conley, Z., Deweese, M. M., Doe, Y. J., Eeds, A., Villanueva, A., Ziegler, H., & Oliver, K. (2024). Redesigning an AI bill of rights with/for young people: Principles for exploring AI ethics with middle and high school students. *Computers and Education: Artificial Intelligence*, *7*, 100317. [CrossRef]

Chi, O. H., Gursoy, D., & Chi, C. G. (2022). Tourists' attitudes toward the use of artificially intelligent (AI) devices in tourism service delivery: Moderating role of service value seeking. *Journal of Travel Research*, *61*(1), 170–185. [CrossRef]

Chin, W. W. (1998). The partial least squares approach to structural equation modeling. In G. A. Marcoulides (Ed.), *Modern methods for business research* (pp. 295–336). Lawrence Erlbaum.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates. [CrossRef]

Doe, J. K., Van de Wetering, R., Honyenuga, B., & Versendaal, J. (2022). Extended contextual validation of stakeholder approach to firm technology adoption: Moderating and mediating relationships in an innovation eco-system. *Society and Business Review*, *17*(4), 506–540. [CrossRef]

Donthu, N., & Gilliland, D. (1996). Observations: The infomercial shopper. *Journal of Advertising Research*, *36*(2), 69–76. [CrossRef]

Farhi, F., Jeljeli, R., Aburezeq, I., Dweikat, F. F., Al-shami, S. A., & Slamene, R. (2023). Analyzing the students' views, concerns, and perceived ethics about chat GPT usage. *Computers and Education: Artificial Intelligence*, *5*, 100180. [CrossRef]

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39–50. [CrossRef]

Freeman, J. (2025). *Student generative AI survey 2025*. HEPI Policy Note 61, Higher Education Policy Institute (HEPI). Available online: https://www.hepi.ac.uk/reports/student-generative-ai-survey-2025/ (accessed on 23 November 2025).

Geng, L., Hui, H., Liang, X., Yan, S., & Xue, Y. (2023). Factors affecting intention toward ICT adoption in rural entrepreneurship: Understanding the differences between business types of organizations and previous experience of entrepreneurs. *Sage Open*, *13*(3), 21582440231197112. [CrossRef]

Gong, Y., Xu, C., Luo, S., & Lin, J. (2025). Modeling teacher education students' adoption of large language models through an extended technology acceptance framework. *Scientific Reports*, *15*(1), 32208. [CrossRef]

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Pearson Prentice Hall. Available online: https://openlibrary.org/works/OL16979906W/Multivariate_data_analysis?edition=key:/books/OL22691711M (accessed on 1 November 2025).

Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2022). *A primer on partial least squares structural equation modeling (PLS-SEM)* (3rd ed.). Sage. Available online: https://uk.sagepub.com/en-gb/eur/a-primer-on-partial-least-squares-structural-equation-modeling-pls-sem/book270548#: :text=Other%20Titles%20in:,popular%20software%20package%20SmartPLS%203 (accessed on 1 November 2025).

Hemdanou, A. L., Sefian, M. L., Achtoun, Y., & Tahiri, I. (2024). Comparative analysis of feature selection and extraction methods for student performance prediction across different machine learning models. *Computers and Education: Artificial Intelligence*, *7*, 100301. [CrossRef]

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, *43*(1), 115–135. [CrossRef]

Huang, D., Hash, N., Cummings, J. J., & Prena, K. (2025). Academic cheating with generative AI: Exploring a moral extension of the theory of planned behavior. *Computers and Education: Artificial Intelligence*, *8*, 100424. [CrossRef]

Johnson, A. (2023, January 18). *ChatGPT in schools: Here's where it's banned—And how it could potentially help students*. Forbes. Available online: https://www.forbes.com/sites/ariannajohnson/2023/01/18/chatgpt-in-schools-heres-where-its-banned-and-how-it-could-potentially-help-students/ (accessed on 23 November 2025).

Kaiser, R., De Benedetto, S., Planing, P., & Müller, P. (2024). What will the delivery robots bring us tomorrow? *International Journal of Consumer Studies*, *48*(5), e13093. [CrossRef]

Kim, E. D., Kuan, K. K., Vaghasiya, M. R., Penm, J., Gunja, N., El Amrani, R., & Poon, S. K. (2023). Passive resistance to health information technology implementation: The case of electronic medication management system. *Behaviour & Information Technology*, *42*(13), 2308–2329. [CrossRef]

Kim, S., Connerton, T. P., & Park, C. (2022). Transforming the automotive retail: Drivers for customers' omnichannel BOPS (Buy Online & Pick up in Store) behavior. *Journal of Business Research*, *139*, 411–425. [CrossRef]

Kolil, V. K., & Achuthan, K. (2023). Longitudinal study of teacher acceptance of mobile virtual labs. *Education and Information Technologies*, *28*(7), 7763–7796. [CrossRef]

Kwak, Y., Seo, Y. H., & Ahn, J. W. (2022). Nursing students' intent to use AI-based healthcare technology: Path analysis using the unified theory of acceptance and use of technology. *Nurse Education Today*, *119*, 105541. [CrossRef] [PubMed]

Li, M. (2025). Integrating artificial intelligence in primary mathematics education: Investigating internal and external influences on teacher adoption. *International Journal of Science and Mathematics Education*, *23*(5), 1283–1308. [CrossRef]

Liébana-Cabanillas, F., Kalinic, Z., Muñoz-Leiva, F., & Higueras-Castillo, E. (2024). Biometric m-payment systems: A multi-analytical approach to determining use intention. *Information & Management*, *61*(2), 103907. [CrossRef]

Liu, C. T., Guo, Y. M., & Huang, S. R. (2023). The factors affecting customers' satisfaction and continuance intention in platform-to-consumer environments: A case of mobile food ordering platforms. *e-Service Journal*, *15*(1), 1–28. [CrossRef]

Marikyan, D., Papagiannidis, S., Rana, O. F., Ranjan, R., & Morgan, G. (2022). "Alexa, let's talk about my productivity": The impact of digital assistants on work productivity. *Journal of Business Research*, *142*, 572–584. [CrossRef]

Marks, L. (2025, September). *What 2025 generative AI trends reveal about student behavior?* Turnitin Blog. Available online: https://www.turnitin.com/blog/what-2025-generative-ai-trends-reveal-about-student-behavior (accessed on 23 November 2025).

Metallo, C., Agrifoglio, R., Lepore, L., & Landriani, L. (2022). Explaing users' technology acceptance through national cultural values in the hospital context. *BMC Health Services Research*, *22*(1), 84. [CrossRef]

Mohammad-Salehi, B., Vaez-Dalili, M., & Heidari Tabrizi, H. (2021). Investigating factors that influence EFL teachers' adoption of web 2.0 technologies: Evidence from applying the UTAUT and TPACK. *Tesl-Ej*, *25*(1), n1. Available online: https://tesl-ej.org/wordpress/issues/volume25/ej97a/ej97a21/ (accessed on 1 November 2025).

Mothibi, N. H., & Malebana, M. J. (2025). Entrepreneurship education, role models, and risk-taking propensity as predictors of entrepreneurial intention and behaviour: Evidence from TVET and university students in Gauteng, South Africa. *Administrative Sciences*, *15*(10), 374. [CrossRef]

Müller, W., Leyer, M., & Gaugel, M. (2025). Trust towards using autonomous taxis: Evidence from Germany. *Transportation Research Part F: Traffic Psychology and Behaviour*, *113*, 357–373. [CrossRef]

Nazaretsky, T., Mejia-Domenzain, P., Swamy, V., Frej, J., & Käser, T. (2025). The critical role of trust in adopting AI-powered educational technology for learning: An instrument for measuring student perceptions. *Computers and Education: Artificial Intelligence*, *8*, 100368. [CrossRef]

Nordhoff, S., Madigan, R., Van Arem, B., Merat, N., & Happee, R. (2021). Interrelationships among predictors of automated vehicle acceptance: A structural equation modelling approach. *Theoretical Issues in Ergonomics Science*, *22*(4), 383–408. [CrossRef]

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill. Available online: https://books.google.com.hk/books/about/Psychometric_Theory.html?id=r0fuAAAAMAAJ&redir_esc=y (accessed on 1 November 2025).

Qin, R., & Yu, Z. (2024). Extending the UTAUT model of tencent meeting for online courses by including community of inquiry and collaborative learning constructs. *International Journal of Human–Computer Interaction*, *40*(18), 5279–5297. [CrossRef]

Queiroz, M. M., Fosso Wamba, S., De Bourmont, M., & Telles, R. (2021). Blockchain adoption in operations and supply chain management: Empirical evidence from an emerging economy. *International Journal of Production Research*, *59*(20), 6087–6103. [CrossRef]

Ribeiro, M. A., Gursoy, D., & Chi, O. H. (2022). Customer acceptance of autonomous vehicles in travel and tourism. *Journal of Travel Research*, *61*(3), 620–636. [CrossRef]

Roy, P. (2024). What drives the adoption of data analytics at Australian university libraries in the perspective of UTAUT2? *The Journal of Academic Librarianship*, *50*(5), 102927. [CrossRef]

Sestino, A., Amatulli, C., Peluso, A. M., & Guido, G. (2024). Integrating internet-of-things technologies in luxury industries: The roles of consumers' openness to technological innovations and status consumption. *Technology Analysis & Strategic Management*, *36*(11), 3577–3591. [CrossRef]

Sitkin, S. B., & Weingart, L. R. (1995). Determinants of risky decision-making behavior: A test of the mediating role of risk perceptions and propensity. *Academy of Management Journal*, *38*(6), 1573–1592. [CrossRef]

Sorwar, G., Aggar, C., Penman, O., Seton, C., & Ward, A. (2023). Factors that predict the acceptance and adoption of smart home technology by seniors in Australia: A structural equation model with longitudinal data. *Informatics for Health and Social Care*, *48*(1), 80–94. [CrossRef]

Terho, H., Giovannetti, M., & Cardinali, S. (2022). Measuring B2B social selling: Key activities, antecedents and performance outcomes. *Industrial Marketing Management*, *101*, 208–222. [CrossRef]

Uymaz, P., Uymaz, A. O., & Akgül, Y. (2024). Assessing the behavioral intention of individuals to use an AI doctor at the primary, secondary, and tertiary care levels. *International Journal of Human–Computer Interaction*, *40*(18), 5229–5246. [CrossRef]

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425–478. [CrossRef]

Xu, H., Wang, H., & Teo, H. H. (2005, January 3–6). *Predicting the usage of P2P sharing software: The role of trust and perceived risk*. 38th IEEE Annual Hawaii International Conference on System Sciences (p. 201a), Big Island, HI, USA.

Yang, C. C., Liu, C., & Wang, Y. S. (2023). The acceptance and use of smartphones among older adults: Differences in UTAUT determinants before and after training. *Library Hi Tech*, *41*(5), 1357–1375. [CrossRef]

Yuswandi, A., & Hamdani, H. (2025). Influence of materialism, financial literacy, and risk perception on propensity to indebtedness among peer-to-peer lending users: A case study of generation Z in Tangerang City. *Eduvest-Journal of Universal Studies*, *5*(9), 11848–11861. [CrossRef]

Zhang, H., Wang, Z., Jiang, R., & Wu, H. (2025). Exploring the influence of technology-enhanced active learning environments on pre-service teachers' TPACK and technology beliefs. *Sage Open*, *15*(3), 21582440251359823. [CrossRef]

Zhao, Y., Li, Y., Xiao, Y., Chang, H., & Liu, B. (2024). Factors influencing the acceptance of ChatGPT in high education: An integrated model with PLS-SEM and fsQCA approach. *Sage Open*, *14*(4), 21582440241289835. [CrossRef]

Zhong, R., Ma, M., Zhou, Y., Lin, Q., Li, L., & Zhang, N. (2024). User acceptance of smart home voice assistant: A comparison among younger, middle-aged, and older adults. *Universal Access in the Information Society*, *23*(1), 275–292. [CrossRef] [PubMed]

education sciences

MDPI

Article

Evaluating the Relationship Between Pre-Service Teachers' Artificial Intelligence Readiness and Professional Self-Efficacy

Kuralay Baimukhambetova [1], Kalibek Ybyraimzhanov [1], Kulakhmet Moldabek [2], Ulsana Borashkyzy Akhatayeva [2], Aliya Zhetkizgenova [3] and Elmira Uaidullakyzy [4,*]

[1] Faculty of Pedagogy and Psychology, Zhetysu University named after Ilyas Zhansugurov, 187a, Zhansugurova Street, Taldykorgan 040009, Kazakhstan
[2] Faculty of Pedagogy and Psychology, South Kazakhstan State Pedagogical University, 13, A. Baitursynova Street, Shymkent 160012, Kazakhstan
[3] Faculty of Pedagogy and Psychology, Kazakh Agrarian University, 62, Zhenis Avenue, Almaty Z05K7B0, Kazakhstan
[4] Faculty of Pedagogy and Psychology, Korkyt Ata Kyzylorda State University, Aiteke bi 29A, Kyzylorda 120000, Kazakhstan
* Correspondence: elmira.uaidullakyzy1988@gmail.com

Abstract

The rapid development of educational technologies requires a deeper understanding of pre-service teachers' readiness for artificial Intelligence and the extent to which their professional self-efficacy beliefs influence this process. Although the integration of emerging technologies has gained increasing attention, the relationship between technological competence and professional confidence among pre-service teachers remains underexplored. This study aims to investigate the interplay between pre-service teachers' readiness for artificial intelligence and their professional self-efficacy. An exploration sequential mixed method design was employed, beginning with a quantitative phase involving 293 pre-service teachers, followed by a qualitative phase to capture deeper insights. Findings revealed that pre-service teachers demonstrated an elevated level of readiness for artificial intelligence and positive self-efficacy beliefs, yet no meaningful relationship emerged between the two variables. The results suggest that professional self-efficacy and technological readiness are influenced by broader contextual and pedagogical factors rather than functioning in a straightforward manner. In the qualitative phase, participants highlighted both opportunities and challenges related to the use of artificial intelligence in primary education. While many emphasized its potential to support personalized learning, reduce workload, and enhance student adaptability, concerns were raised about ethical implications, risks to social-emotional development, cultural values, digital literacy gaps, and infrastructural limitations. The study underscores the necessity for teacher education programs to extend beyond technical training by incorporating pedagogical, ethical, and cultural dimensions to prepare pre-service teachers for meaningful integration of artificial intelligence into educational practice.

Keywords: artificial intelligence; digital literacy; pre-service teachers; self-efficacy; teacher education

1. Introduction

The training of future educators for artificial intelligence (AI) significantly influences their perception of professional self-efficacy in educational settings. The preparation

for AI includes both the knowledge and attitudes that future educators maintain on the integration of AI technologies into their teaching practices. Rajapakse et al. (2024) assert that a substantial link exists between teachers' preparedness for implementing AI and their confidence in efficiently instructing pupils in this area. The analysis of a cohort of pre-service teachers indicated that individuals with enhanced AI preparation exhibited increased confidence in their capacity to instruct on AI-related content, implying that pedagogical trust can be strengthened through targeted training in AI tools and applications. Liu (2025) investigated the mediating influence of AI preparation, suggesting that enhanced preparation for technologies significantly enhances the relationship between the adoption of these technologies and a teacher's self-efficacy beliefs. This emphasizes the necessity for pre-service educators to engage actively in AI training to enhance security in their professional practice.

The convergence of digital literacy and self-efficacy beliefs elucidates the dynamics involved in AI preparation. Lim (2023) identified a positive correlation between the digital competencies of early childhood educators and their perspectives on artificial intelligence instruction for young children. This discovery underscores the significance of digital competence as a prerequisite for the successful incorporation of AI in educational practices, hence impacting educators' self-efficacy. Future educators' digital literacy enhances the seamless integration of AI technologies in their instruction and bolsters their confidence in utilizing these resources to enhance educational outcomes. Furthermore, the study conducted by Yao and Wang (2024) broadens this discussion to encompass special education, illustrating that digital literacy enhances the self-efficacy beliefs of special education teachers prior to their service on the utilization of AI. They found that individuals with a solid foundation in digital abilities felt more competent and secure in utilizing AI to modify their instructional tactics for diverse students, thereby underscoring the essential importance of technology proficiency in shaping self-efficacy.

Existing studies collectively demonstrate that the cultivation of AI preparation among pre-service educators is strongly associated with their beliefs in professional self-efficacy. By contextualizing professional development programs that prioritize AI and digital literacy, teacher training institutions can enhance future educators' confidence in implementing technology advancements in their classrooms (Guettala et al., 2023; Alshorman, 2024; Chiu et al., 2025; Farooq, 2025). This involvement directly impacts overall educational outcomes, as enhanced self-efficacy correlates with increased student participation and achievement. Further exploration of these complex relationships could yield essential insights into the requisite elements of teacher preparation necessary to equip educators for the evolving needs of educational technology (Zainuddin et al., 2024). The ramifications of preparing future educators for artificial intelligence (AI) are substantial for teacher training programs and the resulting educational outcomes. Guan et al. (2025) underscore the imperative of comprehensive training in AI integration for pre-service teachers, claiming that such training is essential for cultivating a sense of preparedness and augmenting beliefs in professional self-efficacy. Guan et al.'s (2025) findings indicate a robust reciprocal association between a teacher's preparation to accept AI and their confidence in efficiently utilizing technology in educational settings. When teacher training programs integrate comprehensive AI training, they furnish educators with essential skills and knowledge, hence enhancing their self-efficacy views.

Moreover, the significance of professional development in enhancing the pedagogical skills of AI is undeniable. Sun et al. (2023) investigated a methodology grounded in the understanding of technological pedagogical content knowledge (TPACK), concentrating on its utility for K-12 computer science educators in the development of AI competitions. This research demonstrates that systematic and ongoing professional development is crucial for

fostering the knowledge and confidence required to incorporate AI into educational practices. When examining the convergence of technology, pedagogy, and knowledge, teacher training programs can enhance the self-efficacy and readiness of pre-service educators to utilize AI tools effectively.

The AI preparation scale by Ramazanoglu and Akın (2025) underscores an immediate necessity for training programs focused on AI competencies. The scale serves as a benchmark to assess instructors' readiness to integrate AI into their teaching methodologies, emphasizing the necessity of tailored training programs. The findings of their study indicate that comprehending the preparedness levels of pre-service teachers can guide the development of training programs that emphasize AI literacy while also enhancing instructors' self-efficacy views.

Martin et al. (2020) investigated various course design attributes that significantly enhance pre-service teachers' perceptions of their instructional competencies, especially for information and communication technology (ICT). Their findings indicate that specific instructional tactics, including experiential learning opportunities and tutoring, are essential for fostering self-efficacy among trained educators. The correlation between course design attributes and enhanced self-efficacy suggests that teacher training programs must emphasize psychological and pedagogical frameworks that foster confidence in employing AI as an educational instrument. Akcil et al. (2021) discovered that integration technology is a complex and multifaceted process characterized by several dynamics, and that complete integration is unattainable. As a result, suggestions were provided concerning diverse models, artificial intelligence, and Google Workspace tools to facilitate technological integration by the challenges outlined in the research.

The collective research underscores a significant connection between the training of pre-service teachers in AI and their attitudes toward professional self-efficacy. Effective teacher training programs that address these characteristics enhance educators' confidence and skill in integrating technology into their work, which eventually supports improved educational outcomes for students. The correlation between the training of pre-service educators in artificial intelligence (AI) and their professional self-efficacy significantly influences educational outcomes. Oran (2023) convincingly illustrates that educators with elevated self-efficacy employ more effective teaching methodologies, resulting in enhanced student performance. These results suggest that the enhancement of self-efficacy is not merely an individual characteristic of educators but a vital element that can transform educational effectiveness and, consequently, student learning outcomes. This link indicates that when pre-service educators possess confidence in their capacity to incorporate AI technologies into their instruction, they are inclined to employ more innovative and successful pedagogical practices. These tactics can enhance student engagement and comprehension, resulting in improved academic outcomes.

Moreover, Zhang et al. (2023) expand this analysis through a multigroup examination of pre-service teachers' acceptance of AI in educational settings. Their findings indicate considerable differences in how various sectors of this population understand and are equipped for AI technologies. Individuals with superior preparation for AI have a heightened inclination to embrace enhanced pedagogical methods involving technology, as well as to indicate elevated self-efficacy. Training programs must consider these disparities in AI preparedness, personalizing techniques to enhance self-efficacy in those who may be more resistant or less equipped for this technological transition. By acknowledging and addressing these disparities, teacher preparation can more efficiently equip educators to confront the challenges of contemporary classrooms, hence enhancing educational outcomes.

Cultural effects significantly shape opinions of AI and its incorporation into educational procedures. Viberg et al. (2023) emphasize that varying cultural contexts can result

in differing opinions regarding the effectiveness of AI in education. For instance, educators from cultures emphasizing social learning may utilize AI tools differently than their counterparts from more individualistic cultures. Cultural differences can influence instructors' perceptions of self-efficacy and their inclination to embrace innovative technologies. The interplay between cultural context and technology readiness creates a dynamic setting that either fosters or inhibits educators' self-efficacy views.

Given these complex relationships, it is evident that equipping pre-service educators with the skills and confidence necessary to effectively interact with AI is essential. Adebagbo (2025) asserts that a comprehensive teacher training program incorporating AI preparation enhances self-efficacy and equips educators to address the diverse issues encountered in contemporary classrooms. Ayanwale et al. (2022) assert the necessity of aligning teacher education priorities with the evolving educational landscape influenced by technology. This alignment has the potential to enhance learning experiences across diverse contexts, leading to improved educational outcomes. The scholarly discussion underscores that education for AI among pre-service educators is crucial, as it strongly correlates with their self-efficacy beliefs and, consequently, with the overall educational outcomes for their pupils.

Based on the findings listed above, the rapid proliferation of artificial intelligence technologies in education makes it crucial to determine the extent to which pre-service teachers are ready for this transformation and their professional efficacy beliefs. Teachers' attitudes toward technology and their perceptions of self-efficacy directly impact how these technologies are implemented in the classroom. In this context, comprehensive studies that address both pre-service teachers' AI readiness and their professional self-efficacy beliefs are quite limited.

This study aims to evaluate the relationship between teacher candidates' readiness for artificial intelligence and their professional self-efficacy beliefs.

To achieve this aim, answers to the following questions were sought.

1. What is the level of preparedness of teacher candidates for artificial intelligence?
2. What is the level of professional self-efficacy beliefs of pre-service teachers?
3. What is the relationship between teacher candidates' readiness for artificial intelligence and their professional self-efficacy beliefs?
4. What are the views of pre-service teachers on the use of artificial intelligence in primary education?

2. Materials and Methods

2.1. Research Model

This research was conducted using an exploratory sequential mixed-method design. Quantitative data were first collected, followed by qualitative data to further explain and support this data. This methodology allowed us to determine the relationship between pre-service teachers' AI readiness levels and their professional self-efficacy beliefs using numerical data, while their views on the use of AI at the primary school level were analyzed in depth using qualitative data. Mixed methods is a research approach that uses both quantitative (closed-ended) and qualitative (open-ended) data to gain a deeper understanding of research questions, and where these data are analyzed and interpreted in a mutually complementary manner. This method is particularly popular in fields based on human interaction, such as the social, health, and behavioral sciences (Creswell, 2021).

This study also provides in-depth qualitative data on pre-service teachers' views on the use of artificial intelligence in primary education, demonstrating that technological integration should be evaluated not only from a technical perspective but also from a pedagogical and ethical perspective. This study aimed to offer concrete recommendations

https://doi.org/10.3390/educsci16010043

for updating the content of teacher education programs and developing strategic plans for artificial intelligence education. In this respect, the study offers guidance for educational policies, teacher education programs, and the integration of technology.

2.2. Participants

The study group consisted of 293 pre-service teachers enrolled in the Faculty of Education at a state university in Kazakhstan during the 2024–2025 academic year. The qualitative study was conducted with 20 pre-service teachers selected from this quantitative sample using a purposive sampling method. Participants were selected from a variety of grade levels and branches, aiming to contribute to the study in a multifaceted manner. This provided a more comprehensive assessment of pre-service teachers' readiness for artificial intelligence and their professional self-efficacy beliefs in the Kazakh context.

2.3. Data Collection Tools

The study yielded both quantitative and qualitative data. Data collection was conducted using three different tools: one qualitative and two quantitative. These tools aimed to quantitatively measure pre-service teachers' readiness levels for artificial intelligence and their professional self-efficacy beliefs, as well as to provide an in-depth, qualitative analysis of their views on the use of artificial intelligence in primary education. The scales used were adapted to Kazakh culture, and validity and reliability studies were conducted.

2.3.1. Artificial Intelligence Readiness Scale

In this study, the Artificial Intelligence Readiness Scale, developed by B. Wang et al. (2023) and adapted into Turkish by Özüdoğru and Yıldız Durak (2024), was used to determine teacher candidates' readiness levels for artificial intelligence. The scale is structured on a 5-point Likert-type scale and consists of 18 items in total. According to the rating scale, 1 is scored as strongly disagree, and 5 is scored as strongly agree. This scale consists of 18 items and 4 sub-dimensions. The sub-dimensions are "cognition, ability, vision, and ethics in teaching". The score ranges for the answer options of the scale are calculated as: 1:00–1:79 Strongly Disagree, 1:80–2.59 Agree, 2.60–3.39 Undecided, 3.40–4.19 Agree, and 4.20–5:00 Strongly Agree.

The "cognition" sub-dimension (5 items) measures participants' understanding of the role of teachers in the age of artificial intelligence. The "Competence" sub-dimension (6 items) assesses their ability to use artificial intelligence effectively in the classroom. The "Vision" sub-dimension (3 items) measures candidates' foresight regarding the potential of artificial intelligence in education. The "Ethics" sub-dimension (4 items) addresses teachers' awareness of their ethical responsibilities in the use of artificial intelligence.

The scale was subjected to linguistic adaptation from Turkish to Kazakh. During the translation process, two linguists translated it back and forth, and the original and Kazakh versions were compared in terms of semantic equivalence. Necessary corrections were made in line with expert opinions to form the definitive version of the scale. In addition, the finalized items were carefully read and approved by two (2) educational technologists and two (2) experts with doctorates in educational psychology. As part of the pilot study, exploratory factor analysis (EFA) was conducted on 132 pre-service teachers. The KMO value was 0.79, the Bartlett test result was $\chi^2 = 548.665$, $p < 0.001$, and the factor structure of the scale was determined to be appropriate. The four (4) factor structure, explaining 75.9% of the total variance, was preserved. Cronbach's Alpha reliability coefficient of the scale was calculated as 0.872, while the sub-dimensions were calculated as "cognition" 0.819, "competence" 0.889, "vision" 0.850, and ethics 0.933.

https://doi.org/10.3390/educsci16010043

2.3.2. Professional Self-Efficacy Beliefs Scale

The Professional Self-Efficacy Beliefs Scale, developed by Çolak et al. (2017), was used to determine pre-service teachers' professional self-efficacy perceptions. The scale consists of four sub-dimensions: "academic," "social," "intellectual," and "professional competence." The scale was rated on a 5-point Likert-type scale. The score ranges for the answer options of the scale are calculated as follows: 1:00–1:79 Strongly Disagree, 1:80–2.59 Agree, 2.60–3.39 Undecided, 3.40–4.19 Agree, and 4.20–5:00 Strongly Agree.

This scale was also adapted from Turkish to Kazakh using a forward-backward translation method, and expert assessments were obtained as part of linguistic equivalence studies. A pilot study was conducted with a separate sample of 132 teacher candidates. The overall reliability coefficient of the scale was found to be Cronbach's alpha = 0.858. The values for the sub-dimensions were calculated as "academic" 0.828, "social" 0.863, "intellectual" 0.851, and "professional competence" 0.893. The findings demonstrated that the scale is valid and dependable in the Kazakh context.

2.3.3. Semi-Structured Interview Form

For the qualitative aspect of the study, a semi-structured interview form was developed to examine pre-service teachers' views on the use of artificial intelligence in primary education. The form included open-ended questions regarding the role of artificial intelligence in education, its impact on teaching processes, potential advantages and disadvantages, ethical and pedagogical implications, and recommendations.

During the form development process, field experts (at least PhD-level, working in primary education programs) were consulted to assess its content validity. Adjustments were made to the form based on feedback from four experts working in primary education programs who hold at least PhD degrees. Additionally, the interview form was piloted with three pre-service teachers to evaluate for clarity and appropriateness of questions. Pilot participants were not included in the main study sample. During the interviews, each participant's responses were clarified when necessary, and the researchers took care to avoid any loss of meaning. Qualitative data were analyzed by dividing them into themes using content analysis.

2.4. Procedure

The data collection for the study was conducted during the 2024–2025 academic year. Before the process began, pre-service teachers were provided with detailed information about the purpose, scope, data confidentiality, and conditions of participation, and it was clearly stated that their participation was voluntary.

Quantitative data were collected primarily through online (Google Form) and in-person surveys. Participants completed the Artificial Intelligence Readiness Scale and the Professional Self-Efficacy Beliefs Scale in an average of 20–30 min. The qualitative data collection process then began, with a semi-structured form administered to 20 pre-service teachers selected through a purposive sampling method. Technical support was provided throughout the process, and the data was securely stored digitally.

2.5. Ethical Principles

All ethical guidelines were adhered to in this study. Approval was obtained from the relevant university's Ethics Committee before data collection. Participants were provided with informed consent, and they were assured that their data would be used solely for scientific purposes and that their identities would be kept confidential. Participants in the study participated voluntarily, and it was clearly stated that they had the right to withdraw from the process at any time if they wished. Furthermore, participant codes and personal

https://doi.org/10.3390/educsci16010043

information used in the qualitative data were kept confidential, and data security was ensured by national legislation.

2.6. Data Analysis

Quantitative data were analyzed using SPSS version 26. Relationships between variables were determined using descriptive statistics (mean, standard deviation) and Pearson correlation analysis. The significance level was set at 0.05. Qualitative data were analyzed using content analysis. In this research, qualitatively collected data were analyzed using thematic analysis techniques. Participant responses were thematically coded and classified into specific categories. As a result of the analysis, pre-service teachers' views on the use of artificial intelligence in primary education were described under the headings of opportunities, limitations, ethical/pedagogical concerns, and recommendations.

3. Results

3.1. What Is the Level of Preparedness of Teacher Candidates for Artificial Intelligence?

Table 1 presents the total and sub-dimensional means and standard deviation values for the readiness levels of pre-service teachers towards artificial intelligence. According to the findings, the general readiness level of the participants was high (M = 68.29; SD = 12.35). When the sub-dimensions were examined, a high level of readiness was observed in the dimensions of cognitive readiness (M = 19.35; SD = 4.27), artificial intelligence use skills (M = 23.42; SD = 5.28), and artificial intelligence perception (M = 11.86; SD = 2.86). These findings indicate that pre-service teachers have a sufficient level of understanding of the basic concepts of artificial intelligence, using technology, and developing a positive perception. In the ethical awareness sub-dimension, the participants' mean score remained at a moderate level (M = 13.65; SD = 3.13). This suggests that pre-service teachers' awareness of the ethical dimensions of AI is more limited than other dimensions and that more educational support is needed in this area. Overall, pre-service teachers in Kazakhstan have an elevated level of cognitive, practical, and perceptual readiness for AI, but there are areas in which ethical awareness needs to be developed.

Table 1. Pre-service teachers' readiness levels for artificial intelligence.

	N	Minimum	Max	*M*	SD	Level
Artificial Total	293	23.00	90.00	68.29	12.35	High
Artificial Cognition	293	5.00	25.00	19.35	4.27	High
Artificial Ability	293	6.00	30.00	23.42	5.28	High
Artificial Vision	293	3.00	15.00	11.86	2.86	High
Artificial ethics	293	6.00	20.00	13.65	3.13	Moderate
Valid N (listwise)	293					High

3.2. What Is the Level of Professional Self-Efficacy Beliefs of Pre-Service Teachers?

Table 2 presents the weighted mean (M) and standard deviation (SD) values for the total and sub-dimension levels of pre-service teachers' professional self-efficacy beliefs. According to the findings, the participants' general professional self-efficacy levels were high (M = 99.74; SD = 17.40). When the sub-dimensions were examined, high levels of self-efficacy were found in all of the sub-dimensions: academic self-efficacy (M = 19.51; SD = 3.63), social self-efficacy (M = 30.17; SD = 6.68), intellectual self-efficacy (M = 25.55; SD = 5.09), and professional self-efficacy (M = 24.49; SD = 5.25). These results demonstrate that pre-service teachers possess academic knowledge and skills related to teaching processes, can communicate effectively with students and stakeholders, are confident in intellectual competencies such as critical thinking and accessing information, and have a

https://doi.org/10.3390/educsci16010043

high level of internalization of professional competencies such as professional ethics and responsibility. Overall, pre-service teachers in Kazakhstan have strong self-efficacy for the teaching profession across all dimensions and are highly confident in their ability to fulfill their professional roles.

Table 2. Professional self-efficacy beliefs of pre-service teachers.

	N	Minimum	Max	*M*	SD	Level
Self-Total	293	59.00	135.00	99.74	17.40	High
Self-Academic	293	11.00	25.00	19.51	3.63	High
Self-Social	293	10.00	40.00	30.17	6.68	High
Self-intellectual	293	14.00	35.00	25.55	5.09	High
Self-Professional	293	14.00	35.00	24.49	5.25	High
Valid N (listwise)	293					

3.3. What Is the Relationship Between Teacher Candidates' Readiness for Artificial Intelligence and Their Professional Self-Efficacy Beliefs?

Table 3 examines the relationship between AI readiness and self-efficacy. The correlation analysis revealed a low-level relationship between the total AI readiness score and the total self-efficacy score ($R = -0.053$, $p > 0.05$), and this relationship was not statistically significant. This result suggests that pre-service teachers' AI readiness levels and their general self-efficacy perceptions do not influence each other.

Table 3. The relationship between teacher candidates' readiness for artificial intelligence and professional self-efficacy.

		Artificial Intelligence Total	Self-Efficacy Total
Artificial Intelligence Total	Pearson Correlation	1	−0.053
	Sig. (2-tailed)		0.363
	N	293	293
Self-Efficacy Total	Pearson Correlation	−0.053	1
	Sig. (2-tailed)	0.363	
	N	293	293

3.4. What Are the Views of Pre-Service Teachers on the Use of Artificial Intelligence in the Primary Education Teaching Process?

Table 4 presents the themes related to prospective teachers' views on the use of artificial intelligence in primary education.

Table 4. Themes regarding pre-service teachers' views on the use of artificial intelligence in primary education.

Theme	Subtheme	f (n = 20)	%	Examples of Participant Opinions
Opportunities	• Individualized learning • Saving time for the teacher	15	75%	"AI can deliver content based on students' learning styles." (K9)
Limitations	• Lack of infrastructure • Digital competence issues	9	45%	"There are still basic equipment shortages in schools." (K14)

Table 4. *Cont.*

Theme	Subtheme	f (n = 20)	%	Examples of Participant Opinions
Ethical/Pedagogical Concerns	• Decreased teacher-student interaction • Data security • Exposure to technology at an early age	11	55%	"The teacher figure is very important for primary school children." (K6)
Suggestions	• The need for teacher training • Development of guide materials • Continuous teacher support	13	65%	"First of all, teachers should receive good training." (K2)

3.5. Qualitative Results

3.5.1. Opportunities

In Table 4, under this theme, pre-service teachers emphasized the potential of artificial intelligence technologies to support individualized learning processes and reduce teacher workload.

In the personalized learning subtheme, many participants noted that AI can provide content tailored to students' learning styles, paces, and needs. One participant expressed this as follows:

> "AI can present content according to students' learning styles. It makes learning easier." (K9)

3.5.2. Limitations

As seen in Table 4,under this theme, participants stated that there are some technical and individual obstacles to the effective use of artificial intelligence at the primary education level. The subtheme of infrastructure deficiency highlighted the lack of adequate technology equipment in schools and inadequate internet access. One participant expressed this problem as follows:

> "There are still basic equipment shortages in schools. These equipment deficiencies need to be addressed." (K14)

3.5.3. Ethical/Pedagogical Concerns

As seen in Table 4, a sizable portion of the participants stated that some ethical and pedagogical risks may arise with the introduction of artificial intelligence technologies into the classroom environment.

The subtheme "decreased teacher-student interaction" emphasized that the teacher, particularly at the primary school level, plays a role not only of imparting knowledge but also of providing emotional and moral guidance. In this context, one participant emphasized the irreplaceable importance of the teacher with the following words:

> "The teacher figure is especially important for primary school children. Primary school is one of the most important periods for children." (K6)

Under the data security subtheme, ethical concerns were raised due to the lack of transparency in how AI systems collect and process student data. It was emphasized that this could create trust issues in the educational environment.

3.5.4. Recommendations

In Table 4, participants made some concrete suggestions for the healthy and efficient integration of artificial intelligence into the educational environment. The subtheme "need for teacher training" emphasized that teachers should undergo comprehensive training,

https://doi.org/10.3390/educsci16010043

both technically and pedagogically, to ensure the informed use of artificial intelligence technologies. This view is evident in one participant's statement:

"First of all, teachers should receive good training." (K2)

The subtheme "Developing Guidance Materials" emphasized the importance of creating sample lesson plans, user guides, and ethical frameworks that teachers can refer to when implementing AI in the classroom.

4. Discussion

The findings of the study indicate that pre-service teachers' readiness levels for artificial intelligence are high. The high means in the sub-dimensions of cognitive competence, practical skills, and ethical sensitivity reveal that preserved teachers possess basic knowledge of artificial intelligence and have a cheerful outlook toward using this technology in educational settings. This indicates that preserved teachers have strong potential in the integration process of pedagogical technologies in the age of digital transformation. There are studies in the literature that reach similar conclusions, such as the study of Ayanwale et al. (2022). Similarly, in a study conducted by Örücü and Hasırcı (2024), assessments conducted using the Artificial Intelligence Readiness Scale revealed that pre-service teachers demonstrated an elevated level of cognitive and practical readiness for artificial intelligence. Ramazanoglu and Akın (2025) and Hopcan et al. (2024) emphasize that pre-service teachers' knowledge and skills regarding artificial intelligence directly affect the formation of their teacher identities and their perceptions of professional competence. In this context, the findings obtained in our study are consistent with the relevant literature and indicate that pre-service teachers can integrate technological innovations. However, some opinions reflected in the qualitative data indicate that, in addition to positive attitudes towards artificial intelligence, there are also certain reservations. Ethical and pedagogical concerns, differences in digital competence levels, and infrastructure deficiencies raise the question of the extent to which readiness levels can be effectively used in practice. In this regard, it is important that teacher preparation programs do not limit AI training to technical skills alone; they should address it comprehensively, including ethical, pedagogical, and cultural dimensions (Rajapakse et al., 2024; Darmawan et al., 2024).

Pre-service teachers have elevated levels of professional self-efficacy. Participants' high mean scores in academic, social, intellectual, and professional subscales indicate that they have developed a sense of confidence in the teaching profession. This suggests that pre-service teachers have strong beliefs in both their pedagogical skills and their competencies related to professional roles such as classroom management, communication, and problem-solving. This finding aligns with Bandura's (1997) social cognitive theory, which argues that teacher self-efficacy is a multifaceted construct encompassing cognitive, motivational, and affective dimensions. Furthermore, as noted in Zhang et al.'s (2023) study on teacher self-efficacy focused on artificial intelligence, pre-service teachers' professional self-efficacy depends not only on technical skills but also on cognitive and affective factors such as self-confidence, decision-making ability, and pedagogical fit. Professional self-efficacy, one of the fundamental elements of qualified teacher training, directly affects teachers' ability to overcome the challenges they encounter in the teaching process and their impact on student achievement (Tatlieşme & Gürgil, 2025). Therefore, the findings of this study indicate that pre-service teachers are at a positive level in the process of internalizing their professional competencies and are ready to assume effective teaching roles in the future. Furthermore, it is frequently emphasized in the literature that professional self-efficacy perceptions are directly related not only to individual factors but also to the learning opportunities offered in pre-service education (Aydede, 2022), practical experiences, and digital pedagogical content (Berlin et al., 2021; Keppens et al., 2021). In this context, supporting teacher

https://doi.org/10.3390/educsci16010043

training programs with practice-based and technology-supported teaching environments will further strengthen these beliefs of pre-service teachers.

In the study, no statistically significant relationship was found between pre-service teachers' readiness levels for artificial intelligence and their professional self-efficacy beliefs ($r = -0.053$; $p > 0.05$). This finding shows that the expected positive relationship between the two variables did not emerge. While some studies in the literature state that teachers' digital competencies are significantly related to their professional self-efficacy perceptions (Katsarou, 2021; Dai, 2023; Z. Wang & Chu, 2023), the results obtained in the current study suggest that this relationship may be due to more complex and contextual factors. The finding shows that pre-service teachers' high readiness for technology does not directly strengthen their perception of professional self-efficacy. This may suggest that individuals' technical knowledge and skills in artificial intelligence have a limited effect on their professional identity development and pedagogical efficacy perceptions. Indeed, Bandura (1997) emphasizes that the perception of self-efficacy is nourished not only by the level of knowledge but also by the individual's belief that he or she can use this knowledge effectively. In this framework, although pre-service teachers' level of awareness about artificial intelligence technologies is high, the fact that they have not developed sufficient experience or confidence in how to use this awareness in a pedagogical context can be considered as one of the possible reasons for the lack of relationship (Aydede, 2022). In addition, the development of self-efficacy beliefs is influenced by multidimensional factors such as individual characteristics, teaching experience, social support, and the quality of educational environments (Keppens et al., 2021). Therefore, the weak relationship between AI readiness and professional self-efficacy may be due to pre-service teachers viewing this technology solely as a technical tool and not evaluating it within pedagogical integrity. In line with these results, simply conveying AI-related content at a cognitive level in teacher education programs should not be considered sufficient. Applied learning environments that support candidates' skills in integrating technology into pedagogical practices are needed. Thus, a stronger link can be established between technological knowledge and pedagogical self-efficacy.

The qualitative data obtained in the study reveal that pre-service teachers have a cheerful outlook towards the use of artificial intelligence at the primary school level. A substantial portion of participants emphasized that artificial intelligence technologies have the potential to support personalized learning processes, provide materials tailored to students' learning pace and needs, and reduce teachers' workload. In this regard, the view that artificial intelligence can provide flexibility and efficiency in learning environments was widely expressed. Within the subtheme "saving teachers' time," it was asserted that AI will assume time-intensive responsibilities in the educational process, enabling educators to concentrate more on pedagogical endeavors. In this regard, pre-service educators perceive AI as an auxiliary instrument for time management. Moreover, Dung (2025) demonstrated that artificial intelligence offers opportunities for language instruction by emphasizing the necessity for professional development that improves instructors' competencies and delivers systemic support.

However, pre-service teachers also noted various limitations and reservations regarding the integration of artificial intelligence into educational processes. Inadequate technical infrastructure, particularly in rural and disadvantaged areas, hinders the effective use of AI-based applications. Factors such as internet access, hardware deficiencies, and software support could further exacerbate educational inequalities. According to the sub-theme of digital competence issues, teachers and pupils have different capacities for using technology, which may make artificial intelligence applications less successful. Similarly to this study, Kartimi et al. (2023) discovered that school-level digital abilities vary among

science teachers. However, there were no variations in digital competencies based on years of service or gender. Collaborative learning can be enhanced by educators who possess strong digital competencies.

Furthermore, concerns have been expressed that with the widespread adoption of artificial intelligence applications, teacher-student interactions could become mechanized, weakening the pedagogical relationship. In the sub-theme of exposure to technology at an early age, concerns were expressed that shaping the developmental processes of primary school students with artificial intelligence-based digital tools could have negative effects on their social and emotional development. Similarly, Santhakumar and Joseph (2024) revealed in their study that the most significant limitation of emerging learning technologies is their limitations in terms of pedagogical interaction characteristics.

Participants emphasized the importance of not only transferring knowledge but also cultural values, ethical understanding, and social-emotional skills in education, and noted that AI tools may be limited in this context. The subtheme "Continuous Teacher Support" underscored the necessity for ongoing technical, pedagogical, and ethical assistance for teachers throughout the AI integration process, ensuring they are not isolated. In addition to these findings, Alzahrani and Al-malaji (2025) emphasize the significance of professional development programs and legislative actions to improve teachers' proficiency and assurance in employing immersive instruments for innovative learning.

Particularly in the Kazakhstani cultural context, because teachers provide not only academic but also moral and cultural guidance to students, there were views that AI-supported education may not adequately fulfill these roles. These findings reveal that cultural factors significantly shape teachers' attitudes toward technological innovations.

Indeed, Kim and Lee (2024) and Ma et al. (2024) state that teachers' attitudes toward AI vary significantly depending on the cultural context, and that technological integration processes should be shaped not only by universal pedagogical principles but also by local values. Ramazanoglu and Akın (2025) similarly emphasize that the integration of AI into education is not merely a technical process; pedagogical, ethical, and cultural dimensions must also be considered. In this context, simply developing teacher candidates' technical competencies is not sufficient.

A holistic professional development approach is also needed that strengthens their ethical sensitivities, prioritizes cultural awareness, and fosters pedagogical relationships. Effectively integrating AI technologies into educational environments requires addressing these three dimensions (technical, ethical, and pedagogical) simultaneously. Furthermore, the continuity of AI integration in teaching: It should be considered that the differences between AI integration for teaching tasks and the direct use of AI with future students cannot be measured or represented on a single scale when discussing "AI integration in teaching". The continuity of AI applications, especially the effects of teacher-centered and student-centered use, should always be taken into account.

5. Conclusions

The findings demonstrate that teacher candidates in Kazakhstan possess a positive and sufficient level of readiness for artificial intelligence across cognitive knowledge, usage skills, and ethical awareness. This readiness supports the potential for the effective integration of artificial intelligence into educational processes within the cultural context of Kazakhstan. The elevated level of self-efficacy observed in academic, social, intellectual, and professional domains further indicates preparedness for the teaching profession, marked by strong self-confidence and versatile competencies. At the same time, the absence of a meaningful relationship between readiness for artificial intelligence and general self-efficacy suggests that preparedness for technology may not necessarily translate into

https://doi.org/10.3390/educsci16010043

professional self-confidence. These results underline the multifaceted nature of teacher readiness, which requires attention not only to technological competence but also to professional identity and confidence.

Positive attitudes toward the use of artificial intelligence in primary education point to important opportunities, including individualized learning processes and the reduction in teacher workload. Nevertheless, concerns expressed by teacher candidates regarding the preservation of cultural values, the maintenance of teacher–student interaction, and existing infrastructure limitations highlight those cultural and contextual dimensions shape perceptions of technological integration. In particular, the emphasis on Kazakh cultural values demonstrates that the integration of artificial intelligence into education cannot be approached solely from a technological perspective, but must also reflect social, ethical, and cultural considerations.

Several recommendations emerge from the findings. Comprehensive training programs should be designed for the integration of artificial intelligence, incorporating not only cognitive knowledge and technical skills but also ethical and pedagogical dimensions to ensure effective use within educational practice. Applied learning opportunities such as internships, workshops, and simulations should be expanded to enable teacher candidates to engage directly with artificial intelligence technologies in authentic educational environments, thereby strengthening professional self-efficacy. Cultural values and social interaction should remain central in the development and implementation of artificial intelligence applications, given the importance of teacher–student relationships and social roles in Kazakh society. Infra-structure deficiencies must be addressed, particularly in rural and disadvantaged regions, through the improvement of technological resources and the promotion of digital literacy among both teachers and students. In addition, continuous professional development programs should be prioritized to support the pedagogical use of artificial intelligence, with a focus on enhancing teachers' self-confidence and perceptions of competence.

Additionally, the dual roles of pre-service teachers as students and emerging professionals—how their roles as students and future/early teachers can affect their readiness for AI in complex ways—should be considered. They benefit from their readiness to use AI both as students and as professionals, considering how/why they might use AI; this could be addressed as a research topic.

Overall, the study emphasizes that the successful integration of artificial intelligence into education in Kazakhstan requires not only technological readiness but also cultural sensitivity, ethical responsibility, and sustained professional development. Attention to these dimensions will contribute to the creation of an educational environment where artificial intelligence enhances teaching and learning while remaining consistent with cultural values and professional practices.

Author Contributions: Conceptualization, K.B.; Methodology, K.B. and E.U.; Software, K.B.; Validation, K.M.; Formal analysis, K.Y.; Investigation, K.Y.; Resources, K.Y. and A.Z.; Data curation, K.M.; Writing—original draft, K.M. and A.Z.; Writing—review and editing, U.B.A. and A.Z.; Visualization, U.B.A. and E.U.; Supervision, U.B.A. and E.U.; Funding acquisition, E.U. 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 Zhetysu University, Kazakhstan, Scientific Research Ethics Committee Zhetysu University/Higher School of Pedagogy and Psychology/2024/132 24 June 2024.

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

Data Availability Statement: Further inquiries and data can be requested directly to the corresponding author.

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

References

Adebagbo, A. (2025). Artificial intelligence integration in teaching and learning: Investigating retirement-date and self-efficacy of in-service teachers in higher education institutions. *Rima International Journal of Education (RIJE)*, *4*(1), 67–75. Available online: https://rijessu.com/wp-content/uploads/2025/03/007-RIJE-2025-V4-019-1.pdf (accessed on 13 September 2025).

Akcil, U., Uzunboylu, H., & Kinik, E. (2021). Integration of technology into learning-teaching processes and google workspace tools: A literature review. *Sustainability*, *13*(9), 5018. [CrossRef]

Alshorman, S. M. (2024). The readiness to use AI in teaching science: Science teachers' perspective. *Journal of Baltic Science Education*, *23*(3), 432–448. [CrossRef]

Alzahrani, A., & Al-malaji, T. Y. (2025). Augmented and virtual reality in education: A comparative study of teacher perceptions in Saudi Arabia and Jordan. *Contemporary Educational Research Journal*, *15*(4), 186–198. [CrossRef]

Ayanwale, M. A., Sanusi, I. T., Adelana, O. P., Aruleba, K. D., & Oyelere, S. S. (2022). Teachers' readiness and intention to teach artificial intelligence in schools. *Computers and Education: Artificial Intelligence*, *3*, 100099. [CrossRef]

Aydede, M. N. (2022). Examining the primary school teacher candidates' science learning skills in terms of their attitudes towards science and their science teaching self-efficacy beliefs. *International Journal of Educational Methodology*, *8*(4), 853–864. [CrossRef]

Bandura, A. (1997). *Self-efficacy: The exercise of control*. W. H. Freeman.

Berlin, R., Youngs, P., & Cohen, J. (2021). How elementary teaching candidates' learning opportunities are associated with their knowledge, self-efficacy, and beliefs. *Teachers College Record*, *123*(7), 1–30. [CrossRef]

Chiu, T. K., Ahmad, Z., & Çoban, M. (2025). Development and validation of teacher artificial intelligence (AI) competence self-efficacy (TAICS) scale. *Education and Information Technologies*, *30*(5), 6667–6685. [CrossRef]

Creswell, J. W. (2021). *Introduction to mixed methods research* (M. Sözbilir, Ed.; 3rd ed.). Pegem Academy Publishing. [CrossRef]

Çolak, İ., Yorulmaz, Y. İ., & Altınkurt, Y. (2017). The validity and reliability study of teacher self-efficacy beliefs scale. *MSKU Journal of Education*, *4*(1), 20–32.

Dai, W. (2023). An empirical study on English pre-service teachers' digital competence regarding ICT self-efficacy, colle-gial collaboration and infrastructural support. *Heliyon*, *9*(9), e19538. [CrossRef] [PubMed]

Darmawan, E., Rahman, T. K. A., & Thamrin, N. R. (2024). Evaluating readiness and acceptance of artificial intelligence adoption among elementary school teachers. *Jurnal Online Informatika*, *9*(2), 228–237. [CrossRef]

Dung, L. Q. (2025). A structural model of Vietnamese EFL teachers' readiness to integrate AI in English language teaching. *World Journal on Educational Technology: Current Issues*, *17*(2), 89–99. [CrossRef]

Farooq, Y. (2025). Exploring teachers' psychological readiness for effective AI integration in the classroom. *The Critical Review of Social Sciences Studies*, *3*(3), 351–364. [CrossRef]

Guan, L., Zhang, Y., & Gu, M. M. (2025). Pre-service teachers' preparedness for AI-integrated education: An investigation from perceptions, capabilities, and teachers' identity changes. *Computers and Education: Artificial Intelligence*, *8*, 100341. [CrossRef]

Guettala, M., Bourekkache, S., Kazar, O., Harous, S., & Zouai, M. (2023). The design and implementation of intelligent ubiquitous learning multi-agent context-aware system. *World Journal on Educational Technology: Current Issues*, *15*(4), 429–450. [CrossRef]

Hopcan, S., Türkmen, G., & Polat, E. (2024). Exploring the artificial intelligence anxiety and machine learning attitudes of teacher candidates. *Education and Information Technologies*, *29*(6), 7281–7301. [CrossRef]

Kartimi, K., Riyanto, O. R., & Winarso, W. (2023). Digital competence of science teachers in terms of gender, length of work, and school levels of teaching. *Cypriot Journal of Educational Sciences*, *18*(1), 31–42. [CrossRef]

Katsarou, E. (2021). The effects of computer anxiety and self-efficacy on L2 learners' self-perceived digital competence and satisfaction in higher education. *Journal of Education and E-Learning Research*, *8*(2), 158–172. [CrossRef]

Keppens, K., Consuegra, E., De Maeyer, S., & Vanderlinde, R. (2021). Teacher beliefs, self-efficacy, and professional vision: Disentangling their relationship in the context of inclusive teaching. *Journal of Curriculum Studies*, *53*(3), 314–332. [CrossRef]

Kim, S. W., & Lee, Y. (2024). Investigation into the influence of socio-cultural factors on attitudes towards artificial intelligence. *Education and Information Technologies*, *29*(8), 9907–9935. [CrossRef]

Lim, E. M. (2023). The effects of pre-service early childhood teachers' digital literacy and self-efficacy on their perception of AI education for young children. *Education and Information Technologies*, *28*(10), 12969–12995. [CrossRef]

Liu, N. (2025). Exploring the factors influencing the adoption of artificial intelligence technology by university teachers: The mediating role of confidence and AI readiness. *BMC Psychology*, *13*(1), 311. [CrossRef]

Ma, D., Akram, H., & Chen, I. H. (2024). Artificial intelligence in higher education: A cross-cultural examination of students' behavioral intentions and attitudes. *International Review of Research in Open and Distributed Learning*, *25*(3), 134–157. [CrossRef]

Martin, D. A., McMaster, N., & Carey, M. D. (2020). Course design features influencing pre-service teachers' self-efficacy beliefs in their ability to support students' use of ICT. *Journal of Digital Learning in Teacher Education*, *36*(4), 221–236. [CrossRef]

Oran, B. B. (2023). Correlation between artificial intelligence in education and teacher self-efficacy beliefs: A review. *Ru-meliDE Journal of Language and Literature Studies*, *34*, 1354–1365. Available online: https://dergipark.org.tr/en/pub/rumelide/article/1316378 (accessed on 10 September 2025). [CrossRef]

Örücü, E., & Hasırcı, I. (2024). The effects of self-efficacy perception and organizational preparedness on artificial intelligence anxiety: A study. *Sinop University Journal of Social Sciences*, *8*(2), 732–760.

Özüdoğru, G., & Yıldız Durak, H. (2024, July 11–13). *Turkish adaptation of the AI readiness scale for pre-service teachers*. 10th International New York Conference on Academic Studies in Social, Human, Administrative, and Educational Sciences, New York, NY, USA. Available online: https://scales.arabpsychology.com/wp-content/uploads/2024/07/ogretmen-adaylarina-yonelik-yapay-zeka-hazirbulunusluk-olcegi-36544-toad.pdf (accessed on 10 September 2025).

Rajapakse, C., Ariyarathna, W., & Selvakan, S. (2024). A self-efficacy theory-based study on the teachers' readiness to teach artificial intelligence in public schools in Sri Lanka. *ACM Transactions on Computing Education*, *24*(4), 1–25. [CrossRef]

Ramazanoglu, M., & Akın, T. (2025). AI readiness scale for teachers: Development and validation. *Education and Information Technologies*, *30*(6), 6869–6897. [CrossRef]

Santhakumar, L., & Joseph, N. P. (2024). Confluence learning: The new normal and emerging technologies. *International Journal of Learning and Teaching*, *16*(3), 168–175. [CrossRef]

Sun, J., Ma, H., Zeng, Y., Han, D., & Jin, Y. (2023). Promoting the AI teaching competency of K-12 computer science teachers: A TPACK-based professional development approach. *Education and Information Technologies*, *28*(2), 1509–1533. [CrossRef]

Tatlieşme, S., & Gürgil, F. (2025). Investigation of self-efficacy perceptions and professional anxiety levels of prospec-tive social studies teachers. *Turkish Journal of Scientific Research*, *10*(1), 91–105.

Viberg, O., Cukurova, M., Feldman-Maggor, Y., Alexandron, G., Shirai, S., Kanemune, S., Wasson, B., Tømte, C., Spikol, D., Milrad, M., & Kizilcec, R. F. (2023). *Teachers' trust and perceptions of AI in education: The role of culture and AI self-efficacy in six countries*. CoRR. Available online: https://discovery.ucl.ac.uk/id/eprint/10184606/ (accessed on 10 September 2025).

Wang, B., Rau, P. L. P., & Yuan, T. (2023). Measuring user competence in using artificial intelligence: Validity and reliability of artificial intelligence literacy scale. *Behaviour & Information Technology*, *42*(9), 1324–1337. [CrossRef]

Wang, Z., & Chu, Z. (2023). Examination of higher education teachers' self-perception of digital competence, self-efficacy, and facilitating conditions: An empirical study in the context of China. *Sustainability*, *15*(14), 10945. [CrossRef]

Yao, N., & Wang, Q. (2024). Factors influencing pre-service special education teachers' intention toward AI in education: Digital literacy, teacher self-efficacy, perceived ease of use, and perceived usefulness. *Heliyon*, *10*(14), e34894. [CrossRef] [PubMed]

Zainuddin, T., Salam, A. J., Mubarak, H., Fahmi, C., Sulaiman, B. H., & Armia, M. S. (2024). The selection of technology in the learning process: Is that effective enough? *Global Journal of Information Technology: Emerging Technologies*, *14*(1), 36–44. [CrossRef]

Zhang, C., Schießl, J., Plößl, L., Hofmann, F., & Gläser-Zikuda, M. (2023). Acceptance of artificial intelligence among pre-service teachers: A multigroup analysis. *International Journal of Educational Technology in Higher Education*, *20*(1), 49. [CrossRef]

education sciences

MDPI

Article

Non-Semantic Multimodal Fusion for Predicting Segment Access Frequency in Lecture Archives

Ruozhu Sheng [1,2,*], Jinghong Li [1,2] and Shinobu Hasegawa [3,*]

[1] Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, Nomi 923-1292, Ishikawa, Japan
[2] R&D Section, J-MAX Co., Ltd., Kamiishizu, Ogaki 503-1601, Gifu, Japan
[3] Center for Innovative Distance Education and Research, Japan Advanced Institute of Science and Technology, Nomi 923-1292, Ishikawa, Japan
* Correspondence: sheng@jaist.ac.jp (R.S.); hasegawa@jaist.ac.jp (S.H.)

Abstract: This study proposes a non-semantic multimodal approach to predict segment access frequency (SAF) in lecture archives. Such archives, widely used as supplementary resources in modern education, often consist of long, unedited recordings that are difficult to navigate and review efficiently. The predicted SAF, an indicator of student viewing behavior, serves as a practical proxy for student engagement. The increasing volume of recorded material renders manual editing and annotation impractical, making the automatic identification of high-SAF segments crucial for improving accessibility and supporting targeted content review. The approach focuses on lecture archives from a real-world blended learning context, characterized by resource constraints such as no specialized hardware and limited student numbers. The model integrates multimodal features from instructor's actions (via OpenPose and optical flow), audio spectrograms, and slide page progression—a selection of features that makes the approach applicable regardless of lecture language. The model was evaluated on 665 labeled one-minute segments from one such course. Experiments show that the best-performing model achieves a Pearson correlation of 0.5143 in 7-fold cross-validation and 61.05% average accuracy in a downstream three-class classification task. These results demonstrate the system's capacity to enhance lecture archives by automatically identifying key segments, which aids students in efficient, targeted review and provides instructors with valuable data for pedagogical feedback.

Keywords: online education; segment access frequency; lecture archives; non-semantic multimodal fusion; deep learning

1. Introduction

Online learning originated in the 19th century through correspondence education and has evolved significantly in today's digital age through advances in computer and Internet technologies. The emergence of Open Education Resources (OER) and Massive Open Online Courses (MOOCs) has fundamentally transformed educational accessibility (Saykili, 2018). While the trend toward online education was already growing, the COVID-19 pandemic accelerated this transformation to unprecedented levels, with UNESCO reporting approximately 862 million students, almost half of the world's student population, affected by school closures across 107 countries (Abuhammad, 2020). This global shift prompted higher education institutions worldwide to rapidly adopt online learning platforms. Even after the pandemic has subsided, most institutions continue to rely on widely adopted

online learning platforms, making it increasingly important to enhance the accessibility and effectiveness of recorded lecture archives. Lecture archives are integral to blended learning, enabling students to combine in-person instruction with flexible online review. This study examines a practical blended learning scenario where students attend face-to-face lectures and selectively revisit key segments of unedited recordings to enhance learning efficiency.

Among these blended learning resources, lecture archives, which are complete recordings of face-to-face lectures without editing, have become an increasingly common practice in higher education institutions, especially in countries like the United Kingdom and the United States. For instance, lecture capture is now a common feature in UK universities (Lamb & Ross, 2022), and earlier surveys already showed that a majority of institutions had deployed such solutions by 2014 (Walker et al., 2014). Similarly, high adoption rates are reported in the US, where one study found 95% of responding medical schools regularly record lectures (Khong et al., 2025). This popularity primarily stems from their cost-effectiveness and ease of distribution. In Japan, the Japan Advanced Institute of Science and Technology (JAIST) also exemplifies this trend, having systematically recorded face-to-face lectures through their Learning Management System since 2006, thereby creating an extensive archive for supplemental learning (Hasegawa et al., 2007). However, the unedited, long-form nature of these recordings presents significant hurdles. Beyond the practical difficulties for instructors in manually editing and managing such large volumes of content, they often prove inefficient for student learning. For instance, students can find it difficult to maintain attention throughout extended viewing periods (Guo et al., 2014; Sablić et al., 2021), and contemporary research indicates that the key to enhancing learning with long-form material is not merely shortening it, but providing a meaningful, navigable structure (Seidel, 2024).

This raises the practical question of how to identify which segments learners find most meaningful. While direct measures of cognitive engagement, such as eye tracking or user self-reporting, are often impractical in real-world educational environments, behavioral indicators derived from interaction data offer a promising alternative. Equating "meaningful" segments with "high-engagement" segments, inferred from viewing patterns, provides a scalable approach to improving the accessibility and effectiveness of lecture archives. For example, Guo et al. proposed the use of engagement time—the duration students spend watching a video—as a proxy for interest (Guo et al., 2014).

Building upon this idea, Bulathwela et al. (2020) introduced the VLEngagement dataset, which estimates video-level engagement by computing a normalized viewing duration aggregated across large numbers of users. Specifically, they defined Engagement Score = Average Watch Time/Video Duration, providing a cost-effective and scalable labeling method for large collections of educational videos. However, the engagement labels in this dataset are defined at the whole-video level, resulting in a coarse granularity that limits its applicability to tasks such as segment-level attention modeling or highlight extraction.

Inspired by these approaches, this study proposes to use segment access frequency (SAF) as a more fine-grained and context-appropriate measure of engagement. This metric, calculated from the number of playback events associated with each time segment, provides a practical and scalable solution that does not depend on semantic content or specialized hardware. It is particularly suitable for classroom lecture recordings, which typically lack rich annotations or auxiliary sensors.

This study focuses on lecture archives recorded in real classroom environments in higher educational institutions. These settings are often resource-constrained, which defines the core challenges we address. Specifically, the recordings themselves are typically unedited, capturing the instructor with a fixed, ceiling-mounted camera and microphone, which can result in audio quality that is too noisy or indistinct for reliable automatic tran-

scription. Furthermore, the context of their use is also constrained: the archives serve a small number of learners from a single course, and the viewing data from these learners is collected without any auxiliary hardware such as eye-tracking devices. One representative example of such a setting is the video archive system accumulated at JAIST, where face-to-face lectures are routinely recorded and made available through the institutional LMS. Building on this setting, prior work by Sheng et al. (2022) presented at AIED 2022 initially explored the feasibility of predicting focal periods using access logs, based on a set of manually extracted features and relatively simple models. However, that study lacked systematic comparisons and did not incorporate advanced fusion strategies. Building upon this early validation, the present study significantly extends their prior work through improved feature design, structured fusion methods, and comprehensive model benchmarking.

This study aims to develop a lightweight and efficient prediction method based on non-semantic multimodal features to address the challenges of such resource-limited settings. This approach is designed to avoid semantic dependence, enable cross lingual adaptability, and minimize training costs for estimating durations with high SAF in lecture archives. Furthermore, by generating SAF labels automatically from aggregated playback data, the system is intended to support scalable deployment in practical educational contexts.

To systematically validate the feasibility and optimize the design of such a non-semantic framework, this study is therefore guided by the following research questions:

- Main Research Question: Can segment access frequency (SAF) in lecture archives be accurately predicted using only non-semantic multimodal features, derived from real-world recordings without transcripts or semantic annotation?
- Sub RQ1: Which non-semantic modality—action, voice, or slide—contributes most to SAF prediction accuracy, and how do combinations of these modalities affect performance?
- Sub RQ2: Which fusion strategy and neural network backbone provide the optimal balance of prediction accuracy and computational efficiency in resource-constrained educational settings?

To address these research questions, this study makes the following contributions:

- The proposal of a language-independent prediction framework for estimating Segment Access Frequency (SAF) based on non-semantic features. This lightweight framework functions without relying on semantic understanding or specialized equipment.
- A comparative analysis of fusion strategies, which demonstrates the superiority of early feature fusion for achieving higher prediction accuracy and training efficiency in resource-limited scenarios.
- A comprehensive ablation and backbone analysis that identifies the dominant contribution of instructor action features and confirms the effectiveness of ResNet-based architectures for this task.

2. Literature Review

This study is situated at the intersection of two areas: video summarization and student engagement modeling. While video summarization focuses on selecting key content segments, engagement modeling aims to estimate which parts of a lecture attract student attention. This section reviews representative works from both directions to clarify the foundation and scope of the non-semantic, engagement-driven approach.

2.1. Video Summarization

Modern educational platforms offer learners extensive access to recorded lecture videos, prompting the need for technologies to enhance the efficiency of video-based learning (Benedetto et al., 2024). Video summarization addresses this need by enabling educators and students to quickly discern the educational value within lengthy recordings. It generates concise representations of video content through combinations of still images, short segments, visual diagrams, or textual annotations. Early research proposed rule-based methods for extractive summaries, typically selecting representative keyframes, segments, or transcript snippets (Alaa et al., 2024; Dimitrova, 2004).

With advancements in machine learning, deep learning models have gained prominence in video summarization. Recent studies have developed sophisticated approaches tailored to diverse video types. For instance, Singh and Kumar (Singh & Kumar, 2024) introduced a deep learning framework integrating Bayesian fuzzy clustering with a Deep Convolutional Neural Network (Deep CNN), optimized via a hybrid Lion-Deer Hunting (LDH) algorithm. Their approach significantly improved crowd video datasets' precision, recall, and F1-score. However, it is designed for dynamic scenes with dense motion and clear foreground and background separation, contrasting sharply with the static camera angles, minimal motion, and subtle engagement cues typical of lecture archives. These differences necessitate distinct design considerations for educational contexts.

On the other hand, Kim et al. emphasize the need to move beyond traditional semantic or content-only analyses, advocating for the use of interaction data to optimize video design and highlighting the pedagogical value of editing videos for brevity and improved engagement (Kim et al., 2014).

Various time-series models leveraging recurrent architectures have been explored for video summarization. Agyeman et al. (2019) developed a hybrid model combining three-dimensional Convolutional Neural Networks (3D-CNN) with Long Short-Term Memory (LSTM) layers to classify events in soccer videos, achieving 96.8% accuracy. While effective for sports and surveillance videos with rapid scene changes, such models are less suitable for lecture archives, which feature limited visual variation, sparse motion, and extended durations. Moreover, the high computational cost of training recurrent neural networks (RNNs) on long videos poses practical challenges for resource-constrained educational settings. To improve summarization performance without heavy temporal modeling, some recent methods have focused on enhancing input representation: Tan et al. (2024) used adaptive clustering to extract more meaningful keyframes. In contrast, Khan et al. (2024) introduced multi-scale feature maps to capture both detailed and semantic-level information.

In the education field, Andra and Usagawa (Andra & Usagawa, 2019) summarized lecture videos through an Attention-based Recurrent Neural Network (RNN) that combines segmentation with the summarization process. The RNN architecture generates a natural summary by capturing critical words and conveying a lecture's central message through attention-based weighting and linguistic features. However, their method significantly depends on semantic analysis of lecture content. Such transcripts are not always available for lecture archives, and the accuracy of existing Automatic Speech Recognition (ASR) techniques is not always satisfactory for audio data with noise and precise terminology. Moreover, the limited availability of annotated data for lecture videos poses another challenge for such methods.

To address the issue of limited annotated data, Vimalaksha et al. (2018) provided a mechanism to segment lecture videos into multiple parts based on crowdsourcing. While this approach alleviates the annotation problem to some extent, manual crowdsourcing methods have their own limitations. They require real-time recording, are labor-intensive, and have strict time synchronization requirements, making them prone to bias. There-

fore, despite the contribution of crowdsourcing methods, there remains a need for more generalized approaches to effectively tackle the annotation problem for lecture videos, particularly in contexts with limited resources and multilingual content as addressed in the current research.

2.2. Student Engagement and Attention

Student engagement is crucial in higher education, yet fostering active participation in online learning environments remains a persistent challenge (Vermeulen & Volman, 2024). To address this issue, researchers have examined various instructional and content-related factors that may influence how students interact with lecture materials.

For instance, embedded semantic approaches have been developed to enhance student interaction within classroom settings. Deng and Gao explore how embedding questions within pre-class instructional videos influences learners' experiences and outcomes in a flipped classroom context (Deng & Gao, 2024). Interestingly, their study revealed no discernible effect on learning performance but did find that the embedded questions significantly reduced students' total viewing time. Indeed, this illustrates that semantic analysis primarily focuses on textual content and overlooks learner interactions with the video, such as pausing, rewinding, or skipping sections. These behaviors provide crucial insights into attention patterns and engagement levels, which semantic methods cannot capture.

On the other hand, several studies have explored the design of video lectures without treating semantic features as a key element for promoting engagement. For example, Chen and Wu investigated the impact of different video lecture formats on learning outcomes, cognitive load, and emotional responses (Chen & Wu, 2015). Their findings suggest that featuring instructors on screen not only enhances students' sense of connection but also reduces cognitive overload. This highlights the importance of visual presence and presentation style in maintaining student attention and fostering participation in online learning environments.

With regard to visual presence and presentation style, Shi et al. examined how instructors' visual attention and lecture delivery styles influence students' perceived engagement and academic performance across various instructional formats (Shi et al., 2024). Similarly, Zhang et al. employed eye-tracking and visualization technologies to investigate the effects of different instructional delivery styles on student viewing behavior (Zhang et al., 2018). Their analysis showed that students were more responsive to auditory cues—such as pauses and vocal emphasis—than to visual elements like gestures or slide transitions. Collectively, these studies underscore the importance of instructor-related features—both visual and auditory—in shaping student engagement with lecture content. In particular, they highlight the pivotal role of auditory cues in sustaining attention during recorded lectures, even when visual presence is emphasized. These findings suggest that integrating non-semantic instructional features, such as visual presence and vocal modulation, is instrumental in directing learners' attention toward key concepts in a lecture.

In addition to instructor behavior, students' behavioral data has also been used to examine engagement at scale. Kim et al. (Kim et al., 2014) conducted a large-scale analysis of video interaction patterns in MOOCs and identified several recurring behaviors—such as replaying specific moments and revisiting explanation-heavy segments—as signals of focused engagement. Such naturally occurring patterns provide a foundation for scalable engagement modeling based on interaction data, without requiring semantic understanding or manual annotation.

2.3. The Position of This Work

This study focuses on estimating high-SAF segments in lecture videos by analyzing non-semantic multimodal features. While SAF is not a direct measurement of cognitive engagement or attention, it serves as a practical behavioral proxy that reflects students' selective rewatching behavior with minimal overhead. Unlike most existing approaches that rely on semantic content analysis, this work emphasizes how students interact with lecture materials through their actual viewing behaviors. This approach offers three main advantages: it is applicable across different languages as it does not require semantic understanding, it reduces the need for manual annotation by utilizing access logs, and it identifies lecture segments with high access frequency as derived from real usage patterns. By combining features extracted from instructor actions, voice, and slides with patterns of student interaction, the method provides a lightweight and flexible solution for identifying high-SAF segments in real educational settings.

3. Research Design and Methodology

This study proposes a lightweight and scalable system to predict SAF in real classroom lecture archives. The overall workflow, illustrated in Figure 1, consists of four main stages. First, the lecture archives are segmented into uniform time intervals. Second, multimodal features not relying on semantic content—including instructor actions, voice spectrograms, and slide transitions—are extracted from each segment. Third, a deep neural network based backbone model fuses and processes these features. Finally, the model outputs predictions of SAF, supervised by labels automatically generated from aggregated access logs. Notably, the entire pipeline, from feature extraction to label generation can be fully automated without manual intervention. The system is designed to operate efficiently in resource-constrained environments without relying on semantic analysis or specialized hardware.

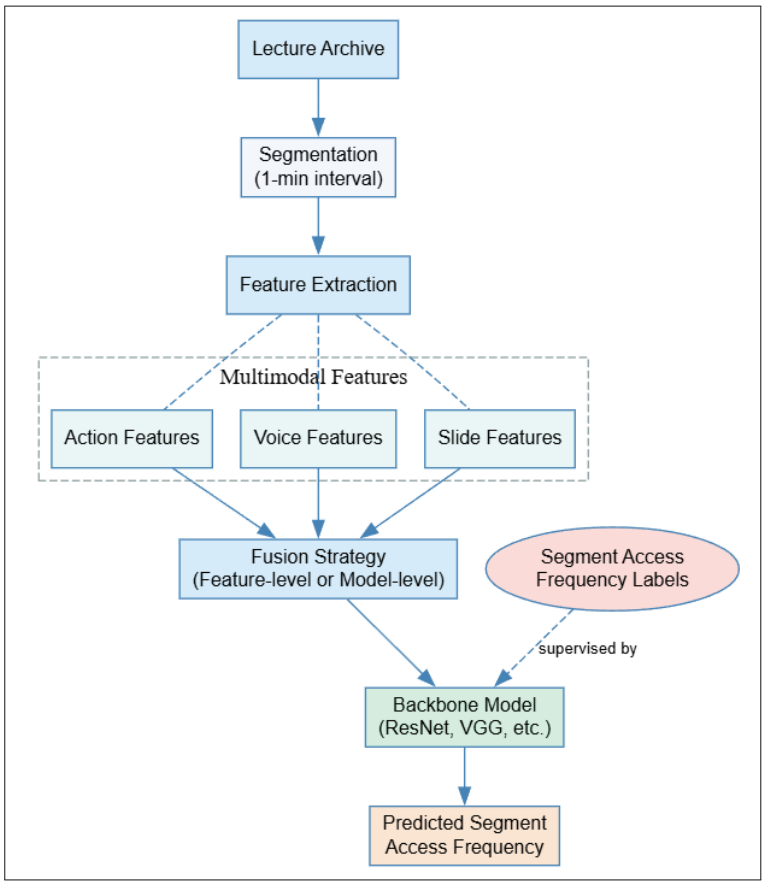


Figure 1. System pipeline for predicting segment access frequency (SAF) from lecture archives.

3.1. Dataset

3.1.1. Lecture Archives

The lecture archives used in this research were recorded from the I239 Machine Learning course offered through the JAIST Learning Management System (JAIST-LMS). The course consisted of seven lessons, each recorded to support students' reflection and supplemental learning following face-to-face instruction. The archives were distributed via the LMS on the campus network within a few hours after the course, and students were able to skip and watch the lectures freely. A ceiling-mounted camera with a fixed angle and a ceiling microphone captured both the instructor's and students' voices. The video files were recorded at 1920 × 1080 resolution and 30 frames per second. Each lesson lasted approximately 100 min. The archives included the podium area, whiteboard, and instructor. In addition, the slide content was integrated into the right-bottom corner of the archives, as shown in Figure 2.

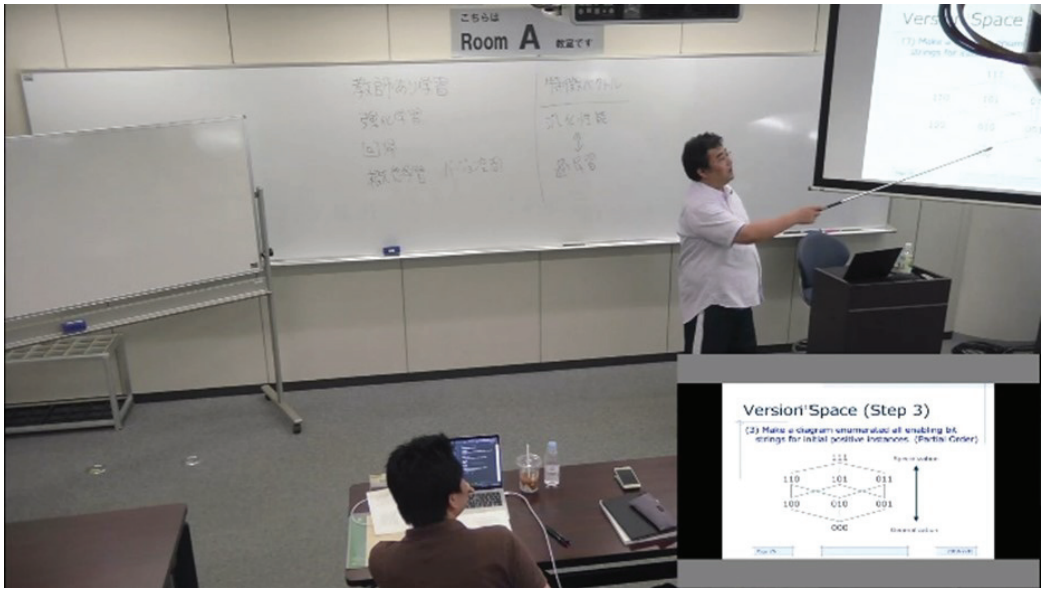


Figure 2. Original Lecture Archive, I239 Machine Learning.

3.1.2. Label Generation

The JAIST-LMS extends Video.js to track students' access to specific durations of lecture archives. Based on this detailed playback history, SAF labels were generated, defined as the number of times each one-minute segment was accessed across all users.

When finding the important durations for students to watch, they repeatedly skip and briefly view segments of the archives. Such behavior causes noise in the labeling of high-SAF segments. To reduce such noise and ensure data reliability, raw viewing records shorter than one minute were first excluded. The total valid viewing time per student was then computed, and only students who watched more than five minutes of a given lecture were retained and labeled as valid viewers. Segment access frequencies were calculated by aggregating only these valid viewers' logs.

After data filtering, Table 1 shows that the label dataset is extremely limited in size. On average, each lecture has only 8.71 valid viewers. This is likely because students had already attended the face-to-face class sessions, and the archives were mainly used as a review resource. In such cases, students may choose to revisit only selected parts of the lecture or rely on other materials, such as textbooks or slides, for review. As a result, the overall number of archive viewers is low.

However, this type of review behavior may also indicate stronger learning intent. According to Kim et al. (2014), students who engage in repeated viewing tend to show more

focused and high-peak interaction patterns compared to first-time viewers. These behaviors are often goal-driven, as students selectively locate and revisit important parts of the content. Therefore, although the dataset is small, the SAF signals it captures may be more concentrated and meaningful, providing a robust basis for identifying high-SAF segments.

To further adapt the data for model training, the non-instructional portions at the beginning and end of each video were removed. All lecture archives were trimmed or padded to a standardized length of 95 min, focusing exclusively on instructional content. Each archive was then divided into one-minute segments. The SAF for each segment was calculated by summing the number of valid viewers who accessed it.

Table 1. Valid viewers and total viewing time for each lecture.

Lesson	Valid Students	Total Watching Time (min)
Lesson-1	10	532
Lesson-2	6	605
Lesson-3	11	647
Lesson-4	8	408
Lesson-5	8	607
Lesson-6	9	643
Lesson-7	9	939
Average	8.71	625.86

A centered moving average with a five-segment window was applied to the raw frequency sequence to suppress short term fluctuations. This method averages two preceding and two succeeding values around each time point, thereby enhancing local trend stability while preserving temporal structure. Such symmetric moving averages are widely adopted in time-series analysis to extract smooth trend-cycle components from noisy data (Hyndman & Athanasopoulos, 2018).

The resulting values were then normalized within each lecture by dividing by the maximum segment frequency, yielding SAF labels in the range of [0, 1]. In total, 665 labeled segments were obtained across the seven lectures for subsequent model training and evaluation. The SAF labels were automatically generated from raw csv format access logs through a batch-processing script, eliminating the need for manual annotation.

To illustrate how SAF patterns are aligned with different lecture structures, the distributions for Lesson 1 and Lesson 4 are presented as contrasting examples. Figures 3 and 4 illustrate the processed SAF distributions for Lesson 1 and Lesson 4, respectively. In Lesson 1, the SAF remains low during the initial 30 min, this part consists of basic concepts and course schedule. A sharp increase is observed between minutes 40 and 60, corresponding to the explanation of the Version Space Algorithm—a relatively complex and central topic of the lecture. The SAF then drops around the 70-min mark, coinciding with a break period, before rising again as the lecture resumes. Toward the end of the session, SAF decreases as the class concludes.

In contrast, Lesson 4 exhibits a different pattern. The SAF peaks within the first 15 min, during which the instructor explains the steps for solving example questions. In the remainder of the lecture, a live demonstration using Google Colab is presented. As this portion focuses on practical execution, students are more likely to engage with the shared Colab notebook directly, rather than watching the video again. Consequently, the SAF gradually declines and remains low until the end.

These examples confirm that SAF patterns are closely aligned with lecture content and structure. Peaks and drops in SAF correspond to conceptually intensive segments, breaks, or procedural demonstrations, supporting the validity of SAF as a proxy for student

engagement. Based on these observations, the thresholds for the downstream classification task were set: values above 0.6 as high, and below 0.4 as low.

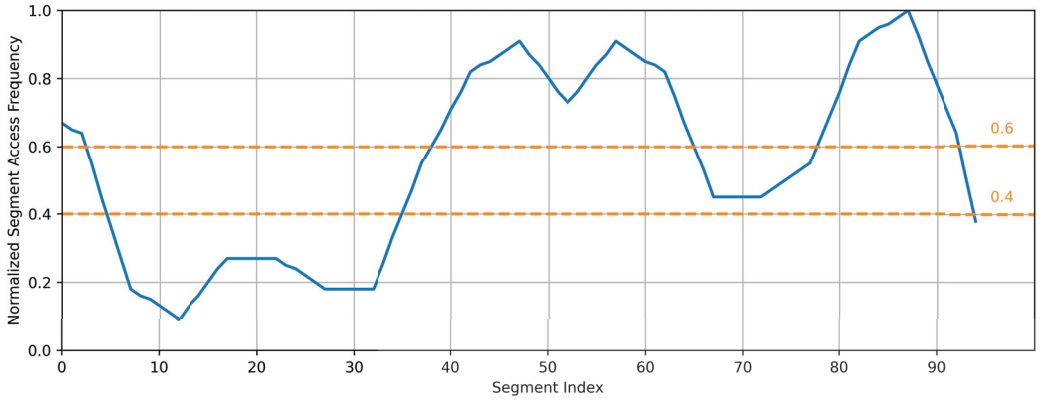


Figure 3. SAF distribution for Lesson 1.

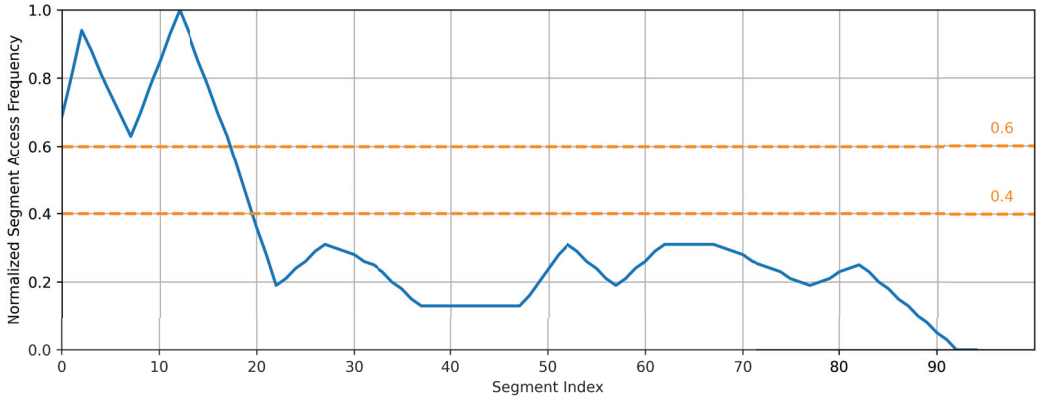


Figure 4. SAF distribution for Lesson 4.

3.1.3. Evaluation Metrics

To assess the performance of the proposed system, both regression-based metrics and 3-classification accuracy are adopted for a comprehensive evaluation. For the primary task of predicting a continuous attention level, represented by normalized SAF values between 0 and 1, we employ the following standard regression metrics:

- Mean Squared Error (MSE): Measures the average squared difference between the predicted and ground-truth values.
- Mean Absolute Error (MAE): Computes the average magnitude of absolute prediction errors.
- Coefficient of Determination (R^2): Indicates the proportion of variance in the ground truth that is explained by the predictions.
- Pearson Correlation Coefficient (PCC): Evaluates the linear correlation between predicted and ground-truth sequences, reflecting trend similarity.

Additionally, for auxiliary evaluation, we compute the 3-classification accuracy by categorizing SAF values as: High (>0.6), Medium (0.4–0.6), and Low (<0.4).

These SAF thresholds follow the observations discussed in the previous subsection. While the division remains relatively coarse, it offers a simple and interpretable way to evaluate model performance as an auxiliary metric. Given that the primary task is regression, the three-class accuracy serves solely as a supplementary indicator, supporting downstream tasks such as heatmap generation. Future work will explore more adaptive thresholding strategies and finer-grained classification schemes.

3.2. Feature Extraction and Preprocessing

3.2.1. Action Features (A)

According to the previous study by Zhang et al. (2018), the behavior of the instructor influences the attention of students. Therefore, the instructor's action in the archive segments was obtained by the optical flow (Burton & Radford, 1978), the pattern of apparent motion of objects, surfaces, and edges in each segment caused by the relative motion between observer and scene.

However, the optical flow could not work well because the corner point is always generated on the slide rather than the instructor in a default setting. To solve this problem, the students' seating area is first masked, and then the instructor's body structure feature is captured by OpenPose (Cao et al., 2019), the first open-source real-time system available to detect multi-person 2D poses, including body, feet, hands, and facial key points. Next, the optical flow for the instructor's action is calculated based on the captured body structure. The Lucas-Kanade method is used in this research to calculate the optical flow for every segment (Lucas & Kanade, 1981). Figure 5 shows an action feature map from a one-minute archive segment extracted by this method.

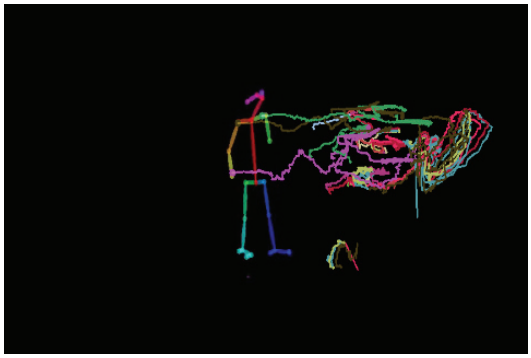


Figure 5. Action Feature.

3.2.2. Voice Features (V)

According to a previous study by Wyse (2017), neural networks used in classification or regression can benefit from spectrograms which are a visual representation of the spectrum of signal frequencies as it varies with time (Flanagan, 1972). In addition, they retain more information than most hand-crafted features traditionally used to analyze voice or sound and have a lower dimension than raw data. A SciPy-based approach utilizing the spectrum function was implemented to generate spectrograms from the lecture audio. The spectrograms were computed using Fast Fourier Transform (FFT) with a window size of 1024 samples and an overlap of 128 samples between adjacent windows at a 44.1 kHz sampling rate. This configuration provides sufficient frequency resolution while maintaining temporal precision necessary for analyzing speech patterns in lecture recordings. The spectrogram generation process effectively converts the time-domain audio signal into a two-dimensional time-frequency representation, capturing both temporal and frequency characteristics of the instructor's voice. The resulting spectrograms were then processed as feature maps for the deep learning model, as shown in Figure 6. This approach enables the model to learn from both the frequency content and temporal dynamics of the instructor's speech, while maintaining computational efficiency.

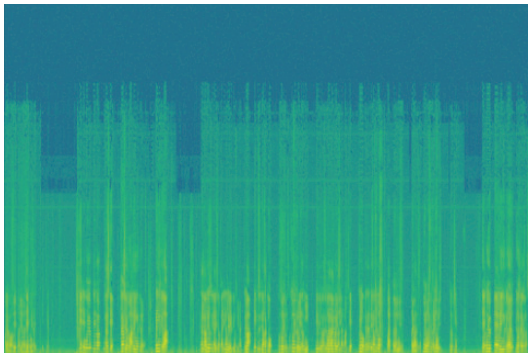


Figure 6. Voice Feature.

3.2.3. Slide Features (S)

Slide transitions indicate lecture pacing and content structure, potentially influencing student engagement. To capture this, numerical features representing net slide progression are extracted, processed through a five-step pipeline to ensure temporal alignment and compatibility with other modalities.

The processing pipeline consists of five steps:

1. Net progression: For each 5-min segment i, the net forward movement is computed as:

$$P_{\text{raw}}[i] = \max\{x_i\} - \max\{x_{i-1}\}, \quad \text{where } P_{\text{raw}}[i] = 0 \text{ if the difference is negative or zero.}$$

2. Temporal resolution adjustment: Each $P_{\text{raw}}[i]$ is repeated 5 times to form a 1-min resolution sequence:

$$P_{1\ \text{min}} = \text{repeat}(P_{\text{raw}}, 5)$$

3. Smoothing: A moving average filter with window size 5 is applied:

$$P_{\text{smooth}}[i] = \frac{1}{5} \sum_{j=i-2}^{i+2} P_{1\ \text{min}}[j]$$

4. Normalization: Within each lesson, the values are normalized by the lesson-wise maximum:

$$P_{\text{norm}}[i] = \frac{P_{\text{smooth}}[i]}{\max(P_{\text{smooth}})}$$

5. Matrix construction: Each normalized value is scaled to 8-bit and expanded into a uniform 2D matrix:

$$S[i] = \mathbf{1}_{h \times w} \cdot \text{int}(P_{\text{norm}}[i] \times 255)$$

where $\mathbf{1}_{h \times w}$ denotes a matrix of ones with spatial dimensions matching other modalities.

This design ensures that the slide feature is temporally aligned and dimensionally compatible with the action and voice feature matrices for multimodal fusion.

3.2.4. Temporal Smoothing

To enhance the temporal consistency of predictions and suppress short-term fluctuations, smoothing techniques are applied to the attention level sequences produced by the regression model. These smoothed values are subsequently used for interpretation, visualization, and threshold-based classification.

Let $y[i]$ denote the predicted attention level at minute i. The following three smoothing methods are applied:

- Moving Average: A centered rolling mean applied over a window of 5 min:

$$\hat{y}[i] = \frac{1}{w} \sum_{j=i-\lfloor w/2 \rfloor}^{i+\lfloor w/2 \rfloor} y[j], \quad w = 5$$

  This method is simple yet effective in suppressing short-term fluctuations.
- Savitzky–Golay Filter: Implemented with a window length of 7 and a second-order polynomial, this filter performs local polynomial regression to preserve peak shapes while smoothing the signal.
- Kalman Filter: A one-dimensional recursive Bayesian estimator was implemented using the FilterPy library, where the state transition and observation matrices were set as $F = H = [1]$, with an initial covariance $P = 500$, measurement noise $R = 0.05$, and process noise $Q = 10^{-4}$.

These methods are applied post hoc to the predicted sequences and do not affect the model training process. Among them, the moving average achieves the best trade-off between simplicity and performance in the experiments. Detailed comparisons are presented in Section 2. The smoothed sequences are also used to derive three-class attention zones via thresholding, as described in Section 3.1.3.

3.3. Model Architecture and Experimental Settings

3.3.1. Feature Fusion Strategies

Two deep learning architectures are prepared to detect the focal points of the above mentioned features. The first option, called "Feature Stacking," converts different feature maps of each archive segment into RGB channels of a single image file. An example of such a stacked input is shown in Figure 7. Specifically, the action features extracted by OpenPose and optical flow are assigned to the R channel, the slide transition features to the G channel, and the spectrogram voice features to the B channel. This method enables the use of well-established deep learning models like VGG-16, VGG-19, ResNet-50, and ResNet-101, which have proven effective in image classification tasks. The primary advantage of this approach is its computational efficiency—by processing all features through a single network path, memory usage and training time can be significantly reduced compared to parallel processing approaches. However, this integration introduces a notable limitation: the compression of feature maps into single-channel images leads to information loss. This is particularly problematic for action features, where lines in different colors represent distinct movement trajectories of tracked corner points. When these colored trajectories are compressed into a single channel, the spatial and temporal relationships between different movement patterns may become less distinguishable, potentially degrading the model's ability to learn complex motion patterns. The whole architecture of the feature-level fusion strategy is illustrated in Figure 8.

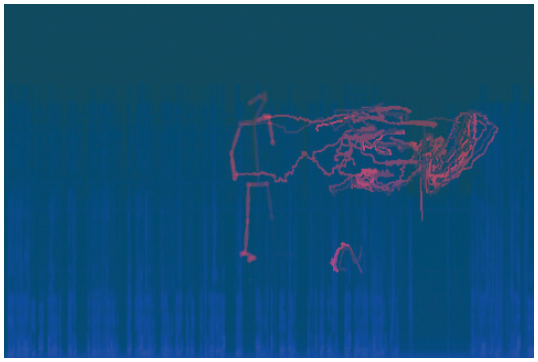


Figure 7. Feature fusion input example combining action, voice, and slide features into RGB channels.

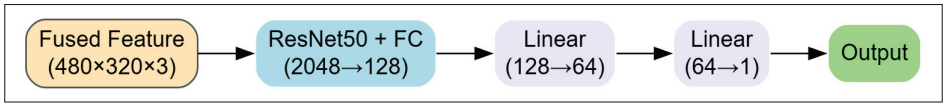


Figure 8. Process of feature-level fusion (example based on ResNet).

Another option, called "Model Stacking," employs multiple parallel deep learning models to process different input features independently before combining, as shown in Figure 9. In this architecture, each feature type (action, slide, and voice) is processed by its own dedicated neural network, preserving the complete dimensionality and characteristics of each feature type. The outputs from these individual networks are then concatenated at their fully connected layers to produce a final prediction. While this strategy maintains the complete information of each feature map and potentially allows for feature-specific network optimization, it comes with increased computational costs. These trade-offs will be further evaluated in the experiment section, where the performance and efficiency of this approach are compared with alternative architectures.

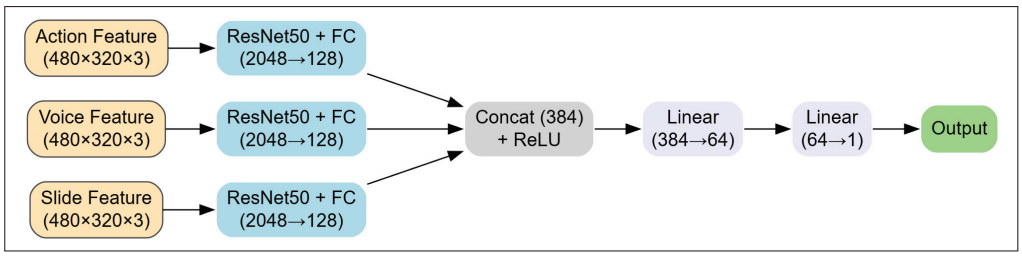


Figure 9. Process of model-level fusion (example based on ResNet).

3.3.2. Backbone Model Selection

After confirming the superiority of feature-level fusion over model-level fusion in earlier experiments, all subsequent model evaluations are conducted under the feature-level fusion setting. Specifically, several representative backbone architectures are compared to identify the most suitable model for the regression task.

The evaluated backbones include:

- VGG-based models: VGG16 and VGG19, known for their simplicity and deep convolutional stacks without residual connections.
- Residual networks: ResNet50 and ResNet101, which introduce skip connections to enable deeper and more stable training.
- Transformer-based model: Vision Transformer (ViT-16), which leverages self-attention to capture global dependencies.
- Temporal model: CNN + LSTM, combining spatial feature extraction and sequential modeling.

All models take as input a fused feature map constructed by stacking action, voice and slide modality features into a 3-channel image (480 × 320 × 3), and share a common training configuration. The evaluation is carried out using both regression metrics (MSE, MAE, R^2, PCC) and the accuracy of the 3-classification.

3.3.3. Experimental Protocol

All models were trained and evaluated using consistent procedures to ensure fair comparison across architectures and fusion strategies. The dataset consisted of seven lecture sessions. For the initial baseline experiment involving only action features and temporal smoothing, a fixed split was used: Lesson 1–5 for training, Lesson 6 for validation, and Lesson 7 for testing. All subsequent experiments—including multimodal fusion, ablation study and backbone comparisons—adopted a 7-fold cross-validation protocol at the lesson

level. In each fold, one lesson was used for testing, while the remaining six were used for training and validation.

The Adam optimizer was employed with a fixed learning rate of 1×10^{-5} and the mean squared error (MSE) was used as the loss function. The batch size was set to 16, and the maximum number of training epochs was 200. Early stopping was applied with a patience of 25 epochs based on validation loss. A fixed random seed of 42 was used to ensure reproducibility.

All experiments were implemented in PyTorch 2.6.0 and conducted on a workstation running Ubuntu 24.04. The system was equipped with an Intel Core i9-12900K processor, 128GB DDR4 RAM, and an NVIDIA RTX A6000 GPU.

4. Experiment

This section presents a series of experiments conducted to evaluate the effectiveness of the proposed prediction framework. First, the predictive capacity of instructor action features used in isolation is assessed. Then, two multimodal fusion strategies are investigated: one that combines features before input into the network (feature fusion), and another that processes each modality separately before combining outputs (model fusion). An ablation study is further conducted to examine the relative contribution of each modality. Finally, several backbone architectures are compared to identify the most effective configuration under resource constraints. All experiments are carried out using seven-fold cross-validation, and evaluated using both regression metrics and three-class accuracy.

4.1. Effectiveness of Action Features

Before introducing multimodal fusion, it was first evaluated whether action features alone could serve as a reliable predictor of student attention. As described in Section 3.2.1, the instructor's motion patterns were extracted from each lecture segment using OpenPose and optical flow, resulting in time-series visual representations. These action features were then used as the sole input to a ResNet50 model.

To enhance temporal stability, three smoothing strategies were applied to the predicted attention values: Moving Average, Savitzky–Golay filter, and Kalman filter. These were compared against unsmoothed (raw) predictions. For this experiment, a fixed data split was used: Lesson 1–5 for training, Lesson 6 for validation, and Lesson 7 for testing.

The results in Table 2 confirm that the visual motion features extracted from the instructor's body movements contain sufficient predictive signals. All four smoothing conditions achieved positive R^2 values and moderate Pearson correlation coefficients (PCC), demonstrating that the model could learn meaningful patterns from the action features even without additional modalities.

Table 2. Performance of action feature regression under different smoothing strategies.

Method	MSE	MAE	R^2	PCC
Raw	0.0292	0.1274	0.2684	0.5464
Moving Average	0.0239	0.1206	0.4026	0.6505
Savitzky–Golay	0.0244	0.1236	0.3897	0.6389
Kalman Filter	0.0242	0.1212	0.3933	0.6413

Among the smoothing methods, the moving average performed best across all metrics, suggesting it effectively suppresses noise while preserving temporal trends. Compared to more complex alternatives such as the Savitzky–Golay filter and Kalman filter, the moving average has a significantly lower computational cost and is extremely simple to implement. Its robustness, interpretability, and real-time applicability make it a strong default choice for smoothing time-series predictions in practical educational settings.

4.2. Fusion Strategies

To evaluate how different fusion strategies impact model performance, two approaches were compared: feature-level fusion and model-level fusion. Both strategies utilized all three modalities—action, voice, and slide—and employed ResNet50 as the backbone to ensure fair comparison.

In the feature-level fusion strategy, all modality features were resized and stacked along the channel dimension to form a single RGB-like image (480 × 320 × 3), which was then passed through a single ResNet50 model. In contrast, the model-level fusion strategy assigned each modality its own dedicated ResNet50 network. The output features from each branch were concatenated and passed through a joint prediction head.

All models were trained using identical cross-validation protocols and evaluated using regression metrics (MSE, MAE, R^2, PCC). Table 3 shows the results.

Table 3. Comparison of fusion strategies using ResNet50 with all three modalities (A + V + S).

Fusion Strategy	MSE	MAE	R^2	PCC	Training Time (min)
Feature-level Fusion	0.0330	0.1424	0.1278	0.5143	22.5
Model-level Fusion	0.0382	0.1562	−0.0136	0.3873	98.4

Table 3 shows that feature-level fusion outperforms model-level fusion across all regression metrics while requiring significantly less training time. This combination of higher accuracy and computational efficiency makes feature-level fusion a practical choice for real-time lecture archive analysis, enabling scalable deployment in educational platforms.

Despite model-level fusion's theoretical advantages—preserving the full resolution of each feature map and enabling modality-specific encoding—it suffered from increased parameter count and training instability. These drawbacks outweigh its theoretical flexibility in this setting, where training data is limited and computational efficiency is an important consideration. Based on these comprehensive findings, feature-level fusion was adopted as the default strategy for all subsequent experiments.

4.3. Ablation Study

To further understand the contribution of each modality to the overall performance, an ablation study was conducted by systematically removing one or more modalities from the input. All experiments in this section were performed under the feature-level fusion setting using ResNet50 as the backbone model. The same cross-validation protocol and smoothing method (moving average) were applied across all conditions.

The tested combinations include the full model (A + V + S), all possible two-modality pairs (A + V, A + S, S + V), and individual modalities (A, V, S). The evaluation results are shown in Table 4.

Table 4. Ablation study on modality combinations using feature-level fusion with ResNet50.

Modality Combination	MSE	MAE	R^2	PCC
Action + Voice + Slide (A + V + S)	0.0330	0.1424	0.1278	0.5143
Action + Voice (A + V)	0.0374	0.1568	0.0125	0.3656
Action + Slide (A + S)	0.0343	0.1456	0.0537	0.4677
Slide + Voice (S + V)	0.0365	0.1544	0.0198	0.4617
Action only (A)	0.0382	0.1575	−0.0064	0.3549
Voice only (V)	0.0419	0.1709	−0.0311	0.4735
Slide only (S)	0.0498	0.1832	−0.2060	0.2819

The results show that the full model using all three modalities (A + V + S) achieved the best performance across all metrics, indicating that each modality contributes complementary information to the prediction. Among the single-modality models, action features alone performed the best, while slide features showed the weakest predictive power when used in isolation. This suggests that instructor motion contains the most informative cues, consistent with the findings in Section 4.1.

Interestingly, the combinations A + S and S + V both outperformed their constituent single-modality models, implying that even relatively weak features like slides can enhance the model when combined with stronger signals. These findings highlight the synergistic effect of multimodal fusion and support the inclusion of all three modalities in the final system design.

4.4. Backbone Model Comparison

Six backbone architectures were systematically evaluated to identify the most effective neural architecture for classroom attention prediction. All models were assessed under identical conditions: feature-level fusion with action, voice, and slide inputs, smoothing via moving average, and training with 7-fold cross-validation. The evaluated architectures span three model families: convolutional networks (VGG16, VGG19), residual networks (ResNet50, ResNet101), a transformer-based model (ViT), and a sequential hybrid (CNN + LSTM).

As shown in Table 5, ResNet50 consistently outperformed all other architectures across metrics, demonstrating the best balance between prediction accuracy and training stability. VGG16 followed closely with good consistency across folds, while VGG19 showed slightly diminished performance. Interestingly, ResNet101 underperformed despite its greater depth, likely due to overfitting on the relatively small dataset. The ViT model exhibited promising regression metrics but slightly lower classification accuracy, suggesting its global attention mechanism may require more data to realize its full potential. The CNN + LSTM architecture provided no clear advantages over purely spatial models, indicating that explicit temporal modeling offers limited benefits at the one-minute feature resolution.

Table 5. Comparison of backbone architectures using feature-level fusion (A + V + S) with 7-fold cross-validation.

Backbone Architecture	MSE	MAE	R^2	PCC	3-Class ACC
ResNet50	0.0330	0.1424	0.1278	0.5143	61.05%
ResNet101	0.0408	0.1595	−0.0641	0.3545	50.23%
VGG16	0.0342	0.1476	0.0841	0.4600	54.14%
VGG19	0.0376	0.1528	−0.0013	0.4128	51.28%
ViT	0.0345	0.1471	0.0969	0.4377	54.44%
CNN + LSTM	0.0372	0.1510	−0.0787	0.4600	51.33%

Given ResNet50's superior performance, a more detailed analysis of its fold-wise behavior was conducted to assess reliability and generalization capabilities.

Table 6 reveals considerable performance variation across the seven validation folds for ResNet50. Fold 7 demonstrated exceptional performance with the lowest MSE (0.0145), MAE (0.0951), highest R^2 (0.6376), and strongest correlation (PCC = 0.8491). However, some folds (particularly 2 and 5) yielded negative R^2 values, indicating challenges in capturing variance for certain lesson contexts. Despite these variations, the average metrics across all folds show a moderate positive correlation (PCC = 0.5143), suggesting the model can capture meaningful attention patterns even with limited training data.

Figure 10 visually compares the predicted Segment Access Frequency (SAF) with the ground-truth labels across the seven validation folds. In each plot, the orange line represents the ground-truth SAF, the light blue line shows the raw predicted values, and

the dark blue line indicates the smoothed predictions. Beneath the curves, the rectangular colored blocks display the three-class classification result for each one-minute segment: green for High-SAF, gray for Medium-SAF, and red for Low-SAF. A key observation is that the model's predictions consistently capture the overall temporal trends of student attention—the rising and falling patterns—even when the absolute numerical accuracy varies. This is evident in folds with poor statistical metrics (e.g., Fold 5, with an R^2 of −0.3305 and 3-class accuracy of 48.42%), where the predicted curve still mirrors the general shape of the ground-truth.

Table 6. Validation results of ResNet50 on each lesson (7-fold cross-validation).

Validation Lesson	Best Epoch	MSE	MAE	R^2	PCC
1	13	0.0623	0.2113	0.1693	0.5323
2	2	0.0264	0.1143	−0.1390	0.2535
3	4	0.0227	0.1217	0.0958	0.3366
4	28	0.0452	0.1850	0.2596	0.7772
5	1	0.0303	0.1406	−0.3305	0.2860
6	6	0.0298	0.1286	0.2020	0.5655
7	91	0.0145	0.0951	0.6376	0.8491
Average	–	0.0330	0.1424	0.1278	0.5143

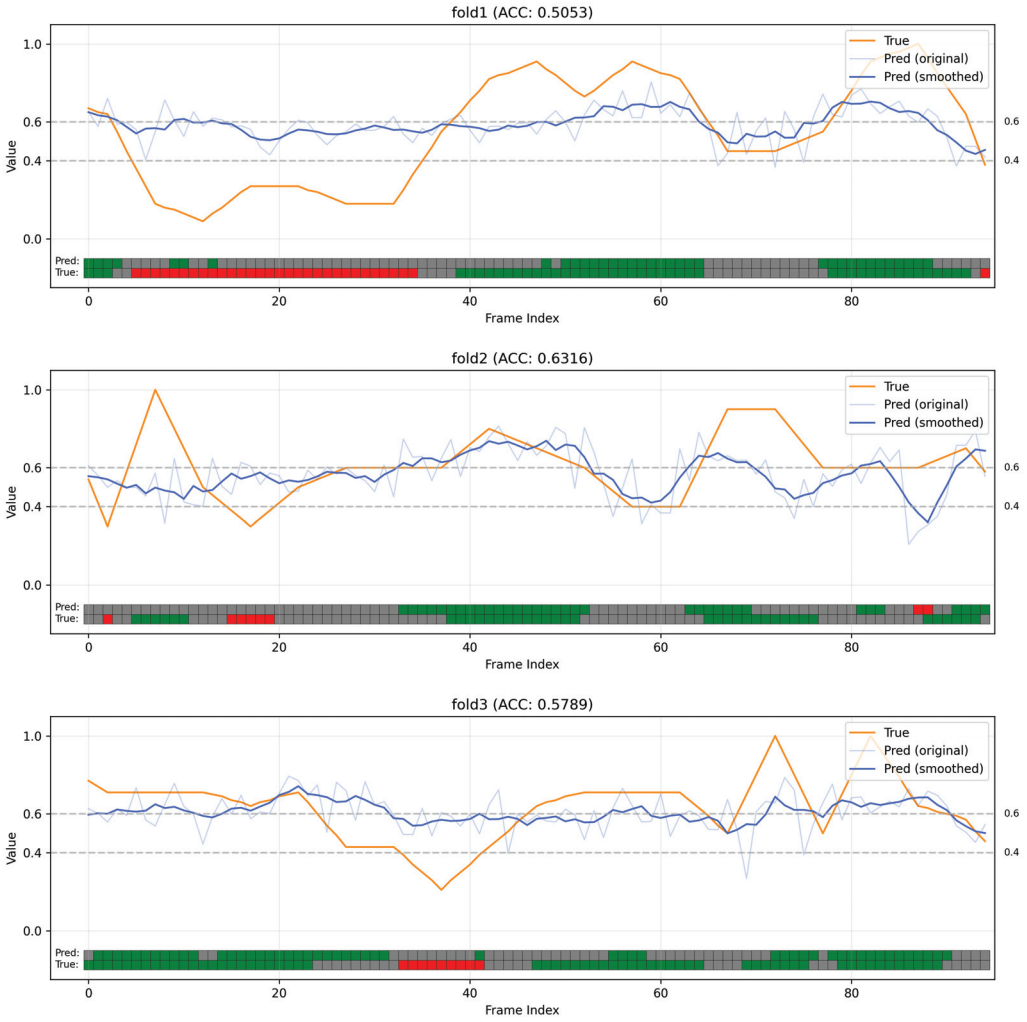


Figure 10. *Cont.*

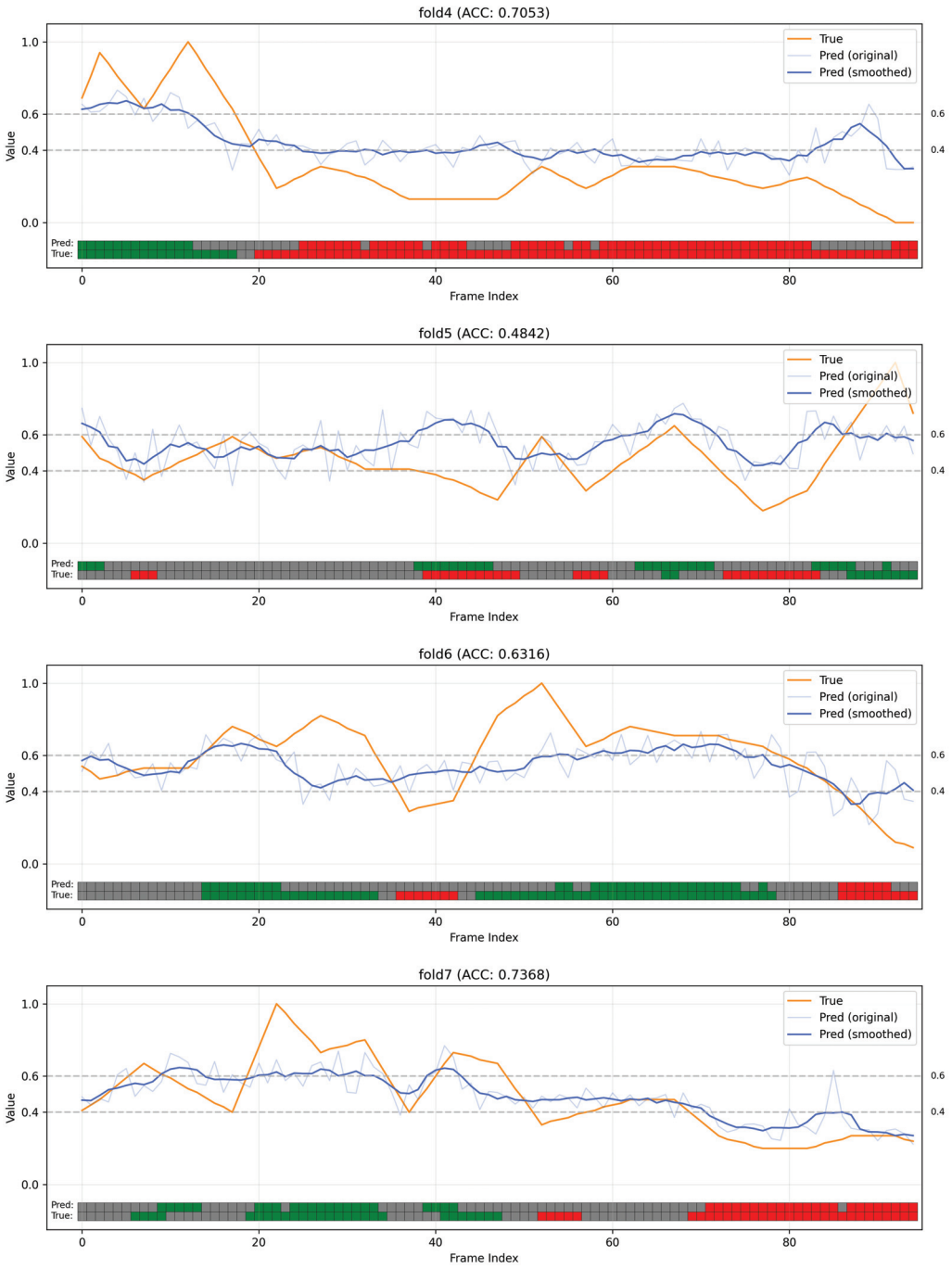


Figure 10. Visualizations of 7-fold cross-validation using ResNet50.

This trend-capturing capability, rather than absolute value precision, is the most critical quality for the intended application. It allows the system to reliably generate an 'attention heatmap' that guides students to conceptually dense segments. The practical utility of this non-semantic approach is underscored by its performance on lessons with known content peaks. For instance, the predicted peak in Fold 1 aligns with the ground truth for Lesson 1, corresponding to the explanation of the Version Space Algorithm, while the peak in Fold 4 correctly identifies the initial segment of Lesson 4, where the instructor demonstrates problem-solving steps. This alignment demonstrates that the model can effectively distinguish high-engagement segments from lulls in the lecture, which is the primary goal for enhancing archive navigation.

These results confirm that ResNet50 offers the most reliable foundation for the multimodal attention prediction framework, providing an optimal balance between prediction accuracy, generalization capability, and computational efficiency. Traditional convolutional architectures—particularly ResNet50 and VGG16—appear well-suited for this task, while more complex models showed no clear advantages under the experimental constraints.

5. Discussion

5.1. Addressing the Research Questions

In this section, we revisit the research questions introduced in the Introduction and evaluate how our findings address each.

> Main Research Question: Can segment access frequency (SAF) in lecture archives be accurately predicted using only non-semantic multimodal features, derived from real-world recordings without transcripts or semantic annotation?

The results confirm that non-semantic features can effectively predict SAF despite limited data. The full multimodal approach achieved a Pearson correlation of 0.5143 and 61.05% three-class classification accuracy in 7-fold cross-validation (Tables 3 and 5). Even using only instructor action features yielded a significant correlation (PCC = 0.5464, Table 2). These findings validate the hypothesis that SAF can be meaningfully predicted without semantic content understanding, even with an extremely limited dataset averaging only 8.71 valid viewers per lecture (Table 1).

> Sub RQ1: Which non-semantic modality—action, voice, or slide—contributes most to SAF prediction accuracy, and how do combinations of these modalities affect performance?

The ablation study (Table 4) revealed that instructor action features performed best in isolation, while slide features performed worst. However, any dual-modality combination outperformed its constituent single-modality models, with the full tri-modal fusion achieving optimal results. This confirms the complementary nature of the selected modalities and highlights the primary contribution of instructor actions.

> Sub RQ2: Which fusion strategy and neural network backbone provide the optimal balance of prediction accuracy and computational efficiency in resource-constrained educational settings?

The experiments indicate that a combination of feature-level fusion and a ResNet50 backbone provides the optimal trade-off. Feature-level fusion significantly outperformed model-level fusion across all metrics while requiring only 23% of the training time (Table 3). Among the tested backbone architectures, ResNet50 consistently outperformed alternatives across all metrics (Table 5), providing the best balance between accuracy and computational efficiency. Deeper networks like ResNet101 showed worse performance due to overfitting. Similarly, the Vision Transformer (ViT) model did not realize its full potential, and the CNN + LSTM architecture's explicit temporal modeling offered no significant advantages over purely spatial models at the feature resolution, demonstrating their unsuitability for the resource-constrained context.

These findings provide practical design guidelines for SAF prediction systems in educational environments with resource constraints, demonstrating the potential of non-semantic approaches for improving lecture archive accessibility. By identifying high-SAF segments, the framework enhances lecture archive usability in blended learning, supporting students' self-directed review after face-to-face instruction and improving integration of online and in-person learning.

5.2. Practical Implications and Potential Applications

To concretely illustrate the practical utility of the framework in an authentic blended learning environment, an in-depth analysis of the results from fold 7 (corresponding to Lesson 7) is conducted, as shown in Figure 10. This lecture exhibits three distinct phases: a high-SAF zone from 0–50 min, corresponding to the explanation of complex example problems; a medium-SAF zone from 50–70 min for conceptual review; and a low-SAF zone from 70–95 min, which features a live programming demonstration. By applying the system to this 95-min lecture, a highlight summary containing only the 22 min of high-SAF segments can be generated. This compresses the content to just 23.16% of its original length, demonstrating significant information compression efficiency.

From the students' perspective, the system's most direct value lies in the substantial improvement of learning efficiency. This 22-min summary enables students preparing for exams to bypass lengthy review and demonstration segments, allowing them to directly access the most critical parts of the example explanations for targeted and efficient review. Furthermore, having a clear learning map helps alleviate the sense of intimidation students often feel when confronted with long lecture videos, fostering a more positive and proactive learning experience.

From the instructor's perspective, the framework serves as a powerful tool for pedagogical diagnosis and intervention. By analyzing the SAF heatmap of existing lectures, instructors can accurately identify common student difficulties and points of confusion. More importantly, the predictive capability of the framework transcends this retrospective analysis to address the cold-start problem in pedagogical feedback. For a new lecture without any viewing data, the model can proactively generate a predicted SAF heatmap. This helps instructors anticipate potential bottlenecks and adjust their teaching materials accordingly, transforming pedagogical assessment from a reactive response into a proactive planning process.

5.3. Limitations

While the approach demonstrates the feasibility of non-semantic multimodal prediction for SAF, several limitations should be acknowledged:

1. Dataset Scope and Diversity. The experiments relied on seven lectures from a single Machine Learning course taught by one instructor, leading to performance variability across validation folds (Table 6). This constrained scope limits the model's generalizability to diverse educational contexts, such as humanities courses or interactive teaching formats, posing challenges for broader applicability in real-world settings.
2. Engagement Measurement Indirectness. SAF serves as an indirect proxy for engagement, primarily capturing revisitation patterns rather than immediate engagement states. This metric may not fully represent the multifaceted nature of student engagement, particularly during first-time viewing, as it primarily reflects post-hoc revisitation behaviors.
3. Temporal Resolution Constraints. The one-minute segment resolution, adopted to balance granularity and computational efficiency, overlooks transient engagement peaks, such as those triggered by key explanations or student questions. This coarse temporal scale restricts the model's precision in identifying brief, high-impact segments critical for applications like highlight extraction.
4. Non-semantic Feature Limitations. While the non-semantic approach offers cross-lingual applicability and low training costs, it inherently limits the model's ability to capture content-driven factors that may influence segment access frequency but are not explicitly reflected in non-semantic features. For example, in the prediction results for Lesson 1 in Figure 10, the model generated moderately high SAF predictions for

the first 30 min. This likely occurred because during this segment, the instructor was introducing himself, explaining the course structure and schedule—activities involving continuous speaking, writing, and movement which, without semantic understanding, appear similar to the delivery of information-dense concepts. However, in reality, this portion held little importance for students' review purposes, leading most students to skip it and resulting in consistently low actual SAF levels.

6. Conclusions

Building on the findings summarized in Table 7, this study has successfully established a lightweight, non-semantic framework for predicting Segment Access Frequency (SAF) in real-world lecture archives. Our results confirm that it is feasible to estimate student engagement patterns from multimodal features without relying on semantic cues. The comprehensive analysis identified the optimal combination of features and model architecture for this task under resource-constrained conditions.

Table 7. Summary of research questions, answers, and supporting results.

Research Question	Answer Summary	Supporting Results
Main RQ	SAF can be predicted using non-semantic features with moderate correlation (PCC = 0.5143).	Tables 3 and 5
Sub RQ1	Action features contribute most; all three modalities improve results.	Table 4
Sub RQ2	Feature-level fusion + ResNet50 best balance of performance and cost.	Tables 3 and 5

Compared to existing methods, the approach offers three key advantages, summarized as follows:

- Language independence: The non-semantic feature design allows the model to be applied across different languages without requiring content understanding.
- Suitability for educational settings with limited scale: The lightweight architecture achieves high computational efficiency and can be trained with limited dataset, such as university lecture archive.
- Automatic label generation: SAF labels are automatically derived from access logs, eliminating the need for manual annotation or specialized hardware.

To further improve model generalization and practical value, future work will expand to more diverse instructional contexts, integrate fine-grained behavioral cues (e.g., facial expressions and gaze), and explore lightweight semantic augmentation such as OCR-based slide content. The development of downstream applications—such as SAF heatmaps and automated highlight extraction—is also aimed, which will support both learners and instructors by enhancing content navigability, instructional feedback, and lecture archive usability.

Author Contributions: R.S. and S.H. conceived the research; R.S. was responsible for data processing and experimental implementation. All authors contributed to the writing and revision of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by JSPS KAKENHI Grant Number 23K28196.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are not publicly available but can be obtained from the corresponding author upon reasonable request, as they are part of an ongoing study.

Conflicts of Interest: Authors Ruozhu Sheng and Jinghong Li were employed by the company J-MAX Co., Ltd. The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Abuhammad, S. (2020). Barriers to distance learning during the COVID-19 outbreak: A qualitative review from parents' perspective. *Heliyon*, *6*(11), e05482. [CrossRef] [PubMed]

Agyeman, R., Muhammad, R., & Choi, G. S. (2019, March 28–30). *Soccer video summarization using deep learning* [Paper presentation]. 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), San Jose, CA, USA.

Alaa, T., Mongy, A., Bakr, A., Diab, M., & Gomaa, W. (2024). Video summarization techniques: A comprehensive review. *arXiv*. [CrossRef]

Andra, M. B., & Usagawa, T. (2019, March 12–14). *Automatic lecture video content summarization with attention-based recurrent neural network* [Paper presentation]. 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), Yogyakarta, Indonesia.

Benedetto, I., La Quatra, M., Cagliero, L., Canale, L., & Farinetti, L. (2024). Abstractive video lecture summarization: Applications and future prospects. *Education and Information Technologies*, *29*(3), 2951–2971. [CrossRef]

Bulathwela, S., Perez-Ortiz, M., Yilmaz, E., & Shawe-Taylor, J. (2020). Vlengagement: A dataset of scientific video lectures for evaluating population-based engagement. *arXiv*. [CrossRef]

Burton, A., & Radford, J. (1978). *Thinking in perspective: Critical essays in the study of thought processes*. Routledge.

Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2019). OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(1), 172–186. [CrossRef] [PubMed]

Chen, C.-M., & Wu, C.-H. (2015). Effects of different video lecture types on sustained attention, emotion, cognitive load, and learning performance. *Computers & Education*, *80*, 108–121. [CrossRef]

Deng, R., & Gao, Y. (2024). Effects of embedded questions in pre-class videos on learner perceptions, video engagement, and learning performance in flipped classrooms. *Active Learning in Higher Education*, *25*(3), 473–487. [CrossRef]

Dimitrova, N. (2004). Context and memory in multimedia content analysis. *IEEE Multimedia*, *11*(3), 7–11. [CrossRef]

Flanagan, J. L. (1972). Speech synthesis. In J. L. Flanagan (Ed.), *Speech analysis synthesis and perception* (pp. 204–276). Springer.

Guo, P. J., Kim, J., & Rubin, R. (2014, March 4–5). *How video production affects student engagement: An empirical study of MOOC videos* [Paper presentation]. First ACM Conference on Learning @ Scale (L@S '14), Atlanta, GA, USA.

Hasegawa, S., Tajima, Y., Matou, M., Futatsudera, M., & Ando, T. (2007, March 14–16). *Case studies for self-directed learning environment using lecture archives* [Paper presentation]. The Sixth IASTED International Conference on Web-based Education (WBE 2007), Chamonix, France.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. OTexts.

Khan, H., Hussain, T., Khan, S. U., Khan, Z. A., & Baik, S. W. (2024). Deep multi-scale pyramidal features network for supervised video summarization. *Expert Systems with Applications*, *237*, 121288. [CrossRef]

Khong, P., Holmes, D., Masoudian, B., Lund, G. C., & Garwood, S. (2025). Lecture capture, transcripts, and captioning in US colleges of osteopathic medicine: Descriptive cross-sectional survey. *Medical Science Educator*, *35*, 625–628. [CrossRef] [PubMed]

Kim, J., Guo, P. J., Seaton, D. T., Mitros, P., Gajos, K. Z., & Miller, R. C. (2014, March 4–5). *Understanding in-video dropouts and interaction peaks in online lecture videos* [Paper presentation]. First ACM Conference on Learning @ Scale (L@S '14), Atlanta, GA, USA.

Lamb, J., & Ross, J. (2022). Lecture capture, social topology, and the spatial and temporal arrangements of UK universities. *European Educational Research Journal*, *21*(6), 961–982. [CrossRef]

Lucas, B. D., & Kanade, T. (1981, August 24–28). *An iterative image registration technique with an application to stereo vision* [Paper presentation]. 7th International Joint Conference on Artificial Intelligence (IJCAI'81), Vancouver, BC, Canada.

Sablić, M., Mirosavljević, A., & Škugor, A. (2021). Video-based learning (VBL)—Past, present and future: An overview of the research published from 2008 to 2019. *Technology, Knowledge and Learning*, *26*(4), 1061–1077. [CrossRef]

Saykili, A. (2018). Distance education: Definitions, generations and key concepts and future directions. *International Journal of Contemporary Educational Research*, *5*(1), 2–17.

Seidel, N. (2024). Short, long, and segmented learning videos: From YouTube practice to enhanced video players. *Technology, Knowledge and Learning*, *29*(4), 1965–1991. [CrossRef]

Sheng, R., Ota, K., & Hasegawa, S. (2022, July 27–31). *An automatic focal period detection architecture for lecture archives* [Poster presentation]. 23rd International Conference on Artificial Intelligence in Education (AIED 2022), Durham, UK.

Shi, Y., Wang, M., Chen, Z., Hou, G., Wang, Z., Zheng, Q., & Sun, J. (2024). The impacts of instructor's visual attention and lecture type on students' learning performance and perceptions. *Education and Information Technologies*, *29*, 16469–16497. [CrossRef]

Singh, A., & Kumar, M. (2024). Bayesian fuzzy clustering and deep CNN-based automatic video summarization. *Multimedia Tools and Applications*, *83*(1), 963–1000. [CrossRef]

Tan, K., Zhou, Y., Xia, Q., Liu, R., & Chen, Y. (2024, August 7–9). *Large model based sequential keyframe extraction for video summarization* [Paper presentation]. International Conference on Computing, Machine Learning and Data Science (CMLDS), Singapore. [CrossRef]

Vermeulen, E. J., & Volman, M. L. (2024). Promoting student engagement in online education: Online learning experiences of Dutch university students. *Technology, Knowledge and Learning*, *29*(2), 941–961. [CrossRef]

Vimalaksha, A., Vinay, S., Prekash, A., & Kumar, N. S. (2018, December 10–13). *Automated summarization of lecture videos* [Paper presentation]. 2018 IEEE Tenth International Conference on Technology for Education (T4E), Chennai, India.

Walker, R., Voce, J., Ahmed, J., Nicholls, J., Swift, E., Horrigan, S., & Vincent, P. (2014). *A survey of technology enhanced learning: Case studies*. Universities and Colleges Information Systems Association.

Wyse, L. (2017). Audio spectrogram representations for processing with convolutional neural networks. *arXiv*. [CrossRef]

Zhang, J., Bourguet, M.-L., & Venture, G. (2018, July 4–6). *The effects of video instructor's body language on students' distribution of visual attention: An eye-tracking study* [Paper presentation]. 32nd International BCS Human Computer Interaction Conference (HCI 2018), Belfast, UK.

education sciences

MDPI

Article

Integrating AI-Driven Wearable Metaverse Technologies into Ubiquitous Blended Learning: A Framework Based on Embodied Interaction and Multi-Agent Collaboration

Jiaqi Xu [1], Xuesong Zhai [2,3,*], Nian-Shing Chen [4], Usman Ghani [5], Andreja Istenic [6,7] and Junyi Xin [8,*]

1 Graduate School of Education, Peking University, Beijing 100871, China; xujiaqi627@gmail.com
2 College of Education, Zhejiang University, Hangzhou 310058, China
3 School of Education, City University of Macau, Macau 999078, China
4 Program of Learning Sciences, National Taiwan Normal University, Taipei 100610, Taiwan; nianshing@gmail.com
5 Department of Business Administration, Iqra University, Karachi 75500, Pakistan; usman.ghani@iqra.edu.pk
6 Faculty of Education, University of Primorska, Cankarjeva 5, 6000 Koper, Slovenia; andreja.istenic7@gmail.com
7 Faculty of Civil and Geodetic Engineering, University of Ljubljana, Jamova 2, 1000 Ljubljana, Slovenia
8 School of Information Engineering, Hangzhou Medical College, Hangzhou 311399, China
* Correspondence: xszhai@zju.edu.cn (X.Z.); xinjunyi@hmc.edu.cn (J.X.)

Abstract: Ubiquitous blended learning, leveraging mobile devices, has democratized education by enabling autonomous and readily accessible knowledge acquisition. However, its reliance on traditional interfaces often limits learner immersion and meaningful interaction. The emergence of the wearable metaverse offers a compelling solution, promising enhanced multisensory experiences and adaptable learning environments that transcend the constraints of conventional ubiquitous learning. This research proposes a novel framework for ubiquitous blended learning in the wearable metaverse, aiming to address critical challenges, such as multi-source data fusion, effective human–computer collaboration, and efficient rendering on resource-constrained wearable devices, through the integration of embodied interaction and multi-agent collaboration. This framework leverages a real-time multi-modal data analysis architecture, powered by the MobileNetV4 and xLSTM neural networks, to facilitate the dynamic understanding of the learner's context and environment. Furthermore, we introduced a multi-agent interaction model, utilizing CrewAI and spatio-temporal graph neural networks, to orchestrate collaborative learning experiences and provide personalized guidance. Finally, we incorporated lightweight SLAM algorithms, augmented using visual perception techniques, to enable accurate spatial awareness and seamless navigation within the metaverse environment. This innovative framework aims to create immersive, scalable, and cost-effective learning spaces within the wearable metaverse.

Keywords: metaverse; embodied interaction; wearable; multi-agent; artificial intelligence; ubiquitous blended learning

1. Introduction

In recent years, the rise of the metaverse has opened up immense opportunities for the field of education. As an advanced immersive environment that blends virtual and physical realities, the metaverse has the potential to revolutionize learning methodologies and reshape educational paradigms (Phakamach et al., 2022). Especially for K-16 learners, this technology promises to unlock educational experiences that are otherwise

impossible, impractical, or unsafe (López-Belmonte et al., 2023).Wearable technology stands out as a key enabler of this vision, facilitating rich, multi-modal interactions within immersive ubiquitous learning environments. By supporting seamless, real-time interaction, wearable devices promise to deliver highly personalized, context-aware educational experiences that go beyond the limitations of traditional learning approaches (Zhou et al., 2024).

The wearable metaverse holds vast application potential while also facing challenges. Leading technology companies, such as Apple and Meta, have launched sophisticated wearable devices, like the Vision Pro and Orion, which integrate multisensory interactions and provide immersive experiences across industries, including education and tourism (Pan, 2024). Nevertheless, while the current research has extensively examined standalone applications of wearable devices in education, less attention has been paid to their potential as part of a cohesive ecosystem. The multi-source data collected by these wearable devices, such as physiological signals, environmental information, and user interaction data, remains underutilized due to a lack of effective integration and analysis (Chakma et al., 2021), hindering the development of high-level embodied interactions that rely on full-body perception. The current research has conducted preliminary explorations into the application of AI agents in the metaverse (X. Kang et al., 2024). However, the existing studies primarily focus on algorithm optimization or task simulation (Feng et al., 2025; Yu, 2023), while the specific mechanisms of human–agent collaboration remain largely unaddressed. As a result, intelligent agents tend to play a relatively passive role in metaverse learning environments, making it difficult to achieve truly intelligent and collaborative interactions. Furthermore, processing the multi-source data required to support complex embodied interactions demands substantial computational resources, which often exceed the capabilities of most wearable devices (X. Wang et al., 2024). Consequently, enabling low-computation-cost yet high-performance data analysis on these wearable devices has become a critical challenge.

This study proposes a conceptual framework for multi-source data analysis, a low calculation cost, and human–machine cooperation in wearable metauniverse environments. By integrating embodied interaction and multi-agent collaboration, this study proposes a comprehensive framework for designing wearable metauniverse learning environments. This framework combines a lightweight real-time multi-modal data processing framework, a multi-agent cooperation framework, and a rendering framework combining a lightweight SLAM algorithm with visual perception. This research aims to provide theoretical and technical support for the construction of wearable metaverse learning spaces and large-scale immersive and ubiquitous learning.

2. Literature Review

2.1. Wearable Devices in Ubiquitous Blended Learning

Wearable devices, with their portability, interactivity, and versatility, are an effective technological means for creating blended environments (Frisoli & Leonardis, 2024; Palermo et al., 2025). Wearable devices allow students to immerse themselves in a virtual space, mobilizing their vision, hearing, and even touch to achieve comprehensive interactions between themselves and the virtual environment. This learning method is consistent with the core principles of immersive and ubiquitous blended learning (Cárdenas-Robledo & Peña-Ayala, 2018). Furthermore, these devices provide students with continuous access to knowledge in various environments, offering greater flexibility, efficiency, and participation in the learning process. In recent years, the rapid development of wearable technologies such as smart watches, smart glasses, and smart clothing has further expanded the possibilities of metaverse-based learning. These wear-

able items help students to stay engaged by allowing them to be educated in different conditions instantaneously and continuously. To this end, smart glasses facilitate school children's multi-modal story creation by combining 3D virtual objects and hologram elements to enable the children to visualize their invented stories (Mills & Brown, 2023). A study quantified the learner experience and usability of a VR game using data from smartwatch gestures, finding that participants felt comfortable with the system, used it easily, and felt empowered (Nascimento et al., 2023).

Wearable devices play an active role in analyzing and enhancing learners' learning process. They can analyze learners' behaviors and emotional states in real time, such as by using smart bracelets to record students' attention levels and emotions and help build adaptive learning systems that dynamically respond to individual needs (Ba & Hu, 2023). This enables the development of personalized learning spaces in the education metaverse using heart rate signals to assess students' emotional engagement and cognitive activity levels (Z. Zhao et al., 2022). In addition, wearable devices have been integrated into various learning activities to enhance their immersion and interactivity. For example, at a museum's dinosaur exhibition presented in English, smart glasses were shown to significantly improve the learning efficiency and motivation compared to tablets (Chen et al., 2023). Similarly, wearable AR and hybrid AR/VR learning materials were also found to significantly improve high school students' situational interest, engagement, and learning performance in physics laboratories, with hybrid AR/VR outperforming traditional learning methods (J. C. Y. Sun et al., 2023).

Despite the numerous advantages of using wearable devices in educational applications, their further development still faces some technical challenges. One of the major hurdles is effectively processing and analyzing the heterogeneous multi-source data collected by wearable devices. Wearable technologies typically contain a variety of sensors that capture various data types such as physiological signals (e.g., eye-tracking data, heart rates, electroencephalograms) and environmental information (e.g., location, temperature) (Heikenfeld et al., 2018). Integrating and making sense of these disparate data sources requires developing sophisticated data fusion techniques and mining algorithms specifically tailored to the unique characteristics of wearable data. Moreover, the limited computational power, storage capacity, and communication capabilities of wearable devices pose significant barriers to their ability to support advanced learning analytics and interactive functionalities (Nahavandi et al., 2022). This limitation hinders the implementation of real-time and immersive interactive features (Hazarika & Rahmati, 2023). Research on low-computation-cost technologies for wearable devices used within immersive metaverse learning environments has become an urgent direction to address these challenges.

2.2. Embodied Interaction in Ubiquitous Blended Learning

With the advancement of cognitive science, embodied cognition theory has attracted widespread attention in education research. The theoretical framework emphasizes the basic role of the body in cognitive processes. It is believed that cognition not only depends on the function of the brain but also occurs through the dynamic interaction between the body and its environment (Foglia & Wilson, 2013). This view has opened up new research and practical application methods for use in immersive learning environments. Embodied interaction, as an emerging paradigm in human–computer interaction, integrates whole-body sensory and motor systems to create more natural and intuitive interactive experiences (Crowell et al., 2018). In immersive education environments, the application of embodied interaction mainly occurs across three dimensions.

First, designing and applying diversified interactive devices, such as motion capture systems, tactile feedback equipment, and brain–computer interfaces, provides technical

support for ubiquitous learning. These devices can track learners' physical movement, physiological state, and nerve signals in real time, thereby providing corresponding immersive feedback (Crowell et al., 2018; Fleury et al., 2020). For example, VR and motion capture have been used to offer an interactive Tai Chi learning system with a virtual coach, real-time feedback, and avatar control, enhancing self-learning by overcoming the limitations of traditional and video-based methods (J. Liu et al., 2020). Innovative wearable rings with multi-modal sensors and haptic feedback have enhanced immersive social interactions in metaverse-based education by enabling tactile and thermal perception (Z. Sun et al., 2022). Additionally, utilizing a brain–computer interface (BCI) to monitor a student's brain activity, an embodied robot can detect attention lapses in real time and provide immediate, adaptive responses, thereby improving learning efficacy (Vrins et al., 2022).

Second, embodied interaction has been implemented across various academic disciplines with various application patterns. For instance, Kinect sensors and gesture-based interactions have been used in physics education to create mixed-reality environments where students learn about electric fields through bodily movements and interactive gestures (Johnson-Glenberg & Megowan-Romanowicz, 2017). A study comparing traditional controls and 3D-printed haptic devices in a mixed-reality chemistry lesson found that while both groups exhibited improved knowledge, highly embodied interaction enhanced science identity and efficacy (Johnson-Glenberg et al., 2023). Virtual museums used in science education have been studied using eye-tracking technology to analyze students' performance and mental effort, with the goal of enhancing virtual museum design and resource development (Wu et al., 2024).

Third, embodied interaction can greatly improve students' motivation (Lindgren et al., 2016). An embodied interactive teaching model can be personalized and differentiated to meet students' needs, attracting their attention through multisensory presentation methods and thereby enabling more effective knowledge transfer and skill development. Embodied interaction has taken root in gamified learning. In such environments, embodied interaction technology supports interactive learning by integrating educational content into virtual contexts, stimulating students' intrinsic motivation and sparking their curiosity and interest (Abrahamson et al., 2020). The positive impact of embodied interaction on students' learning outcomes is well-established (Mira et al., 2024). It not only enhances students' cognitive abilities and motor skills but also fosters their emotional development and learning motivation (Kosmas & Zaphiris, 2023). However, in spite of the promising prospects of using embodied interaction in ubiquitous learning, its large-scale promotion and practical application still face multiple challenges. These include high costs and computational resource requirements, the need for more comprehensive principles and standards in interaction design, and the difficulty of accurately quantifying the comprehensive impact of embodied interaction on students' cognition, emotions, and behavior.

2.3. AI Agents in Metaverse

The rapid advancements in Large Language Models (LLMs) have led to significant breakthroughs in natural language understanding and generation by LLM-based agents, bringing revolutionary changes to education (Xi et al., 2023). LLMs empower AI agents with multidimensional capabilities such as perception, tool invocation, reasoning, planning, interaction, and self-evolution, enabling them to autonomously learn, make decisions, and act in complex, blended-reality environments (Gao et al., 2024). Through real-time interactions with the environment or humans, agents continuously optimize their behavioral strategies by receiving feedback (González-Briones et al., 2018), allow-

ing them to learn continuously in real-world scenarios and enhance their intelligence, interactivity, and collaboration. However, because a single agent struggles with a high cognitive load and inefficient task division in complex educational settings, multi-agent systems (MASs) represent a promising solution (Amirkhani & Barshooi, 2022). By incorporating social attributes and defining roles and communication mechanisms, MASs can engage in cooperative and competitive social interactions to handle more complex educational tasks (Song et al., 2024). MASs can share parameters, knowledge, and decisions, enhancing the robustness and scalability of algorithms through communication (Janbi et al., 2023). Additionally, through interactive collaboration, MASs simulate complex social scenarios that reflect group cooperation dynamics, helping learners understand the behaviors and emotions associated with different roles, thereby enhancing social perception. To simplify the development of MASs, researchers have created frameworks based on LLMs, such as AutoGen, CrewAI, CAMEL, and MetaGPT (Arslan et al., 2024). These frameworks provide powerful tools for facilitating collaboration and competition among agents.

Leveraging MAS frameworks will enhance the performance, efficiency, robustness, and scalability of metaverse educational systems. In metaverse educational practice, LLM-based MASs demonstrate excellent human–computer collaboration capabilities (Xia et al., 2024). Unlike traditional human–computer cooperation processes, MASs can manage human resources proactively by designing socially interactive virtual–physical roles. These systems simulate complex social role interactions, understand learners' social behavior, and dynamically adjust according to the social norms implied by users' actions and the environment (Gatto et al., 2022). This adaptive flexibility enables MASs to be widely applied in immersive educational scenarios such as video games, virtual reality, and training simulations. Examples include Stanford University's AI Agent Town (Park et al., 2023) and agent-based hospitals (Li et al., 2024). Despite the promising prospects of using LLM-based MASs in metaverse education, their development still faces numerous challenges. Specifically, these include the need to advance multi-agent collaboration algorithms, develop robust frameworks, and improve agents' recognition of metaverse elements (Gatto et al., 2022).

3. A Conceptual Framework for Wearable Metaverse Environments

3.1. The Overall Framework of the Model

This study constructed a wearable metaverse learning environment framework, as shown in Figure 1, to enhance the learning experience in immersive ubiquitous learning. The model consists of four key modules: (1) an Embodied Interaction Module; (2) Multi-Agent Collaboration Module; (3) Multi-Source Data Fusion Module; and (4) Low-Computational-Cost Optimization Module. Through the interconnection of these modules and their interaction with various components of the system, an immersive ubiquitous learning system is formed.

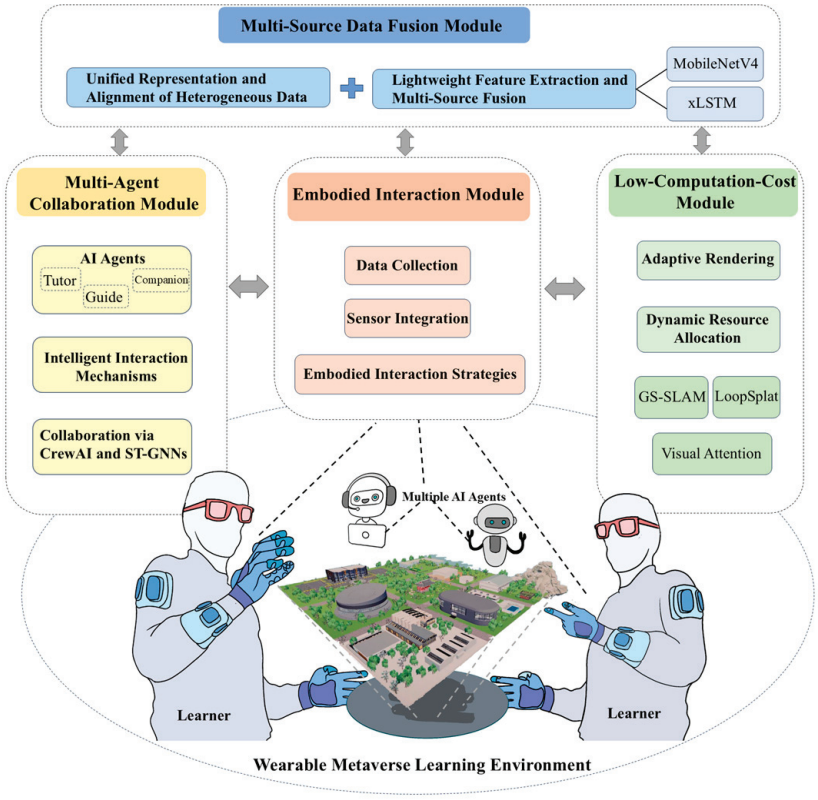


Figure 1. The wearable metaverse learning environment conceptual framework.

3.2. Embodied Interaction Module

3.2.1. Data Collection and Sensor Integration

In a wearable metaverse learning environment, the comprehensive and real-time collection of learners' multi-modal data forms the foundation for achieving embodied interaction (Closser et al., 2022). In this study, we developed a modular system for embodied data collection, integrating diverse sensors and wearable devices. The key components include an eye-tracking sensor embedded in smart glasses to capture metrics like fixation points, the gaze duration, and the blink frequency for analyzing learners' attention, cognitive load, and visual health; EEG sensors to assess learners' emotional states through their brainwave patterns; and a positioning wristband with inertial sensors to monitor real-time positions, movement trajectories, and gestures. Additionally, haptic sensors measure environmental parameters such as the temperature and pressure, enabling context-aware haptic simulations.

3.2.2. Embodied Interaction Strategies

In the proposed wearable metaverse learning environment, we suggest implementing embodied interaction strategies organized across three dimensions: interaction between learners, interaction between learners and the metaverse environment, and interaction between learners and the real environment (Table 1). These strategies collectively create an integrated learning experience that bridges the virtual and real worlds.

Table 1. Embodied interaction strategies in wearable metaverse environments.

Interaction Dimension	Contents	Description
Learner-to-Learner	Gesture Recognition	Capturing natural gestures for non-verbal communication and object manipulation in virtual spaces.
	Synchronized Activities	Mapping shared physical activities (e.g., virtual sports) in real time to foster collaboration.
	Haptic Feedback	Simulating remote physical touch, enhancing social presence and emotional connection.
Learner-to-Metaverse	Immersive Operations	Enabling direct and intuitive interaction with virtual objects through body movements.
	Multisensory Feedback	Providing rich experiences through integrated visual, auditory, and haptic feedback.
	Spatial Navigation	Allowing for natural navigation of virtual spaces using physical movements to enhance exploration.
Learner-to-Real-Environment	AR Annotations	Overlaying real-world objects with contextual learning information.
	Interaction Mapping	Mapping real-world actions to virtual environments for seamless learning.
	Environmental Adaptation	Dynamically adjusting learning content based on environmental data.

3.3. Multi-Agent Collaboration Module

To enhance learners' immersive interaction experience in wearable metaverse environments, the Multi-Agent Collaboration Module relies on multi-agent human–computer collaboration algorithms, enabling different intelligent agents to possess expertise in various domains and adapt to collaborative needs across multiple modalities and scenarios. This study proposes a multi-agent collaboration framework based on CrewAI and spatio-temporal graph neural networks (ST-GNNs).

3.3.1. Functions of Multi-Agent Module

The multi-agent module promotes deep human–machine collaboration, resource optimization, and the co-evolution of intelligence. On one hand, multiple agents can be flexibly defined and dynamically adjusted based on their identity, efficiently functioning in diverse interaction modes such as equal collaboration, the structured hierarchical division of labor, and spontaneous discussion. This not only provides users with multidimensional interactive experiences but also allows users to participate in collaborative problem-solving with intelligent agents, exploring the possibilities of using different cooperation models and responsibility sharing. On the other hand, multiple agents will filter and recommend high-quality resources, facilitate on-demand application and adaptive optimization, generate diverse content, act as virtual companions, and provide heuristic dialogue and metacognitive support through emotion perception and cognitive state adjustment, promoting the realization of deep collaboration modes. Furthermore, this study explored a mutual feeding mechanism involving multiple agents and human collaboration to achieve the co-evolution of technological capabilities and human intelligence. Within this framework, humans engage in knowledge co-construction and task resolution through deep collaboration with multiple agents, providing guidance and correction for the optimization of agent behavior, helping them continuously improve the depth of their professional knowledge in vertical domains. Simultaneously, intelligent agents provide feedback on individuals' cognitive and practical abilities through complex data analysis and providing adaptive behavior feedback.

3.3.2. Intelligent Interaction Mechanisms

Intelligent interaction mechanisms govern how learners and virtual agents engage within the environment. The interaction modes between learners and virtual agents include proactive modes where agents anticipate needs, passive modes where agents respond upon request, hybrid modes combining both approaches, and group collaboration modes involving multiple agents for complex tasks. Additionally, real-time context-based adaptation enhances interactions through the analysis of learners' behavior, awareness of environmental changes, optimization of interaction strategies based on feedback, and personalized tuning of agents' parameters to ensure tailored and effective support.

3.3.3. Collaboration Using CrewAI and ST-GNNs

The framework utilizes the CrewAI framework to structure the collaboration among agents. CrewAI allows us to define distinct roles, responsibilities, and goals for each agent, thereby forming a cohesive and mission-focused team dedicated to a specific learning task. This framework supports sophisticated workflows where agents can operate in parallel, delegate tasks, and communicate sequentially, mirroring real-world team dynamics.

To achieve seamless coordination between agents' actions and learners' movements, ST-GNNs are employed. This technology is essential for processing and interpreting the complex, dynamic relationships between multiple entities (learners and agents) in both space and time.

3.4. Multi-Source Data Fusion Module

A lightweight and scalable solution is critical for efficient real-time multi-modal data processing in resource-constrained wearable metaverse environments. One promising approach is to integrate MobileNetV4-based lightweight feature extraction with xLSTM-based multi-source fusion.

The feature extraction module leverages MobileNetV4's depthwise separable convolutions to balance computational efficiency and high performance (Qin et al., 2025). Each modality-specific branch can be independently trained, and the extracted features are integrated into a shared embedding space, enabling seamless integration. Thanks to the efficient architecture of MobileNetV4, this module can support low-latency, real-time inference on resource-constrained devices such as wearables, providing a solid foundation for immersive interactions.

Building on the extracted multi-modal features, an xLSTM network can be employed to fuse features from different modalities and model their temporal dependencies (Beck et al., 2024). Specifically, feature vectors from different branches are fed in parallel into the corresponding input gates of xLSTM. By introducing modality interaction units, xLSTM can explicitly learn the temporal correlation patterns across different modalities, capturing long-range cross-modal dependencies. Additionally, the gating mechanisms of xLSTM allow it to adaptively decide which modality information should be updated or retained at each time step, enhancing the flexibility and robustness of cross-modal information fusion.

3.5. Low-Computation-Cost Strategy Module

The low-computation-cost strategy module plays a critical role in this project, aiming to reduce computational complexity and deliver a smooth, high-quality visual embodied interaction experience.

Inspired by the spatial resolution distribution of the human visual system, we propose that rendering in wearable metaverse learning environments should adopt an adaptive resolution rendering method based on gaze tracking. Specifically, the density of cone cells in the retina peaks in the foveal region (the area surrounding the gaze point), providing

the highest visual resolution, while progressively decreasing toward the peripheral areas (Reiniger et al., 2021). Based on this characteristic, the rendering engine of wearable metaverse learning environments needs to dynamically track the user's gaze position in real time and adjust the rendering resolution accordingly, with regions closer to the gaze point rendered in higher detail and peripheral regions rendered at lower resolutions to optimize computational efficiency without compromising visual quality. Therefore, a low-computation-cost rendering strategy for wearable metaverse learning environments should follow these principles:

- High fidelity in gaze-sensitive areas: Regions closer to the gaze point are rendered at higher resolutions to ensure high-fidelity viewing in the user's focus area;
- Optimized peripheral rendering: Regions further away from the viewpoint are rendered at lower resolutions, reducing the computational demands while maintaining acceptable visual quality.

This adaptive-resolution rendering technique balances the computational cost and visual quality, and even under resource-limited conditions, a high-quality visual presentation is maintained.

To enable efficient environment perception and modeling, this framework adopts lightweight SLAM algorithms, such as GS-SLAM + LoopSplat, to construct 3D environment maps and estimate device pose changes in real time on low-power devices. To enhance the learner's experience, the framework includes a visual perception optimization module that dynamically adjusts the rendering parameters based on real-time quality assessment using algorithms such as CrossScore and GR-PSN. Perceptual mapping techniques like VDP (Visual Difference Prediction) are employed to identify areas of higher visual importance, ensuring that system resources are allocated to maximize the perceptual quality.

By employing this dynamic resource allocation strategy, the system achieves smooth interaction performance and improved overall resource utilization.

4. Technical Approaches for Implementing Wearable Metaverse Environments

The construction of a future immersive ubiquitous learning environment aims to promote educational equity, improve learning outcomes, and optimize resource allocation (X. S. Zhai et al., 2023). However, since the fully immersive ubiquitous learning era has not yet arrived, the concept of the educational metaverse is still in its exploratory phase, and large-scale experimental research is not yet feasible. Therefore, this study addresses specific educational challenges that may arise in future immersive ubiquitous learning environments and proposes four key technical pathways to address these. These pathways aim to overcome the limitations of traditional educational methods through embodied interaction and multi-agent collaboration, providing innovative solutions to challenges such as data collection precision in enhancing immersive learning experiences.

4.1. Enhancing Precision Through Multi-Source Data Fusion

In future immersive ubiquitous learning environments, traditional data collection methods, such as surveys and interviews, may prove insufficient for capturing the dynamic and embodied interactions between learners, intelligent agents, and the virtual–physical environment. The real-time analysis of multi-modal learner data is essential for supporting personalized and adaptive learning (Di Mitri et al., 2022; X. Zhai et al., 2023). However, the heterogeneous nature of data sources in wearable metaverse learning environments poses significant challenges in terms of data integration, synchronization, and interpretation.

This study introduces a lightweight neural network architecture combining MobileNetV4 and xLSTM for multi-source heterogeneous data fusion and analysis, as shown

in Figure 2. This approach enables the efficient extraction and integration of features from various data modalities, such as text, images, speech, and sensor data, while maintaining the real-time performance on resource-constrained wearable devices.

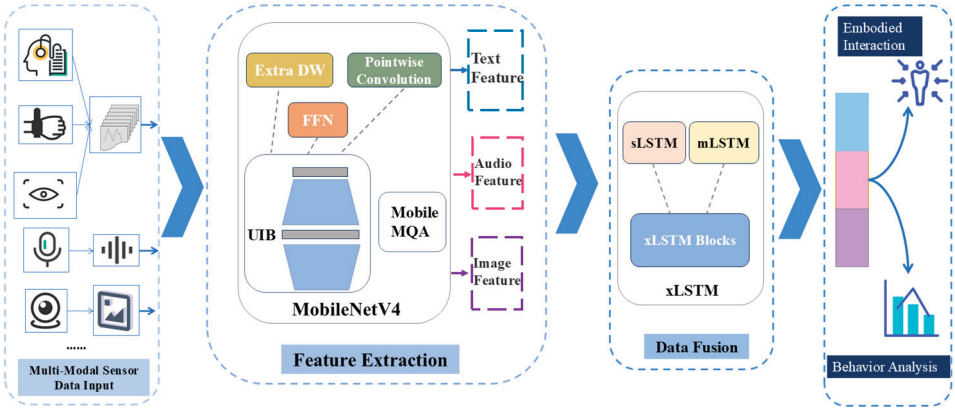


Figure 2. Multi-source data fusion framework.

4.1.1. Feature Extraction with MobileNetV4

The lightweight feature extraction module for use in wearable metaverse environments leverages MobileNetV4, which effectively balances computational efficiency and model performance using its universally efficient architecture designs for mobile devices. MobileNetV4 introduces the Universal Inverted Bottleneck (UIB) search block, a unified and flexible structure that merges an Inverted Bottleneck (IB), ConvNext, Feed-Forward Network (FFN), a novel Extra-Depthwise (ExtraDW) variant, and Mobile Multi-Query Attention (Mobile MQA) (Qin et al., 2025).

- Text Data: Discrete text data is transformed into continuous low-dimensional semantic vectors using an embedding layer. These vectors are then processed through modified UIB blocks, specifically adapted for text data, to extract high-level semantic features.
- Image and Video Data: MobileNetV4's depthwise separable convolution, as a key element of the UIB block, is leveraged to efficiently extract spatial features from image data. For video data, these spatial features are temporally aggregated using temporal modeling layers, such as the Mobile MQA attention block, enabling the capture of dynamic temporal dependencies.
- Speech Data: High-level acoustic features are initially extracted using a pre-trained acoustic model (e.g., Wav2Vec or HuBERT). These features are subsequently compressed using MobileNetV4's UIB blocks, which are fine-tuned for speech data, to reduce the dimensionality without losing essential information. The Mobile MQA attention block is then applied to capture long-range dependencies within the speech sequences.

Each modality-specific branch is independently trained to maximize its individual performance. The extracted features are then projected into a shared embedding space for cross-modality integration. By leveraging the efficient architecture of MobileNetV4, this module ensures real-time feature extraction and compatibility with resource-constrained devices, such as wearable hardware.

4.1.2. Dynamic Cross-Modality Fusion with xLSTM

After extracting feature maps from multi-modal data using MobileNetV4, fusion employing xLSTM involves the following steps:

- Input Transformation: The extracted feature maps are pre-processed (e.g., normalization, dimensionality alignment) to ensure compatibility across the modalities.

- Temporal Alignment Using Modality Interaction Units (MIUs): MIUs in xLSTM explicitly model the temporal relationships between the modalities.
- Dynamic Modality Weighting: At each time step, xLSTM calculates the relative importance of each modality using learned weighting parameters.
- Output Fusion for Downstream Tasks: The fused multi-modal representation is passed to task-specific layers (e.g., classification, regression, or decision-making modules).

This architecture supports a highly adaptive and scalable data fusion process by combining MobileNetV4's efficient feature extraction with xLSTM's advanced temporal modeling capabilities.

4.2. Agents' Collaboration Based on Multi-Agent Framework and Graph Neural Networks

Effective collaboration between learners, virtual agents, and the environment is critical for achieving an interactive and adaptive learning experience in the wearable metaverse. However, ensuring seamless coordination between multiple intelligent agents and the dynamic virtual–physical environment is a significant challenge, particularly in terms of agent communication, task allocation, and context awareness.

This study proposes a multi-agent collaborative model based on the CrewAI and spatio-temporal graph neural networks (ST-GNNs), as shown in Figure 3, integrating learners, intelligent agents, virtual environments, and real-world environments. Crew AI offers a decentralized network of communicating agents with decentralized architectures and coordination, providing flexibility and scalability for such systems. ST-GNNs can consider the fine details of the spacetime relationship of the agent with its surroundings, which may support context-aware decision-making and adaptations to the environment. A hybrid learning setup that includes this technology can make the proposed model particularly useful for education in a hyper-virtual world incorporating collective studying, such as through problem-solving situations, role-playing, and interactive simulations.

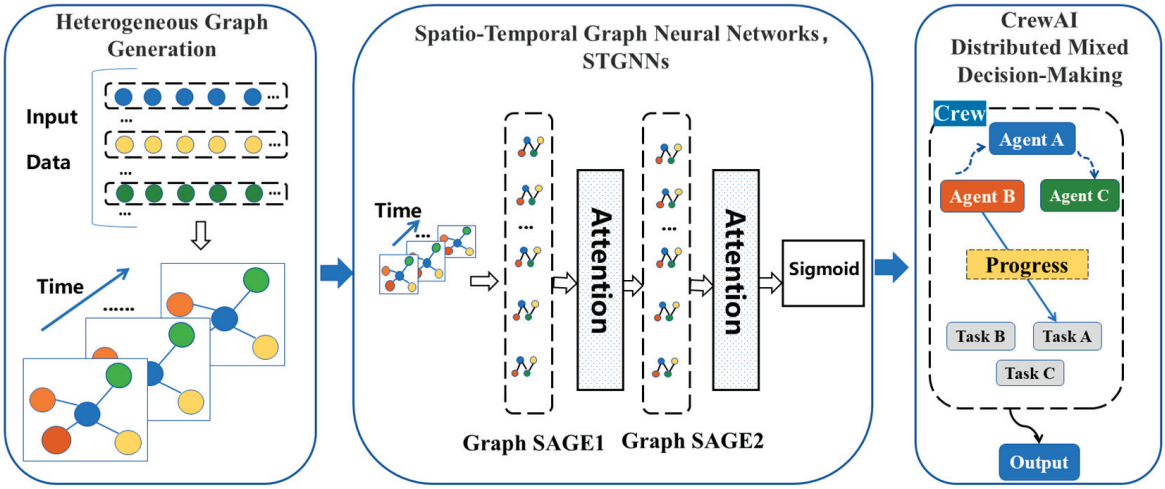


Figure 3. Multi-agent collaborative model.

4.2.1. Spatio-Temporal Collaboration Modeling with ST-GNNs

To capture and model the complex interactions among learners, intelligent agents, and their environments, the framework employs spatio-temporal graph neural networks (ST-GNNs) (Sahili & Awad, 2023). This approach first constructs a spatio-temporal heterogeneous behavior graph, where learners, agents, and environments are represented as nodes with both static and dynamic features and their timestamped interactions form the edges. Within this graph, the framework models the spatial dimension using heterogeneous graph attention networks to aggregate information from neighboring nodes,

while simultaneously addressing the temporal dimension using temporal convolutional networks to capture the evolution of features over time. Through this integrated analysis, the ST-GNNs generate low-dimensional collaborative embeddings for each agent. These embeddings encode their roles, states, and mutual dependencies, providing a rich feature space for downstream decision-making. By leveraging ST-GNNs, this framework identifies and models complex spatio-temporal dependencies, enabling adaptive and cooperative interactions that enhance learning outcomes.

4.2.2. Distributed and Hybrid Decision-Making with CrewAI

This framework adopts the CrewAI (Barbarroxa et al., 2025) paradigm to organize distributed decision-making and coordination among agents at both the macro and micro levels:

- Macro-Level Coordination: A central platform agent serves as a global coordinator, aggregating information from all the agents and generating high-level decisions using graph neural networks. The platform agent evaluates the states of learners, virtual environments, and real-world contexts to identify optimal task–agent matches. For example, it might assign a specific virtual tutor to a struggling student or coordinate collaborative tasks among urban and rural students.
- Micro-Level Distributed Decisions: Individual agents (e.g., virtual tutors, learning companions, or environment agents) independently generate localized decisions based on their private states. Using deep reinforcement learning, the agents express personalized preferences for scheduling or task execution, which are communicated back to the platform agent through CrewAI's interaction mechanisms. This two-way communication ensures that global decisions are informed by local needs while maintaining the overall system coherence.

This hybrid decision-making approach balances centralized coordination with decentralized adaptability, ensuring scalability and responsiveness in complex learning scenarios.

4.3. Optimization of Visual Experiences Based on Low Computation Cost

As metaverse technologies have continued to evolve, the demand for immersive user experiences has grown significantly. From early Three-Degrees-of-Freedom (3DOF) head tracking to the current Six-Degrees-of-Freedom (6DOF) head and hand tracking, VR/AR devices are increasingly simulating interactions that closely resemble the real world (Manawadu & Park, 2024). However, achieving high-resolution, wide-field-of-view, and low-latency rendering in wearable devices poses significant computational challenges, particularly in resource-constrained environments.

This study argues that wearable metaverse learning environments should be based on low-computation-cost rendering technologies. By leveraging the characteristics of the human visual system, the framework dynamically adjusts the rendering strategies based on the learner's gaze position, ensuring high-fidelity rendering in visually sensitive regions while reducing the computational load for peripheral areas. This approach not only improves the visual fidelity of wearable metaverse learning environments but also enhances user comfort and reduces power consumption, making it more feasible for prolonged use in educational settings. The framework integrates four key technical components, as shown in Figure 4.

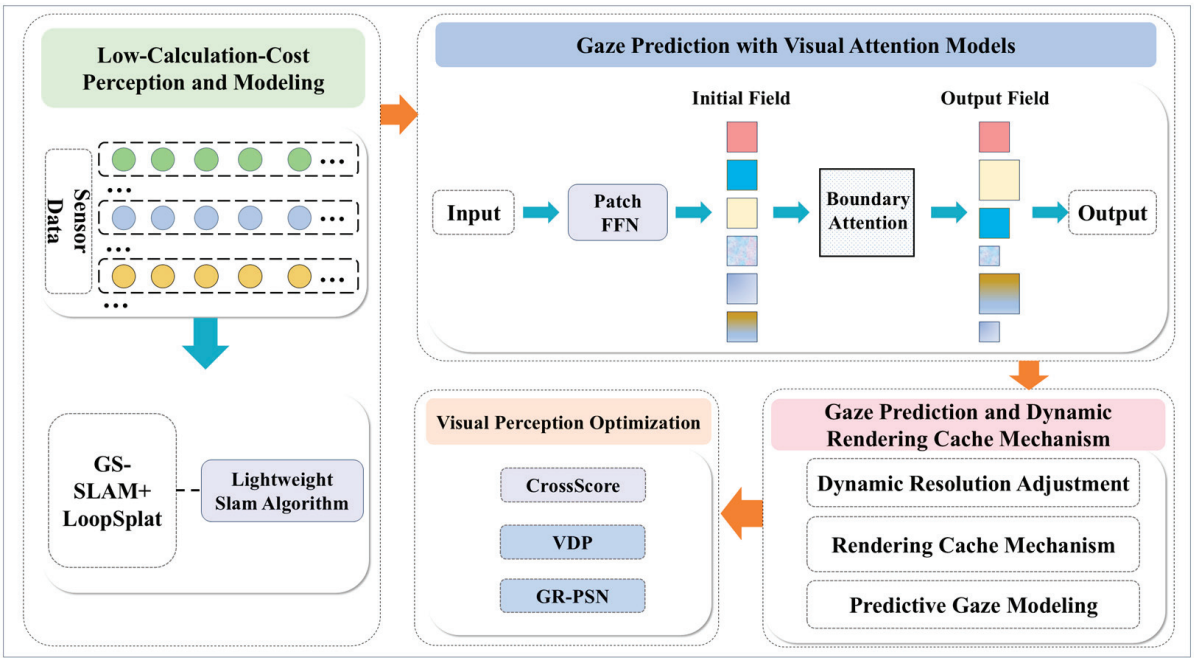


Figure 4. Low-computation-cost rendering framework.

4.3.1. Low-Computation-Cost Environment Perception and Modeling

To enable efficient environment perception and modeling, this framework adopts optimized, lightweight SLAM (Simultaneous Localization and Mapping) algorithms, such as GS-SLAM + LoopSplat (Zhu et al., 2024), to construct 3D environment maps and estimate device pose changes in real time on low-power devices. By utilizing sparse feature extraction and efficient graph optimization techniques, the framework significantly reduces the computational costs while maintaining effective performance.

4.3.2. Gaze Prediction Using Visual Attention Models

To improve the rendering efficiency and user experience, the framework incorporates real-time gaze tracking and visual attention modeling:

- Visual Attention Models: Inspired by the human visual system, lightweight convolutional neural networks (e.g., boundary attention models) are used to predict potential regions of interest in images or videos (Polansky et al., 2024). These predictions guide rendering optimizations by focusing computational resources on areas the user is likely to attend to.
- Real-Time Gaze Tracking: The system utilizes low-computation-cost gaze-tracking algorithms to identify the learner's gaze position in real time, ensuring that the rendering priorities align with the user's visual attention.

This gaze prediction mechanism provides precise data to dynamically optimize the rendering strategies while reducing unnecessary computational overheads.

4.3.3. Gaze Prediction and Dynamic Rendering Cache Mechanism

A core module of the framework integrates gaze prediction with a dynamic rendering cache mechanism to optimize the rendering efficiency:

- Dynamic Resolution Adjustment: Based on gaze prediction, the rendering engine dynamically adjusts the resolution of different regions. Higher resolutions are prioritized for gaze-sensitive areas, while peripheral regions are rendered at lower resolutions. Techniques such as the Level of Detail (LOD) method and frustum culling (Su et al., 2017) are used to allocate resources effectively.

- Rendering Cache Mechanism: Leveraging temporal coherence, previously rendered frames are stored and reused to avoid redundant computations. Frame difference encoding and result compression techniques are further applied to reduce the computational cost for static or minimally changing regions.
- Predictive Gaze Modeling: Recurrent neural networks (e.g., RNNs) predict potential gaze shifts, allowing the system to pre-render areas of future interest and minimize the latency.

This module ensures the efficient utilization of computational resources while maintaining high-quality visual experiences in key areas of user attention.

4.3.4. Visual Perception Optimization

To enhance the learner experience, this framework incorporates a visual perception optimization module that dynamically adjusts the rendering parameters based on real-time quality assessment:

- Image Quality Evaluation: Algorithms such as CrossScore (Z. Wang et al., 2025) and GR-PSN (Ju et al., 2024) are used to assess the visual quality of the rendered frames in real time. These evaluations guide adjustments to the rendering parameters, such as the resolution and texture detail, to balance visual fidelity and computational efficiency.
- Perceptual Mapping Techniques: Techniques like VDP (Visual Difference Prediction) (Mantiuk et al., 2023) are employed to identify areas of higher visual importance, ensuring that the system resources are allocated in a way that maximizes the perceptual quality.

5. Discussion

This study proposes a lightweight data processing solution tailored to the needs of ubiquitous wearable metaverse environments, effectively facilitating human–computer interaction and supporting ubiquitous learning. The integration of the MobileNetV4 and xLSTM algorithms helps improve the computational efficiency on resource-constrained wearable devices, thereby enhancing the model's performance. On the one hand, MobileNetV4's efficient architecture, including the Universal Inverted Bottleneck (UIB) block and Mobile MQA mechanism, provides very high real-time feature extraction accuracy and computational efficiency (Qin et al., 2025). Such model developments acknowledge the developing requirements for light yet powerful models to be implemented in learning systems (G. Zhao et al., 2021). The xLSTM model, on the other hand, captures the long-term dependencies and models the temporal dynamics of different modalities, making it particularly suitable for processing the various time series data generated in metaverse learning environments (Alharthi & Mahmood, 2024). By reducing the need for computing resources, this framework enhances the feasibility of allowing learners to receive education on resource-limited wearable devices anytime and anywhere, promoting equitable access to metaverse-based education across different hardware platforms.

Secondly, this study proposes a human–machine collaboration framework that integrates multiple agents. The multi-agent system-based framework relies on coordination methods to link students, agents, the virtual reality environment, and reality together to form a vibrant, collaborative learning system. Meanwhile, it applies spatio-temporal heterogeneous behavior graphs, allowing the varying behavior parameters of learners, agents, and their environments to be witnessed and researched because they are important for both analyzing and optimizing interactive behaviors. Inside this mixed-learning metaverse space, besides providing novelty and depth to traditional cooperative learning activities, multi-agent collaboration is also a step toward a human–computer education model (Lin,

2015). This framework not only focuses on traditional collaborative learning between students but also explores the interaction between learners and agents and between agents, making it more suitable for ubiquitous learning in the metaverse environment. Multi-agent systems, supported by large-scale language models, can bolster equity in learning, make education more personalized and context-aware, and make the possibility of individualized and resource-rich education more realistic (Cheng et al., 2024). That is, the joint efforts of agents will enable personalized, context-aware, and emotional learning processes, where the recommendations will be changed dynamically and tasks' difficulty will be adjusted based on the learner's trajectory.

Finally, this study proposes the adoption of a low-computation-cost rendering strategy in wearable metaverse learning environments to achieve high-quality visual rendering under resource-constrained conditions. This strategy aims to solve the performance bottlenecks faced by mobile and wearable devices when processing complex metaverse scenes, ensuring that learners can experience smooth and realistic interactions in a dynamic, immersive environment. Specifically, the research focused on rendering methods based on lightweight SLAM algorithms and boundary attention frameworks. This method improves the computing efficiency to a certain extent while optimizing resource allocation and ensures the smoothness and realism of educational scene rendering to the greatest possible extent. This method is consistent with cutting-edge research in the field of mobile and wearable devices, focusing on enhancing performance and the image quality using deep learning algorithms (Suo et al., 2023). In educational applications, this strategy ensures the feasibility of running complex metaverse learning applications on resource-constrained wearable devices. In addition, this solution helps reduce the reliance of metaverse resources on high-performance hardware devices, thereby promoting the large-scale adoption of wearable ubiquitous learning.

6. Conclusions

This study proposes a framework and technical solution for developing wearable metaverse learning environments, aiming to achieve immersive and ubiquitous learning experiences through the innovative combination of lightweight data processing, multi-agent collaboration, and low-computation-cost technologies. While significant progress has been made in theoretical exploration and technical design, there remain certain limitations that need to be addressed in future research. On the one hand, as metaverse technology is still in its infancy, this study primarily focused on providing a conceptual framework and technical solution. However, large-scale empirical studies have yet to be conducted to validate the effectiveness and feasibility of the proposed framework and solutions in real-world scenarios. The framework also needs to be further integrated with educational practice by designing and testing specific learning scenarios for different K-16 subjects. On the other hand, aspects such as the computational efficiency and scalability of the lightweight data processing framework, the interaction modeling capabilities of the multi-agent collaboration framework, and the environmental adaptability of the low-computation-cost rendering strategy require further evaluation and optimization through practical system implementation and user studies.

As metaverse technology continues to evolve and mature, we anticipate the emergence of more prototype systems and application scenarios. It is important to explore the applicability and acceptability of metaverse technologies across different teaching applications, instructional models, and educational strategies to ensure their global inclusivity, universality, and sustainability (Y. Liu & Fu, 2024). Through interdisciplinary collaboration and iterative improvements, the wearable metaverse learning environment

can be continuously optimized and refined, ultimately enabling ubiquitous and intelligent learning transformations.

Author Contributions: Conceptualization, J.X. (Jiaqi Xu), X.Z. and A.I.; Methodology, J.X. (Jiaqi Xu), X.Z. and N.-S.C.; Investigation, J.X. (Jiaqi Xu) and U.G.; Resources, X.Z., N.-S.C. and A.I.; Writing—original draft, J.X. (Jiaqi Xu); Writing—review & editing, J.X. (Jiaqi Xu), X.Z., N.-S.C., U.G., A.I. and J.X. (Junyi Xin); Visualization, J.X. (Jiaqi Xu); Supervision, X.Z. and J.X. (Junyi Xin); Project administration, X.Z.; Funding acquisition, X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Zhejiang Provincial Natural Science Foundation of China, grant number: Y24F020009; Zhejiang Provincial Education Science Planning Project, grant number: 2024SCG247; China Association for Science and Technology (CAST) 2024 Graduate Student Science Popularization Competence Enhancement Program, grant number: KXYJS2024008; Major Project of Humanities and Social Sciences in Higher Education Institutions of Zhejiang Province, grant number: 2023QN075.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

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

References

Abrahamson, D., Nathan, M. J., Williams-Pierce, C., Walkington, C., Ottmar, E. R., Soto, H., & Alibali, M. W. (2020). The future of embodied design for mathematics teaching and learning. *Frontiers in Education*, *5*, 147. [CrossRef]

Alharthi, M., & Mahmood, A. (2024). xLSTMTime: Long-term time series forecasting with xLSTM. *AI*, *5*(3), 1482–1495. [CrossRef]

Amirkhani, A., & Barshooi, A. H. (2022). Consensus in multi-agent systems: A review. *Artificial Intelligence Review*, *55*(5), 3897–3935. [CrossRef]

Arslan, M., Munawar, S., & Cruz, C. (2024). Sustainable digitalization of business with multi-agent RAG and LLM. *Procedia Computer Science*, *246*, 4722–4731. [CrossRef]

Ba, S., & Hu, X. (2023). Measuring emotions in education using wearable devices: A systematic review. *Computers & Education*, *200*, 104797. [CrossRef]

Barbarroxa, R., Gomes, L., & Vale, Z. (2025). Benchmarking large language models for multi-agent systems: A comparative analysis of AutoGen, CrewAI, and TaskWeaver. In P. Mathieu, & F. De La Prieta (Eds.), *Advances in practical applications of agents, multi-agent systems, and digital twins: The PAAMS collection* (Vol. 15157, pp. 39–48). Springer Nature. [CrossRef]

Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., Klambauer, G., Brandstetter, J., & Hochreiter, S. (2024). xLSTM: Extended long short-term memory. *arXiv*, arXiv:2405.04517. [CrossRef]

Cárdenas-Robledo, L. A., & Peña-Ayala, A. (2018). Ubiquitous learning: A systematic review. *Telematics and Informatics*, *35*(5), 1097–1132. [CrossRef]

Chakma, A., Faridee, A. Z. M., Khan, M. A. A. H., & Roy, N. (2021). Activity recognition in wearables using adversarial multi-source domain adaptation. *Smart Health*, *19*, 100174. [CrossRef]

Chen, H. R., Lin, W. S., Hsu, T. Y., Lin, T. C., & Chen, N. S. (2023). Applying smart glasses in situated exploration for learning English in a national science museum. *IEEE Transactions on Learning Technologies*, *16*(5), 820–830. [CrossRef]

Cheng, Y., Zhang, C., Zhang, Z., Meng, X., Hong, S., Li, W., Wang, Z., Wang, Z., Yin, F., Zhao, J., & He, X. (2024). Exploring large language model based intelligent agents: Definitions, methods, and prospects. *arXiv*, arXiv:2401.03428. [CrossRef]

Closser, A. H., Erickson, J. A., Smith, H., Varatharaj, A., & Botelho, A. F. (2022). Blending learning analytics and embodied design to model students' comprehension of measurement using their actions, speech, and gestures. *International Journal of Child-Computer Interaction*, *32*, 100391. [CrossRef]

Crowell, C., Mora-Guiard, J., & Pares, N. (2018). Impact of interaction paradigms on full-body interaction collocated experiences for promoting social initiation and collaboration. *Human–Computer Interaction*, *33*(5–6), 422–454. [CrossRef]

Di Mitri, D., Schneider, J., & Drachsler, H. (2022). Keep me in the loop: Real-time feedback with multimodal data. *International Journal of Artificial Intelligence in Education*, *32*(4), 1093–1118. [CrossRef]

Feng, L., Jiang, X., Sun, Y., Niyato, D., Zhou, Y., Gu, S., Yang, Z., Yang, Y., & Zhou, F. (2025). Resource allocation for metaverse experience optimization: A multi-objective multi-agent evolutionary reinforcement learning approach. *IEEE Transactions on Mobile Computing*, *24*(4), 3473–3488. [CrossRef]

Fleury, M., Lioi, G., Barillot, C., & Lécuyer, A. (2020). A survey on the use of haptic feedback for brain-computer interfaces and neurofeedback. *Frontiers in Neuroscience*, *14*, 528. [CrossRef] [PubMed]

Foglia, L., & Wilson, R. A. (2013). Embodied cognition. *WIREs Cognitive Science*, *4*(3), 319–325. [CrossRef] [PubMed]

Frisoli, A., & Leonardis, D. (2024). Wearable haptics for virtual reality and beyond. *Nature Reviews Electrical Engineering*, *1*(10), 666–679. [CrossRef]

Gao, C., Lan, X., Li, N., Yuan, Y., Ding, J., Zhou, Z., Xu, F., & Li, Y. (2024). Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, *11*(1), 1259. [CrossRef]

Gatto, L., Fulvio Gaglio, G., Augello, A., Caggianese, G., Gallo, L., & La Cascia, M. (2022, October 19–21). *MET-iquette: Enabling virtual agents to have a social compliant behavior in the Metaverse*. 2022 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS) (pp. 394–401), Dijon, France. [CrossRef]

González-Briones, A., De La Prieta, F., Mohamad, M. S., Omatu, S., & Corchado, J. M. (2018). Multi-agent systems applications in energy optimization problems: A state-of-the-art review. *Energies*, *11*(8), 1928. [CrossRef]

Hazarika, A., & Rahmati, M. (2023). Towards an evolved immersive experience: Exploring 5G-and beyond-enabled ultra-low-latency communications for augmented and virtual reality. *Sensors*, *23*(7), 3682. [CrossRef] [PubMed]

Heikenfeld, J., Jajack, A., Rogers, J., Gutruf, P., Tian, L., Pan, T., Li, R., Khine, M., Kim, J., Wang, J., & Kim, J. (2018). Wearable sensors: Modalities, challenges, and prospects. *Lab on a Chip*, *18*(2), 217–248. [CrossRef] [PubMed]

Janbi, N., Katib, I., & Mehmood, R. (2023). Distributed artificial intelligence: Taxonomy, review, framework, and reference architecture. *Intelligent Systems with Applications*, *18*, 200231. [CrossRef]

Johnson-Glenberg, M. C., & Megowan-Romanowicz, C. (2017). Embodied science and mixed reality: How gesture and motion capture affect physics education. *Cognitive Research: Principles and Implications*, *2*(1), 24. [CrossRef] [PubMed]

Johnson-Glenberg, M. C., Yu, C. S. P., Liu, F., Amador, C., Bao, Y., Yu, S., & LiKamWa, R. (2023). Embodied mixed reality with passive haptics in STEM education: Randomized control study with chemistry titration. *Frontiers in Virtual Reality*, *4*, 1047833. [CrossRef]

Ju, Y., Shi, B., Chen, Y., Zhou, H., Dong, J., & Lam, K. M. (2024). GR-PSN: Learning to estimate surface normal and reconstruct photometric stereo images. *IEEE Transactions on Visualization and Computer Graphics*, *30*(9), 6192–6207. [CrossRef] [PubMed]

Kang, J., Chen, J., Xu, M., Xiong, Z., Jiao, Y., Han, L., Niyato, D., Tong, Y., & Xie, S. (2024). UAV-assisted dynamic avatar task migration for vehicular metaverse services: A multi-agent deep reinforcement learning approach. *IEEE/CAA Journal of Automatica Sinica*, *11*(2), 430–445. [CrossRef]

Kosmas, P., & Zaphiris, P. (2023). Improving students' learning performance through Technology-Enhanced Embodied Learning: A four-year investigation in classrooms. *Education and Information Technologies*, *28*(9), 11051–11074. [CrossRef]

Li, J., Wang, S., Zhang, M., Li, W., Lai, Y., Kang, X., Ma, W., & Liu, Y. (2024). Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv*, arXiv:2405.02957. [CrossRef]

Lin, L. (2015). Exploring collaborative learning: Theoretical and conceptual perspectives. In L. Lin (Ed.), *Investigating Chinese HE EFL classrooms* (pp. 11–28). Springer. [CrossRef]

Lindgren, R., Tscholl, M., Wang, S., & Johnson, E. (2016). Enhancing learning and engagement through embodied interaction within a mixed reality simulation. *Computers & Education*, *95*, 174–187. [CrossRef]

Liu, J., Zheng, Y., Wang, K., Bian, Y., Gai, W., & Gao, D. (2020). A real-time interactive tai chi learning system based on VR and motion capture technology. *Procedia Computer Science*, *174*, 712–719. [CrossRef]

Liu, Y., & Fu, Z. (2024). Hybrid intelligence: Design for sustainable multiverse via integrative cognitive creation model through human–computer collaboration. *Applied Sciences*, *14*(11), 4662. [CrossRef]

López-Belmonte, J., Pozo-Sánchez, S., Moreno-Guerrero, A.-J., & Lampropoulos, G. (2023). Metaverse in education: A systematic review. *Revista De Educación a Distancia (RED)*, *23*(73), 2252656. [CrossRef]

Manawadu, M., & Park, S. Y. (2024). 6DoF object pose and focal length estimation from single rgb images in uncontrolled environments. *Sensors*, *24*(17), 5474. [CrossRef] [PubMed]

Mantiuk, R. K., Hammou, D., & Hanji, P. (2023). HDR-VDP-3: A multi-metric for predicting image differences, quality and contrast distortions in high dynamic range and regular content. *arXiv*, arXiv:2304.13625. [CrossRef]

Mills, K. A., & Brown, A. (2023). Smart glasses for 3D multimodal composition. *Learning, Media and Technology*, *50*(2), 156–177. [CrossRef]

Mira, H. H., Chaker, R., Maria, I., & Nady, H. (2024). Review of research on the outcomes of embodied and collaborative learning in STEM in higher education with immersive technologies. *Journal of Computing in Higher Education*, 1–38. [CrossRef]

Nahavandi, D., Alizadehsani, R., Khosravi, A., & Acharya, U. R. (2022). Application of artificial intelligence in wearable devices: Opportunities and challenges. *Computer Methods and Programs in Biomedicine*, *213*, 106541. [CrossRef] [PubMed]

Nascimento, T. H., Fernandes, D., Vieira, G., Felix, J., Castro, M., & Soares, F. (2023, October 9–11). *MazeVR: Immersion and interaction using google cardboard and continuous gesture recognition on smartwatches*. 28th International ACM Conference on 3D Web Technology (pp. 1–5), San Sebastian, Spain. [CrossRef]

Palermo, F., Casciano, L., Demagh, L., Teliti, A., Antonello, N., Gervasoni, G., Shalby, H. H. Y., Paracchini, M. B., Mentasti, S., Quan, H., Santambrogio, R., Gilbert, C., Roveri, M., Matteucci, M., Marcon, M., & Trojaniello, D. (2025). Advancements in context recognition for edge devices and smart eyewear: Sensors and applications. *IEEE Access*, *13*, 57062–57100. [CrossRef]

Pan, A. (2024). *How wearables like apple vision pro and orion are transforming human interactions with interfaces—*. . .. Medium. Available online: https://medium.com/@alexanderpanboy/how-wearables-like-apple-vision-pro-and-orion-are-transforming-human-interactions-with-interfaces-95f3c390a77d (accessed on 10 December 2024).

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *arXiv*, arXiv:2304.03442. [CrossRef]

Phakamach, P., Senarith, P., & Wachirawongpaisarn, S. (2022). The metaverse in education: The future of immersive teaching & learning. *RICE Journal of Creative Entrepreneurship and Management*, *3*(2), 75–88. [CrossRef]

Polansky, M. G., Herrmann, C., Hur, J., Sun, D., Verbin, D., & Zickler, T. (2024). Boundary attention: Learning curves, corners, junctions and grouping. *arXiv*, arXiv:2401.00935. [CrossRef]

Qin, D., Leichner, C., Delakis, M., Fornoni, M., Luo, S., Yang, F., Wang, W., Banbury, C., Ye, C., Akin, B., Aggarwal, V., Zhu, T., Moro, D., & Howard, A. (2025). MobileNetV4: Universal models for the mobile ecosystem. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), *Computer vision—ECCV 2024* (Vol. 15098, pp. 78–96). Springer Nature. [CrossRef]

Reiniger, J. L., Domdei, N., Holz, F. G., & Harmening, W. M. (2021). Human gaze is systematically offset from the center of cone topography. *Current Biology*, *31*(18), 4188–4193.e3. [CrossRef] [PubMed]

Sahili, Z. A., & Awad, M. (2023). Spatio-temporal graph neural networks: A survey. *arXiv*, arXiv:2301.10569. [CrossRef]

Song, T., Tan, Y., Zhu, Z., Feng, Y., & Lee, Y. C. (2024). Multi-agents are social groups: Investigating social influence of multiple agents in human-agent interactions. *arXiv*, arXiv:2411.04578. [CrossRef]

Su, M., Guo, R., Wang, H., Wang, S., & Niu, P. (2017, July 18–20). *View frustum culling algorithm based on optimized scene management structure*. 2017 IEEE International Conference on Information and Automation (ICIA) (pp. 838–842), Macau, China. [CrossRef]

Sun, J. C. Y., Ye, S. L., Yu, S. J., & Chiu, T. K. F. (2023). Effects of wearable hybrid AR/VR learning material on high school students' situational interest, engagement, and learning performance: The case of a physics laboratory learning environment. *Journal of Science Education and Technology*, *32*(1), 1–12. [CrossRef]

Sun, Z., Zhu, M., Shan, X., & Lee, C. (2022). Augmented tactile-perception and haptic-feedback rings as human-machine interfaces aiming for immersive interactions. *Nature Communications*, *13*(1), 5224. [CrossRef] [PubMed]

Suo, J., Zhang, W., Gong, J., Yuan, X., Brady, D. J., & Dai, Q. (2023). Computational imaging and artificial intelligence: The next revolution of mobile vision. *Proceedings of the IEEE*, *111*(12), 1607–1639. [CrossRef]

Vrins, A., Pruss, E., Prinsen, J., Ceccato, C., & Alimardani, M. (2022). Are you paying attention? The effect of embodied interaction with an adaptive robot tutor on user engagement and learning performance. In F. Cavallo, J.-J. Cabibihan, L. Fiorini, A. Sorrentino, H. He, X. Liu, Y. Matsumoto, & S. S. Ge (Eds.), *Social robotics* (Vol. 13818, pp. 135–145). Springer Nature. [CrossRef]

Wang, X., Wang, Y., Yang, J., Jia, X., Li, L., Ding, W., & Wang, F. Y. (2024). The survey on multi-source data fusion in cyber-physical-social systems: Foundational infrastructure for industrial metaverses and industries 5.0. *Information Fusion*, *107*, 102321. [CrossRef]

Wang, Z., Bian, W., & Prisacariu, V. A. (2025). CrossScore: Towards multi-view image evaluation and scoring. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), *Computer vision—ECCV 2024* (Vol. 15067, pp. 492–510). Springer Nature. [CrossRef]

Wu, X., Chen, X., Zhao, J., & Xie, Y. (2024). Influences of design and knowledge type of interactive virtual museums on learning outcomes: An eye-tracking evidence-based study. *Education and Information Technologies*, *29*(6), 7223–7258. [CrossRef]

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., . . . Gui, T. (2023). The rise and potential of large language model based agents: A survey. *arXiv*, arXiv:2309.07864. [CrossRef]

Xia, Y., Shin, S.-Y., & Lee, H.-A. (2024). Adaptive learning in AI agents for the metaverse: The ALMAA framework. *Applied Sciences*, *14*(23), 11410. [CrossRef]

Yu, D. (2023, September 18–20). *AI-empowered metaverse learning simulation technology application*. 2023 International Conference on Intelligent Metaverse Technologies & Applications (iMETA) (pp. 1–6), Tartu, Estonia. [CrossRef]

Zhai, X., Xu, J., Chen, N. S., Shen, J., Li, Y., Wang, Y., Chu, X., & Zhu, Y. (2023). The syncretic effect of dual-source data on affective computing in online learning contexts: A perspective from convolutional neural network with attention mechanism. *Journal of Educational Computing Research*, *61*(2), 466–493. [CrossRef]

Zhai, X. S., Chu, X. Y., Chen, M., Shen, J., & Lou, F. L. (2023). Can edu-metaverse reshape virtual teaching community (VTC) to promote educational equity? An exploratory study. *IEEE Transactions on Learning Technologies*, *16*(6), 1130–1140. [CrossRef]

Zhao, G., Liu, S., Zhu, W. J., & Qi, Y. H. (2021). A lightweight mobile outdoor augmented reality method using deep learning and knowledge modeling for scene perception to improve learning experience. *International Journal of Human–Computer Interaction*, *37*(9), 884–901. [CrossRef]

Zhao, Z., Zhao, B., Ji, Z., & Liang, Z. (2022). On the personalized learning space in educational metaverse based on heart rate signal. *International Journal of Information and Communication Technology Education (IJICTE)*, *18*(2), 1–12. [CrossRef]

Zhou, X., Yang, Q., Zheng, X., Liang, W., Wang, K. I. K., Ma, J., Pan, Y., & Jin, Q. (2024). Personalized federated learning with model-contrastive learning for multi-modal user modeling in human-centric metaverse. *IEEE Journal on Selected Areas in Communications*, *42*(4), 817–831. [CrossRef]

Zhu, L., Li, Y., Sandström, E., Huang, S., Schindler, K., & Armeni, I. (2024). LoopSplat: Loop closure by registering 3D gaussian splats. *arXiv*, arXiv:2408.10154. [CrossRef]

education sciences

MDPI

Article

Evaluating an Artificial Intelligence (AI) Model Designed for Education to Identify Its Accuracy: Establishing the Need for Continuous AI Model Updates

Navdeep Verma *, Seyum Getenet, Christopher Dann and Thanveer Shaik

School of Education, University of Southern Queensland, Springfield, QLD 4300, Australia; seyum.getenet@unisq.edu.au (S.G.); chris.dann@unisq.edu.au (C.D.); thanveer.shaik@unisq.edu.au (T.S.)
* Correspondence: navdeepverma2032@gmail.com

Abstract: The growing popularity of online learning brings with it inherent challenges that must be addressed, particularly in enhancing teaching effectiveness. Artificial intelligence (AI) offers potential solutions by identifying learning gaps and providing targeted improvements. However, to ensure their reliability and effectiveness in educational contexts, AI models must be rigorously evaluated. This study aimed to evaluate the performance and reliability of an AI model designed to identify the characteristics and indicators of engaging teaching videos. The research employed a design-based approach, incorporating statistical analysis to evaluate the AI model's accuracy by comparing its assessments with expert evaluations of teaching videos. Multiple metrics were employed, including Cohen's Kappa, Bland–Altman analysis, the Intraclass Correlation Coefficient (ICC), and Pearson/Spearman correlation coefficients, to compare the AI model's results with those of the experts. The findings indicated low agreement between the AI model's assessments and those of the experts. Cohen's Kappa values were low, suggesting minimal categorical agreement. Bland–Altman analysis showed moderate variability with substantial differences in results, and both Pearson and Spearman correlations revealed weak relationships, with values close to zero. The ICC indicated moderate reliability in quantitative measurements. Overall, these results suggest that the AI model requires continuous updates to improve its accuracy and effectiveness. Future work should focus on expanding the dataset and utilise continual learning methods to enhance the model's ability to learn from new data and improve its performance over time.

Keywords: AI; video conferencing; online student engagement; teachers' behaviours; teachers' movements; design-based research

1. Introduction

Over the past decade, there has been substantial growth in online education within higher education institutions. This growth is due to its flexibility, accessibility, and cost efficiency (Castro & Tumibay, 2021; Dhawan, 2020). Further, COVID-19 has compelled higher education institutes worldwide to transition to online learning (Xie et al., 2021). Due to this sudden change, teachers encounter notable challenges in adapting to online learning, with student engagement emerging as the most prominent challenge (Alenezi et al., 2022). Studies have highlighted that fostering online student engagement is more complex than engaging students in traditional face-to-face learning (Gillett-Swan, 2017; Hew, 2016). The potential of online learning and its trends brings forth new opportunities but also poses various challenges (Liang & Chen, 2012).

Incorporating AI can assist in addressing these challenges by identifying and evaluating discrepancies and offering suggestions for enhancing teaching effectiveness. AI opens up new avenues for learning and teaching (Limna et al., 2022). AI technologies' abilities to quickly analyse large datasets, recognise patterns, and make predictions support more personalised and effective learning experiences (Harry & Sayudin, 2023; Shaikh et al., 2022; Tahiru, 2021). For instance, AI-powered systems can recommend personalised learning paths, automate grading, and enhance educational resources (Nguyen, 2023). However, a critical challenge lies in evaluating the accuracy of AI models, especially when they are tasked with assessing complex human behaviours and movements, such as those of teachers, aimed at encouraging student engagement. Despite its potential, there is still much to learn about how accurately AI can interpret and predict the behaviours that enhance student engagement in online learning environments.

This study employed design-based research (DBR) to address these gaps by designing an AI model to identify engagement-enhancing teacher behaviours and movements during video conferences. During the initial phase of this DBR, the authors conducted a systematic literature review to determine the characteristics and indicators of engaging teaching videos Verma et al. (2023b). In the second phase, the authors, with the assistance of an AI expert, trained an AI model to replace the manual annotation of teaching videos based on teachers' behaviours and movements (Verma et al., 2023a), which expedites the process as manual annotation was identified as time-consuming (Beaver & Mueen, 2022). The identified characteristics and indicators were then applied to train the AI model using deep learning as an AI methodology. The current phase focuses on evaluating the AI model to ensure its accuracy and determine whether continuous AI model updates are necessary. Specifically, this study seeks to address the following research questions:

> "How accurately can an AI model generate a report for characteristics and indicators of engaging teaching videos based on teachers' behaviours and movements?" (RQ1)

> "Why is it important to continuously update the AI model designed to enhance online learning and teaching?" (RQ2)

By addressing these questions, this research aims to contribute to the ongoing effort to accurately and sustainably integrate AI into online learning.

2. Background

This section consists of three subsections. Section 2.1 presents the three distinct phases of the DBR, with a special focus on the current phase. Section 2.2 explores existing studies on evaluation methods in the field of education. Finally, Section 2.3 delves into studies that discuss evaluation methods within AI. Each section provides valuable insights and analysis into these important topics, highlighting their significance and implications in their respective domains.

2.1. Previous Phases

This study is the third phase of a DBR where the authors evaluate an AI model to ensure its accuracy and to determine whether continuous model updates are necessary. In the first phase, the authors conducted a systematic literature review to identify the characteristics and indicators of engaging teaching videos. The authors reviewed 34 studies and identified 11 characteristics crucial for enhancing student engagement in video conferencing based on teachers' behaviours and movements Verma et al. (2023b). Further, 47 indicators that can describe each characteristic were identified. The identification and categorisation of these indicators into the 11 main characteristics are backed by the significant findings from the reviewed studies and research concerning online student

engagement. These characteristics were organised into three overarching domains: Teachers' behaviours, movements, and use of technology Verma et al. (2023b). Appendix A.1 illustrates the main theme, characteristics, and indicators of engaging teaching videos.

Researchers have demonstrated significant interest in examining the influence of teachers' behaviours and movements on online student engagement (Cents-Boonstra et al., 2021; J. Ma et al., 2015). Verma et al. (2023b) strongly believe that the characteristics and indicators outlined in Appendix A.1 can be used as a benchmark for improving teachers' performance in online learning. Educational institutions can implement these indicators and characteristics of engaging teaching videos to enhance and regulate online teaching practices. Educational institutions worldwide can use this information to develop and offer training for teachers aimed at refining their skills in creating teaching videos that effectively boost online student engagement. However, identifying these engaging characteristics and indicators within recorded lecture videos requires human participation (Verma et al., 2023a). This manual identification and analysis process demands a significant amount of time and resources (Beaver & Mueen, 2022). Additionally, this approach may introduce human bias into the analysis. Therefore, in order to mitigate human bias and maintain efficiency in identifying engaging teaching videos, the authors collaborated with an AI expert to develop an AI model in phase 2. This tool generates a report on the characteristics and indicators of engaging teaching videos (Verma et al., 2023a).

In the second phase, the educational experts annotated 25 recorded lecture videos. The recorded lecture videos were presented to higher education students by lecturers from a university in Australia. The videos encompass a range of fields, including law, business, health, education, arts, and sciences, with an average length of 01:28:37 (Verma et al., 2023a). There were 13 female and 12 male speakers featured in the videos, and the authors secured ethical approval from the local university under the ethics approval number H20REA185. The manual annotation of these videos was performed individually using the Visual Geometry Group (VGG) Image Annotator (VIA) (Version 3) tool accessible from https://www.robots.ox.ac.uk/~vgg/software/via/app/via_video_annotator.html (accessed on 11 January 2024). The manual annotation was carried out at the indicator level. Through the manual annotation of 25 recorded lecture videos, the authors identified 7 characteristics and 15 descriptive indicators, as detailed in Table 1. Based on the outcomes of this manual annotation, the AI expert assisted the authors during the development and training of an AI model designed to identify the characteristics and indicators of engaging teaching videos each time a video is processed.

The engaging characteristics and indicators identified through manual video annotation were utilised to train prototype 1. Recognising challenges like misleading metrics and class imbalance, the model underwent refinement in prototype 2 by implementing the oversampling technique. By implementing the oversampling technique, the model was further improvised and demonstrated promising results, achieving an average precision, recall, F1-score, and balanced accuracy of 68%, 75%, 73%, and 79%, respectively, in categorising the annotated videos at the indicator level (Verma et al., 2023a).

The developed model has the potential to support higher education institutions in establishing moderation in lecture delivery. Moreover, it can significantly influence teaching and learning by providing teachers with reports on their technology utilisation effectiveness and identifying engagement-enhancing behaviours and movements present or lacking during their lecture delivery. To ensure the AI model's effectiveness and accuracy in generating reports, the current study evaluates its performance using a range of metrics.

Table 1. Characteristics and indicators identified in manual annotation (Verma et al., 2023a, p. 7).

Characteristics	Indicators
Encouraging Active Participation	• Encouraging students' participation in discussion • Encouraging students to share their knowledge and ideas • Encouraging students to ask questions • Encouraging collaborative learning activities • Encouraging meaningful interaction
Establishing Teacher Presence	• Providing learning resources • Giving clear instructions
Establishing Clear Expectations	• Outlining the learning objectives
Demonstrating Empathy	• Using appropriate changes in tone of voice
Using Nonverbal Cues	• Facial expressions • Eye contact • Appropriate body language
Using Technology Effectively	• Enabling class recording for later review • Screen sharing and enabling chat, camera, and microphone • Varying the presentation media

2.2. Evaluation Methods in Education

Researchers have used various evaluation methods to evaluate the available instruments for measuring student engagement in education (Apicella et al., 2022; Giang et al., 2022; Shekhar et al., 2018).

Giang et al. (2022) conducted a validation of their proposed model to measure student engagement, which includes four sub-components, emotional engagement, cognitive engagement, participatory engagement, and agentic engagement, by employing a qualitative analysis approach, conducting interviews and focus group sessions as part of their data collection process. An interview in research is a data collection method where a researcher asks participants questions to gather information about their experiences, opinions, and perspectives (Kvale, 1996). Frequently, interviews are combined with other data collection methods to ensure a comprehensive and diverse range of information for analysis purposes (Turner, 2010).

In their recent study, Apicella et al. (2022) carried out an experimental case study to verify the effectiveness and validity of the tool they introduced to assess and monitor student engagement. A case study is commonly defined as a thorough and methodical examination of an individual, group, community, or another entity where the researcher carefully analyses detailed information about various factors or variables (Heale & Twycross, 2018).

Shekhar et al. (2018) employed a mixed-methods approach, combining quantitative and qualitative methods to assess the effectiveness and validity of the instruments they developed for observing active learning, instructor participation, student resistance, and student engagement. This combination of methods allowed for the validation of broader frameworks through qualitative analysis and the identification of specific elements to incorporate into quantitative tools during the developmental stage, as Sandelowski (2000) suggested.

Chiu (2021) applied questionnaires in their study and adopted a quantitative analysis method to evaluate the model they provided, where they leveraged digital tools to fulfil the requirements of competence, relatedness, and autonomy, leading to active student engagement in online learning. A questionnaire serves as a methodical approach for gathering primary quantitative data in the literature. It typically consists of a sequence of written inquiries to which respondents are required to provide responses (Bell, 1999).

Lee et al. (2019) incorporated expert opinions and conducted reliability and validity analyses to ensure the accuracy and consistency of the model they proposed to enhance student engagement in e-learning environments. Expert opinion refers to a judgment by an individual with superior knowledge in a specific domain. It encompasses two key components: expertise and domain specificity (Pingenot & Shanteau, 2009).

2.3. Evaluation Methods in AI

Several studies have explored using deep learning and computer vision techniques to evaluate AI-enabled tools that identify engagement-enhancing teacher behaviours and movements in video conferencing.

X. Ma et al. (2021) presented a deep learning-based approach to recognise online student engagement, employing both convolutional and recurrent neural networks. They analysed facial expressions, body movements, and gaze patterns to predict engagement levels.

Behera et al. (2020) focused on automatically analysing teachers' nonverbal behaviours in online learning settings. They employed computer vision techniques such as face detection, tracking, gesture recognition, and body pose estimation to extract meaningful features from video data. AI algorithms were applied to classify nonverbal behaviours and assess their impact on student engagement. In their research, Weng et al. (2023) conducted a systematic literature review on video-based learning analytics in online education. The review highlighted the importance of utilising computer vision techniques to analyse teachers' behaviours and their influence on online student engagement and learning outcomes. Ashwin and Guddeti (2019) explored the utilisation of deep learning techniques for automatic emotion recognition in educational videos. They used convolutional neural networks and recurrent neural networks to analyse teachers' and students' facial expressions and body movements, demonstrating the potential of deep learning models in capturing emotional cues and evaluating their impact on student engagement.

A handful of studies (Ashwin & Guddeti, 2019; Behera et al., 2020; X. Ma et al., 2021; Weng et al., 2023) highlight the use of deep learning and computer vision techniques in evaluating AI-enabled tools for identifying engagement-enhancing teacher behaviours and movements in video conferencing. They offer valuable perspectives on the capacity of these techniques to enhance student engagement and improve the quality of online learning experiences.

Existing research in education lacks evaluation methods specifically designed for measuring online student engagement using AI-enabled tools (Huang et al., 2023). Previous studies have focused on developing instruments and models for traditional face-to-face settings, utilising methods such as interviews, case studies, mixed-methods approaches, and questionnaires. The evaluation methods used to validate the instruments in education might not be suitable for the AI model created by the authors as these methods require human analysis, which can lead to bias (Heeg & Avraamidou, 2023).

This paper seeks to evaluate the AI model developed in the preceding phase through the use of various metrics such as Cohen's Kappa, Bland–Altman analysis, the Intraclass Correlation Coefficient (ICC), and Pearson/Spearman correlation coefficients to assess its accuracy and identify whether it is necessary to perform continuous AI model updates.

3. Methods

The authors utilised a DBR approach to develop an AI model that generates reports on teachers' behaviours and movements whenever it processes a recorded lecture video. The DBR methodology has gained recognition in educational research, with many researchers highlighting its ability to support the development of practical research processes (Tinoca et al., 2022). Following the principles of the DBR methodology, this study has unfolded in three distinct phases. The phases of the DBR process are summarised in Figure 1.

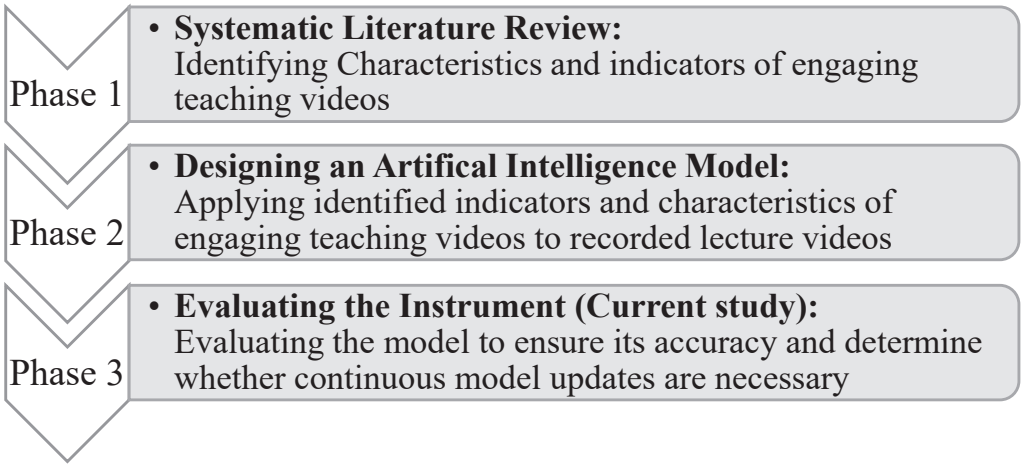


Figure 1. Research phases.

Phase 1, systematic literature review: This phase involves a systematic review of the existing literature to identify the characteristics and indicators of engaging teaching videos. By analysing previous research, a foundational understanding of what constitutes effective teacher behaviours and movements in online teaching environments is established. In this study, the authors identified 47 indicators and 11 characteristics categorised into three main themes (see Appendix A). These identified indicators then guided the development of the AI model in subsequent phases.

Phase 2, designing an AI model, involves video annotation to create an AI model capable of analysing the characteristics and indicators identified in Phase I, to recognise and evaluate teachers' engagement-enhancing behaviours and movements in recorded lecture videos using Zoom. The model was designed through two prototypes.

AI process

The authors developed a deep learning model to learn a teacher's movements in a recording with the support of an AI expert. This is achieved by recording the temporal coordinates extracted from the tool's manual video annotation. Temporal coordinates are markers in the video timeline that help identify specific points in time. Selected lecture videos were split based on these coordinates, and we transformed them into a stack of image frames. The pre-processed frames were then labelled with corresponding teaching indicators, and we prepared the data model for training. Next, the data were split into two sets—training and testing—for model training and evaluation. An AI expert fed the training set to the convolutional neural network (CNN) model to learn the actions in the image frames and their corresponding labels. Finally, the test set was used to evaluate the performance of the CNN model.

Data pre-processing

During the data pre-processing step, the AI expert captured the temporal coordinates provided by the video annotation tool. For example, suppose a lecture recording displayed

the teaching indicator "Clear and concise explanation of information" at the temporal coordinates (3051.315, 3053.256). In that case, the recorded lecture was divided into video segments highlighting and extracting the teaching indicator. Then, each video was split into segments of image frames and annotated each frame with the "Clear and concise explanation of information" teaching indicator. These annotated image frames are represented as 2D matrices and serve as inputs for the convolution layer of the deep learning model, as described in the subsequent subsection.

Deep learning model

The AI expert developed the CNN model as a deep learning approach for classifying two-dimensional (2D) data images. The CNN model offers the advantage of reducing the high dimensionality of images while preserving their information. Figure A2 illustrates the learning process of the CNN model. First, the input image frames, pre-processed in the previous step, are passed to a two-dimensional (2D) convolution layer, which uses a set of filters to divide the image frame into smaller sub-images and analyse them individually. The convolution layer's output is then passed to the pooling layer, which estimates the maximum value for a feature set and creates a down-sampled group feature. The pooled features can be flattened into a 2D array and then processed in the output layer of the CNN model. The output layer provides a probability for each label classification, which can be optimised using a threshold value to classify the features into a label.

As shown in Figure 1, the present study, Phase 3, focuses on the third phase of this DBR, where authors have evaluated the AI model to ensure its accuracy and determine whether continuous updates are required. The authors have used multiple statistical methods to ensure the model's accuracy. As part of the evaluation process, the model processed two recorded lecture videos and then generated results, identifying indicators of engaging teaching videos. Meanwhile, human experts who are well-versed in the domain independently analysed the same set of videos and provided their findings. The AI model was evaluated using multiple statistical methods to identify the statistical agreement and consistency between the findings of an AI model and two human experts in evaluating specific segments of video data.

3.1. Data Collection

The evaluation of the AI model's ability to identify engagement-enhancing teacher behaviours and movements in video conferencing involved the participation of two human experts who manually annotated two videos and the AI-generated reports. The results obtained from the AI model and the two human experts were carefully analysed using various metrics.

Two videos of varying durations were utilised, one lasting 49 min and 3 s with 11 segments and the other lasting 58 min and 40 s with 23 segments, featuring presenters with different camera settings. The research was carried out with ethical clearance obtained from a regional university in Australia (ethics approval number H20REA185). However, demographic information about the lecturers, such as age, location, and academic background, was not collected.

3.2. Video Analysis

This section explores two distinct approaches for processing and analysing a set of videos to identify teachers' engagement-enhancing behaviours and movements. It highlights the annotation process carried out by human experts and the use of an AI model designed by the authors in the previous phase to achieve a similar objective.

3.2.1. Expert Involvement

The two human experts conducted an annotation process guided by the 7 characteristics and 15 descriptive indicators of engaging teaching videos identified in the previous phase (refer to Table 1). Having two experts for comparison brings in diverse perspectives and broader insights and potentially leads to more comprehensive solutions or decisions. Additionally, it reduces the chances of individual bias influencing the outcomes, leading to a more balanced and reliable evaluation. To complete the manual annotation process, the Visual Geometry Group Image Annotator (VIA) tool was used (refer to Appendix A.2).

3.2.2. AI Reports

The AI model employed a deep learning model known as a convolutional neural network (CNN) to process the same set of recorded lecture videos. Its main goal was to identify the teachers' engagement-enhancing behaviours and movements based on the characteristics and indicators it had been trained with, similar to what the human experts utilised for manual annotation. By examining visual cues and patterns, the model generated detailed reports highlighting the teachers' behaviours and movements that enhance student engagement.

3.3. Data Analysis

The analysis involved multiple statistical methods to assess the agreement and consistency between the findings of an AI tool and two human experts in evaluating specific segments of video data. Cohen's Kappa was used to measure the inter-rater agreement for categorical items, considering the possibility of agreement occurring by chance. To analyse the differences between their assessments, Bland–Altman analysis was employed to explore the agreement between the AI tool and the experts. The Intraclass Correlation Coefficient (ICC) was calculated to assess the reliability and agreement of the quantitative measurements between the AI tool and both experts. Lastly, the Pearson and Spearman correlation coefficients were computed to measure the linear and rank-order relationships between the AI tool's assessments and those of the experts.

4. Results

Tables 2 and 3 serve as invaluable resources, offering a clear outline of the analyses conducted on each video and facilitating a deeper understanding of the comparative evaluations undertaken by both human experts and the AI model. Table 4 presents the statistical agreement and consistency analysis between the AI model and experts evaluating video 1 and video 2 data. The combined analysis results are discussed in detail, pointing out the findings for each statistical method used.

Table 2. AI and experts' findings from video 1.

Video 1	AI Model	Expert 1	Expert 2
Segment 0	1	1	14
Segment 1	6	8	6
Segment 2	6	8	6
Segment 3	14	8	14
Segment 4	1	14	8
Segment 5	15	7	15
Segment 6	7	7	No identified indicator
Segment 7	5	9	No identified indicator
Segment 8	2	8	No identified indicator
Segment 9	5	9	No identified indicator
Segment 10	9	9	No identified indicator
Segment 11	5	9	No identified indicator

Table 3. AI and experts' findings from video 2.

Video 2	AI Model	Expert 1	Expert 2
Segment 0	1	14	15
Segment 1	10	8	15
Segment 2	5	7	5
Segment 3	5	4	5
Segment 4	1	7	2
Segment 5	12	12	4
Segment 6	5	7	2
Segment 7	10	12	10
Segment 8	5	7	7
Segment 9	7	12	7
Segment 10	1	12	1
Segment 11	1	12	No identified indicator
Segment 12	5	9	No identified indicator
Segment 13	1	12	No identified indicator
Segment 14	1	12	No identified indicator
Segment 15	9	7	No identified indicator
Segment 16	5	7	No identified indicator
Segment 17	14	15	No identified indicator
Segment 18	5	12	No identified indicator
Segment 19	14	12	No identified indicator
Segment 20	1	9	No identified indicator
Segment 21	14	15	No identified indicator
Segment 22	1	1	No identified indicator
Segment 23	5	12	No identified indicator

Table 4. Statistical agreement and consistency analysis between the AI tool and experts.

Statistical Measure	AI Tool vs. Expert 1	AI Tool vs. Expert 2	Interpretation
Cohen's Kappa	0.09	0.07	Slight agreement
Bland–Altman Analysis			
-Mean Difference	4.92	2.24	Moderate variability in differences
-Standard Deviation of Differences	4.55	6.18	
-95% Limits of Agreement	(−4.00, 13.84)	(−9.87, 14.35)	
Intraclass Correlation Coefficient (ICC2k)	0.45	0.45	Moderate reliability
Pearson Correlation Coefficient	0.09	−0.02	Weak linear relationship
Spearman Correlation Coefficient	0.09	−0.10	Weak rank-order relationship

4.1. Explanation of Findings

This section analyses the findings for the two distinct videos at each level. Tables (List the tables) showcase the outcomes of AI processing and expert analysis, forming the foundation for further exploration and discussion.

4.1.1. Video 1 Results

Table 2 highlights video 1 segments (0 to 11) and the results obtained from the AI model and Experts 1 and 2.

The findings from video 1, as analysed by both the AI model and experts, are organised into four columns. The first column displays the video segments. The second column lists the indicators identified by the AI model. The third column presents the indicators identified by Expert 1, while the fourth column outlines the indicators identified by Expert 2. (Refer to Figure A4 in Appendix A.2 for the complete list of indicators.)

4.1.2. Video 2 Results

Further, Table 3 presents video 2 segments (0 to 23) and the results from the AI model and Experts 1 and 2.

The findings from video 2 follow the same format, with four columns. The first column displays the video segments, the second contains the indicators identified by the AI model, the third presents the indicators identified by Expert 1, and the fourth outlines those identified by Expert 2. (Refer to Figure A4 in Appendix A.2 for the complete list of indicators.)

Table 4 summarises the result of the statistical agreement and consistency analysis between the AI model and expert findings, followed by a detailed explanation of the results.

In this study, multiple statistical methods were employed to assess the agreement and consistency between the findings of an AI model and two human experts in evaluating specific segments of video data. The analysis involved the calculation of Cohen's Kappa, Bland–Altman analysis, the Intraclass Correlation Coefficient (ICC), and Pearson/Spearman correlation coefficients to comprehensively explore the degree of similarity between the AI-generated results and the expert assessments.

Cohen's Kappa was used to measure the inter-rater agreement for categorical items, taking into account the possibility of agreement occurring by chance. The results indicated slight agreement between the AI model and the experts, with Cohen's Kappa values of 0.09 for Expert 1 and 0.07 for Expert 2. These low Kappa values suggest that the AI model's categorical assessments are only marginally aligned with those of the human experts, with a considerable amount of disagreement present.

When analysing the differences between their assessments, Bland–Altman analysis was employed to explore the agreement between the AI model and the experts. For the comparison between the AI model and Expert 1, the mean difference was 4.92, with a standard deviation of 4.55. The 95% limits of agreement ranged from -4.00 to 13.84. Similarly, the comparison with Expert 2 yielded a mean difference of 2.24, with a standard deviation of 6.18 and 95% limits of agreement from -9.87 to 14.35. These results reveal a moderate degree of variability in the differences between the AI model and the experts, indicating that while there is some level of agreement, the variability is substantial enough to warrant further refinement of the AI model.

The Intraclass Correlation Coefficient (ICC) was calculated to assess the reliability and agreement of the quantitative measurements between the AI model and both experts. The ICC value (ICC2k) for the comparison was 0.45, indicating moderate reliability. This suggests that while there is some consistency in the measurements between the AI model and the experts, the level of agreement is not strong enough to be considered highly reliable.

Finally, the Pearson and Spearman correlation coefficients were computed to measure the linear and rank-order relationships between the AI model's assessments and those of the experts. The Pearson correlation coefficient for the AI model and Expert 1 was 0.09, indicating a weak positive linear relationship, while the correlation with Expert 2 was -0.02, reflecting a weak negative linear relationship. Similarly, the Spearman correlation coefficients showed a weak positive rank-order correlation of 0.09 with Expert 1 and a weak negative rank-order correlation of -0.10 with Expert 2. These results suggest that the AI model's findings have a minimal linear or monotonic relationship with the expert assessments.

The statistical analyses reveal that the AI model's assessments exhibit slight to moderate agreement and consistency with those of the human experts. While there is some level of alignment, the relatively low agreement metrics indicate that there is significant room for improvement in the AI model's performance. Enhancing the AI model, perhaps through additional training with a more diverse dataset or by refining its algorithms, could

potentially increase its reliability and consistency with expert evaluations. This would be crucial for ensuring the AI tool's effectiveness and accuracy in real-world applications.

5. Discussion

Researchers (e.g., Apicella et al., 2022; Giang et al., 2022; Shekhar et al., 2018) have developed various evaluation methods such as interviews, case studies, mixed-methods approaches, and questionnaires to validate instruments and ensure their effectiveness in education. However, existing research in education lacks evaluation methods specifically designed for measuring online student engagement using AI models (Heeg & Avraamidou, 2023; Huang et al., 2023). Therefore, the authors employed multiple statistical methods to measure the developed AI model's accuracy and identify whether it requires continuous model updates.

5.1. Exploration of Research Findings

Upon evaluating the model trained in 2022 by annotating 25 recorded lecture videos by education experts, the results revealed that the model requires updating. This is mainly due to the significant increase in expert knowledge concerning human characteristics over the past two years, while the model's knowledge has not changed. Further, research in this field indicates that AI models require regular updates to maintain their effectiveness (Li et al., 2023; Murtaza et al., 2022; Ocaña & Opdahl, 2023; Roshanaei et al., 2024).

In relation to the RQ1: How accurately can an AI model generate a report on the characteristics and indicators of engaging teaching videos based on teachers' behaviours and movements? The findings revealed that the AI model's ability to identify the characteristics and indicators of engaging teaching videos was only marginally aligned with expert analyses. The main reason for these results is the evolving nature of the human mind. From the development of the model to its evaluation, the experts' understanding has evolved significantly, enabling them to recognise more characteristics and indicators from the recorded lecture sessions, while the knowledge embedded in the AI model remains static. If the AI model was trained on more data, such as more videos manually annotated by experts, these results would likely reflect a stronger alignment between the experts' assessments and the AI model, indicating a significant improvement in the model's performance and accuracy. This overall result was drawn from multiple statistical methods, including Cohen's Kappa, Bland–Altman analysis, the ICC, and Pearson and Spearman correlation coefficients, which indicated limited agreement between the AI model and the human experts. Specifically, Cohen's Kappa values were low at 0.09 for Expert 1 and 0.07 for Expert 2, suggesting minimal alignment with expert findings. Bland–Altman analysis showed a mean difference of 4.92 (SD = 4.55) for Expert 1 and 2.24 (SD = 6.18) for Expert 2, with 95% limits of agreement ranging from -4.00 to 13.84 and -9.87 to 14.35, respectively, demonstrating moderate variability in differences. The ICC value (ICC2k) of 0.45 indicated moderate reliability, while Pearson and Spearman correlation coefficients revealed weak relationships: 0.09 with Expert 1 and -0.02 with Expert 2 for Pearson, and 0.09 and -0.10 for Spearman, respectively. These findings highlight significant room for improvement in the AI model's performance, suggesting that a further update is needed to enhance its accuracy and consistency with expert evaluations.

In relation to the RQ2: Why is it important to continuously update the AI model designed to enhance online learning and teaching? The evaluation findings indicate only a slight to moderate alignment of the AI model's performance outcome with the experts' analysis results, emphasising the need for further improvement through continuous model updates. Apart from the findings of this study, various factors support the importance of continuously updating AI models. AI models are trained and rely on historical data, which

may become outdated as the data environment evolves. Such changes can significantly impact the AI model's performance, making regular updates necessary to keep the model's performance from declining (Li et al., 2023). Roshanaei et al. (2024) describe regular updates and patches for AI models as the process of refreshing them to address any weaknesses in their design or data handling processes. AI models need to be regularly updated to keep up with new information (Ocaña & Opdahl, 2023). Pianykh et al. (2020) recommend the incorporation of feedback from match results and adjusting algorithms as part of the continuous training and updating of AI models to improve their predictive accuracy over time. Further, model updates can be influenced by other factors such as the availability of new or higher-quality training data, user feedback, learning algorithm advancements, and the need to ensure fairness in the model (X. Wang & Yin, 2023). Murtaza et al. (2022) highlight that continuously updating AI learning models with new training data can enhance the learning experience. Therefore, keeping models up to date ensures that AI models can continuously offer relevant, effective, and fair support in online learning environments.

5.2. Implications

This study holds significant implications for the use of AI models in education. Firstly, this three-phase research project provides the characteristics and indicators of engaging teaching videos that can improve online student engagement. These characteristics and indicators can help teachers and educational institutions enhance their pedagogical approaches.

Secondly, this study provides a procedure to train AI models for education. Further, by creating an AI model in phase 2, this research proves that AI can be used to create models and tools to replace the manual identification process. This can avoid challenges such as time consumption, cost, and potential human bias. According to De Silva et al. (2024), one of the multifaceted benefits of AI is its ability to automate processes, leading to increased efficiency in terms of both time and cost.

Thirdly, this study highlights the importance of model monitoring and validation. Monitoring and validating AI systems to ensure accuracy and fairness are crucial. Aldoseri et al. (2023) highlighted that inaccurate, biased, or irrelevant outcomes derived from low-quality data can have adverse effects on decision-making processes grounded in AI outputs, emphasising the importance of validation to enable AI systems to generate dependable and valuable outcomes. Thus, this study employed various metrics to guarantee the reliability of the evaluation results for the developed AI model, assessing its accuracy and identifying the importance of continuous AI model updates. This establishes the need for a policy that requires educational institutions to regularly enhance and update AI models to maintain accuracy and reliability and ensure the models remain relevant.

Moreover, if the AI model accurately identifies these characteristics and indicators of engaging teaching videos effectively, it can provide teachers with significant support in various aspects, such as saving time, enhancing learning, and reinforcing professional development. Regarding professional growth and continual improvement, AI-generated reports are instrumental in aiding teachers in recognising both the strong points and areas needing improvement in their lecture delivery concerning engagement. Similarly, processing engaging recorded lecture videos using the AI model provides teachers with valuable insights into what resonates most effectively with their students. This empowers them to make well-informed decisions for future learning experiences, ultimately resulting in improved teaching and learning outcomes. Further, this research also provides a manual annotation procedure that can assist AI engineers in developing similar AI models.

6. Limitations and Future Directions

While the authors have developed an AI model to understand student engagement based on teachers' behaviours and movements in video conferencing, certain limitations must be recognised. Firstly, significant differences in outcomes have been identified, attributed to factors such as human bias, evolution, and the limited training of the AI model due to a small dataset containing few indicators and variations. These factors underscore the need to enhance the AI model's performance to better align with the analyses conducted by human experts. Additionally, the reliance on a small dataset for evaluation emphasises the need for assessments on larger datasets by processing and analysing more lecture videos to comprehensively evaluate the model's performance.

In future research, the findings from this final phase may be incorporated for improvement. The results reveal that the AI model developed in this study to identify engagement-enhancing behaviours and movements needs continuous updates to address the challenges posed by evolving data. This study also establishes the importance of continuous model updates. As noted by Žliobaite et al. (2015) and Roshanaei et al. (2024), the performance of predictive models can degrade if they lack mechanisms for regular updates and adaptation to new data, highlighting the importance of continuous updates in preventing such vulnerability in AI models. In their study, C. Wang et al. (2024) suggested various triggers to perform model updates. Firstly, they introduced periodic updates, in which model updates are performed at intervals such as quarterly, monthly, or weekly. Secondly, they suggested performance-driven updates, where models are refreshed when their accuracy metrics fall below a predefined threshold. Lastly, they suggested a data-driven approach, where models are updated upon accumulating significant data. Another recommended approach is continual learning (CL), which enables AI models to be updated with new data without the need to retrain them from the beginning (Nikoloutsopoulos et al., 2024). Continual learning refers to an AI model's ability to continuously learn from new data streams while retaining its previous knowledge. In this process, the model improves its performance by adapting to new data and updating its knowledge base as new information becomes available.

7. Conclusions

As detailed in the explanation of findings, the AI model evaluation involved various statistical methods used to perform a statistical agreement and consistency analysis, comparing the AI model's findings with those of human experts. The results showed relatively low agreement between the AI model's ability to identify the characteristics and indicators of engaging teaching videos and the experts' analysis. While the AI model shows potential, the results highlight significant room for improvement, suggesting further updates are needed to improve the model's accuracy and achieve strong to excellent alignment with expert evaluations.

Author Contributions: N.V.: Conceptualization, Methodology, Formal Analysis, Writing—Original Draft and Review and Editing. S.G.: Conceptualization, Writing—Original Draft and Review and Editing. C.D.: Conceptualization, Writing—Original Draft and Review and Editing. T.S.: AI Methodology, Formal Analysis, Review and Editing. 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 research obtained ethics approval from the local university under the ethics approval number H20REA185, approval date 19 February 2021.

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

Data Availability Statement: Please contact the authors for a data request.

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

Abbreviations

Abbreviation	Definition
AI	Artificial Intelligence
CNN	Convoluted Neural Network
COVID-19	Coronavirus Disease 2019
DBR	Design-based Research
VIA	VGG Image Annotator

Appendix A

Appendix A.1

Main theme, characteristics, and indicators of engaging teaching videos (Verma et al., 2023a, p. 11).

Main Theme	Characteristics	Indicators
Teachers' Behaviours	Encouraging Active Participation	• Encouraging students' participation in discussion • Encouraging students to share their knowledge and ideas • Encouraging students to ask questions • Encouraging collaborative learning activities • Encouraging meaningful interaction • Encouraging students to turn on their webcams
	Establishing Teacher Presence	• Clear and concise explanations of information • Recognising and considering learners' individual differences • Using an appropriate style of presentation • Allowing sufficient time for students' information processing • Providing learning resources • Giving clear instructions • Using a range of teaching strategies • Appropriate speed of lecture delivery
	Establishing Social Presence	• Maintaining constant teacher–student interaction • Encouraging student–student interaction (peer collaboration) • Active and constructive communication • Taking on multiple roles
	Establishing Cognitive Presence	• Giving students a sense of puzzlement (trigger) • Providing opportunities for students to reflect (exploration) • Leading students to think and learn through discussion with others (integration) • Helping students apply knowledge to solve issues (resolution)
	Questions and Feedback	• Addressing students' questions and providing prompt feedback • Asking for questions and feedback • Clarifying misunderstanding
	Displaying Enthusiasm	• Motivating students • Displaying positive emotion

Main Theme	Characteristics	Indicators
	Establishing Clear Expectations	• Outlining the learning objectives • Outlining teachers' expectations of students' behaviours and responsibilities
	Demonstrating Empathy	• Using appropriate changes in tone of voice • Ensuring the learning environment is a respectful, safe, and supportive one • Showing concern
	Demonstrating Professionalism	• Demonstrating in-depth and up-to-date knowledge • Displaying appropriate behaviours
Teachers' Movements	Using Nonverbal Cues	• Facial expressions • Gestures • Eye gazes • Silence • Eye contact • Physical proximity • Appropriate body language
Use of Technology	Using Technology Effectively	• Screen sharing and enabling chat, camera, and microphone • Varying the presentation media • Providing technical support to students • Providing multiple communication channels • Providing interactive software tools • Enabling class recording for later review

Appendix A.2. Manual Video Annotation Procedure

VGG Image Annotator (VIA) software (Version 3) was used in this manual video annotation process to annotate Zoom-based lecture recordings. VIA is an open-source project-based annotation software for annotating images, audio, and videos, available at https://www.robots.ox.ac.uk/~vgg/software/via/app/via_video_annotator.html (accessed on 11 January 2024).

In this project, the researchers used the following steps to annotate the videos:

Step 1: Creating a new project: Open the VIA annotation tool by clicking the link above. Add the project name on the top left-hand side (refer to Figure 1). The project name should be the same as the recorded lecture name.

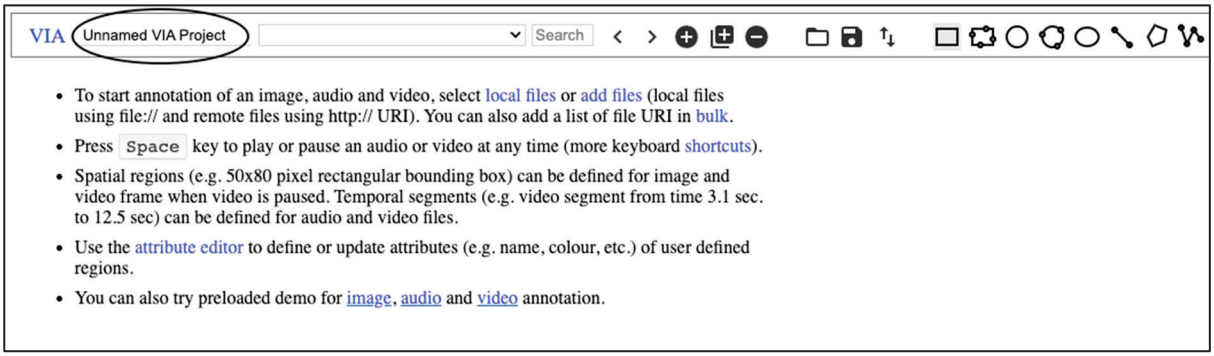


Figure A1. Create a new project.

Step 2: Adding a video file: The second step is to add a video by clicking the plus icon (refer to Figure A1). Select the video to be annotated from the desktop or cloud storage.

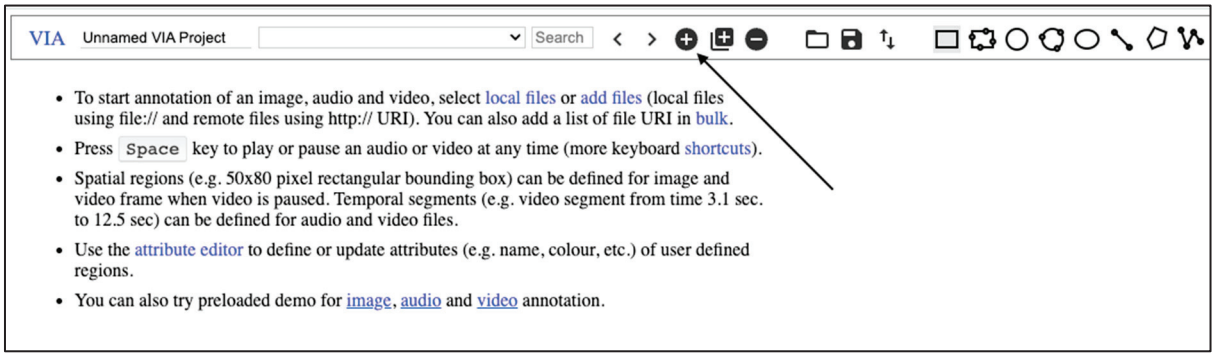


Figure A2. Add a video.

Step 3: Define the attributes: Once the video is added, define the attributes by clicking on 1 (refer to Figure A2). In this step, two attributes have been created by typing the attribute name in 2 (refer to Figure A2) and clicking Create. In this project, the first attribute was created to identify the engaging teaching video indicators and the second to highlight the presenter's location in the video.

While defining the attributes, the following information was inserted (refer to Figure A3):

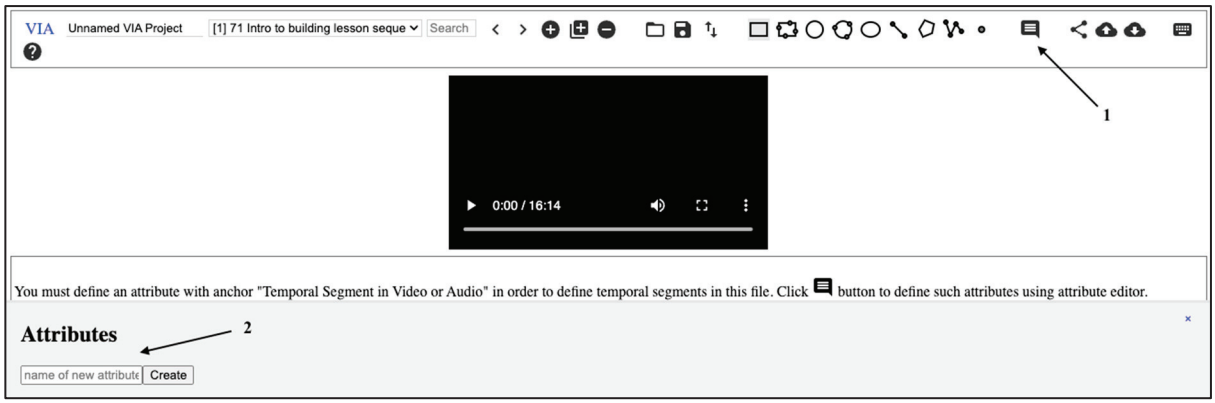


Figure A3. Define the attributes.

Attribute 1: The name of the first attribute is "Engaging teaching video indicators". The anchor is set to "Temporal Segment in Video or Audio" as researchers identified the indicators in small video segments. The text function is selected for the input type(refer to Figure A4).

Attribute 2: The name of the second attribute is "Presenter location". The attribute is created to signal the presenter's location in the video. The anchor is set to "Spatial region in a video frame" as an area is highlighted to indicate the presenter's location. The input type is set as Select. In the options section, the researchers have typed "presenter" to

Name = Presenter location

Anchor = Spatial region in a video frame

Input Type = Select

Options = *Presenter (Note: if there are multiple presenters in a video, we can add *presenter 1, presenter 2)

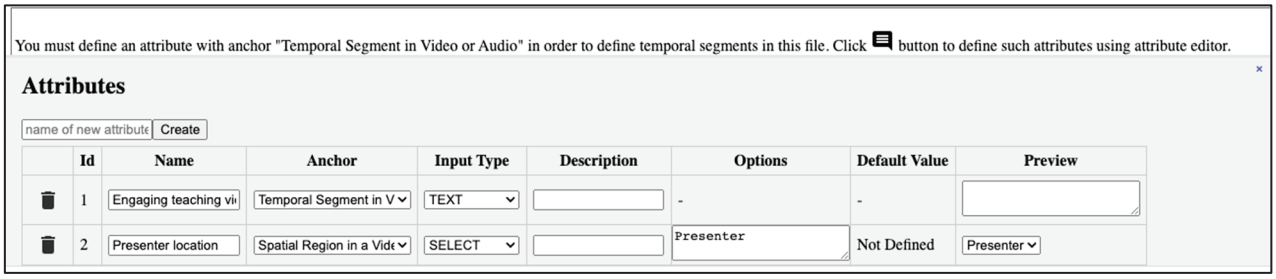


Figure A4. Attribute 1 and 2.

Step 4: Adding indicators to Attribute 1 (engaging teaching video indicators): After defining the attributes, the next step is adding the indicators. The researchers added the indicators at the bottom left-hand side by writing the indicator name and then clicking Add (refer to Figure A5). The following indicators have been added.

Indicators		Description
1.	Encouraging students' participation in discussion	Teachers to engage students in discussions or debates to attract their interest and motivate a deeper understanding
2.	Encouraging students to share their knowledge and ideas	Teachers to ask for students' participation in active learning methods by sharing their perceptions, knowledge, and ideas
3.	Encouraging students to ask questions	Teachers to create a safe and open environment that allows students to ask their questions, to enhance the student interaction experience
4.	Encouraging collaborative learning activities	Teachers to create opportunities for students to interact with each other through group activities or collaborative work
5.	Encouraging meaningful interaction	Teachers to construct a welcoming and efficient online learning environment by fostering regular and meaningful communication with students and providing meaningful answers to students' enquiries
6.	Providing learning resources	Teachers to provide students with various learning resources, videos, etc., to increase students' active participation
7.	Giving clear instructions	Teachers to be clear and detailed in communicating the instructions, expectations, roles, and responsibilities, to show commitment to meeting the course goals
8.	Outlining the learning objectives	Teachers to clearly outline and communicate the topics and instructions to increase student engagement in online learning
9.	Using appropriate changes in tone of voice	Teachers to read and respond to perceived restlessness by using appropriate changes in tone of voice or changes in direction
10.	Facial expressions	Teachers to maintain appropriate facial expressions such as smiling and nodding
11.	Eye contact	Teachers to maintain eye contact with students in online learning
12.	Appropriate body language	Teachers to maintain appropriate body language in the online classroom
13.	Enabling class recording for later review	Teachers to increase the value of the online learning experience by enabling class recording, which allows students access to classroom sessions from the comfort of their home and if they want to review afterwards

Indicators	Description
14. Screen sharing and enabling chat, camera, and microphone	Teachers to assure students of their presence and positively impact student engagement and satisfaction by communicating in real-time through a chat, camera, microphone, and screen sharing
15. Varying the presentation media	Teachers to vary the presentation media (e.g., videos, slides, note sharing, etc.) to capture students' attention and foster engagement

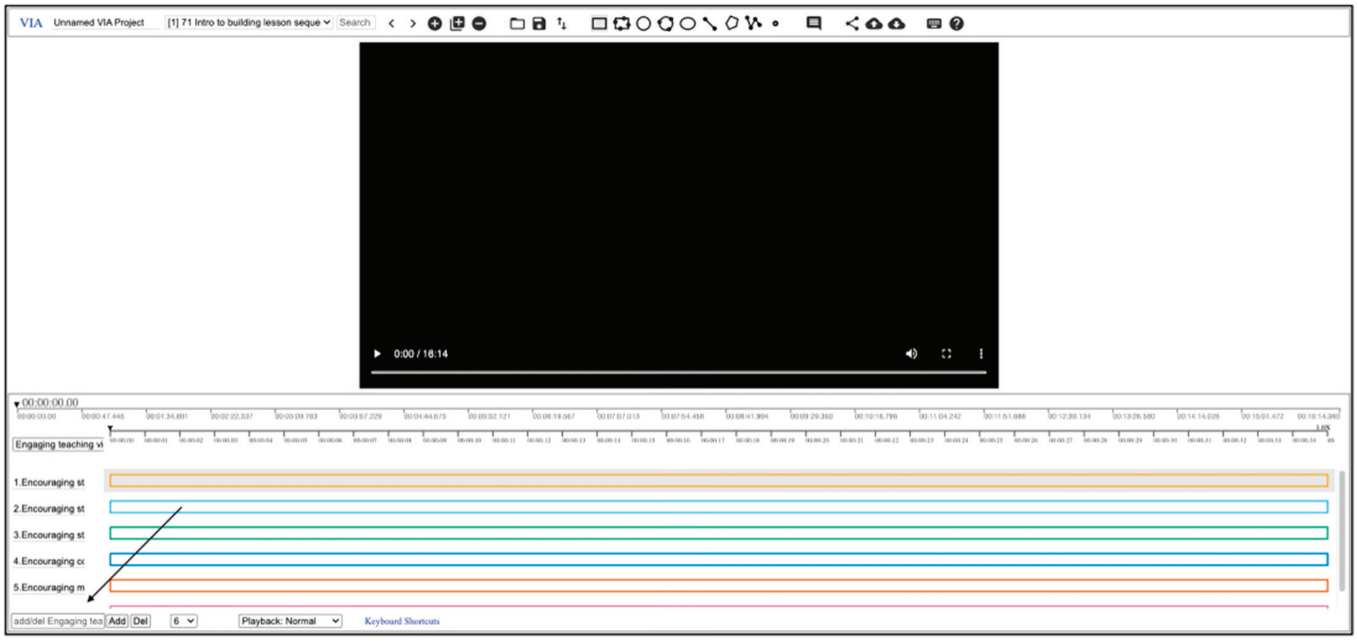


Figure A5. Adding indicators.

Step 5: Drawing a boundary box by clicking on 1 to signal the presenter's location by clicking on 2 (Attribute 2: presenter location): The researchers drew a boundary box to indicate the presenter's location (refer to Figure A6).

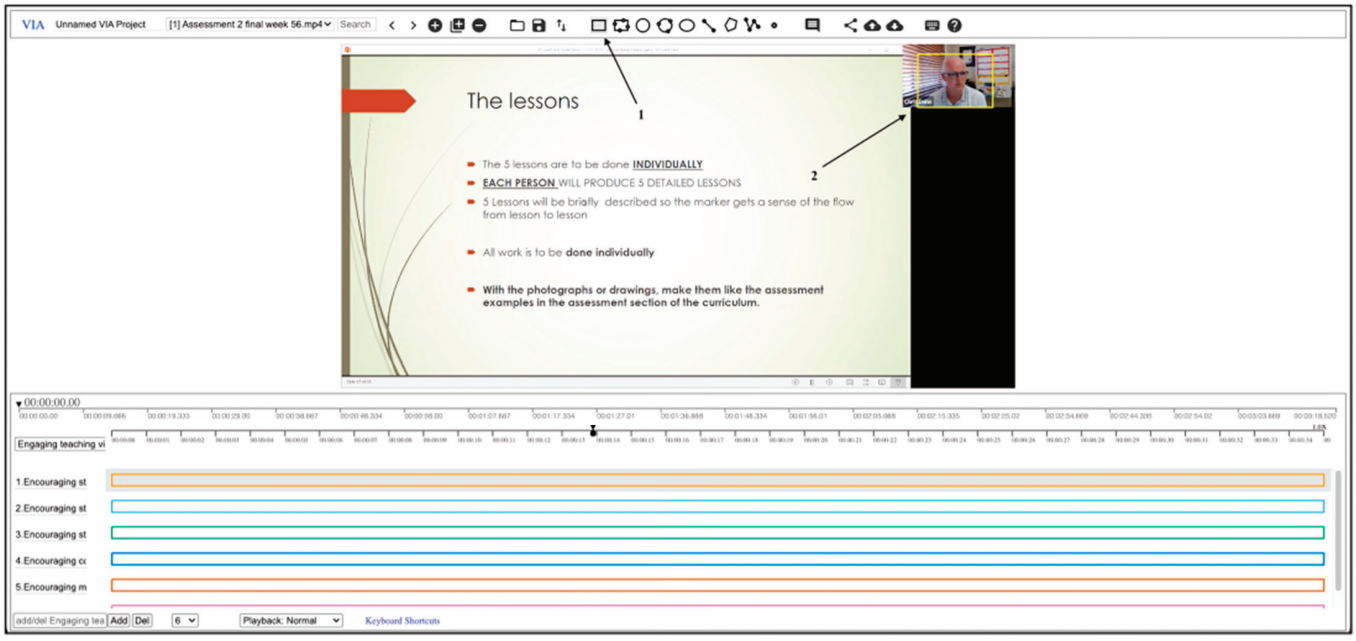


Figure A6. Drawing boundary box.

Step 6: Identifying the indicators from the video: Manual annotation is performed after defining the attributes and indicating the presenter's location. In this process, the video is played, and indicators are identified in small segments (refer to arrows in Figure A7). To start the temporal segment, click "a", and to stop it, click "Shift" + "a".

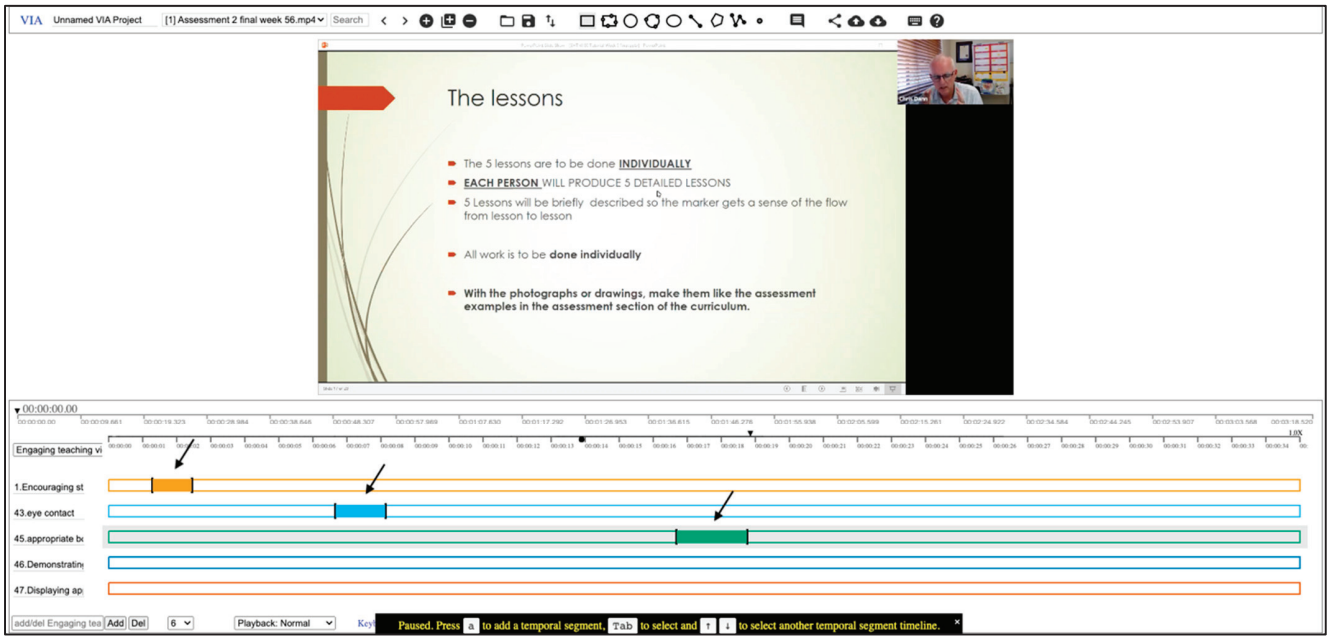


Figure A7. Identifying the indicators.

Step 7: Saving and Exporting the Project for Machine Learning: Once the annotation is complete, save the project by clicking on 1 and selecting the project's location. Similarly, click on 2 to export the project (refer to Figure A8).

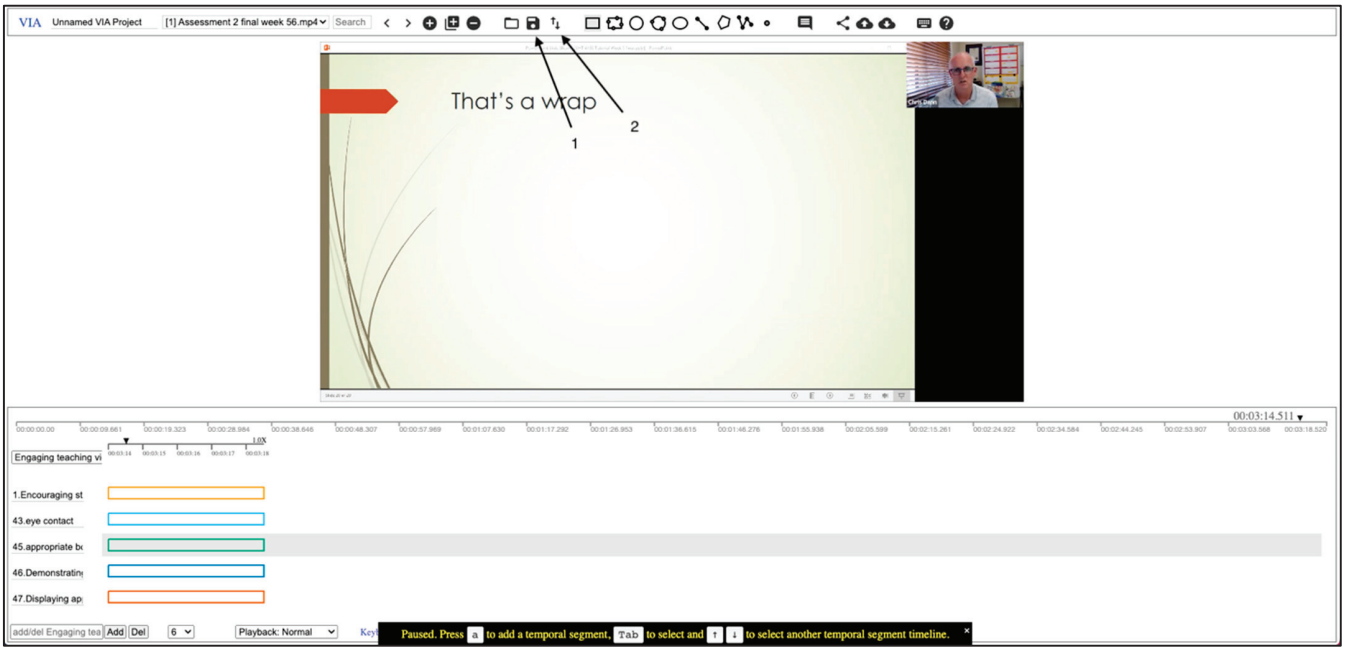


Figure A8. Save and export.

References

Aldoseri, A., Al-Khalifa, K. N., & Hamouda, A. M. (2023). Re-thinking data strategy and integration for Artificial Intelligence: Concepts, opportunities, and challenges. *Applied Sciences*, *13*(12), 7082. [CrossRef]

Alenezi, E., Alfadley, A. A., Alenezi, D. F., & Alenezi, Y. H. (2022). The sudden shift to distance learning: Challenges facing teachers. *Journal of Education and Learning*, *11*(3), 14. [CrossRef]

Apicella, A., Arpaïa, P., Frosolone, M., Improta, G., Moccaldi, N., & Pollastro, A. (2022). EEG-based measurement system for monitoring student engagement in learning 4.0. *Scientific Reports*, *12*(1), 5857. [CrossRef]

Ashwin, T. S., & Guddeti, R. M. R. (2019). Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks. *Education and Information Technologies*, *25*(2), 1387–1415. [CrossRef]

Beaver, I., & Mueen, A. (2022). On the care and feeding of virtual assistants: Automating conversation review with AI. *AI Magazine*, *42*(4), 29–42. [CrossRef]

Behera, A., Matthew, P., Keidel, A., Vangorp, P., Fang, H., & Canning, S. (2020). Associating facial expressions and upper-body gestures with learning tasks for enhancing intelligent tutoring systems. *International Journal of Artificial Intelligence in Education*, *30*(2), 236–270. [CrossRef]

Bell, J. (1999). *Doing your research project: A guide for first-time researchers in education and social science* (3rd ed.). Open University Press.

Castro, M. D. B., & Tumibay, G. M. (2021). A literature review: Efficacy of online learning courses for higher education institution using meta-analysis. *Education and Information Technologies*, *26*, 1367–1385. [CrossRef]

Cents-Boonstra, M., Lichtwarck-Aschoff, A., Lara, M. M., & Denessen, E. (2021). Patterns of motivating teaching behaviour and student engagement: A microanalytic approach. *European Journal of Psychology of Education*, *37*, 227–255. [CrossRef]

Chiu, T. K. F. (2021). Applying the self-determination theory (SDT) to explain student engagement in online learning during the COVID-19 pandemic. *Journal of Research on Technology in Education*, *54*(Suppl. S1), S14–S30. [CrossRef]

De Silva, D., Kaynak, O., El-Ayoubi, M., Mills, N., Alahakoon, D., & Manic, M. (2024). Opportunities and challenges of Generative artificial intelligence: Research, education, industry engagement, and social impact. *IEEE Industrial Electronics Magazine*, 2–17. [CrossRef]

Dhawan, S. (2020). Online learning: A panacea in the time of COVID-19 crisis. *Journal of Educational Technology Systems*, *49*(1), 5–22. [CrossRef]

Giang, T. T. T., Andre, J., & Lan, H. H. (2022). Student engagement: Validating a model to unify in-class and out-of-class contexts. *Journal of Education and Learning*, *8*(4), 1–14. [CrossRef]

Gillett-Swan, J. (2017). The challenges of online learning: Supporting and engaging the isolated learner. *Journal of Learning Design*, *10*(1), 20–30. [CrossRef]

Harry, A., & Sayudin, S. (2023). Role of AI in education. *Interdiciplinary Journal and Humanity (Injurity)*, *2*(3), 260–268. [CrossRef]

Heale, R., & Twycross, A. (2018). What is a case study? *Evidence-Based Nursing*, *21*(1), 7–8. [CrossRef]

Heeg, D. M., & Avraamidou, L. (2023). The use of Artificial intelligence in school science: A systematic literature review. *Educational Media International*, *60*(2), 125–150. [CrossRef]

Hew, K. F. (2016). Promoting engagement in online courses: What strategies can we learn from three highly rated MOOCS. *British Journal of Educational Technology*, *47*(2), 320–341. [CrossRef]

Huang, A. Y. Q., Lu, O. H. T., & Yang, S. J. H. (2023). Effects of artificial Intelligence–Enabled personalised recommendations on learners' learning engagement, motivation, and outcomes in a flipped classroom. *Computers & Education*, *194*, 104684. [CrossRef]

Kvale, S. (1996). *Interview views: An Introduction to qualitative research interviewing*. Sage Publications.

Lee, J., Song, H., & Hong, A. J. (2019). Exploring factors, and indicators for measuring students' sustainable engagement in e-Learning. *Sustainability*, *11*(4), 985. [CrossRef]

Li, J., Lin, F., Yang, L., & Huang, D. (2023). AI service placement for Multi-Access Edge Intelligence systems in 6G. *IEEE Transactions on Network Science and Engineering*, *10*(3), 1405–1416. [CrossRef]

Liang, R., & Chen, D. T. V. (2012). Online learning: Trends, potential and challenges. *Creative Education*, *3*(8), 1332. [CrossRef]

Limna, P., Jakwatanatham, S., Siripipattanakul, S., Kaewpuang, P., & Sriboonruang, P. (2022). A review of artificial intelligence (AI) in education during the digital era. *Advance Knowledge for Executives*, *1*(1), 1–9. Available online: https://ssrn.com/abstract=4160798 (accessed on 5 January 2024).

Ma, J., Han, X., Yang, J., & Cheng, J. (2015). Examining the necessary condition for engagement in an online learning environment based on learning analytics approach: The role of the instructor. *The Internet and Higher Education*, *24*, 26–34. [CrossRef]

Ma, X., Xu, M., Dong, Y., & Sun, Z. (2021). Automatic student engagement in online learning environment based on Neural Turing Machine. *International Journal of Information and Education Technology*, *11*(3), 107–111. [CrossRef]

Murtaza, M., Ahmed, Y., Shamsi, J. A., Sherwani, F., & Usman, M. (2022). AI-Based personalised E-Learning systems: Issues, challenges, and solutions. *IEEE Access*, *10*, 81323–81342. [CrossRef]

Nguyen, N. D. (2023). Exploring the role of AI in education. *London Journal of Social Sciences*, *6*, 84–95. [CrossRef]

Nikoloutsopoulos, S., Koutsopoulos, I., & Titsias, M. K. (2024, May 5–8). *Kullback-Leibler reservoir sampling for fairness in continual learning*. 2024 IEEE International Conference on Machine Learning for Communication and Networking (ICMLCN) (pp. 460–466), Stockholm, Sweden. [CrossRef]

Ocaña, M. G., & Opdahl, A. L. (2023). A software reference architecture for journalistic knowledge platforms. *Knowledge-Based Systems*, *276*, 110750. [CrossRef]

Pianykh, O. S., Langs, G., Dewey, M., Enzmann, D. R., Herold, C. J., Schoenberg, S. O., & Brink, J. A. (2020). Continuous Learning AI in radiology: Implementation principles and early applications. *Radiology*, *297*(1), 6–14. [CrossRef]

Pingenot, A., & Shanteau, J. (2009). Expert opinion. In M. W. Kattan (Ed.), *Encyclopedia of medical decision making*. Sage Publications, Inc. Available online: https://www.researchgate.net/publication/263471207_Expert_Opinion (accessed on 2 January 2024).

Roshanaei, M., Khan, M. R., & Sylvester, N. N. (2024). Enhancing Cybersecurity through AI and ML: Strategies, challenges, and future directions. *Journal of Information Security*, *15*(3), 320–339. [CrossRef]

Sandelowski, M. (2000). Combining qualitative and quantitative sampling, data collection, and analysis techniques. *Research in Nursing & Health*, *23*(3), 246–255. [CrossRef]

Shaikh, A. A., Kumar, A., Jani, K., Mitra, S., García-Tadeo, D. A., & Devarajan, A. (2022). The role of Machine Learning and Artificial Intelligence for making a digital classroom and its sustainable impact on education during COVID-19. *Materials Today Proceedings*, *56*, 3211–3215. [CrossRef] [PubMed]

Shekhar, P., Prince, M. J., Finelli, C. J., DeMonbrun, M., & Waters, C. (2018). Integrating quantitative and qualitative research methods to examine student resistance to active learning. *European Journal of Engineering Education*, *44*(1–2), 6–18. [CrossRef]

Tahiru, F. (2021). AI in education. *Journal of Cases on Information Technology*, *23*(1), 1–20. [CrossRef]

Tinoca, L., Piedade, J., Santos, S., Pedro, A., & Gomes, S. (2022). Design-Based research in the educational field: A systematic literature review. *Education Sciences*, *12*(6), 410. [CrossRef]

Turner, D. J. (2010). Qualitative interview design: A practical guide for novice investigators. *The Qualitative Report*, *15*(3), 754–760. [CrossRef]

Verma, N., Getenet, S., Dann, C., & Shaik, T. (2023a). Characteristics of engaging teaching videos in higher education: A systematic literature review of teachers' behaviours and movements in video conferencing. *Research and Practice in Technology Enhanced Learning*, *18*, 040. [CrossRef]

Verma, N., Getenet, S., Dann, C., & Shaik, T. (2023b). Designing an artificial intelligence tool to understand student engagement based on teacher's behaviours and movements in video conferencing. *Computers & Education: Artificial Intelligence*, *5*, 100187. [CrossRef]

Wang, C., Yang, Z., Li, Z. S., Damian, D., & Lo, D. (2024). Quality assurance for Artificial intelligence: A study of industrial concerns, challenges and best practices. *arXiv*, arXiv:2402.16391. [CrossRef]

Wang, X., & Yin, M. (2023, April 23–28). *Watch out for updates: Understanding the effects of model explanation updates in ai-assisted decision making*. 2023 CHI Conference on Human Factors in Computing Systems (pp. 1–19), Hamburg, Germany. [CrossRef]

Weng, X., Ng, O.-L., & Chiu, T. K. F. (2023). Competency development of pre-service teachers during video-based learning: A systematic literature review and meta-analysis. *Computers & Education*, *199*, 104790. [CrossRef]

Xie, J., A, G., Rice, M. F., & Griswold, D. E. (2021). Instructional designers' shifting thinking about supporting teaching during and post-COVID-19. *Distance Education*, *42*, 1–21. [CrossRef]

Žliobaite, I., Budka, M., & Stahl, F. (2015). Towards cost-sensitive adaptation: When is it worth updating your predictive model? *Neurocomputing*, *150*, 240–249. [CrossRef]

education sciences

MDPI

Article

Subject-Specialized Chatbot in Higher Education as a Tutor for Autonomous Exam Preparation: Analysis of the Impact on Academic Performance and Students' Perception of Its Usefulness

Fulgencio Sánchez-Vera

Department of Didactics and Educational Research, University of La Laguna, 38200 San Cristóbal de La Laguna, Spain; fsanchev@ull.edu.es

Abstract: This study evaluates the impact of an AI chatbot as a support tool for second-year students in the Bachelor's Degree in Early Childhood Education program during final exam preparation. Over 1-month, 42 students used the chatbot, generating 704 interactions across 186 conversations. The study aimed to assess the chatbot's effectiveness in resolving specific questions, enhancing concept comprehension, and preparing for exams. Methods included surveys, in-depth interviews, and analysis of chatbot interactions. Results showed that the chatbot was highly effective in clarifying doubts (91.4%) and aiding concept understanding (95.7%), although its perceived usefulness was lower in content review (42.9%) and exam simulations (45.4%). Students with moderate chatbot use achieved better academic outcomes, while excessive use did not lead to further improvements. The study also identified challenges in students' ability to formulate effective questions, limiting the chatbot's potential in some areas. Overall, the chatbot was valued for fostering study autonomy, though improvements are needed in features supporting motivation and study organization. These findings highlight the potential of chatbots as complementary learning tools but underscore the need for better user training in "prompt engineering" to maximize their effectiveness.

Keywords: educational technology; educational chatbot; artificial intelligence; higher education; prompt engineering

1. Introduction

In recent years, the integration of digital technologies in higher education has grown exponentially. The COVID-19 health crisis accelerated the adoption of technological solutions in teaching and learning processes, highlighting the potential of online learning to facilitate self-directed and flexible environments. As a result, universities have been forced to develop the ability to quickly adapt to unforeseen circumstances (Cramarenco et al., 2023). Furthermore, the Fourth Industrial Revolution has driven a profound digital transformation in higher education institutions, linked to the need to align with labor market demands (Schwab, 2017).

The digitalization of educational institutions has promoted the expansion of online courses and hybrid or blended-learning teaching, allowing greater accessibility and educational flexibility (Alenezi, 2023). However, digital technologies have also been integrated into face-to-face teaching through virtual classrooms and learning platforms, complementing traditional education, and tailoring learning experiences to students' individual

needs. Among the benefits of this process, we can highlight more equitable access to education by making learning more flexible, personalized, and accessible (Dhawan, 2022; Valverde-Berrocoso et al., 2022).

One of the most recent advances in the use of digital technologies in higher education has been the integration of artificial intelligence (AI) in educational processes. In particular, AI-based chatbots have gained popularity due to their ability to interact dynamically with students, providing real-time feedback and fostering personalized learning. These systems allow students to access immediate answers to specific questions, which not only reduces exam-related anxiety but also enhances their ability to self-regulate the learning process (Ng & Leung, 2024; Su & Yang, 2023). Moreover, LLMs have demonstrated significant potential to enhance innovative teaching methods, such as the flipped classroom. By enabling the creation of interactive materials for students to review before face-to-face lessons, they promote active and personalized learning at home, fostering deeper engagement during class (Tan, 2023). In addition, automated quiz-generation tools powered by LLMs allow educators to provide personalized assessments aligned with individual learning progress (Hang et al., 2024).

The emergence of tools like ChatGPT, launched in 2022, along with other systems like Gemini, Perplexity, Claude, and others, has revolutionized the way we interact with information. These platforms automate complex cognitive tasks and encourage creativity, presenting new opportunities and challenges in education and research. Their widespread adoption among higher education students has been facilitated by the numerous advantages they offer, such as immediate and efficient access to knowledge and increased productivity in academic tasks. However, these tools are not always used appropriately, transparently, and ethically (Dabis & Csáki, 2024; Chen et al., 2024; Mironova et al., 2024).

The educational potential of chatbots is considerable. One of the most notable benefits is the personalization of learning (Abas et al., 2023; Agbong-Coates, 2024). Unlike traditional forms of teaching, which are often more static and generalized, chatbots enable direct and adaptive interaction with students, adjusting both the content and the pace of teaching to each individual's specific needs. This personalization has been shown to have a positive impact on student motivation, as it provides a less stressful study environment more focused on their interests and learning pace (Holmes et al., 2019). Additionally, by offering immediate responses and allowing for content repetition, chatbots help students delve deeper into topics of greater interest or those they find more challenging, thereby improving knowledge retention (Deng & Yu, 2023).

Personalization also has significant implications for self-directed learning. By allowing students to manage their own study pace, chatbots reduce the pressure associated with traditional academic deadlines, fostering greater flexibility in the learning process. This not only contributes to improved academic performance but also strengthens autonomy and self-management, key skills in today's educational environment. Recent studies indicate that students using chatbots tend to show greater satisfaction with their learning experiences and achieve better academic outcomes compared to those without access to these tools (Zhang et al., 2024; Guttierrez-Aguilar et al., 2024).

Self-regulation of learning is another aspect that chatbots can enhance. Students' ability to plan, monitor, and adjust their own learning process is essential for academic success, especially in self-directed learning environments. Chatbots offer real-time feedback, allowing students to modify their study strategies as needed. This continuous and adaptive feedback fosters more efficient and self-directed learning, reducing dependence on teacher intervention and enabling students to take control of their own educational process (Ng & Leung, 2024; Labadze et al., 2023).

The adoption of chatbots in higher education largely depends on students' perception of their usefulness and ease of use. Davis' Technology Acceptance Model (TAM) (Davis, 1989) provides a theoretical foundation to evaluate these perceptions, emphasizing that the adoption of new technologies is driven by two main factors: perceived usefulness (the extent to which a tool enhances performance) and perceived ease of use (the effort required to use the tool). In the case of chatbots, research has shown that students find them intuitive and accessible, facilitating their adoption (Schei et al., 2024).

Despite the aforementioned benefits, the use of educational chatbots raises several important ethical challenges, particularly regarding student privacy. Chatbots collect large amounts of data on students' study habits and academic performance, raising concerns about how such data is managed and protected (Kasneci et al., 2023). Therefore, educational institutions must develop clear and transparent policies to ensure the ethical and responsible use of student data, respecting their privacy rights (Holmes & Tuomi, 2022). This is crucial in a context of increasing AI use in educational settings, where the risk of data misuse is growing.

Another important concern is access to technology. Although chatbots can offer significant benefits for personalized learning, it is essential to ensure that all students, regardless of their socioeconomic status, have access to these tools. The digital divide remains a problem in many regions, and unequal access to advanced technologies can exacerbate existing inequalities in the education system (Holmes & Tuomi, 2022; Rahman & Watanobe, 2023). To mitigate this risk, universities must develop strategies to promote equitable access to technology and provide resources for students who may be at a technological disadvantage (Mat Dangi & Mohamed Saat, 2021).

The impact of chatbots on learning outcomes has been extensively researched in recent years. Studies have shown that students using these technologies experience significant improvements in academic performance, both in content comprehension and information retention, as well as in learning self-regulation (Ng & Leung, 2024; Deng & Yu, 2023; Mungai et al., 2024; Martins et al., 2024; Yigci et al., 2024). These improvements are primarily due to the personalization of learning and the constant feedback that chatbots provide, allowing students to adjust their study strategies more effectively and resolve doubts immediately (Chang et al., 2023). Nonetheless, we are still in the early stages of integrating these technologies, and further studies are needed to support these apparent advantages.

In any case, to maximize the potential of chatbots, their implementation must be supported by a solid pedagogical framework. Educators must strategically integrate these tools into their teaching methodologies, promoting not only self-directed learning but also active learning and critical engagement (Deng & Yu, 2023; Halaweh, 2023; Chukwuere, 2024). A well-structured pedagogical approach that combines the use of chatbots with other teaching strategies can improve not only academic outcomes but also student motivation and engagement with their learning process (Lin & Chang, 2023).

This study evaluates the educational impact of a subject-specialized chatbot implemented as a support tool for second-year students enrolled in the course "The School of Early Childhood Education" in the Bachelor's Degree in Early Childhood Education program at the University of La Laguna (Spain) during the 2023–2024 academic year. Through this implementation, the aim is to provide an accessible and effective resource for resolving specific doubts, contributing to more autonomous learning during the exam preparation period.

Research Questions:

- RQ1: How effective is the specialized chatbot in resolving students' doubts and delivering clear and precise explanations to support exam preparation?

- RQ2: What is the frequency and patterns of chatbot use among students, and what types of questions do they typically ask?
- RQ3: How do students and teachers perceive the effectiveness and usefulness of the chatbot as a complementary learning tool?

2. Materials and Methods

This study evaluated the impact of using a specialized chatbot in the course "The School of Early Childhood Education" as a personal tutor during the study process and final exam preparation for second-year students in the Bachelor's Degree in Early Childhood Education program during the June 2024 exam period. The research was conducted using a combination of surveys, in-depth interviews, and an analysis of the students' interactions with the chatbot.

2.1. Participants

Participation in the use of the chatbot was entirely voluntary. All students enrolled in the course (92 in total) were invited to participate in the study, and the final sample consisted of the 42 students who agreed to take part. For 1 month, participants used the chatbot as a support tool for their final exam preparation. The course was led by a single instructor, who was also directly involved in the design and implementation of the research.

2.2. Research Procedure

2.2.1. Pre-Intervention Phase

The first step involved selecting and designing a chatbot. The Chatbase platform (https://chatbase.co, accessed on 17 October 2024) was chosen as the tool for creating and managing the educational conversational assistant. This selection was based on several key features that made Chatbase an appropriate choice for the study's objectives:

- Ease of use: Chatbase provides an intuitive interface, allowing chatbot creation without advanced programming knowledge. This facilitated the quick setup and customization of the chatbot for the educational context.
- Compatibility with advanced AI: The tool uses natural language processing (NLP) technologies based on advanced models, such as GPT, which enabled the chatbot to answer complex questions coherently and accurately.
- Integration with knowledge bases: Chatbase allows the chatbot to be trained with specific course files and documents.
- Interaction recording: The platform logs and analyzes all user interactions with the chatbot, which proved valuable for subsequent data analysis.

The chatbot was designed and programmed to answer questions related to the course content, including definitions, detailed explanations, and practical examples. It also featured closed and open questions, using a knowledge base built from course materials and previous exams. To ensure pedagogical alignment, the chatbot design incorporated principles of active and self-directed learning. Specifically, it was programmed to guide students through incremental problem-solving tasks, encouraging independent inquiry through targeted questions to improve critical thinking skills. These pedagogical principles were further reinforced during training sessions, where students were taught strategies to maximize the chatbot's educational potential. An evaluation framework was developed to assess the chatbot, focusing on key constructs such as accuracy, consistency, fluency, and perceived usefulness, all aligned with the goal of supporting self-directed learning. These constructs were assessed through tests conducted by the authors, including predefined questions to assess the chatbot's ability to deliver correct answers, maintain consistent responses across similar queries, and demonstrate fluency across interactions. Findings

from these evaluations informed iterative adjustments to the chatbot design, ensuring alignment with intended educational goals. Under these conditions, the chatbot's performance was implemented and tested following a systematic procedure until optimal results were achieved before offering it to students (Sánchez-Vera, 2024). At the same time, an initial survey was conducted among participating students to collect demographic data, assess their familiarity with digital tools, and gather their expectations regarding the chatbot and its educational usefulness.

2.2.2. Intervention Phase

A training session was conducted to familiarize students with the use of the chatbot. During this session, the tool's purpose, main features, and basic prompting techniques were explained, teaching students how to interact effectively with the chatbot to obtain useful responses (Knoth et al., 2024). Three prompting strategies were introduced: zero-shot prompting, few-shot prompting, and chain-of-thought prompting. Zero-shot prompting involved the chatbot responding to queries without prior examples, focusing on retrieving general knowledge effectively. Few-shot prompting included providing the chatbot with a small number of examples to guide its responses, enhancing its ability to address context-specific questions. Chain-of-thought prompting encouraged the chatbot to articulate its reasoning process step by step, making it particularly useful for solving complex or multi-step problems.

These strategies were illustrated through practical examples and live demonstrations, ensuring students could understand and apply them effectively. Additional guidance was offered to refine their ability to formulate clear, well-structured questions, maximizing the chatbot's potential to provide relevant and useful responses.

For 1 month, students used the chatbot as a support tool for their final exam preparation. All interactions were recorded for later analysis to assess the tool's impact on the learning process.

2.2.3. Post-Intervention Phase

Once the intervention period ended, several actions were taken to measure the results obtained. First, a final survey replicating the initial survey's items was administered to identify changes in students' perceptions. This survey also included specific questions to evaluate their experience with the chatbot. Additionally, six in-depth qualitative interviews were conducted with selected students to gain a more detailed understanding of their experiences. These interviews focused on exploring the chatbot's perceived strengths and weaknesses and its usefulness as a study aid.

In terms of analysis, both quantitative and qualitative approaches were applied. Data from the pre- and post-intervention surveys were analyzed statistically to detect significant variations in students' perceptions and skills. Additionally, a qualitative analysis of the interview responses and chatbot interactions was conducted, identifying key themes and patterns that provided a more comprehensive understanding of the chatbot's impact on the learning process.

2.3. Data Collection Instruments

2.3.1. Instruments for Evaluating the Chatbot's Quality

The chatbot's performance was evaluated through an iterative process of testing and refinement. Adjustments were made to the chatbot's prompts and knowledge base to ensure optimal performance in terms of accuracy, consistency, fluency, problem-solving ability, exam preparation effectiveness, and reliability. These constructs were selected for their relevance to both the pedagogical and technical goals of the chatbot. The selected indicators for evaluation reflect realistic standards suitable for a university environment.

The evaluation was conducted through systematic tests in a controlled environment. While a 100% performance would be ideal, current technological limitations prevent such perfection. Therefore, the proposed values seek to balance pedagogical utility with the chatbot's technical capabilities. These limits were established by consensus between the researchers and chatbot designers, representing conventional ranges designed to ensure that the chatbot is robust enough to meet educational needs without demanding impossible performance for current technology:

- Response Accuracy ($\geq$90%): The chatbot must provide correct answers 90% of the time, ensuring that responses are largely accurate and contribute to quality learning.
- Response Consistency ($\geq$95%): A 95% consistency rate ensures that responses are aligned with repeated concepts in different contexts, avoiding student confusion.
- Interaction Fluency ($\geq$85%): Fluent conversation is critical for a positive experience, with 85% representing a comprehensible and natural interaction in most cases.
- Problem-Solving Ability ($\geq$90%): The chatbot should effectively resolve complex questions at least 90% of the time, ensuring it can guide students in understanding difficult concepts.
- Exam Preparation Effectiveness ($\geq$85%): To be perceived as useful, the chatbot must effectively help students prepare for exams, achieving an 85% user satisfaction rate.
- Reliability (100%): The chatbot is expected to have 100% availability, ensuring uninterrupted operation throughout the usage period.

The evaluation process included two sets of tests. The first set consisted of 60 questions taken from real exams, divided into six categories: concept definitions, true/false, multiple choice, concept association, short answer, and short essay. The second set evaluated the chatbot's ability to adapt to open-ended and unstructured conversations, simulating more fluid and spontaneous interactions to measure its adaptability.

The chatbot was subjected to multiple rounds of these tests, and its responses were evaluated according to the criteria defined above. After each round of testing, adjustments were made to both the prompts and the chatbot's knowledge base to improve the accuracy and adaptability of the responses. This continuous refinement cycle was maintained until the chatbot reached optimal performance levels according to the proposed indicators, ensuring its ability to meet pedagogical and technological expectations.

2.3.2. Pre- and Post-Intervention Surveys

Two surveys were designed and administered to assess the impact of the intervention. The constructs measured in the questionnaires were developed specifically for this study (ad hoc) to capture perceived usefulness, ease of use, interaction frequency, and overall satisfaction with the chatbot. These constructs were chosen to align with the study's objectives of evaluating the chatbot's role in supporting exam preparation and self-directed learning. The questions were created by the research team based on the specific context of the course and the chatbot's intended functionalities. While the ad hoc nature of the survey limits the possibility of external validation, the questions were reviewed by three educational technology experts to ensure clarity, relevance, and alignment with the study's goals. The first survey, conducted before the intervention, consisted of nine Likert-type questions addressing demographic data (age and gender), frequency of technology use, perceived technological proficiency, expectations regarding the chatbot, and prior experience with chatbots or AI in educational contexts. This set of questions provided baseline data on the participants.

The second survey, conducted after the intervention, included a total of 16 questions. Nine of these replicated items from the initial survey, allowing for direct comparisons between pre- and post-intervention results. The seven additional items focused on evaluat-

ing the chatbot experience, including interaction frequency, overall satisfaction, perceived usefulness in exam preparation, ease of use, areas where the chatbot was most helpful, interaction naturalness, and willingness to recommend the chatbot to other students.

The survey design, based on Likert scales, allowed for descriptive analysis of frequencies and percentages to identify changes in students' perceptions and evaluate the intervention's effectiveness, following validated methodologies in previous studies on educational interventions (Boone & Boone, 2012; Joshi et al., 2015).

2.3.3. In-Depth Interviews

In-depth interviews were conducted with six students after the intervention. These interviews provided a qualitative understanding of students' experiences with the chatbot. This type of interview is particularly useful when seeking information about complex or subjective processes that cannot be captured by traditional quantitative methods. In-depth interviews are a valuable tool for exploring participants' perceptions, experiences, and attitudes in detail, generating a richer, more contextualized understanding of the phenomenon under study (Kvale, 2007; Patton, 2015). The interviews focused on the perceived effectiveness of the chatbot, areas for improvement, and perceived challenges.

The six students were selected based on their level of interaction with the chatbot, determined by analyzing the number of conversations they had during the intervention phase. Participants were categorized into three groups: low interaction (1–10 interactions), moderate interaction (11–20 interactions), and high interaction (more than 20 interactions). Two students were chosen from each category to ensure a balanced representation of chatbot usage patterns. This selection strategy aimed to capture a diverse range of experiences, from minimal to intensive use of the chatbot. By including students from all interaction levels, the interviews offered a comprehensive understanding of how varying degrees of engagement influenced their perceptions of the chatbot's usefulness, ease of use, and its impact on exam preparation.

Furthermore, during the analysis of the interviews, data saturation was achieved, as no new themes or insights emerged after conducting the six interviews. This suggests that the selected participants provided a sufficiently comprehensive representation of the experiences and perceptions within the study cohort.

2.3.4. Chatbot Interaction Analysis

The conversations between students and the chatbot were analyzed to identify the most frequent types of questions. The conversations were categorized thematically, allowing for an analysis of the main types of interaction and, consequently, how students used the chatbot. This approach provided a clear understanding of their tutoring needs and identified areas where gaps or deficiencies were observed. This method corresponds to what the literature calls human-machine interaction analysis, a technique that studies interactions between humans and automated systems, evaluating both the content of questions and usage patterns (Følstad et al., 2018; Diederich et al., 2021). This type of analysis is common in research on chatbots in educational settings, where conversations are examined to reflect user understanding and identify areas where technological interventions can improve the educational experience (Winkler & Söllner, 2018).

2.4. Ethical Considerations

This study was conducted in compliance with current regulations on data protection and ethics in educational research. Specifically, the principles established by the General Data Protection Regulation (GDPR) (Regulation EU 2016/679) (Winkler & Söllner, 2018) and the Spanish Organic Law 3/2018, of December 5, on Personal Data Protection and

Guarantee of Digital Rights (LOPDGDD) (España, 2018), were followed to ensure the appropriate handling of participants' personal information.

Informed Consent: All participants were fully informed about the study's objectives, procedures, and scope before participating. They were told about the voluntary nature of the study and their right to withdraw at any time without consequences. Participants signed an informed consent form, which guaranteed their understanding and acceptance of the study's terms.

Confidentiality and Anonymity: Participants' personal and academic data were treated with strict confidentiality. Anonymization measures were applied by assigning unique codes to avoid the direct identification of participants. Data handling complied with GDPR and LOPDGDD provisions, ensuring that only the research team had access to the information, which was stored on secure, encrypted servers. It was also established that the data would be deleted after the study's completion and the publication of the results.

Transparency and Results: It was ensured that the study's results would be reported in aggregate form, without including information that could identify participants. Additionally, students were offered the option to receive a summary of the results, promoting transparency and the return of information to the participants.

In conclusion, the research rigorously adhered to data protection and digital rights regulations, ensuring the ethical and secure handling of participants' personal information and respecting their privacy at all times.

3. Results

3.1. Overall Chatbot Performance

The research team conducted an iterative testing process in a controlled environment to evaluate and refine the chatbot's performance before releasing it to students. These lab tests were based on predefined methodological indicators: accuracy, consistency, fluency, problem-solving ability, exam preparation effectiveness, and reliability.

Below are the results obtained in each category after successive adjustment cycles:

1. Response Accuracy: The chatbot achieved 98% accuracy, surpassing the methodological threshold of $\geq$90%. This indicator assessed the system's ability to provide correct answers to simple questions such as true/false or concept definitions. Zero-shot prompting was used for these types of questions, a technique where the chatbot generated answers without prior examples. This approach was effective for direct inquiries and required no additional adjustments to improve accuracy in this context.
2. Response Consistency: Response consistency was measured based on the uniformity of responses in different contexts, especially for repeated questions or those involving related concepts. The chatbot achieved 96% consistency, exceeding the minimum standard of $\geq$95%. This result was attained through the application of few-shot prompting, where specific examples were provided to guide responses. This technique significantly improved response alignment in multiple-choice scenarios and questions that required concept matching.
3. Interaction Fluency: The chatbot scored 89% in interaction fluency, meeting the methodological threshold of $\geq$85%. This metric was assessed based on the chatbot's ability to maintain understandable and natural interactions, particularly in open-ended or unstructured conversations. Chain-of-thought prompting (CoT) was used in the evaluation, allowing the chatbot to break down its reasoning into clear steps, improving the structure and clarity of responses in essay-type or more complex questions.
4. Problem-Solving Ability: The chatbot's problem-solving ability, assessed in terms of its skill in answering complex questions and guiding the understanding of difficult

concepts, reached 92%, surpassing the threshold of ≥90%. This result was achieved through a combination of few-shot prompting and CoT prompting, techniques that enabled the chatbot to provide more detailed and explanatory answers. These iterative adjustments to the chatbot's configuration enhanced its ability to address deeper conceptual issues, aligning with the established pedagogical goals.

5. Exam-Preparation Effectiveness: The research team also evaluated the chatbot's effectiveness in exam preparation, achieving an 88% satisfaction rate, exceeding the ≥85% standard. This indicator measured the perceived usefulness of the chatbot in content review and exam simulations. Few-shot prompting and CoT prompting techniques were essential to ensure the chatbot provided precise and relevant feedback, helping students better prepare for their academic assessments.
6. Reliability: The chatbot demonstrated 100% reliability during the lab testing period, meeting the proposed standard of full availability. This metric evaluated the system's ability to remain operational without interruptions, even under high-demand conditions. The absence of failures or downtime during testing ensured that the system could respond continuously and stably, solidifying its technical robustness.

Under these conditions, the chatbot was deemed well-prepared for implementation in a real educational environment, meeting the minimum requirements established in the methodological framework.

3.2. Analysis of Chatbot Interactions: Most Consulted Content Types and Relationship with Learning Outcomes

During the exam preparation period, students interacted with the chatbot through 186 conversations, generating a total of 704 interactions. The average usage was 3.7 interactions per conversation and 16 interactions per student, although there was a high standard deviation of 11.35, indicating considerable variability in usage patterns. This suggests that some students made few inquiries, while others engaged more intensively, using the tool in various ways.

The interactions with the chatbot were categorized into six main types, reflecting the range of topics explored by students during their exam preparation. These categories closely align with the course content. For example, the Regulations and Legislation category includes inquiries related to educational laws, a component of the course that focuses on understanding regulatory frameworks governing various aspects of education. Table 1 below presents a detailed breakdown of these categories, their frequency, and the corresponding percentage of total interactions:

The table shows that students primarily used the tool to clarify basic concepts and delve into theories, with the categories of Definitions (34.66%) and Detailed Explanations (22.30%) accounting for more than half of the total interactions. This was confirmed by several students in interviews, who indicated they turned to the chatbot when they needed to better understand a concept or theory, as "the chatbot explained things clearly and quickly, without needing to search through notes".

The categories of Practical Applications (14.91%) and Comparisons and Differences (12.36%) were less frequent. These types of questions were relevant for connecting theory with practice and contrasting pedagogical approaches. However, students mentioned in interviews that they used the chatbot less for these topics because "they did not know how to formulate questions that would generate more applied or comparative responses." Queries about Regulations and Legislation (8.66%) were limited, reflecting a lower interest in this area among students.

Table 1. Most consulted content types and distribution of the interactions.

Category	Description	Number of Interactions	Percentage (%)
Definitions	Queries to clarify key terms and concepts	244	34.66%
Detailed Explanations	Requests for in-depth information on theories and frameworks	157	22.30%
Practical Applications	Questions about applying pedagogical theories in real contexts	105	14.91%
Comparisons and Differences	Queries to contrast concepts or methodologies	87	12.36%
Regulations and Legislation	Questions about educational laws and regulations	61	8.66%
Other Queries	Technical questions or requests for additional resources	50	7.10%

To analyze the relationship between chatbot usage and academic outcomes, students were grouped into four categories based on the number of interactions they conducted. This categorization was guided by pedagogical principles rather than statistical distribution, aiming to capture meaningful differences in how students engaged with the chatbot as a learning tool. The interaction ranges—" Low", "Moderate", "High", and "Very High"—were defined based on observed patterns of usage, reflecting varying levels of engagement. For instance: The Low (1–10 interactions) category captures sporadic or minimal use, typical of students who relied on the chatbot occasionally as a supplementary resource. The Moderate (11–20 interactions) category represents balanced usage, where the chatbot complemented other study methods without dominating the learning process. The High (21–30 interactions) and Very High (>31 interactions) categories reflect increasingly intensive usage, with the latter potentially indicative of over-reliance on the tool (Table 2).

Table 2. Relationship Between Chatbot Usage and Academic Performance.

Usage (Interaction Range)	Number of Students	Average Interactions	Final Exam Average Score (Range 1 to 10)
Low (1–10)	12	9.83	4.92
Moderate (11–20)	26	13.15	6.58
High (21–30)	2	22.00	8.00
Very High (>31)	4	50.00	6.00

This pedagogical framework prioritizes practical relevance, providing a nuanced understanding of how varying levels of engagement influence academic outcomes. Each group's average final exam score was then compared to evaluate the potential impact of chatbot usage on academic performance.

Students who made moderate use of the chatbot (11–20 interactions) achieved an average score of 6.58, which was significantly higher than those with low use (1–10 interactions), whose average score was 4.92. Students with high use (21–30 interactions) reached the highest average score (8.00), although this group was small, making the results less generalizable. The group with very high use (>31 interactions) obtained an average score of 6.00, suggesting that excessive chatbot use does not guarantee better academic outcomes.

This pattern indicates that moderate and strategic use of the chatbot aligns with optimal learning outcomes, while over-reliance may hinder independent study efforts.

Qualitative interviews provided deeper insights. Some students in the very high use group acknowledged that they relied too heavily on the chatbot for preparation, neglecting other forms of study and deeper engagement with materials. One student noted: "I think I relied too much on the chatbot to prepare, and in the end, I neglected other study methods, like reviewing my notes or digging deeper into the materials, which may have affected my results." This excessive dependence on the chatbot seems to have limited their academic performance, as they failed to integrate it as a complement to more traditional study strategies.

The Pearson correlation coefficient analysis yielded a value of r = 0.186, $p = 0.05$, indicating a weak positive relationship between the number of interactions and exam scores. This analysis used continuous data for both variables (number of interactions and exam scores), which aligns with the assumptions of the Pearson correlation. Given the data's normality and continuous nature, Pearson was deemed appropriate. Future studies with ordinal or non-normal data may consider the Spearman correlation as an alternative.

These findings suggest that while the chatbot can enhance learning when used strategically, its effectiveness diminishes with excessive reliance. Qualitative interviews reinforce this interpretation, as many students noted that although the chatbot was useful as a complementary tool, "it couldn't replace reviewing notes or more in-depth study". These insights emphasize the importance of integrating AI tools strategically to support rather than replace traditional study methods.

In summary, the quantitative and qualitative data show that the chatbot was mostly used to consolidate theoretical knowledge, particularly in the categories of definitions and detailed explanations. Moderate and balanced use of the tool appears to be associated with better academic outcomes, while excessive use offers no additional advantages and may reflect a reliance on the chatbot that limits other study methods. Overall, the chatbot is valuable as a complementary resource, but it cannot replace other study strategies.

3.3. Perceptions of the Chatbot's Effectiveness and Usefulness

A key finding from the initial survey revealed that 51.2% of students had no prior experience using this technology in an educational setting. Therefore, it was crucial to analyze students' perceptions after using the chatbot in light of their initial expectations. The expectations of a significant portion of users were based on a lack of familiarity, which may influence their final perception. This analysis helps identify possible discrepancies between what was expected and what was experienced, which in turn affects overall satisfaction with the tool. If expectations are met or exceeded, students are more likely to continue using it. On the other hand, a disconnect between expectations and perceived usefulness can lead to frustration, negatively impacting their evaluation.

Thus, we will first address the students' initial expectations before the intervention and the perceived usefulness after using the chatbot. "Expectation" refers to students' anticipated usefulness of the chatbot before the intervention phase, as measured through a pre-intervention survey. This included their thoughts on how the tool could support their learning and exam preparation. "Perceived usefulness", in contrast, represents their evaluation of the chatbot's effectiveness after the intervention, as captured in a post-intervention survey.

The impact of the chatbot on critical learning areas such as concept comprehension, content review, and exam preparation was analyzed. The following Table 3 summarizes the results of student opinions gathered before and after the intervention:

Table 3. Comparison of students' expectations and perceived usefulness of the chatbot.

Chatbot Usefulness Aspect	Expectations (%)	Perceived Usefulness (%)
Specific Question Resolution	84.2	91.4
Concept Comprehension	100	95.7
Practical Examples	42.1	61.4
Exam Training	52.6	45.4
Content Review	73.7	42.9
Motivation and Study Tips	10.5	0

The area of specific question resolution received the highest ratings, with a perceived usefulness of 91.4%, surpassing initial expectations by 7.2%. This indicates that the chatbot was particularly effective when students asked clear and precise questions, allowing for direct and satisfactory responses. Comments such as "the chatbot provided the explanation I needed to understand a question and keep studying" highlight its ability to resolve specific issues and facilitate the study process.

In terms of concept comprehension, although there was a slight decrease from 100% to 95.7%, the chatbot still maintained a positive rating. Students praised the clarity of the responses, stating that the tool was helpful in defining key concepts. However, the slight drop suggests that some users expected a higher level of depth in the responses, likely influenced by the lack of precision in the questions asked, which impacted the interaction and the quality of the responses received.

The area of practical examples experienced a significant increase in perceived usefulness, rising from 42.1% to 61.4%. This increase reflects that when students formulated their questions correctly, the chatbot provided practical examples that helped consolidate learning. As one student explained: "The good thing about the chatbot is that if you have doubts about a concept, you can ask for examples, and it becomes much clearer."

On the other hand, in exam simulation, the perceived usefulness decreased from 52.6% to 45.4%. This result contrasts with the technical tests, which showed an 88% effectiveness in exam simulations. Upon reviewing interactions, it was found that the issue lay in the low quality of the questions asked and the ineffective use of interaction strategies. The lack of precision in the queries prevented the chatbot from generating effective exam simulations, limiting its potential in this area. This highlights the importance of students' proficiency in prompt engineering to maximize the chatbot's performance.

In content review, perceived usefulness also saw a significant drop, from 73.7% to 42.9%. One student commented: "I expected the chatbot to help me review more, but the responses weren't as useful." Analyzing the interactions confirmed that the main obstacle was the inability to formulate questions that delved deeply into the topics, which affected the quality of the responses. Additionally, some students did not understand that the chatbot was a complementary tool to be used alongside more traditional study strategies. This lack of understanding may have limited the chatbot's effectiveness in the review process.

Perceptions of motivation and study tips fell from an already low initial expectation of 10.5% to 0% after using the chatbot. This result shows that the tool did not meet expectations in this area, though it is important to note that the chatbot was not specifically designed to provide motivation or advice on study habits. This gap represents an area for improvement in educational chatbots. Incorporating features aimed at motivating and supporting study organization could significantly improve student engagement. By

providing more comprehensive support, including both emotional and strategic guidance, chatbots could become even more valuable learning tools.

Overall, the results show that the chatbot was most effective in areas such as solving specific questions, understanding concepts, and exemplifications, as long as the questions were clear and well-structured. However, declines in areas such as content review and exam simulation reveal that a lack of precision in the questions negatively affected perceived usefulness. Furthermore, the lack of functionality related to motivation and study tips showed that the chatbot did not meet these important needs for some students, which could be improved in future versions.

In conclusion, the results highlight that students' perceptions of the chatbot's usefulness are heavily influenced by their ability to formulate effective questions, meaning their mastery of prompt engineering. Although students received a preparatory session, it was not enough to fully maximize the chatbot's potential. The analysis of interactions reveals a high number of poorly formulated questions, which negatively impacted the quality of responses received and limited the tool's effectiveness in several key areas.

Overall Perception of the Chatbot's Usefulness

When asked whether they considered the chatbot to be an advantage for course preparation, the majority of students expressed a positive perception. Using a Likert scale (1 = not useful at all, 5 = extremely useful), 42.9% gave a score of 4, indicating that the chatbot was seen as a valuable, though not indispensable, resource. Another 35.7% rated the tool with the maximum score (5), suggesting that a considerable portion of the students found the chatbot very useful for their learning. However, 21.4% gave it an intermediate score (3), and 14.2% gave low scores (1 or 2), indicating that a minority did not find the chatbot particularly useful.

Interviews support these results. Some students highlighted the quick response time as one of the tool's main strengths. One student noted: "It's a very useful tool for resolving doubts quickly." Another positive aspect mentioned was the autonomy the chatbot provided, allowing them to resolve doubts without relying directly on teachers. One student remarked: "The chatbot allows me to organize my study on my own without depending so much on the teachers and tutoring." Additionally, several students said the chatbot enabled them to ask questions they wouldn't feel comfortable asking in class due to embarrassment. One said: "Sometimes I don't dare ask questions in class, but with the chatbot, I can ask anything without feeling bad", indicating that the tool reduced emotional barriers in the learning process.

Another important finding is that both teachers and students observed a reduction in the need for personal tutoring and email queries. This highlights the chatbot's role in fostering student autonomy, allowing them to manage their own learning and seek immediate answers to their questions without constant teacher intervention. This greater independence contributed to students feeling more confident in their exam preparation, allowing teachers to focus their tutoring on more complex topics or students who needed more personalized support.

However, some students pointed out limitations, such as the depth of responses on more complex topics. As one student stated: "The chatbot isn't very useful for more difficult questions, where I need a deeper explanation." This type of feedback reflects that while the chatbot was seen as a valuable tool, its usefulness was limited when it came to questions that required more detailed explanations or a more personalized pedagogical approach.

4. Discussion and Conclusions

This study evaluated the impact of a subject-specialized educational chatbot on exam preparation for second-year students in the Bachelor's Degree in Early Childhood Education program. The results reveal that the chatbot was highly effective in resolving specific questions and improving concept comprehension, exceeding initial expectations in these areas. These findings align with previous studies that highlight the usefulness of chatbots in personalizing learning and providing immediate feedback, which reduces anxiety and improves academic performance (Ng & Leung, 2024; Holmes et al., 2019; Deng & Yu, 2023).

However, areas for improvement were identified, particularly in exam simulation and content review, where initial expectations were not met. The perceived usefulness in these areas was lower (42.9% for content review and 45.4% for exam simulation), suggesting a disconnect between what students expected and what the chatbot could offer. This limitation appears to be related to a lack of proficiency in prompt engineering, as students were unable to formulate questions effectively to obtain deeper and more applied responses. As Knoth et al. (2024) point out, success in interacting with AI systems like chatbots depends significantly on the user's ability to generate clear and precise queries. In this educational intervention, the initial training in these skills was not sufficient to maximize the chatbot's use. Future studies should provide more comprehensive training in prompting techniques and evaluate its impact on chatbot interactions.

Another important aspect is the relationship between chatbot use and academic performance. Students who made moderate use of the tool achieved better exam results (average score of 6.58) than those with low use (4.92). However, excessive use of the chatbot (> 31 interactions) did not lead to additional improvements, suggesting an over-reliance on the tool without a complementary focus on more traditional study strategies. These findings are consistent with research by Zhang et al. (2024), which shows that intensive use of educational technology does not always lead to better outcomes if not combined with traditional learning methods.

Additionally, the study highlighted the chatbot's ability to foster study autonomy. Students valued the tool's immediacy and accessibility, which reduced their dependence on tutoring and email queries. This increase in autonomy aligns with the principles of self-directed learning, as observed in previous studies on the use of chatbots in higher education (Labadze et al., 2023). However, the lack of functionalities that support motivation and study organization was noted as a significant limitation by students, suggesting that educational chatbots still have room for improvement in providing more comprehensive support.

The Technology Acceptance Model (TAM) serves as an insightful framework for analyzing these findings. According to TAM, the adoption of new technologies hinges on two key factors: perceived usefulness and perceived ease of use (Davis, 1989). In this study, students demonstrated a high level of satisfaction with the chatbot's ability to address specific queries and enhance their understanding of concepts. However, the identified limitations highlight areas for enhancing perceived ease of use. Addressing this challenge could involve two complementary strategies: first, refining large language models (LLMs) to better interpret and respond to student requests with higher accuracy; and second, equipping users with improved skills in "prompt engineering" to effectively interact with the chatbot. These efforts can jointly contribute to a more effective and productive user experience.

5. Limitations and Future Research

While this study provides valuable insights into the impact of specialized educational chatbots, certain limitations should be acknowledged. The relatively small sample size,

confined to a single academic context, may limit the generalizability of the findings to other courses, institutions, or educational levels. Furthermore, the students' proficiency in question formulation (prompt engineering) significantly influenced the chatbot's effectiveness, underscoring the need to explore more comprehensive training strategies in this area. Future research could focus on evaluating the impact of more extensive training programs in AI interaction, as well as developing chatbots with enhanced capabilities for interpreting less structured queries. Additionally, exploring the integration of features that support motivation, study organization, and students' emotional well-being would be relevant. Finally, expanding the analysis to multicultural contexts and more diverse cohorts could provide a broader understanding of the educational applications of chatbots.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Research Ethics and Animal Welfare Committee (CEIBA) of the University of La Laguna (protocol code: CEIBA2024-3499; date of approval: 25 October 2024).

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

Data Availability Statement: The data supporting the findings of this study are not publicly available due to privacy and ethical restrictions. For further inquiries, please contact the corresponding author.

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

References

Abas, M. A., Arumugam, S. E., Yunus, M. M., & Rafiq, K. R. M. (2023). ChatGPT and personalized learning: Opportunities and challenges in higher education. *International Journal of Academic Research in Business and Social Sciences*, *13*, 3936–3945. [CrossRef] [PubMed]

Agbong-Coates, I. J. (2024). ChatGPT integration significantly boosts personalized learning outcomes: A Philippine study. *International Journal of Educational Management and Development Studies*, *5*, 165–186. [CrossRef]

Alenezi, M. (2023). Digital learning and digital institution in higher education. *Education Sciences*, *13*, 88. [CrossRef]

Boone, H. N., & Boone, D. A. (2012). Analyzing likert data. *Journal of Extension*, *50*(2), 48. [CrossRef]

Chang, D. H., Lin, M. P., Hajian, S., & Wang, Q. Q. (2023). Educational design principles of using AI chatbot that supports self-regulated learning in education: Goal setting, feedback, and personalization. *Sustainability*, *15*, 12921. [CrossRef]

Chen, K., Tallant, A. C., & Selig, I. (2024). Exploring generative AI literacy in higher education: Student adoption, interaction, evaluation, and ethical perceptions. *Information and Learning Science*, *ahead-of-print*. [CrossRef]

Chukwuere, J. E. (2024). The future of generative AI chatbots in higher education. *arXiv*. [CrossRef]

Cramarenco, R., Burcă-Voicu, M., & Dabija, D. (2023). Student perceptions of online education and digital technologies during the COVID-19 pandemic: A systematic review. *Electronics*, *12*, 319. [CrossRef]

Dabis, A., & Csáki, C. (2024). AI and ethics: Investigating the first policy responses of higher education institutions to the challenge of generative AI. *Humanities & Social Sciences Communications*, *11*, 1006. [CrossRef]

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *Management Information Systems Research Center*, *13*(3), 319–340. [CrossRef]

Deng, X., & Yu, Z. (2023). A meta-analysis and systematic review of the effect of chatbot technology use in sustainable education. *Sustainability*, *15*, 2940. [CrossRef]

Dhawan, S. (2022). Higher education quality and student satisfaction: Meta-analysis subgroup analysis and meta-regression. *Metamorphosis: A Journal of Management Research*, *21*(1), 48–66. [CrossRef]

Diederich, S., Lembcke, T., Brendel, A. B., & Kolbe, L. M. (2021). Understanding the impact that response failure has on how users perceive anthropomorphic conversational service agents: Insights from an online experiment. *AIS Transactions on Human-Computer Interaction*, *13*(1), 82–103. [CrossRef]

España. (2018). Ley orgánica 3/2018, de 5 de diciembre, de protección de datos personales y garantía de los derechos digitales. *Boletín Oficial del Estado*, *294*, 119788–119857. Available online: https://www.boe.es/buscar/act.php?id=BOE-A-2018-16673 (accessed on 1 September 2024).

Følstad, A., Nordheim, C. B., & Bjørkli, C. A. (2018, September 24). *What makes users trust a chatbot for customer service? An exploratory interview study*. International Conference on Internet Science (pp. 194–208), St. Petersburg, Russia. [CrossRef]

Guttierrez-Aguilar, O., Huarsaya-Rodriguez, E., & Duche-Pérez, A. (2024). The mediating effect of academic performance on ChatGPT satisfaction in university students. In G. F. Olmedo Cifuentes, D. G. Arcos Avilés, & H. V. Lara Padilla (Eds.), *Emerging research in intelligent systems* (Volume 903, pp. 155–168). Springer. [CrossRef]

Halaweh, M. (2023). ChatGPT in education: Strategies for responsible implementation. *Contemporary Educational Technology*, *15*, ep421. [CrossRef]

Hang, C. N., Wei Tan, C., & Yu, P. -D. (2024). MCQGen: A large language model-driven mcq generator for personalized learning. *IEEE Access*, *12*, 102261–102273. [CrossRef]

Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. *European Journal of Education*, *57*(4), 542–570. [CrossRef]

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Centre for Curriculum Redesign.

Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). Likert scale: Explored and explained. *Current Journal of Applied Science and Technology*, *7*(4), 396–403. [CrossRef]

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., & Gasser, U. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. [CrossRef]

Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computer and Education: Artificial Intelligence*, *6*, 100225. [CrossRef]

Kvale, S. (2007). *Doing interviews*. SAGE Publications. [CrossRef]

Labadze, L., Grigolia, M., & Machaidze, L. (2023). Role of AI chatbots in education: Systematic literature review. *International Journal of Educational Technology in Higher Education*, *20*, 56. [CrossRef]

Lin, M. P., & Chang, D. (2023). CHAT-ACTS: A pedagogical framework for personalized chatbot to enhance active learning and self-regulated learning. *Computer and Education: Artificial Intelligence*, *5*, 100167. [CrossRef]

Martins, L., Fernández-Ferrer, M., & Puertas, E. (2024). Analysing conversation pathways with a chatbot tutor to enhance self-regulation in higher education. *Education Sciences*, *14*, 590. [CrossRef]

Mat Dangi, M. R., & Mohamed Saat, M. (2021). 21st-century educational technology adoption in accounting education: Does institutional support moderate accounting educators' acceptance behavior? *International Journal Academic Research in Business and Social Sciences*, *11*(1), 8288. [CrossRef] [PubMed]

Mironova, J., Riashchenko, V., Kinderis, R., Djakona, V., & Dimitrova, S. (2024). Ethical concerns in using generative tools in higher education: Cross-country study. *Environment Technology Resources*, *2*, 444–447. [CrossRef]

Mungai, B. K., Omieno, P. K. K., Egessa, M., & Manyara, P. N. (2024). AI chatbots in LMS: A pedagogical review of cognitive, constructivist, and adaptive principles. *Engineering and Technology Journal*, *9*(8), 4709–4715. [CrossRef]

Ng, D. T. K., & Leung, J. K. L. (2024). Empowering student self-regulated learning and science education through ChatGPT: A pioneering pilot study. *British Journal of Educational Technology*, *55*(4), 1328–1353. [CrossRef]

Patton, M. Q. (2015). *Qualitative research & evaluation methods* (4th ed.). SAGE Publications.

Rahman, M. M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences*, *13*(9), 5783. [CrossRef]

Sánchez-Vera, F. (2024). Developing effective educational chatbots with GPT: Insights from a pilot study in a university subject. *Trends in Higher Education*, *3*(1), 155–168. [CrossRef]

Schei, O. M., Møgelvang, A., & Ludvigsen, K. (2024). Perceptions and use of AI chatbots among students in higher education: A scoping review of empirical studies. *Education Science*, *14*, 922. [CrossRef]

Schwab, K. (2017). *The fourth industrial revolution*. World Economic Forum.

Su, J. H., & Yang, W. P. (2023). Unlocking the power of ChatGPT: A framework for applying generative AI in education. *ECNU Review of Education*, *6*(3), 355–366. [CrossRef]

Tan, C. W. (2023). Large language model-driven classroom flipping: Empowering student-centric peer questioning with flipped interaction. *arXiv*. [CrossRef]

Valverde-Berrocoso, J., Acevedo-Borrega, J., & Cerezo-Pizarro, M. (2022). Educational technology and student performance: A systematic review. *Frontier in Education*, *7*, 916502. [CrossRef]

Winkler, R., & Söllner, M. (2018). Unleashing the potential of chatbots in education: A state-of-the-art analysis. *Academy of Management Annual Meeting (AOM)*, *15*, 44. [CrossRef]

Yigci, D., Eryilmaz, M., Yetisen, A. K., Tasoglu, S., & Ozcan, A. (2024). Large language model-based chatbots in higher education. *Advanced Intelligent Systems*, 2400429. [CrossRef]

Zhang, X., Jiang, H., Qiao, Z., & Li, P. (2024). Students' response to ChatGPT: An adaptive technology-to-performance model. *Journal of Computer Information Systems*, 1–18. [CrossRef]

education sciences

MDPI

Systematic Review

Integrating Artificial Intelligence into the Cybersecurity Curriculum in Higher Education: A Systematic Literature Review

Jing Tian

NUS-ISS, National University of Singapore, Singapore 119615, Singapore; tianjing@nus.edu.sg

Abstract: Background: To understand the state of the art of how artificial intelligence (AI) and cybersecurity are taught together, this paper conducts a systematic literature review on integrating AI into the cybersecurity curriculum in higher education. Methods: The peer-reviewed works were screened from major databases published between 2020 and 2025. Integrating AI and cybersecurity typically requires new learning designs. To address this gap in higher education, this review is organized by three categories of research questions: (1) who we teach (audiences and delivery modes), (2) what we teach (related AI topics and cybersecurity topics and how they are integrated), and (3) how we teach (instructional activities and tools used in teaching). Results: The course delivery is mostly face-to-face. The course curricula focus mostly on perception AI. Teaching methods are active and practical, with hands-on labs, interactive tasks, and game-based activities, supported by hardware, programming notebooks, and interactive visualizations. Conclusion: This paper provides the state of the art of integrating AI into the cybersecurity curriculum in higher education, actionable recommendations, and implications for further research. Therefore, it is relevant and transferable for instructors in the field of artificial intelligence education and cybersecurity education.

Keywords: artificial intelligence education; cybersecurity education; systematic literature review

1. Introduction

Cybersecurity education must evolve alongside the rapid evolution of artificial intelligence (AI) in a practice-oriented curriculum that develops both AI expertise and security expertise (Beuran et al., 2022; Cusak, 2023; Jimenez & O'Neill, 2023). In industrial deployments, professionals need competencies in machine learning and secure AI deployment, not only to defend AI-enabled systems but also to leverage AI for threat detection (Bhuiyan & Park, 2025; Zivanovic et al., 2024). This has increased the gap between university outcomes and workplace expectations, particularly in hands-on skills and cross-disciplinary knowledge (Bendler & Felderer, 2023; Tian, 2025). To close this gap, an integration is needed to embed AI into the cybersecurity curriculum.

There are three major types of AI techniques, including *perception* AI, *generative* AI, and *agentic* AI, each of which has distinct capabilities and risk profiles that require different mitigation. Perception AI analyzes sensor data in critical systems such as autonomous driving. It recognizes road context (e.g., traffic signs, road conditions, and obstacles) in real time and triggers actions such as proceeding or urgent braking. These models are particularly vulnerable to adversarial examples deliberately crafted to induce misclassification (Afolabi & Adewale Akinola, 2024). Generative AI produces new content in response to user prompts (e.g., an e-commerce chatbot that handles customer inquiries). It faces

unique threats such as jailbreaks, which elicit harmful outputs, and prompt injection, which triggers intended behavior and instructions (Das et al., 2025). Agentic AI orchestrates end-to-end workflows through collaborating agents that sense, reason, and act. In enterprise settings, such systems may manage orders, make purchases, and coordinate supply chains. Their attack surface and failure modes differ fundamentally from those of perceptive and generative systems, introducing new cybersecurity challenges around tool use, autonomy, and authorization (Deng et al., 2025).

The differences across various AI paradigms motivate a tighter integration of AI and cybersecurity in higher education. There are two complementary strategies: *security for AI* and *AI for security*. Lessons learned from securing AI systems inform how we responsibly embed models into security operations. In turn, operational use in defense surfaces new attacks and governance needs, tightening the feedback loop between security for AI and AI for security. *Security for AI* emphasizes safeguarding AI systems through governance, policy, and technical controls that mitigate risks and manage threats across data, model development, deployment, and operations (Jaffal et al., 2025). *AI for security* applies machine learning and deep learning methods to strengthen protective technologies (e.g., network defense, endpoint protection, and email filtering), accelerating detection and response and augmenting analyst capacity (Okdem & Okdem, 2024).

Effective course delivery relies on instructional methods and digital tooling (Ali et al., 2024; Lozano & Blanco Fontao, 2023; Michel-Villarreal et al., 2023). Integrating AI and cybersecurity typically requires new learning designs, especially hands-on activities, and appropriate tools such as programming environments, curated datasets, sandboxes or simulation platforms, and visualization utilities that make model behavior and security mechanisms transparent.

To address the above-identified gaps in higher education, this paper conducts a systematic literature review on integrating AI into the curriculum of cybersecurity. The review is organized around three groups of research questions focusing on course context, course curriculum design, and the course's instructional activities and tools. This paper makes two key contributions to the literature on AI education and cybersecurity education.

- First, it systematically synthesizes studies from multiple major databases (Scopus, IEEE Xplore, and Web of Science), offering a broader and more representative view than prior reviews that were limited to specific sources or course formats. Furthermore, it provides the most up-to-date perspective on the field by covering the period from 2020 to 2025.
- Second, it adopts an integrated lens that examines three categories of six research questions, covering course context, course curriculum, and course instructional activities and tools.

The rest of this paper is organized as follows. Section 2 introduces the relevant research works and highlights the difference between them and this paper. Then, Section 3 presents the three categories of six research questions covered in this study, including course context, course curriculum, and course instruction. It also presents the systematic literature search process using a PRISMA framework (Page et al., 2021). The research findings are presented in Section 4, followed by discussions on the key observations, recommendations, and limitations of this study in Section 5. Finally, Section 6 provides the conclusion of this paper.

2. Related Works

This section briefly describes relevant review studies on AI and cybersecurity and then highlights the difference between them and this paper. Laato et al. (2020) investigate how cybersecurity has been taught in online courses by conducting a systematic review of

prior works on *massive open online courses* (MOOCs). They find only a limited number of peer-reviewed evaluations of individual cybersecurity MOOCs and highlight the absence of focused treatment of AI applications in cybersecurity education. The article by Dewi et al. (2024) provides a bibliometric analysis of 637 articles; it maps the research landscape on AI, cybersecurity, and education. It identifies thematic clusters and concludes that the use of AI in cybersecurity education and awareness programs remains underdeveloped. The study by Svabensky et al. (2020) synthesizes a decade of cybersecurity education research presented at major computing education conferences. It shows that while many technical and human-centric topics are addressed, few studies provide reusable materials or datasets. The study by Aris et al. (2022) addresses the challenge of updating curricula by proposing a structured method for integrating AI into cybersecurity education. By analyzing around 300 papers from major cybersecurity-related conferences, it demonstrates the growing efforts of AI in security research and argues for its inclusion in teaching. The article by Lasisi et al. (2022) reviews undergraduate cybersecurity programs to assess the presence of AI-related content. It reports that despite the increasing role of AI in enabling advanced cyberattacks, AI courses are lacking, revealing a gap in preparing working professionals for this emerging skill gap. The study by Weitl-Harms et al. (2023) provides a review of 74 papers applying a gamification strategy in cybersecurity operations education in undergraduate coursework.

Unlike these existing works, this paper conducts a literature review with a focus on the integration of AI and cybersecurity in education. Its fundamental difference with existing works is summarized in Table 1. Firstly, while earlier studies typically limited themselves to a single type of data source, such as selected computer science conferences, this paper systematically studies three databases (Scopus, IEEE Xplore, and Web of Science). Secondly, previous reviews often covered past decades or a narrower duration. By spanning 2020 to 2025, this paper captures the most up-to-date research works, including the integration of emerging AI technologies into cybersecurity education. Lastly, earlier works tended to emphasize specific focus, such as MOOC evaluations, bibliometric mapping, or curriculum design. In contrast, this paper provides a comprehensive study on how AI and cybersecurity are taught together, including the course curriculum designs, instructional activities, and the use of digital tools in teaching.

Table 1. A comparison between relevant works and this review paper.

Reference	Year	Source	Number of Studies	Coverage (Year)	Remark
Laato et al. (2020)	2020	ACM, IEEE Xplore, Springer, DBLP	15	2003–2019	MOOC course only
Svabensky et al. (2020)	2020	Selected computer science conferences SIGCSE and ITiCSE	71	2010–2019	General cybersecurity education
Aris et al. (2022)	2022	Selected cybersecurity conferences only	300	2016–2021	Curriculum only
Lasisi et al. (2022)	2022	Undergraduate courses in USA	24	2020	Curriculum only
Weitl-Harms et al. (2023)	2023	ACM, Taylor & Francis, Scopus, IEEE Xplore	74	2007–2022	Gamification only
Dewi et al. (2024)	2024	Scopus	637	2019–2024	A bibliometric analysis
Ours		Scopus, IEEE Xplore, Web of Science	16	2020–2025	A study on course context, curriculum, instruction, and tools

3. Methodology

3.1. Literature Search Process

We conducted a systematic literature search following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework (Page et al., 2021). The search targeted relevant research in the field of integrated AI and cybersecurity teaching and was performed across three major academic databases: Scopus, IEEE Xplore, and Web of Science. These databases were selected for their comprehensive coverage and relevance to this study.

Due to differing search syntax across databases, customized queries were crafted for each database. To ensure relevance and quality, we applied the following inclusion criteria: articles had to be (i) published between 2020 and 2025, (ii) written in English, and (iii) published in peer-reviewed journals or conference proceedings. We selected the 2020–2025 period to capture studies published during the period of fastest methodological and curricular change in AI-enabled cybersecurity. Table 2 provides a detailed breakdown of the search strings used for each database.

Table 2. A list of search syntax used in various databases.

Database	Search Syntax
Scopus	ABS ((Cybersecurity) AND (AI OR "artificial intelligence" OR "machine learning") AND (teaching OR education))
IEEE Xplore	("Abstract":Cybersecurity) AND ("Abstract":AI OR "Abstract":"artificial intelligence" OR "Abstract":"machine learning") AND ("Abstract":teaching OR "Abstract":education)
Web of science	AB = (cybersecurity AND (AI OR "artificial intelligence" OR "machine learning") AND (teaching OR education))

The initial search in August 2025 returned a total of 263 records after removing duplicates. We employed a multi-phase screening process to determine the final selection of studies, as illustrated in Figure 1. This involved (i) scope review (e.g., relevance to teaching and education) and (ii) manual abstract screening and full-text screening (e.g., focus on the integration of AI and cybersecurity). In the manual abstract screening process, we excluded 184 papers that were not teaching studies, 36 papers that were teaching papers but not related to cybersecurity, and 26 papers that were traditional cybersecurity teaching papers without the integration of AI. Then, in the manual full-text screening process, we excluded 5 review papers and 4 studies focusing on K-12 education. On the other hand, with the additional web search, we managed to find 6 papers. We further complemented database searching with citation chasing (snowballing) to find 5 papers. Then, we excluded 3 review papers from them. In summary, 16 papers that met all inclusion criteria were selected for in-depth analysis in this study. The annual distribution of these 16 papers is presented in Table 3, and their brief description is provided in Table 4.

Table 3. Annual distribution of papers (2020–2025) covered in this paper.

Year	2020	2021	2022	2023	2024	2025	Total
Journal papers	0	1	0	0	4	3	8
Conference papers	2	0	2	2	1	1	8

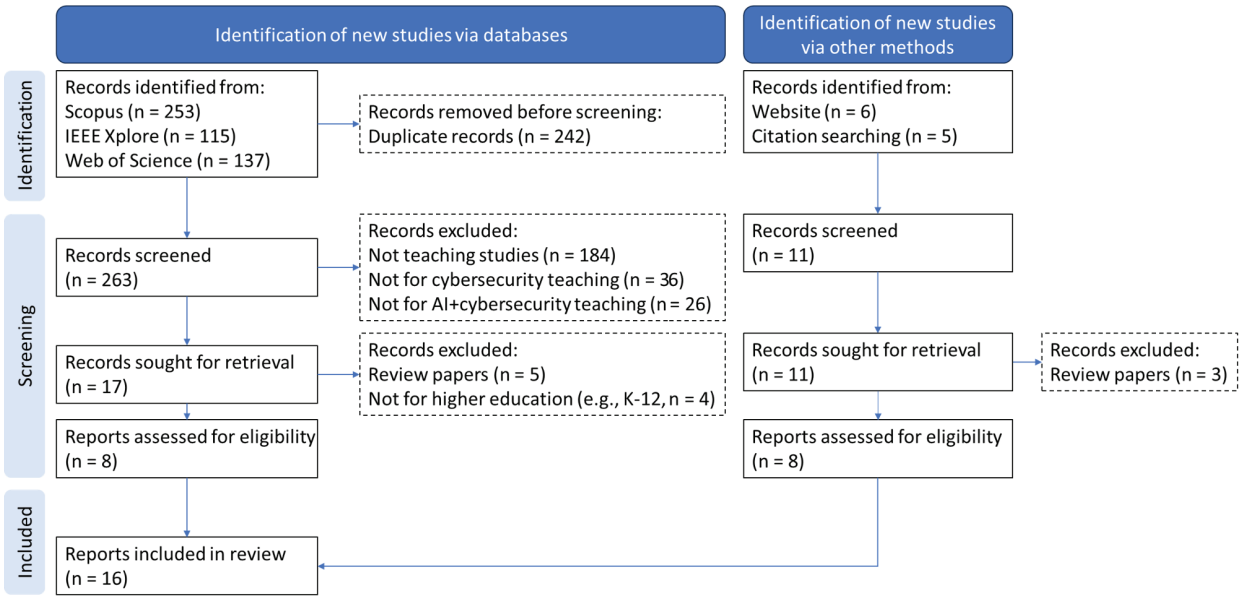


Figure 1. The PRISMA flow diagram used in this paper.

Table 4. A brief overview of papers covered in this literature review.

Reference	A Short Description
Alexander et al. (2024)	It teaches vulnerabilities in AI systems via hands-on, inquiry-based labs.
Apruzzese et al. (2023)	It provides teaching and practitioner guidance that is focused on threat models and evaluation.
Arai et al. (2024)	It teaches AI security concepts by leveraging game design and mechanics that embed adversarial machine learning ideas into engaging play.
Brito et al. (2025)	It provides a set of recommended topics, sequencing, and course structures that integrate machine learning methods into security education.
Calhoun et al. (2022)	Focuses on teaching hardware-level threats and defenses affecting machine learning systems. It provides a suite of hands-on lab modules that expose students to issues such as accelerator vulnerabilities.
Debello et al. (2023)	A practical way to integrate artificial intelligence into cybersecurity programs so students encounter real, security-relevant AI problems.
Farahmand (2021)	It integrates AI cybersecurity research into engineering and computer science education by structuring interdisciplinary courses and projects that connect students with active research problems and security-relevant AI work.
Lo et al. (2022)	It provides a low-cost portable lab kit that can support authentic machine learning workflows in cybersecurity contexts, where students collect data, train models, and evaluate results on security-relevant problems.
Mathews et al. (2025)	It designs a course to equip students with a foundational understanding of Generative AI and explore their applications in cybersecurity, by a combination of lectures, hands-on projects, and industry guest lectures.
Okpala et al. (2025)	A practical lab platform is created to offer experiential learning for students outside computing disciplines. Participants explore foundational AI ideas and how these can be used to identify online harassment.
Payne and Glantz (2020)	It shares how to teach adversarial machine learning to security professionals through a course design with learning objectives and hands-on exercises.
Pusey et al. (2024)	It studies which prior knowledge indicators can assess student preparedness for modules involving AI-supported malware investigation, aiding educators in shaping effective teaching strategies.

Table 4. *Cont.*

Reference	A Short Description
Salman (2024)	A pedagogical framework for blending AI topics into cybersecurity education. It provides a set of design principles and case studies that show how to implement the framework across contexts, including learning outcomes and assessment choices.
Shahriar et al. (2020)	A portable machine learning for cybersecurity lab platform. It provides a set of realistic cases and exercises that guide students from problem framing through model building and reflection, enabling consistent, hands-on practice.
Wei-Kocsis et al. (2024)	A proactive, collaborative pedagogy that connects AI concepts with cybersecurity practice via a learning model combining teamwork, community engagement, and early risk awareness.
You et al. (2025)	An interactive visualization that reveals how tiny, structured perturbations can mislead image classifiers. It provides a learning tool that helps students grasp adversarial examples, decision boundaries, and feature sensitivity through direct manipulation and immediate visual feedback.

3.2. Research Questions

This paper examines six research questions from three dimensions. The statements, motivations, and pedagogical gaps and aims of these research questions are provided as follows.

- Course context-related research questions.
  - RQ1. Who are the target audiences of courses?
    Motivation: Identifying the intended learners clarifies the background knowledge, skill gaps, and professional needs the curriculum is designed to address.
    Pedagogical gap and aim: Calibrate learning objectives, scaffolding, and assessment to learner readiness and context.
  - RQ2. What delivery modes are adopted in teaching?
    Motivation: Understanding whether courses are offered face-to-face, online, or in hybrid formats provides insight into the accessibility and scalability of instruction.
    Pedagogical gap and aim: Match the delivery modality to learning outcomes (e.g., labs needing hands-on time vs. asynchronous theory), while considering the resource constraints.
- Course curriculum-related research questions.
  - RQ3. What AI topics are included in the curriculum?
    Motivation: Mapping the range of AI content helps reveal the breadth of technical coverage in current educational practice, particularly on the emerging AI technologies.
    Pedagogical gap and aim: Ensure up-to-date topic sequences that build from fundamentals to advanced methods aligned with current practice.
  - RQ4. How are AI and cybersecurity concepts integrated in teaching?
    Motivation: Exploring integration strategies shows whether courses treat AI and cybersecurity separately or promote interdisciplinary learning.
    Pedagogical gap and aim: Promote interdisciplinarity via aligned learning outcomes and iterative tasks that connect AI methods to concrete security problems.
- Course instruction-related research questions.
  - RQ5. What instructional activities and pedagogical approaches are used?
    Motivation: Examining teaching activities (e.g., lectures, labs, and projects) highlights how learning objectives are implemented in practice.
    Pedagogical gap and aim: Adopt evidence-informed designs (scaffolded labs and project-based learning) that cultivate problem-solving and professional practices.

– RQ6. What digital tools support the course delivery?
Motivation: Investigating the tools used (e.g., simulation environments and security platforms) reveals how digital tools facilitate effective learning.
Pedagogical gap and aim: Select and integrate tools that are accessible and aligned with tasks and simulate real-world workflows to enhance the learning outcome.

4. Results

This section presents the results synthesized from the sixteen selected papers in this literature review. Each subsection corresponds to the three categories of research questions outlined in Section 3.2.

4.1. Course Context-Related Research Questions

The first research question is as follows: *RQ1. Who are the target audiences of courses?* To address this question, we examined the learner groups reported in the sixteen papers. This question is important because identifying the intended audience helps to clarify the expected prior knowledge and professional needs that the curriculum is designed to meet. As summarized in Table 5, the most frequently mentioned target learner group is university students, including both undergraduate and postgraduate learners, discussed in eleven studies. One study emphasizes non-computing major students, while another study emphasizes the cybersecurity major. The remaining five papers did not provide explicit information on the target audience. These patterns suggest that AI-based curricula should include various pathways, such as foundational AI literacy for non-computing learners and deeper, practice-oriented tracks for cybersecurity majors, so that prerequisites align with learners' backgrounds.

Table 5. The summarized findings for RQ1 and RQ2. The symbol – indicates that no explicit information was provided in the paper.

Reference	RQ1. Who Are the Target Audiences of Courses?	RQ2. What Delivery Modes Are Adopted in Teaching?
Farahmand (2021)	University undergraduate	Face-to-face
Calhoun et al. (2022)	University undergraduate	Face-to-face
Debello et al. (2023)	University undergraduate	Face-to-face
Lo et al. (2022)	University postgraduate	–
Payne and Glantz (2020)	University undergraduate; university postgraduate	Face-to-face
Pusey et al. (2024)	University undergraduate; university postgraduate	Face-to-face
Wei-Kocsis et al. (2024)	University undergraduate; university postgraduate	Face-to-face
Brito et al. (2025)	University undergraduate; university postgraduate	Face-to-face
You et al. (2025)	University undergraduate; university postgraduate	Face-to-face
Mathews et al. (2025)	University undergraduate; university postgraduate (cybersecurity major)	Face-to-face
Okpala et al. (2025)	University undergraduate; university postgraduate (non-computing major)	Face-to-face
Arai et al. (2024)	–	Online
Shahriar et al. (2020)	–	–
Apruzzese et al. (2023)	–	–
Alexander et al. (2024)	–	–
Salman (2024)	–	–

The second research question is as follows: *RQ2. What delivery modes are adopted in teaching?* This question is equally significant, as the chosen delivery mode (face-to-face, online, or hybrid) affects accessibility and scalability of instruction. Among the sixteen papers reviewed and summarized in Table 5, eleven papers reported face-to-face teaching, one paper described an online course, and the remaining four did not specify the delivery modes. Given the findings that most courses are face-to-face offerings, AI-based curricula should adopt modality-agnostic designs, such as cloud notebooks and virtual labs, to preserve hands-on practice.

4.2. Course Curriculum-Related Research Questions

The third research question is as follows: *RQ3. What AI topics are included in the curriculum?* We consider three major classes of AI, including perception, generative, and agentic, each with distinct capabilities and risk profiles that call for different mitigation. Mapping the range of AI technologies alongside their security implications reveals the breadth of technical coverage in current practice and equips learners with risk-appropriate defenses by design. Across the sixteen papers, the majority of them, fifteen papers, addressed perception AI, while only one paper explicitly taught generative AI, as summarized in Table 6. This imbalance suggests that curricula should be rebalanced beyond perception systems by adding core modules on generative (e.g., prompt injection, jailbreaks, and data leakage) and agentic systems (e.g., tool-use safety and human-in-the-loop).

The fourth research question is as follows: *RQ4. How are AI and cybersecurity concepts integrated into teaching?* We examine two complementary integration strategies: *cybersecurity for AI* and *AI for cybersecurity*. Among the sixteen papers, the coverage was fairly balanced: five papers addressed cybersecurity for AI only, eight focused on AI for cybersecurity only, and the remaining three covered both, as summarized in Table 6. Programs can run the two modules in parallel with integrative capstones. For example, students harden a model and then implement it in a real exercise to exercise both assurance competencies and applied defensive effectiveness.

Table 6. The summarized findings for RQ3 and RQ4. The symbol − indicates that it was not covered in the paper.

Reference	RQ3. What AI Topics Are Included in the Curriculum?	RQ4. How Are AI and Cybersecurity Concepts Integrated into Teaching?	
		Cybersecurity for AI	AI for Cybersecurity
Payne and Glantz (2020)	Perception AI	✓	−
Calhoun et al. (2022)	Perception AI	✓	−
Apruzzese et al. (2023)	Perception AI	✓	−
Alexander et al. (2024)	Perception AI	✓	−
You et al. (2025)	Perception AI	✓	−
Shahriar et al. (2020)	Perception AI	−	✓
Lo et al. (2022)	Perception AI	−	✓
Debello et al. (2023)	Perception AI	−	✓
Pusey et al. (2024)	Perception AI	−	✓
Salman (2024)	Perception AI	−	✓
Brito et al. (2025)	Perception AI	−	✓
Okpala et al. (2025)	Perception AI	−	✓
Farahmand (2021)	Perception AI	✓	✓
Arai et al. (2024)	Perception AI	✓	✓
Wei-Kocsis et al. (2024)	Perception AI	✓	✓
Mathews et al. (2025)	Generative AI	−	✓

4.3. Course Instruction-Related Research Questions

The fifth research question is as follows: *RQ5 What instructional activities and pedagogical approaches are used?* It is important to examine teaching activities (e.g., lectures, labs, and projects) to understand how learning objectives are implemented in practice. Table 7 summarizes instructional activities used across the reviewed papers, which reveals a clear emphasis on active, practice-oriented designs (e.g., hands-on, interactive, game-based, and experiential) with selective use of case studies, scaffolding, and project experiences. Hands-on and project-based learning is prominent, provided as either standalone lab or project work in three studies (Calhoun et al., 2022; Mathews et al., 2025; Payne & Glantz, 2020) and as a combined laboratory plus an independent project in one course (Debello et al., 2023). These activities prioritize skill acquisition in integrative demonstrations. Interactive activities are reported in two papers (Pusey et al., 2024; You et al., 2025), typically to sustain engagement, surface misconceptions early, and support rapid feedback. Game-based learning features both full-course game structures and targeted gamification elements (Arai et al., 2024; Debello et al., 2023; Wei-Kocsis et al., 2024); these approaches aim to increase motivation and provide safe environments for experimenting with offensive/defensive tactics or AI model behaviors. Experiential learning is explicitly implemented in three papers (Alexander et al., 2024; Okpala et al., 2025; Salman, 2024) to support learners' progression from guided exercises to open tasks. Case studies are used to integrate technical detail with contextual judgment (Apruzzese et al., 2023; Shahriar et al., 2020). One course emphasizes authentic learning (Lo et al., 2022), situating activities in realistic professional contexts to bridge classroom and practice. Traditional elements are also effective; homework serves as structured reinforcement (Farahmand, 2021), while a course that combines theoretical and practical components (Brito et al., 2025) illustrates a blended model that pairs conceptual grounding with implementation. Finally, an independent capstone project (Debello et al., 2023) provides a culminating experience for synthesis and evaluation of learning outcomes. These findings suggest that the curriculum should emphasize authentic practice and interactive game/simulation tasks to expose misconceptions, while providing standardized artifacts (reproducible notebooks and datasets) as assessable learning materials.

The sixth research question is as follows: *RQ6. What digital tools support the course delivery?* To effectively deliver the technical content, it is critical to select appropriate tools (e.g., simulation environments and security platforms) to facilitate effective learning. Table 7 summarizes the tools and platforms reported across the papers, spanning hardware setups, programming environments with datasets, visualization, and game-oriented delivery. Several papers offer anchor learning in physical systems to surface real-world constraints and attack surfaces, such as circuit hardware in a very large-scale integration context (Calhoun et al., 2022). A few papers rely on accessible software stacks that enable reproducible work. One study combines a Python (version 3) programming tool with open-source datasets to scaffold implementation and evaluation (Alexander et al., 2024; Debello et al., 2023; Okpala et al., 2025). Open-source tools and online programming platforms are used in Lo et al. (2022); Payne and Glantz (2020); Shahriar et al. (2020), facilitating convenient development and simplified classroom logistics. AI tools, such as ChatGPT, are also studied in Mathews et al. (2025). To unlock the complicated topics, multiple courses adopt block-based programming paired with online visualization webpages (You et al., 2025), which provide interpretable outputs or interactive dashboards that support formative feedback. Game-based or simulation-centric platforms appear as an online web game (Arai et al., 2024) and as an immersive learning setup (Wei-Kocsis et al., 2024). These findings indicate that the curricula should standardize on portable, reproducible stacks, such as cloud notebooks and labs, so learners can practice end-to-end workflows with minimal setup friction.

Table 7. The summarized findings for RQ5 and RQ6. The symbol – indicates that no explicit information was provided in the paper.

Reference	RQ5. What Instructional Activities and Pedagogical Approaches Are Used?	RQ6. What Digital Tools Support the Course Delivery?
Alexander et al. (2024)	Experiential learning	Programming platform (e.g., Colab, Anaconda)
Apruzzese et al. (2023)	Case study	Robot car
Arai et al. (2024)	Game-based learning	Python programming tool; open-source dataset
Brito et al. (2025)	Theoretical materials; practical materials	Block-based programming; online visualization webpage
Calhoun et al. (2022)	Hands-on activity	–
Debello et al. (2023)	Hands-on gamified labs; Capstone project	Drones; Raspberry Pi
Farahmand (2021)	Homeworks	Circuit hardware
Lo et al. (2022)	Authentic learning	–
Mathews et al. (2025)	Project-based learning	AI tools (e.g., ChatGPT)
Okpala et al. (2025)	Experiential learning	Online programming platform
Payne and Glantz (2020)	Hands-on activity	Open-source tool
Pusey et al. (2024)	Interactive workshop	–
Salman (2024)	Scaffolding; experiential learning	Visualization tool
Shahriar et al. (2020)	Case study	–
Wei-Kocsis et al. (2024)	Game-based learning	Online programming platform
You et al. (2025)	Interactive activity	Online web game

5. Discussion

This section summarizes the key observations and provides recommendations for both research and practice.

5.1. Summary of the Key Observations and Actionable Recommendations

Three observations emerge from the three categories of research questions addressed in this study.

5.1.1. Course Context-Related Findings

- *Finding:* Integrating AI and cybersecurity across learner populations is important, with offerings targeting university students; this pattern reveals the current educational landscape and underscores the need for audience-appropriate scaffolding.
- *Educational framework:* This finding supports the constructivist learning framework, where learners build their understanding by connecting new information with their existing knowledge and experiences. Intended learning outcomes and activities should be aligned to distinct learner profiles. For example, a practical lab platform is created to offer experiential learning for non-computing students (Okpala et al., 2025), while cybersecurity students are equipped with a foundational understanding of generative AI to further explore their applications (Mathews et al., 2025).
- *Actionable recommendation:* Instructors should consider extending constructive alignment to the AI–cybersecurity intersection, provide scaffolded prerequisites, and adopt blended delivery so that varied learners can reach aligned outcomes.

5.1.2. Course Curriculum-Related Findings

- *Finding:* Current studies exhibit a balanced emphasis on the two integration strategies, security for AI and AI for security, highlighting cross-disciplinary integration rather than independent treatment.
- *Educational framework:* Constructivist learning treats the two lenses, security for AI and AI for security, as paired problems that support knowledge construction through cognitive conflict and resolution (e.g., risk vs. mitigation and attack vs. defense). This finding complements (Arai et al., 2024), where learners experience damage caused by attacks and the advantages of their countermeasures. In addition, an immersive learning environment is designed in Wei-Kocsis et al. (2024) to motivate the students to explore AI development in the context of real-world cybersecurity scenarios, where AI techniques can be manipulated and evaded, resulting in new security implications.
- *Actionable recommendation:* Instructors could design lab structures that bind security for AI to AI for security. For each topic, we can design mirrored labs (e.g., prompt injection vs. guardrail; data poisoning vs. governance) so that learners can experience the impact of the AI technique and the inherent risk of the AI technique itself.

5.1.3. Course Instruction-Related Findings

- *Finding:* Active pedagogy is prevalent (e.g., hands-on labs/projects, experiential and case-based activities, and visualization to unpack complex concepts), which indicates a need for learning by doing with structured supports that build transferable competencies for AI and cybersecurity practice.
- *Educational framework:* Our finding about active pedagogy aligns with the connectivist learning framework, where learners build understanding through manipulation of tools, datasets, and reflection on experience. This is consistent with immersive and visualization-centric designs in Salman (2024); You et al. (2025), hands-on programming design (Alexander et al., 2024), and even the hardware implementation (Apruzzese et al., 2023; Debello et al., 2023).
- *Actionable recommendation:* Instructors should ground theory in practice. For example, they can start each lecture with a brief real-world artifact (e.g., a prompt injection transcript), state the intended learning outcomes, and then introduce the concepts that explain the artifact. Furthermore, they can provide one-click, sandboxed environments (e.g., Docker/Colab) so learners can run paired attack–defend labs and safely explore AI techniques. Lastly, they can conclude each hands-on activity with a guided reflection, prompting students to articulate what worked, what failed, and how they would improve their approach.

5.2. Limitations

There are a few limitations in this study. This review may be affected by search bias arising from database coverage, indexing delays, English-language restrictions, and the evolving terminology of AI and cybersecurity that could cause relevant studies to be missed by our keywords. Moreover, this review is focused on peer-reviewed journal and conference papers; consequently, it excludes course websites and practitioner reports that may capture cutting-edge practice. This coverage limits generalizability to other settings (e.g., professional training). Given the rapid pace of AI (especially generative AI and emerging agentic AI systems), this paper might only provide insights for very recent innovations and practices.

5.3. Future Research

Through this literature review, we propose three actionable recommendations for instructors in AI education and cybersecurity education. First, we need to align curriculum activities with two lenses, including security for AI (e.g., threat modeling and red-teaming) and AI for security (e.g., anomaly detection and phishing classifiers), and pair them in hands-on exercises to ensure balanced coverage. Second, we need to standardize on accessible and reproducible tooling, such as online programming notebooks, curated open datasets, visualization dashboards, and one-click environments (e.g., Docker or Colab) with starter kits so students focus on learning rather than setup. Last but not least, we need to provide case studies that reflect real threat scenarios to connect technical work to real-world decision-making in this fast-moving domain.

6. Conclusions

This paper presents a systematic literature review on a focused topic of integrating artificial intelligence into cybersecurity education. The findings show that the current practices reach multiple learner groups (from university undergraduate to postgraduate). However, online delivery and hybrid (online and face-to-face) delivery remain underused. The course curricula currently emphasize perception AI only, while emerging areas, like generative and agentic AI systems, are rarely addressed. To effectively integrate AI technology and the cybersecurity content, hands-on activities (e.g., online programming notebooks) and visual explanations are needed to make concepts interactive and explainable. This paper offers a practical reference for instructors seeking to enhance their courses by embedding AI content into the cybersecurity curriculum.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Dataset available on request from the author.

Acknowledgments: The author would like to express their sincere gratitude to the editor and the four reviewers for their insightful suggestions on revising this manuscript.

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

References

Afolabi, A. S., & Adewale Akinola, O. (2024, September 18–20). *Vulnerable AI: A survey*. IEEE International Symposium on Technology and Society, Puebla, Mexico. [CrossRef]

Alexander, R., Ma, L., Dou, Z.-L., Cai, Z., & Huang, Y. (2024). Integrity, confidentiality, and equity: Using inquiry-based labs to help students understand AI and cybersecurity. *Journal of Cybersecurity Education Research and Practice*, *2024*(1), 10. [CrossRef]

Ali, D., Fatemi, Y., Boskabadi, E., Nikfar, M., Ugwuoke, J., & Ali, H. (2024). ChatGPT in teaching and learning: A systematic review. *Education Sciences*, *14*(6), 643. [CrossRef]

Apruzzese, G., Anderson, H. S., Dambra, S., Freeman, D., Pierazzi, F., & Roundy, K. (2023, February 8–10). *Real attackers don't compute gradients: Bridging the gap between adversarial ML research and practice*. 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML) (pp. 339–364), Raleigh, NC, USA. [CrossRef]

Arai, M., Tejima, K., Yamada, Y., Miura, T., Yamashita, K., Kado, C., Shimizu, R., Tatsumi, M., Yanai, N., & Hanaoka, G. (2024). REN-AI: A video game for AI security education leveraging episodic memory. *IEEE Access*, *12*, 47359–47372. [CrossRef]

Aris, A., Rondon, L. P., Ortiz, D., Ross, M., & Finlayson, M. (2022, June 26–29). *Integrating artificial intelligence into cybersecurity curriculum: New perspectives*. ASEE Annual Conference and Exposition (pp. 1–15), Minneapolis, MN, USA. [CrossRef]

Bendler, D., & Felderer, M. (2023). Competency models for information security and cybersecurity professionals: Analysis of existing work and a new model. *ACM Transactions on Computing Education*, *23*2, 1–33. [CrossRef]

Beuran, R., Hu, Z., Zeng, Y., & Tan, Y. (2022). *Artificial intelligence for cybersecurity education and training*. Springer. [CrossRef]

Bhuiyan, S., & Park, J. S. (2025). Cybersecurity threats and mitigation strategies in AI applications. *Journal of The Colloquium for Information Systems Security Education*, *12*(1), 1–7. [CrossRef]

Brito, F., Mekdad, Y., Ross, M., Finlayson, M. A., & Uluagac, S. (2025, February 26–March 1). *Enhancing cybersecurity education with artificial intelligence content*. ACM Technical Symposium on Computer Science Education (pp. 158–164), Pittsburgh, PA, USA. [CrossRef]

Calhoun, A., Ortega, E., Yaman, F., Dubey, A., & Aysu, A. (2022, June 6–8). *Hands-on teaching of hardware security for machine learning*. Great Lakes Symposium on VLSI (pp. 455–461), Irvine, CA, USA. [CrossRef]

Cusak, A. (2023). Case study: The impact of emerging technologies on cybersecurity education and workforces. *Journal of Cybersecurity Education Research and Practice*, *1*, 3. [CrossRef]

Das, B. C., Amini, M. H., & Wu, Y. (2025). Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*, *57*(6), 1–39. [CrossRef]

Debello, J. E., Troja, E., & Truong, L. M. (2023, May 1–4). *A framework for infusing cybersecurity programs with real-world artificial intelligence education*. IEEE Global Engineering Education Conference (pp. 1–5), Kuwait, Kuwait. [CrossRef]

Deng, Z., Guo, Y., Han, C., Ma, W., Xiong, J., Wen, S., & Xiang, Y. (2025). AI agents under threat: A survey of key security challenges and future pathways. *ACM Computing Surveys*, *57*(7), 1–36. [CrossRef]

Dewi, H. A., Candiwan, C., & Sari, P. K. (2024, December 17–19). *Artificial intelligence in security education, training and awareness: A bibliometric analysis*. 2024 International Conference on Intelligent Cybernetics Technology & Applications (pp. 914–919), Bali, Indonesia. [CrossRef]

Farahmand, F. (2021). Integrating cybersecurity and artificial intelligence research in engineering and computer science education. *IEEE Security and Privacy*, *19*(6), 104–110. [CrossRef]

Jaffal, N. O., Alkhanafseh, M., & Mohaisen, D. (2025). Large language models in cybersecurity: A survey of applications, vulnerabilities, and defense techniques. *AI*, *6*(9), 216. [CrossRef]

Jimenez, R., & O'Neill, V. E. (2023). *Handbook of research on current trends in cybersecurity and educational technology*. IGI Global Scientific Publishing. [CrossRef]

Laato, S., Farooq, A., Tenhunen, H., Pitkamaki, T., Hakkala, A., & Airola, A. (2020, July 6–9). *AI in cybersecurity education—A systematic literature review of studies on cybersecurity MOOCs*. IEEE 20th International Conference on Advanced Learning Technologies (pp. 6–10), Tartu, Estonia. [CrossRef]

Lasisi, R. O., Menia, M., Farr, Z., & Jones, C. (2022, May 15–18). *Exploration of AI-enabled contents for undergraduate cyber security programs*. International Florida Artificial Intelligence Research Society Conference (pp. 1–4), Hutchinson Island, FL, USA. [CrossRef]

Lo, D. C.-T., Shahriar, H., Qian, K., Whitman, M., Wu, F., & Thomas, C. (2022, March 2–5). *Authentic learning of machine learning in cybersecurity with portable hands-on labware*. ACM Technical Symposium on Computer Science Education (p. 1153), Providence, RI, USA. [CrossRef]

Lozano, A., & Blanco Fontao, C. (2023). Is the education system prepared for the irruption of artificial intelligence? A study on the perceptions of students of primary education degree from a dual perspective: Current pupils and future teachers. *Education Sciences*, *13*(7), 733. [CrossRef]

Mathews, N., Schwartz, C., & Wright, M. (2025). Teaching generative AI for cybersecurity: A project-based learning approach. *Journal of The Colloquium for Information Systems Security Education*, *12*(1), 1–10. [CrossRef]

Michel-Villarreal, R., Vilalta-Perdomo, E., Salinas-Navarro, D. E., Thierry-Aguilera, R., & Gerardou, F. S. (2023). Challenges and opportunities of generative AI for higher education as explained by ChatGPT. *Education Sciences*, *13*(9), 856. [CrossRef]

Okdem, S., & Okdem, S. (2024). Artificial intelligence in cybersecurity: A review and a case study. *Applied Sciences*, *14*(22), 487. [CrossRef]

Okpala, E., Vishwamitra, N., Guo, K., Liao, S., Cheng, L., Hu, H., Yuan, X., Wade, J., & Khorsandroo, S. (2025). AI-cybersecurity education through designing AI-based cyberharassment detection lab. *Journal of The Colloquium for Information Systems Security Education*, *12*(1), 1–8. [CrossRef]

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., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews*, *10*, 89. [CrossRef]

Payne, C., & Glantz, E. J. (2020, October 7–9). *Teaching adversarial machine learning: Educating the next generation of technical and security professionals*. Annual Conference of the Special Interest Group in Information Technology Education (pp. 7–12), Virtual Event. [CrossRef]

Pusey, P., Gupta, M., Mittal, S., & Abdelsalam, M. (2024). An analysis of prerequisites for artificial intelligence machine learning-assisted malware analysis learning modules. *Journal of The Colloquium for Information Systems Security Education*, *11*, 1–5. [CrossRef]

Salman, A. (2024, December 17–19). *Integrating artificial intelligence in cybersecurity education: A pedagogical framework and case studies*. International Conference on Computer and Applications (pp. 1–5), Cairo, Egypt. [CrossRef]

Shahriar, H., Whitman, M., Lo, D., Wu, F., & Thomas, C. (2020, March 11–14). *Case study-based portable hands-on labware for machine learning in cybersecurity*. ACM Technical Symposium on Computer Science Education (p. 1273), Portland, OR, USA. [CrossRef]

Svabensky, V., Vykopal, J., & Celeda, P. (2020, March 11–14). *What are cybersecurity education papers about? A systematic literature review of SIGCSE and ITiCSE conferences*. ACM Technical Symposium on Computer Science Education (pp. 2–8), Portland, OR, USA. [CrossRef]

Tian, J. (2025). A practice-oriented computational thinking framework for teaching neural networks to working professionals. *AI*, *6*(7), 140. [CrossRef]

Wei-Kocsis, J., Sabounchi, M., Mendis, G. J., Fernando, P., Yang, B. J., & Zhang, T. L. (2024). Cybersecurity education in the age of artificial intelligence: A novel proactive and collaborative learning paradigm. *IEEE Transactions on Education*, *67*(3), 395–404. [CrossRef]

Weitl-Harms, S., Spanier, A., Hastings, J., & Rokusek, M. (2023). A systematic mapping study on gamification applications for undergraduate cybersecurity education. *Journal of Cybersecurity Education Research and Practice*, *2023*(1), 9. [CrossRef]

You, Y., Tse, J., & Zhao, J. (2025). Panda or not panda? Understanding adversarial attacks with interactive visualization. *ACM Transactions on Interactive Intelligent Systems*, *15*(2), 11. [CrossRef]

Zivanovic, M., Lendák, I., & Popovic, R. (2024, July 30–August 2). *Tackling the cybersecurity workforce gap with tailored cybersecurity study programs in central and eastern europe*. ACM International Conference on Availability, Reliability and Security, Vienna, Austria. [CrossRef]

education sciences

MDPI

Systematic Review

ChatGPT in Teaching and Learning: A Systematic Review

Duha Ali [1], Yasin Fatemi [2], Elahe Boskabadi [3], Mohsen Nikfar [2], Jude Ugwuoke [4] and Haneen Ali [5,*]

[1] Department of Industrial and Manufacturing Engineering, California Polytechnic State University—San Luis Obispo, San Luis Obispo, CA 93407, USA; duali@calpoly.edu
[2] Department of Industrial and Systems Engineering, Auburn University, Auburn, AL 36849, USA; yzf0024@auburn.edu (Y.F.); mzn0042@auburn.edu (M.N.)
[3] Department of Economics, Le Moyne College, Syracuse, NY 13214, USA; boskabe@lemoyne.edu
[4] Department of Political Science, Auburn University, Auburn, AL 36849, USA; jcu0005@auburn.edu
[5] Health Services Administration Program, Department of Industrial and Systems Engineering, Auburn University, Auburn, AL 36849, USA
* Correspondence: hba0007@auburn.edu

Abstract: The increasing use of artificial intelligence (AI) in education has raised questions about the implications of ChatGPT for teaching and learning. A systematic literature review was conducted to answer these questions, analyzing 112 scholarly articles to identify the potential benefits and challenges related to ChatGPT use in educational settings. The selection process was thorough to ensure a comprehensive analysis of the current academic discourse on AI tools in education. Our research sheds light on the significant impact of ChatGPT on improving student engagement and accessibility and the critical issues that need to be considered, including concerns about the quality and bias of generated responses, the risk of plagiarism, and the authenticity of educational content. The study aims to summarize the utilizations of ChatGPT in teaching and learning by addressing the identified benefits and challenges through targeted strategies. The authors outlined some recommendations that will ensure that the integration of ChatGPT into educational frameworks enhances learning outcomes while safeguarding academic standards.

Keywords: ChatGPT; education; student engagement; plagiarism; bias; academia

1. Introduction

ChatGPT is a generative artificial intelligence model developed by OpenAI. Since its launch at the end of 2022, it has rapidly become a prominent technological phenomenon, especially in education. Research by Lozano and Fontao, 2023 and Waltzer et al., 2023 has demonstrated the ability of ChatGPT to engage in human-like conversations across various topics, a feature that has revolutionized interactive learning [1,2]. As highlighted by Van Slyke et al., 2023, ChatGPT's capacity for interactive learning and practical applications underscores its significant impact on educational methodologies [3]. Other studies have emphasized its expansive training dataset, which enables the model to provide comprehensive and coherent responses, enhancing its utility in educational settings [4].

Beyond providing conventional educational assistance, ChatGPT has been the subject of multiple studies focusing on its multifaceted applications in learning environments [3,4]. Its rapid adoption across various fields reflects its versatile application potential [3,5,6]. The model's ability to process and generate human-like text responses makes it an innovative tool for use in professional settings and diverse educational landscapes. In this context, different authors pointed out that ChatGPT can facilitate teaching and learning by generating quizzes, designing assessment, generating summaries, and translating complex terminologies, illustrating its multifunctional nature and far-reaching impact [4–9].

The integration of ChatGPT into academic settings signifies a major shift in the educational technology landscape [2]. Some studies have highlighted ChatGPT's ability to

redefine learning experiences through its advanced natural language processing capabilities, fostering a level of student engagement that mimics that via human interaction [10,11]. Similarly, Lozano and Fontao, 2023 and Yan et al., 2023 emphasized ChatGPT's ability to enhance communication within educational environments [1,10].

Building on these foundational changes, ChatGPT's academic applications are vast and varied. Research by Yan et al., 2023 on technology-enhanced language learning showcased ChatGPT as a powerful tool for language acquisition, providing interactive and responsive support to learners [10]. Additionally, other authors investigated its ability to personalize educational content, tailoring learning experiences to individual student needs [11–14]. ChatGPT facilitates automated grading and intelligent tutoring, aids educational content creation, and improves accessibility to support diverse learning needs. This transformative potential of ChatGPT in education can make learning more accessible, engaging, and effective. Moreover, ChatGPT can also be used as a student virtual assistant [15,16]. In this capacity, ChatGPT serves as a helpful tool, providing on-demand explanations, answering questions, and offering guidance on various academic topics [2]. Looking towards the future of education with ChatGPT, there is a landscape in which AI is not merely an auxiliary tool but rather a fundamental component of the educational framework.

Some researchers pointed to a future where ChatGPT continually expands the boundaries of education [2,17,18]. E-learning and online self-learning, which have recently gained popularity due to digital education platforms and the effects of the COVID-19 pandemic are areas in which ChatGPT can provide significant support as noted by [7,19–21]. ChatGPT acts as a virtual tutor, assisting learners in navigating online courses and offering insights into course materials [1]. Other researchers have highlighted the importance of ChatGPT in lifelong learning and digital literacy, illustrating how this technology is reshaping both the content and methods of education [12].

Additionally, ChatGPT's adaptability caters to individual learning styles, making it a valuable resource in self-paced online education. ChatGPT interprets complex questions, generates relevant responses, and facilitates complex academic tasks, offering substantial opportunities for personalized learning support in educational settings [22]. Some researchers explored ChatGPT's adaptability across various learning contexts, from traditional classrooms to digital platforms, also in addition to incorporating it for teaching design, underscoring its value in creating versatile and dynamic educational experiences [23,24]. Collectively, these perspectives paint a picture of an academic future in which AI is seamlessly integrated into the fabric of education, enhancing learning experiences while presenting new challenges and opportunities for educators, students, and policymakers.

Despite these promising applications, the integration of ChatGPT in educational settings is challenging. Some researchers have raised concerns about the quality of responses and potential biases within AI-generated content [8,13,25–27]. These issues are crucial, particularly in educational contexts, where the accuracy and reliability of information are paramount. Others, like Deraga et al., 2023 and Strzelecki, 2023 have advocated for a balanced approach to integrating technologies like ChatGPT in education [28,29]. It is essential to leverage the benefits of ChatGPT while mitigating its potential risks and ensuring that ethical standards are upheld. Most researchers have stressed the importance of academic integrity and ethical considerations, especially considering the ease with which students might rely excessively on AI for learning and problem-solving [9,30].

The incorporation of ChatGPT into educational settings has its detractors. Some authors have explored the controversies surrounding the use of ChatGPT in education, highlighting concerns about overreliance on technology and a potential loss of critical thinking skills among students (for example, [27–29,31]. Researchers like Farazouli et al., 2023 and Farrokhnia et al., 2023 have emphasized the potential downsides, such as the automation of academic tasks leading to a lack of deep learning [22,32]. These debates are not only confined to the classroom but also extend to the ethical dimensions of using such advanced technology.

Schön et al., 2023 and Tam et al., 2023 highlighted the importance of maintaining academic integrity and addressing the ethical implications of AI in education [26,32]. Many researchers have contributed to this discussion, emphasizing the need for a balanced approach in integrating AI tools like ChatGPT [22,26,33–37]. The central concern revolves around the potential for these tools to facilitate plagiarism, undermine academic integrity, and erode the traditional student–teacher dynamic [38].

There is an intense debate among scholars about using ChatGPT in educational settings, with diverse and often conflicting viewpoints. This range of perspectives highlights a significant gap in the literature: the need for a comprehensive study synthesizing these varying opinions to provide a cohesive understanding of ChatGPT's role in learning environments (e.g., assignments and tutoring). First, a notable group of authors have focused on the innovative potential of ChatGPT [10,30,39–42]. They have often discussed the advantages of ChatGPT in personalizing learning experiences, automating routine educational tasks, and facilitating interactive learning. However, their enthusiasm is not universally shared. Other scholars have raised concerns about the potential pitfalls, including the impact on students' critical thinking skills, the risk of fostering dependency on AI for academic tasks, and the implications for academic integrity (for example, [43–46]).

Second, the debate has been further intensified by authors highlighting the ethical and pedagogical implications of integrating AI into learning environments [47–49]. They have emphasized the need to carefully consider how ChatGPT is implemented, ensuring it is used to complement rather than to replace traditional teaching methods. Meanwhile, some authors have argued for a balanced approach, highlighting the importance of maintaining academic integrity and an authentic learning experience in the face of rapidly advancing AI technologies [49–52]. Finally, the last set of authors have underscored the need for a comprehensive study that considers the full spectrum of views on ChatGPT's role in education [53–56].

Several review papers focus on the impact of ChatGPT in education. For instance, Castro [57] conducted a critical literature review using electronic databases such as newspapers affiliated with Harvard University, Google Scholar, Springer, and Scopus. However, the paper is quite short and includes only thirteen references, representing a limited sample of the papers analyzed. Another literature review paper by Jameela [58] focused on the impact of ChatGPT in educational and organizational contexts, utilizing twenty-two papers. However, they did not mention the names of the databases from which they retrieved the papers and simply summarized each one. Once again, their lack of a systematic approach and the limited number of papers they included affected the robustness of their results. On the other hand, Faisal conducted a systematic literature review on the benefits of ChatGPT in higher education. He utilized the Web of Science, EBSCO, and ProQuest databases and selected 52 papers. The results included papers published until 5 June 2023 and discussed only the benefits of ChatGPT for higher education. This paper concluded with remarks on the potential use of ChatGPT in higher education in Saudi Arabia.

Our work complements this latter body of literature by reconciling these diverse perspectives and provides insights into best practices for integrating ChatGPT in education. Thus, it contributes significantly to the academic discourse by providing a better understanding of how ChatGPT can enhance learning experiences while addressing legitimate concerns about its potential drawbacks. Through its comprehensive analysis of 112 articles, this study robustly covers the current state of research on ChatGPT in education, addressing both breadth and depth within the scope defined. The results of this comprehensive review will be instrumental in guiding educators, policymakers, and researchers in making informed decisions about the use of ChatGPT in educational settings.

2. Methodology

2.1. Research Questions

The primary aim of this literature review is to investigate the benefits and limitations of ChatGPT use in academia, with a specific focus on its application in teaching and learning. Two research questions guided our inquiry:

1. What are the benefits of ChatGPT use in academia, particularly in teaching and learning contexts?

This question seeks to uncover the positive aspects of ChatGPT in educational contexts. It explores how ChatGPT enhances teaching and learning experiences, its role in engaging students, and its potential to aid educators. By examining its benefits, this research aims to identify how ChatGPT can be effectively integrated into academic practices to improve outcomes and experiences for both teachers and students.

2. What are the limitations associated with the use of ChatGPT in educational settings?

The focus here is to understand the challenges and constraints associated with ChatGPT usage in academia. This includes potential issues like the accuracy of the information provided, ethical concerns such as plagiarism, and the impact on students' learning and critical thinking skills. This question is crucial for identifying potential risks and developing strategies to mitigate them, ensuring a balanced and responsible approach to incorporating ChatGPT in educational environments.

2.2. Search Strategy

We utilized a comprehensive set of keywords to systematically search for the relevant literature in three major databases: Academic Search Premier, Web of Science, and IEEE. The keywords used in our search strategy are presented in Table 1.

Table 1. Search strings and keywords.

Search Strings
((ChatGPT OR Chat gpt) AND (Challenge OR pros OR cons OR benefit OR Advantage OR problem OR disadvantage OR harm OR support))
(((ChatGPT OR Chat gpt) AND (future OR application OR possibility)))
(((ChatGPT OR Chat gpt) AND (opinion OR feeling OR attitude OR user OR professional OR evaluation OR evaluate OR experience OR perception OR misconduct OR ethics OR integrity)))
(((ChatGPT OR Chat gpt) AND (teaching OR assignment OR homework OR education OR school OR student OR teacher OR practice OR project)))

We selected the keywords for the literature search with precision, ensuring a thorough and targeted approach to data collection across the three databases. To expand the scope of our search, we incorporated a series of outcome-oriented keywords, ranging from evaluative terms like "challenge", "benefit", and "advantage" to forward-looking terms like "future" and "possibility". We refined this search further by incorporating terms related to ethical dimensions, including "opinion", "ethics", and "integrity", ensuring a comprehensive perspective on ChatGPT's impact. We also employed educational keywords like "teaching", "assignment", and "homework", focusing the search on the practical educational applications and implications of ChatGPT, thus guaranteeing a rich and relevant dataset for subsequent analysis.

2.3. Inclusion and Exclusion Criteria

Table 2 presents the inclusion and exclusion criteria that were applied in the selection of articles in this study. We excluded conference articles, works that were not peer-reviewed, summaries of other work, book chapters, articles from magazines, theses, and notes to editors.

Table 2. Inclusion and exclusion criteria.

Inclusion Criteria	Exclusion Criteria
Peer-reviewed journal articles. Availability of full texts. Published in the English language. Published any time after 30 November 2022 (the launch day of ChatGPT).	Conference articles, non-peer-reviewed publications, review articles, book chapters, magazine articles, theses, and notes to editors. Articles not directly related to the application of ChatGPT in educational contexts.

2.4. Screening the Articles

To avoid bias, we applied the following process. In each round of screening, each author coded the papers as either "0": The paper should not be included, "1": Not certain about including the paper, or "2": The paper is eligible for inclusion. It is worth mentioning that the authors were not aware of the others' decisions. Discrepancies were discussed and addressed in a focus session with all authors present.

Figure 1 presents a PRISMA flow diagram of the search methodology. Initially, three databases were searched, yielding 999 articles from Academic Search Premier, Web of Science, and IEEE. From these, 133 articles were excluded due to duplication. The remaining 866 articles' abstracts were screened. Based on the inclusion and exclusion criteria, 678 articles were removed. Subsequently, 188 full-text articles were assessed for their relevance in discussing the benefits and challenges of ChatGPT, leading to the further exclusion of 76 articles. The process resulted in a final data set comprising 112 articles that met all the criteria for the review. In this study, we utilized Web of Science, IEEE Xplore, and Academic Search Premier based on the following considerations. Web of Science: We chose Web of Science due to its comprehensive coverage of high-quality, peer-reviewed journals and its strong emphasis on scientific citation indexing, which is crucial for the scope of our analysis. IEEE Xplore: Given that our research intersects significantly with technology and education, IEEE Xplore was selected for its extensive repository of technical literature and proceedings, which are highly relevant and authoritative in the fields of engineering and technology. Academic Search Premier: This database was included to broaden our research scope to capture interdisciplinary perspectives that are not strictly covered by the more specialized databases, thus ensuring a comprehensive exploration of the literature.

2.5. Research Profile

The top countries with the highest number of included articles related to our topic were the United States (n = 22), the United Kingdom and Northern Ireland (n = 11), and Australia (n = 11). The country of the published article was determined based on the first author's country of affiliation. Additionally, many other researchers from other countries conducted research related to ChatGPT in education. This may be interpreted to mean that the application of ChatGPT in teaching and learning has been attracting attention globally. Figure 2 presents the heatmap for the number of papers per country.

Figure 3 displays the number of included articles per publication month in 2023 (hereafter, when we talk about our results in 2023, we mean the results we found up till 12 October 2023). The highest numbers of papers were published in July (n = 25) and August (n = 18), while the lowest numbers of papers were published in February (n = 3) and January (n = 1). This trend suggests a growing academic interest in the field, potentially influenced by academic calendars, grant cycles, or conferences. The data also indicate that the start and end of the year had less research output, which reflects the planning and development phase of research activities. Furthermore, the database search was conducted on 12 October 2023; therefore, no data were available for November or December.

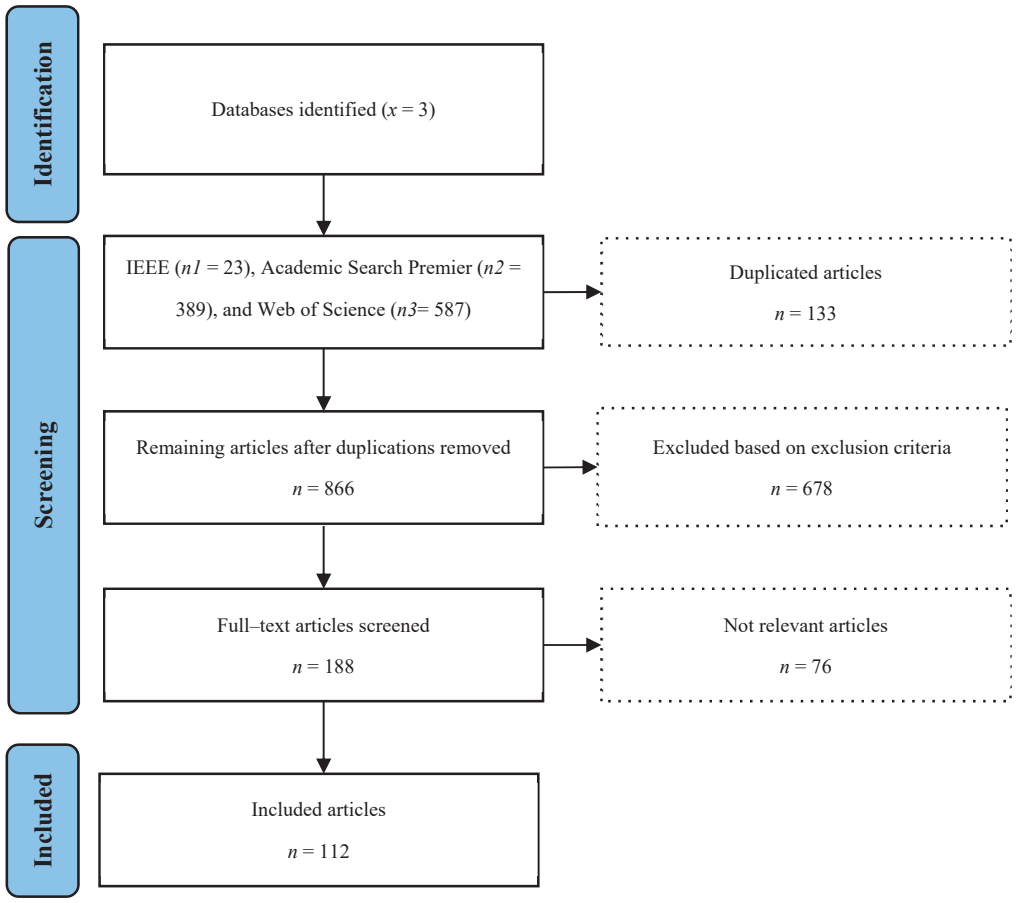


Figure 1. PRISMA flow diagram.

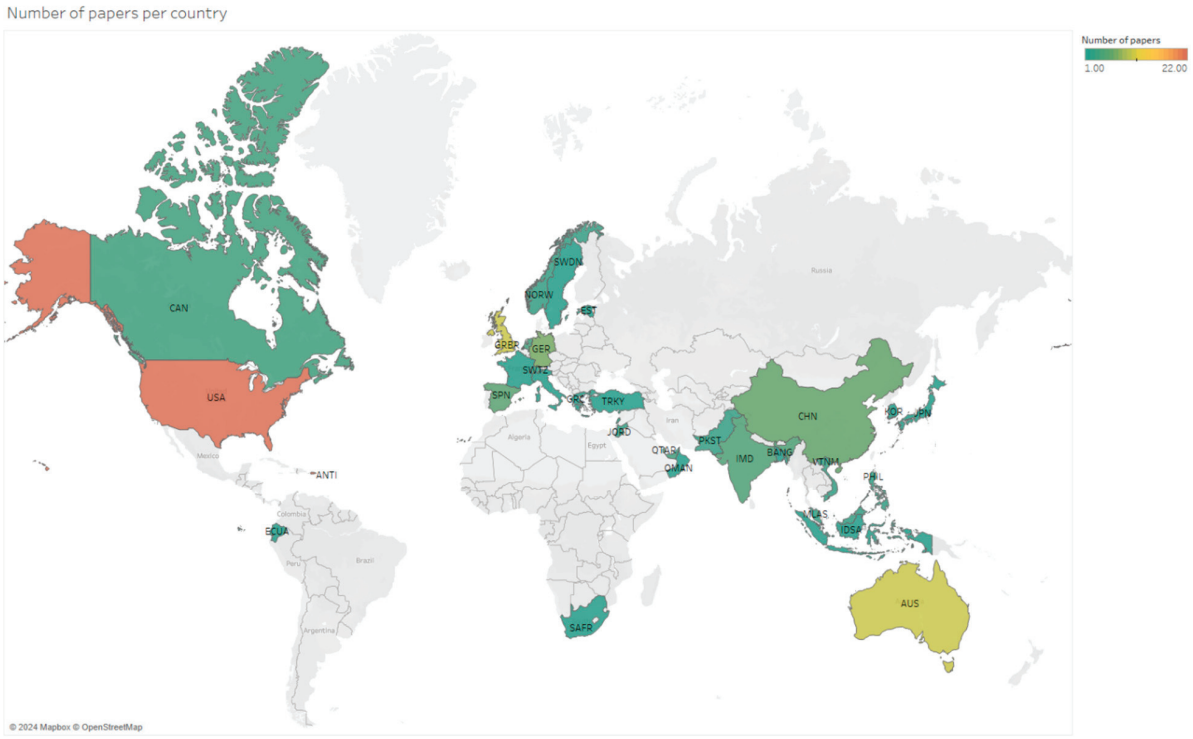


Figure 2. Number of articles published per country.

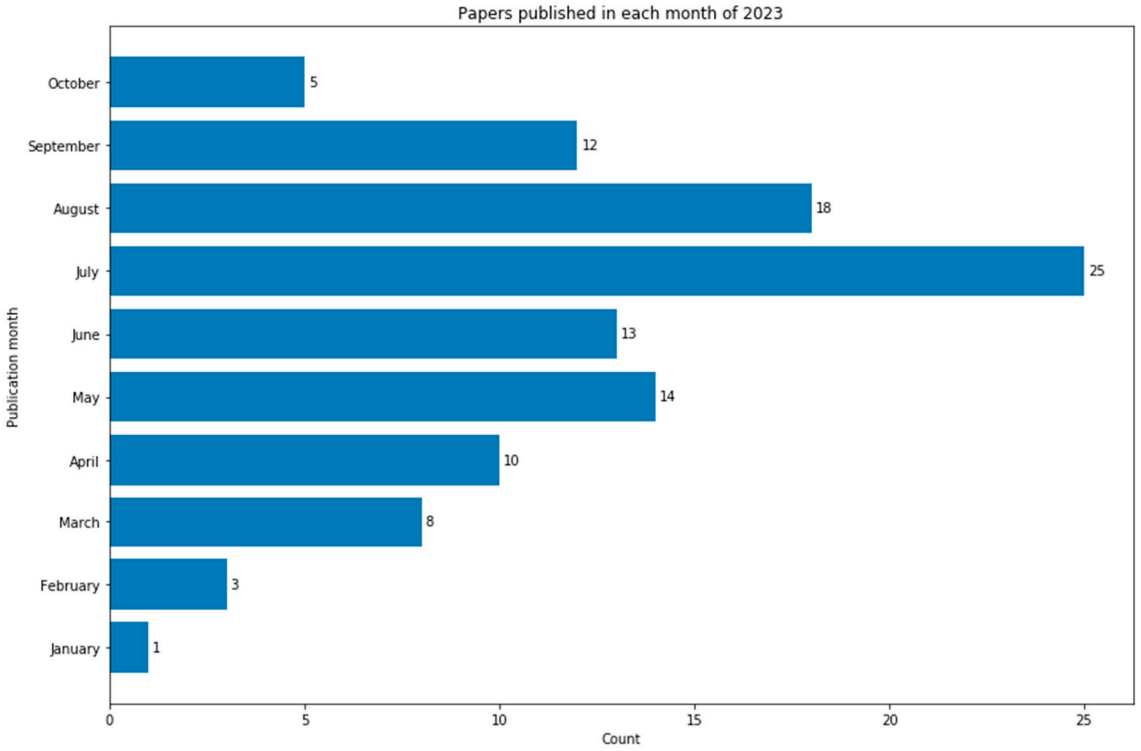


Figure 3. Number of articles per publication month of 2023.

Figure 4 illustrates the distribution of published papers on ChatGPT across a spectrum of academic journals in 2023. *The Journal of Chemical Education* (n = 4) had the most publications, indicating a particular interest or focus on ChatGPT within the chemical education community. This was closely followed by other journals, including *Cureus Journal of Medical Science*, *Sustainability, Education and Information Technologies*, *Scientific Reports*, and *International Journal of Management Education* (n = 3), each contributing three papers to the discourse on ChatGPT. The diversity of journals—from medical sciences to sustainability and information technology—reflects ChatGPT's interdisciplinary impact and the wide-ranging academic curiosity it has sparked. A single paper was found in numerous other journals, highlighting the expansive influence of this AI technology, which extends across various fields and specialties within the academic landscape.

Additionally, we created word clouds for abstracts, ChatGPT benefits, and limitations (Figure 5, Figure 6, and Figure 7, respectively). Figure 5 presents the abstracts of papers related to ChatGPT, highlighting the central themes and concepts emerging from academic discussions. Dominant terms like "education", "student", "learning", and "question" highlight the focus on the educational applications of ChatGPT and its role in student engagement and learning processes. The visibility of words like "study", "research", and "artificial intelligence" reflects the scholarly examination of AI tools in academic settings. Meanwhile, the presence of "tool" and "response" within the cloud suggests an interest in the functional aspects of ChatGPT, such as its responsiveness and utility as a supportive tool in educational practices. Together, these terms sketch a landscape in which ChatGPT's influence on education can be scrutinized from multiple angles, emphasizing its integration into teaching methods, student interaction, and the broader implications of AI in learning environments.

Figure 6 graphically represents the most frequently mentioned benefits of ChatGPT as gleaned from the literature. Words like "learning", "teaching", "feedback", "student", and "teacher" are prominent, highlighting ChatGPT's significant role in educational enhancement. The repeated appearances of "generate" and "support" suggest ChatGPT's capability to produce content and assist in various tasks, indicating its utility as an educational tool. Terms like "medical", "language", "code", and "writing" emphasize the

diverse applications of ChatGPT, from aiding in medical education to supporting language learning and coding. The visual aggregation of these terms paints a picture of ChatGPT as a versatile aid that can potentially transform how students learn and teachers instruct, providing personalized assistance and enriching the learning experience across disciplines. Moreover, many of our papers come from medical fields.

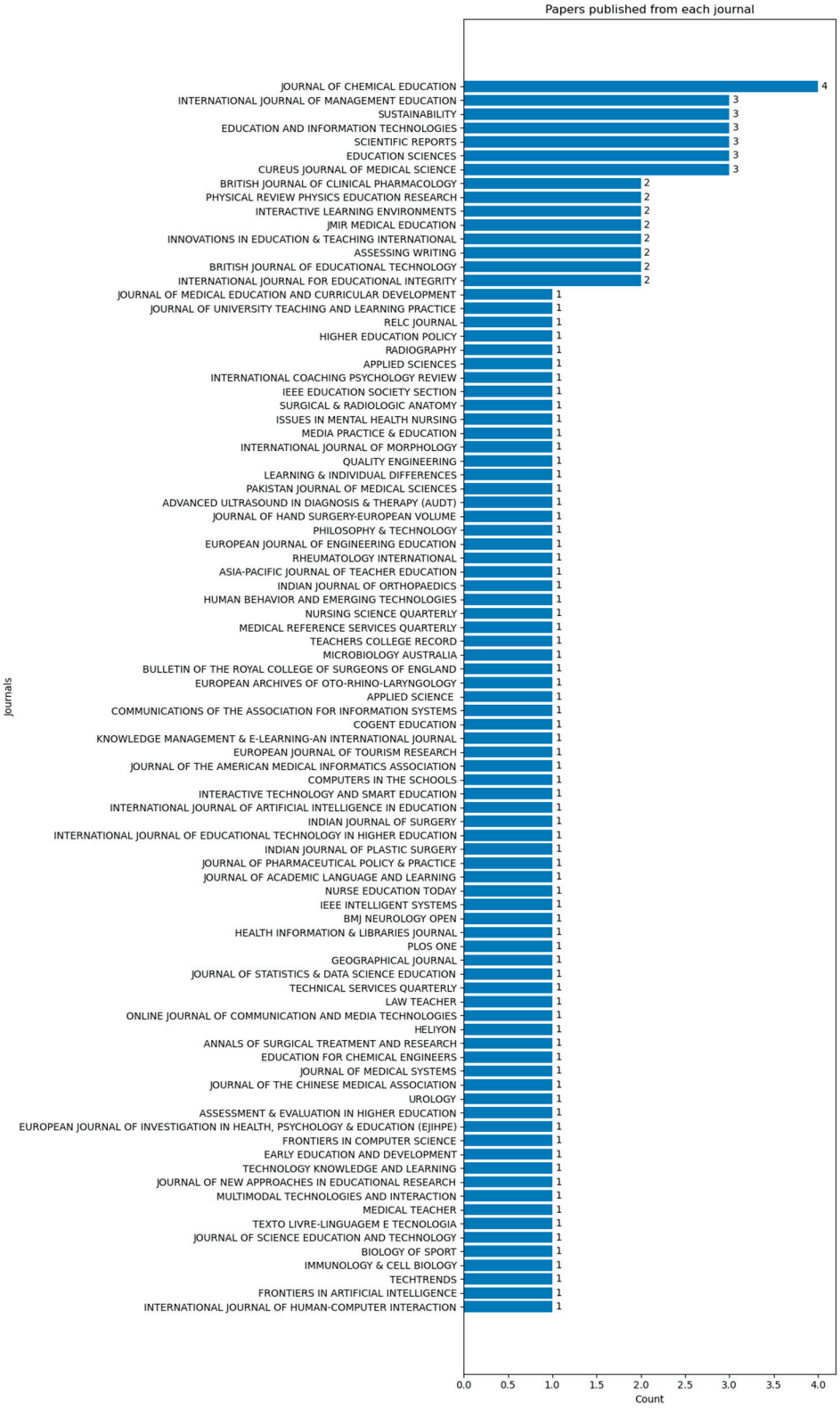


Figure 4. Number of articles published in each journal.

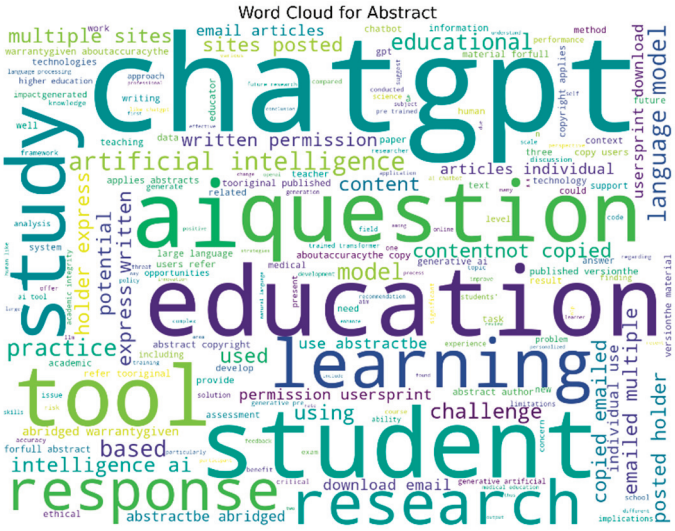


Figure 5. Word cloud for article abstracts.

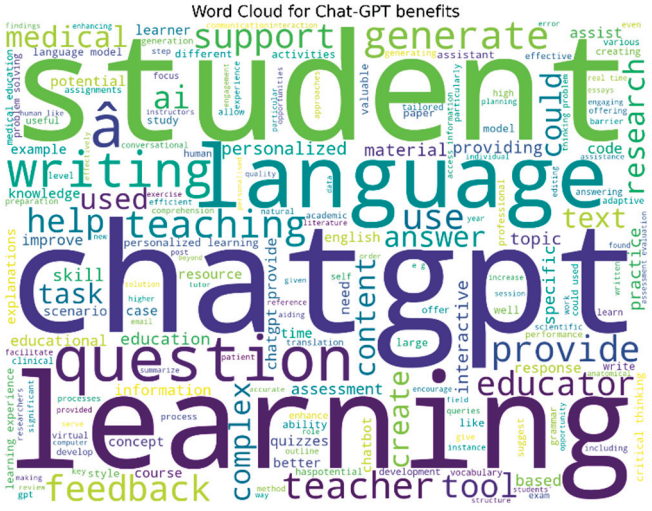


Figure 6. Word cloud for ChatGPT benefits.

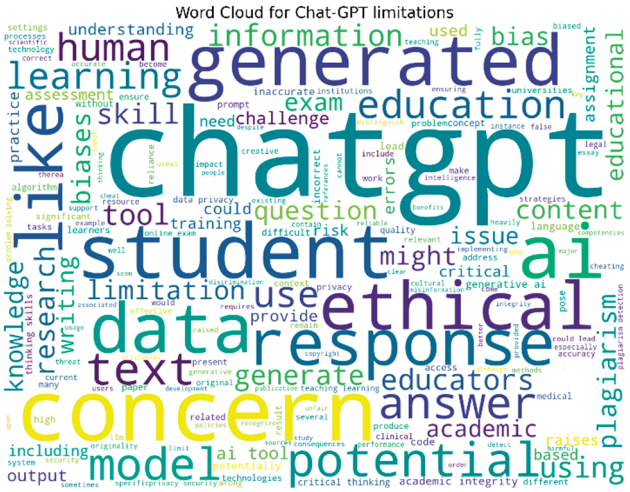


Figure 7. Word cloud for ChatGPT limitations.

Figure 7 captures the key concerns surrounding ChatGPT's limitations within an educational context. Central terms like "concern", "plagiarism", "bias", and "information"

suggest a focus on ethical and quality issues related to the use of ChatGPT in learning environments. Meanwhile, the prominence of the terms "generated" and "content" alongside "limitations" and "potential" implies a critical view of the reliability and originality of the material produced by ChatGPT. "Data" and "bias" appear near one other, highlighting worries about the originality of ChatGPT outputs, and the word "student" at the center reflects the centrality of the learner's experience in these discussions. This collection of terms illustrates the cautionary stance researchers and educators take regarding the adoption of AI tools like ChatGPT, emphasizing the need to address these significant challenges for responsible integration into academic settings.

3. Results and Discussion

The data suggest that while ChatGPT is appreciated for its ability to enhance performance evaluation, natural language processing capabilities, and text generation, there are significant concerns regarding the quality and bias of its responses as well as issues related to plagiarism and content authenticity. The advantages and disadvantages indicate a nuanced view of ChatGPT in higher education, where its potential is recognized but an awareness of its limitations and challenges also exists. All the papers that were screened for the advantages and disadvantages of ChatGPT are cited in Table 3.

Table 3. Included articles in each theme.

Advantages	
Natural Language Processing	[4,10,17,19,22,28,29,35,37,41,58–75]
Enhanced Communication	[1,40,52,60,67,76–82]
Learning Engagement and Accessibility	[1,4,10,11,15,17,18,20,21,26,27,32–34,45,53,58–61,67,69–73,75,77,79,83–96]
Performance Evaluation	[10,19,20,43,51–53,58,67,70,72,79,88,97–102]
Text Generation	[2,3,10,13,15,34,40,41,46,48,50,52,53,59,73,80,82,84,86,88,89,101–107]
Versatility—ability to adapt to many different functions	[4,6,23,31,36,39,43,50,53,54,56,74,89,90,103,108–110]
Other	[10,15,25,47,54,61,76,90,111,112]
Disadvantages	
Error recognition	[2,18,30,46,52,60,61,84–86,89,92,95,99,101,111,113]
Plagiarism and Authenticity	[1,3,4,10,26,33–35,37,47–49,53–55,59,60,62–64,67,69–71,73,77,84,86,88,94,96,99,100,104,106,114–116]
Quality of Responses and Bias in AI	[4,7,13,15,17,20–23,25,27,31,32,37,39–41,43–46,52,53,59,61,64,65,67,68,70,72–74,76,78,80,81,85,86,89,92,95,98,104,106–108,110,117,118]
Other	[4,6,7,10,15,18,28,46,53,66,67,70,90,106,119]
Dependency	[11,15,29,36,53,59,62,78,82,86,87,103,105,107]
Privacy and Data Security	[4,7,36,46,53,67,70,106,113,118]

3.1. Advantages of ChatGPT in Teaching and Learning

- Learning Engagement and Accessibility (45 occurrences): This category includes articles that discuss the role of ChatGPT in engaging students and making learning more accessible, which is the most frequently mentioned advantage in the included articles. ChatGPT can make learning more engaging and accessible, especially for students with disabilities. In their study, Hsu and Ching, 2023 categorized the applications of ChatGPT for learning, stating that teachers can use ChatGPT for support in the following ways: (1) assistance with teaching; (2) help with student assessment; (3) support for student learning; (4) suggestions for improving teaching; and (5) assistance with teacher–student and teacher–parent communication [27]. For students, the chatbot can provide support in the following areas: (1) personalized learning;

(2) creative thinking; (3) assessment; and (4) reading and writing comprehension [27]. Rodrigues and Rodrigues, 2023 emphasized the potential of ChatGPT to facilitate more personalized and adaptive learning due to its interactivity [11]. This enables the execution of effective learning mechanisms, with feedback being a core feature of learner support that is highly effective in supporting learning. In a special case, Houston et al., 2023 showed that libraries can benefit from this solution by improving their reference practices, developing their collections, and transforming and creating metadata [18].

- Natural Language Processing (NLP) (31 occurrences): ChatGPT's ability to understand and generate human-like text is a significant advantage. Students can practice conversational skills in different languages with ChatGPT, enhancing their language proficiency. Farrokhnia et al. 2023 reported that ChatGPT's natural language model is sophisticated and generates plausible, personalized, and real-time answers while self-improving [22]. In another study, Clark, 2023 found that ChatGPT's responses demonstrate strong language processing abilities, performing better on questions that require generalizable information rather than specific skills, particularly the skills taught in lectures [60]. Masters, 2023 noted that ChatGPT can be helpful in the grading process, especially for written assignments [66].
- Text Generation (29 occurrences): The capability of ChatGPT to generate text is highlighted in this category. ChatGPT can help students overcome writer's block by generating ideas or outlines for essays and research papers. NLP covers text generation; however, we created a separate category for text generation because it is an essential aspect of education and has attracted significant attention from authors in this field. It can also assist teachers in creating educational content, such as lesson plans or example texts for class discussions. For example, generative-AI chatbots excel at quickly generating plausible answers for any question [104]. Marquez et al., 2023 conducted a study focusing on biobased materials education, finding that it was possible to use AI-powered text generation to brainstorm biobased materials and products and develop academic written text by applying a scientific method approach [106]. As an AI-powered assistant, ChatGPT can answer questions, generate text, write code, summarize papers, evaluate responses, and more. It offers a range of relevant topics and ideas that can be included in a course's curriculum [80].
- Performance Evaluation (19 occurrences): Teachers can use ChatGPT to provide instant feedback on students' assignments or essays. For example, Ruiz et al., 2023 found that this virtual assistant tool allows teachers to provide real-time personalized support by answering student queries and offering additional information [100]. Additionally, Karabacak et al., 2023 highlighted opportunities to improve medical education through the use of personalized feedback and evaluation methods [43]. Clark et al., 2023 emphasized the potential of chatbots to support student learning in an interactive way by providing real-time feedback, thereby increasing student engagement [60].
- Versatility (20 occurrences): ChatGPT can be applied across various academic fields. In science, it can be used to explain complex concepts, while in history, it can provide historical context. When used in mathematics, ChatGPT can solve problems and explain the steps in the process. Indeed, it can be used for a wide variety of tasks, for example, to evaluate task performance, provide feedback, generate human-like writing, offer expert solutions to complex tasks, and assist in solving mathematical problems [23]. Therefore, ChatGPT is likely to have a major impact on work and education, as it provides quick and easily understandable answers to a variety of questions [39].
- Enhanced Communication (12 occurrences): This category refers to the potential for improved communication and interaction using ChatGPT. For example, Lozano and Fontao, 2023 noted that ChatGPT has great potential for improving communication between teachers and students. It can be used to generate innovative methodologies to improve the teaching–learning process, thereby increasing student performance [1].

AI assistants like ChatGPT can help by explaining complex concepts in simple language. This modern approach can help medical students learn more efficiently and provide better patient care [76]. Students can practice conversational skills in different languages with ChatGPT, enhancing their language proficiency. Moreover, ChatGPT can serve as an intermediary for students who may be shy or reluctant to ask questions in class or for students with disabilities [13]. Integrating ChatGPT with speech-to-text technology can support inclusive education for students with visual impairments or dyslexia.

- Other (12 occurrences): This category includes responses that do not fit neatly into the predefined themes. For example, Schen et al., 2023 focused on ChatGPT voice response automation, which could overcome issues with response behavior or response quality [30]. Other examples include thinking about the current instrumentalization in education [47], debugging code [61], and clinical decision-making [76].
- Not Applicable (10 occurrences): This category includes papers that were fully reviewed but did not discuss any advantages related to the use of ChatGPT in teaching and learning.

3.2. Disadvantages of ChatGPT in Teaching and Learning

- Quality of Responses and Bias in AI (51 occurrences): Concerns about the accuracy of ChatGPT's responses and potential biases in AI models are the most frequently cited disadvantages. The accuracy of ChatGPT's responses may not always be reliable, and there can be biases in the AI model. This requires teachers to double-check information and discuss these biases with students. Rawas, 2023 mentioned that the potential for bias in AI implementation must be approached with caution and a clear understanding of the opportunities and challenges involved [13]. Moreover, Iskender, 2023 argued that ChatGPT could exacerbate existing biases in education, such as socio-economic and racial disparities [7]. Tsang, 2023 expressed concerns about the reliability of ChatGPT due to hallucinations and its training sources, which limit its use as a clinical support resource and evidence-based research tool [74]. Naidu and Sevnaravan, 2023 emphasized that the quality of the responses provided by ChatGPT are contingent on the quality of the input received, and it can generate better answers if the questions and prompts are clear [31].
- Plagiarism and Authenticity (39 occurrences): It is challenging to ensure the originality and authenticity of content generated by ChatGPT. Moreover, there is a risk that students might submit ChatGPT-generated text as their own work. Educators need to emphasize the importance of academic integrity and may need to use plagiarism detection tools. Indeed, Dalalah and Dalalah, 2023 warned that plagiarism could become commonplace and endanger scientific research, leading to the loss of uniqueness and creativity in writing and art [63]. Furthermore, Wilby and Esson, 2023 highlighted ethical concerns regarding academic misconduct, model bias, robustness, and toxic output [96]. Thomas, 2023 reported that educators are concerned about cheating and may resort to oral exams [73]. Many are warning students that the use of ChatGPT will result in a failing grade.
- Error Recognition (17 occurrences): ChatGPT may not always recognize its own errors. Teachers and students should be aware of this limitation and cross-verify information with credible sources. For instance, Houston and Corrado, 2023 showed that ChatGPT's proficiency in generating text is best in the language it has been extensively trained on, namely, English [18]. Its ability to produce quality responses in other languages may not be as good as its responses in English, and there might be inconsistencies or errors in its language generation. Further, the importance of continually updating AI models with the latest medical knowledge has been emphasized to ensure that they remain reliable and accurate in the rapidly evolving field of medicine [111]. Failure to do so could cause ChatGPT to provide inaccurate responses. In the context of programming, Borger et al., 2023 noted that relying on ChatGPT-generated code

requires users to have a fundamental understanding of programming concepts to avoid erroneous outputs [84].

- Dependency (15 occurrences): There is a risk of overreliance on AI tools in learning environments, which can hinder students' ability to think critically and solve problems independently. Educators need to balance the use of AI with traditional teaching methods. For example, Hosseini et al., 2023 warned that clinicians may become overly reliant on ChatGPT-like systems, putting their clinical reasoning skills at risk [87]. Similarly, Marzuki et al., 2023 reported that some educators are concerned that excessive use of these AI tools for language refinement and idea generation may limit students' creative thinking and originality [82]. Ratten and Jones [107] expressed concerns about students relying too heavily on ChatGPT in completing their assignments, impeding their development of intuitive skills and potentially altering assessment practices.
- Privacy and Data Security (11 occurrences): The use of ChatGPT in education raises concerns about data privacy and security. Schools and educational institutions must ensure that they are using AI tools in compliance with data protection laws and regulations. For instance, Marquez et al., 2023 noted using ChatGPT in education raises ethical concerns regarding privacy, data ownership, and algorithmic bias [106]. Michel-Villarreal et al., 2023 argued that is crucial for universities to address data privacy, algorithmic bias, and responsible use of AI-generated content to avoid skepticism around the implementation of ChatGPT [67].
- Other (6 occurrences): This includes miscellaneous concerns that do not fit into the predefined categories. For example, Dergaa et al., 2023 pointed out that this technology has the potential to generate harmful outputs, such as spam and ransomware, which is a cause for concern in modern societies [28]. Other examples include environmental concerns, an inability to be used in highly specialized contexts, and a lack of contextual and nuanced understanding [6,46,90].
- Not Applicable (14 occurrences): This category includes papers that were fully reviewed but did not discuss any disadvantages related to the use of ChatGPT in teaching and learning.

4. Recommendations

Through this extensive analysis, the authors suggest that ChatGPT should be integrated thoughtfully into curricula, complementing traditional teaching methods rather than replacing them [24]. This integration should focus on enhancing student engagement and providing supplementary learning support. To ensure effective integration, educators should receive training on the technical use of ChatGPT as well as its limitations, including potential biases and inaccuracies. Such awareness is crucial for guiding students effectively. Additionally, institutions should establish guidelines for the ethical use of AI tools, emphasizing the importance of academic integrity and the prevention of plagiarism. Courses should include components that teach students how to critically evaluate AI-generated information, which is a vital skill in the modern digital age, where information is abundant and varied in quality.

There are different avenues for future studies on the long-term effects of AI tools on learning processes and outcomes. This research should include diverse educational settings and consider different student demographics. Moreover, it is crucial to develop policies within the educational institution that ensure equitable access to AI technologies like ChatGPT so that all students have equal opportunities to benefit from these advancements. It is important to balance the innovative potential of ChatGPT with caution and awareness of its limitations and challenges to support educators and institutions in harnessing the benefits of AI while mitigating its risks [120].

As an educator, it is important to understand the benefits and limitations of using ChatGPT in teaching. While ChatGPT can enhance student engagement and provide diverse learning materials, it is essential to critically assess its output to avoid misinformation and plagiarism [53,121]. For policymakers, developing policy guidelines for AI tools in

education is crucial, including addressing content authenticity, bias in AI, and promoting critical thinking skills in students who use these technologies [122]. Further empirical research is needed to determine the impact of ChatGPT and other AI tools on learning outcomes, student engagement, and the development of critical thinking skills.

5. Conclusions

The analysis of ChatGPT's benefits and limitations in higher education reveals a complex landscape. On one hand, ChatGPT's capabilities in natural language processing, text generation, and performance evaluation offer significant opportunities to enhance the educational experience. These aspects align with the increasing interest in adaptive learning technologies that can provide personalized education experiences. On the other hand, concerns about the quality and bias of ChatGPT's responses, plagiarism, and content authenticity pose significant challenges. These findings resonate with ongoing debates in educational technology regarding the ethical and practical implications of AI use in learning environments.

Author Contributions: The design of the study was led by D.A. The methodology was developed by all the authors. The task of reading and screening the papers was carried out by all the authors. The initial draft of the paper was written by all the authors. H.A. provided critical revisions to the manuscript. All the authors contributed to and have approved the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability Statement: The datasets presented in this article are not readily available because the data are part of an ongoing study. Requests to access the datasets should be directed to Yasin Fatemi.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

Abbreviations

AI	Artificial Intelligence
NLP	Natural Language Processing

References

1. Lozano, A.; Fontao, C. Is the Education System Prepared for the Irruption of Artificial Intelligence? A Study on the Perceptions of Students of Primary Education Degree from a Dual Perspective: Current Pupils and Future Teachers. *Educ. Sci.* **2023**, *13*, 733. [CrossRef]
2. Waltzer, T.; Cox, R.; Heyman, G. Testing the Ability of Teachers and Students to Differentiate between Essays Generated by ChatGPT and High School Students. *Hum. Behav. Emerg. Technol.* **2023**, *2023*, 1923981. [CrossRef]
3. Van Slyke, C.; Johnson, R.; Sarabadani, J. Generative Artificial Intelligence in Information Systems Education: Challenges, Consequences, and Responses. *Commun. Assoc. Inf. Syst.* **2023**, *53*, 14. [CrossRef]
4. Totlis, T.; Natsis, K.; Filos, D.; Ediaroglou, V.; Mantzou, N.; Duparc, F.; Piagkou, M. The Potential Role of ChatGPT and Artificial Intelligence in Anatomy Education: A Conversation with ChatGPT. *Surg. Radiol. Anat.* **2023**, *45*, 1321–1329. [CrossRef] [PubMed]
5. Bissessar, C. To Use or Not to Use ChatGPT and Assistive Artificial Intelligence Tools in Higher Education Institutions? The Modern-Day Conundrum—Students' and Faculty's Perspectives. *Equity Educ. Soc.* **2023**, 27526461231215083. [CrossRef]
6. Cooper, G. Examining Science Education in ChatGPT: An Exploratory Study of Generative Artificial Intelligence. *J. Sci. Educ. Technol.* **2023**, *32*, 444–452. [CrossRef]
7. Iskender, A. Holy or Unholy? Interview with Open AI's ChatGPT. *Eur. J. Tour. Res.* **2023**, *34*, 3414. [CrossRef]
8. Bringula, R. ChatGPT in a Programming Course: Benefits and Limitations. *Front. Educ.* **2024**, *9*, 1248705. [CrossRef]
9. da Silva, C.A.G.; Ramos, F.N.; de Moraes, R.V.; Santos, E.L. ChatGPT: Challenges and Benefits in Software Programming for Higher Education. *Sustainability* **2024**, *16*, 1245. [CrossRef]
10. Yan, L.; Sha, L.; Zhao, L.; Li, Y.; Martinez-Maldonado, R.; Chen, G.; Li, X.; Jin, Y.; Gasevic, D. Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review. *Br. J. Educ. Technol.* **2023**, *55*, 90–112. [CrossRef]
11. Rodrigues, O.S.; Rodrigues, K.S. Artificial Intelligence in Education: The Challenges of ChatGPT. *Texto Livre* **2023**, *16*, e45997. [CrossRef]

12. Baidoo-Anu, D.; Owusu Ansah, L. Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. *J. AI* **2023**, *7*, 52–62. [CrossRef]
13. Rawas, S. ChatGPT: Empowering Lifelong Learning in the Digital Age of Higher Education. *Educ. Inf. Technol.* **2024**, *29*, 6895–6908. [CrossRef]
14. Klayklung, P.; Chocksathaporn, P.; Limna, P.; Kraiwanit, T.; Jangjarat, K. Revolutionizing education with chatgpt: Enhancing learning through conversational AI. *Univers. J. Educ. Res.* **2023**, *2*, 217–225.
15. Choudhary, O.P.; Saini, J.; Challana, A. ChatGPT for Veterinary Anatomy Education: An Overview of the Prospects and Drawbacks. *Int. J. Morphol.* **2023**, *41*, 1198–1202. [CrossRef]
16. Sok, S.; Heng, K. ChatGPT for Education and Research: A Review of Benefits and Risks. *SSRN Electron. J.* **2023**. *pre-print*. [CrossRef]
17. Cotton, D.R.E.; Cotton, P.A.; Shipway, J.R. Chatting and Cheating: Ensuring Academic Integrity in the Era of ChatGPT. *Innov. Educ. Teach. Int.* **2023**, *61*, 228–239. [CrossRef]
18. Houston, A.; Corrado, E. Embracing ChatGPT: Implications of Emergent Language Models for Academia and Libraries. *Tech. Serv. Q.* **2023**, *40*, 76–91. [CrossRef]
19. Bauer, E.; Greisel, M.; Kuznetsov, I.; Berndt, M.; Kollar, I.; Dresel, M.; Fischer, M.R.; Fischer, F. Using Natural Language Processing to Support Peer-Feedback in the Age of Artificial Intelligence: A Cross-Disciplinary Framework and a Research Agenda. *Brit. J. Educ. Tech.* **2023**, *54*, 1222–1245. [CrossRef]
20. Hoch, C.C.; Wollenberg, B.; Lüers, J.-C.; Knoedler, S.; Knoedler, L.; Frank, K.; Cotofana, S.; Alfertshofer, M. ChatGPT's Quiz Skills in Different Otolaryngology Subspecialties: An analysis of 2576 Single-Choice and Multiple-Choice Board Certification Preparation Questions. *Eur. Arch. Otorhinolaryngol.* **2023**, *280*, 4271–4278. [CrossRef]
21. Jiang, H.; Cheong, K. Developing Teaching Strategies for Rural School Pupils' Concentration in the Distance Music Classroom. *Educ. Inf. Technol.* **2024**, *29*, 5903–5920. [CrossRef]
22. Farrokhnia, M.; Banihashem, S.K.; Noroozi, O.; Wals, A. A SWOT Analysis of ChatGPT: Implications for Educational Practice and Research. *Innov. Educ. Teach. Int.* **2024**, *61*, 460–474. [CrossRef]
23. Zhu, I.C.; Sun, M.; Luo, J.; Li, T.; Wang, M. How to Harness the Potential of ChatGPT in Education? *Knowl. Manag. E-Learn.* **2023**, *15*, 133–152. [CrossRef]
24. do Amaral, I. Reflection on the use of Generative Language Models as a Tool for Teaching Design. In Proceedings of the 2024 IEEE World Engineering Education Conference (EDUNINE), Kos, Greece, 10–13 March 2024; pp. 1–4.
25. Dahlkemper, M.; Lahme, S.; Klein, P. How Do Physics Students Evaluate Artificial Intelligence Responses on Comprehension Questions? A Study on the Perceived Scientific Accuracy and Linguistic Quality of ChatGPT. *Phys. Rev. Phys. Educ. Res.* **2023**, *19*, 010142. [CrossRef]
26. Duong, D. How Effort Expectancy and Performance Expectancy Interact to Trigger Higher Education Students' Uses of ChatGPT for Learning. *Interact. Technol. Smart Educ.* **2023**. *ahead of print*. [CrossRef]
27. Hsu, Y.-C.; Ching, Y.-H. Generative Artificial Intelligence in Education, Part One: The Dynamic Frontier. *TechTrends* **2023**, *67*, 603–607. [CrossRef]
28. Dergaa, I.; Chamari, K.; Zmijewski, P.; Ben Saad, H. From Human Writing to Artificial Intelligence Generated Text: Examining the Prospects and Potential Threats of ChatGPT in Academic Writing. *Biol. Sport* **2023**, *40*, 615–622. [CrossRef] [PubMed]
29. Strzelecki, A. To Use or Not to Use ChatGPT in Higher Education? A Study of Students' Acceptance and Use of Technology. *Interact. Learn. Environ.* **2023**, 1–14. [CrossRef]
30. Schön, E.-M.; Neumann, M.; Hofmann-Stölting, C.; Baeza-Yates, R.; Rauschenberger, M. How Are AI Assistants Changing Higher Education? *Front. Comput. Sci.* **2023**, *5*, 1208550. [CrossRef]
31. Naidu, K.; Sevnarayan, K. ChatGPT: An Ever-Increasing Encroachment of Artificial Intelligence in Online Assessment in Distance Education. *Online J. Commun. Media Technol.* **2023**, *13*, e202336. [CrossRef]
32. Farazouli, A.; Cerratto Pargman, T.; Laksov, K.; McGrath, C. Hello GPT! Goodbye Home Examination? An Exploratory Study of AI Chatbots Impact on University Teachers' Assessment Practices. *Assess. Eval. High. Educ.* **2024**, *49*, 363–375. [CrossRef]
33. Lancaster, T. Artificial Intelligence, Text Generation Tools and ChatGPT—Does Digital Watermarking Offer a Solution? *Int. J. Educ. Integr.* **2023**, *19*, 10. [CrossRef]
34. Tam, W.; Huynh, T.; Tang, A.; Luong, S.; Khatri, Y.; Zhou, W. Nursing Education in the Age of Artificial Intelligence Powered Chatbots (AI-Chatbots): Are We Ready Yet? *Nurse Educ. Today* **2023**, *129*, 105917. [CrossRef]
35. Kumah-Crystal, Y.; Mankowitz, S.; Embi, P.; Lehmann, C. ChatGPT and the Clinical Informatics Board Examination: The End of Unproctored Maintenance of Certification? *J. Am. Med. Inform. Assoc. JAMIA* **2023**, *30*, 1558–1560. [CrossRef] [PubMed]
36. Abd-alrazaq, A.; Alsaad, R.; Alhuwail, D.; Ahmed, A.; Healy, M.; Latifi, S.; Aziz, S.; Damseh, R.; Alrazak, S.; Sheikh, J. Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions. *JMIR Med. Educ.* **2023**, *9*, e48291. [CrossRef] [PubMed]
37. Liu, H.; Azam, M.; Bin Naeem, S.; Faiola, A. An Overview of the Capabilities of ChatGPT for Medical Writing and Its Implications for Academic Integrity. *Health Inf. Libr. J.* **2023**, *40*, 440–446. [CrossRef]
38. Al-Ghonmein, A.M.; Al-Moghrabi, K.G. The Potential of ChatGPT Technology in Education: Advantages, Obstacles and Future Growth. *IAES Int. J. Artif. Intell.* **2024**, *13*, 1206–1213. [CrossRef]

39. Ajevski, M.; Barker, K.; Gilbert, A.; Hardie, L.; Ryan, F. ChatGPT and the Future of Legal Education and Practice. *Law Teach.* **2023**, *57*, 352–364. [CrossRef]
40. Deebel, N.; Terlecki, R. ChatGPT Performance on the American Urological Association (AUA) Self-Assessment Study Program and the Potential Influence of Artificial Intelligence (AI) in Urologic Training. *Urology* **2023**, *177*, 29–33. [CrossRef]
41. Megahed, F.; Chen, Y.-J.; Ferris, J.; Knoth, S.; Jones-Farmer, L.A. How Generative AI Models Such as ChatGPT Can Be (Mis)Used in SPC Practice, Education, and Research? An Exploratory Study. *Qual. Eng.* **2024**, *36*, 287–315. [CrossRef]
42. Zhou, J.; Ke, P.; Qiu, X.; Huang, M.; Zhang, J. ChatGPT: Potential, Prospects, and Limitations. *Front. Inf. Technol. Electron. Eng.* **2023**, *25*, 6–11. [CrossRef]
43. Karabacak, M.; Ozkara, B.B.; Margetis, K.; Wintermark, M.; Bisdas, S. The Advent of Generative Language Models in Medical Education. *JMIR Med. Educ.* **2023**, *9*, e48163. [CrossRef] [PubMed]
44. Cross, J.; Robinson, R.; Devaraju, S.; Vaughans, A.; Hood, R.; Kayalackakom, T.; Honnavar, P.; Naik, S.; Sebastian, R. Transforming Medical Education: Assessing the Integration of ChatGPT Into Faculty Workflows at a Caribbean Medical School. *Cureus* **2023**, *15*, e41399. [CrossRef] [PubMed]
45. Nikolic, S.; Daniel, S.; Haque, R.; Belkina, M.; Hassan, G.M.; Grundy, S.; Lyden, S.; Neal, P.; Sandison, C. ChatGPT Versus Engineering Education Assessment: A Multidisciplinary and Multi-Institutional Benchmarking and Analysis of This Generative Artificial Intelligence Tool to Investigate Assessment Integrity. *Eur. J. Eng. Educ.* **2023**, *48*, 559–614. [CrossRef]
46. Nune, A.; Iyengar, K.; Manzo, C.; Barman, B.; Botchu, R. Chat Generative Pre-Trained Transformer (ChatGPT): Potential Implications for Rheumatology Practice. *Rheumatol. Int.* **2023**, *43*, 3. [CrossRef] [PubMed]
47. Heimans, S.; Biesta, G.; Takayama, K.; Kettle, M. ChatGPT, Subjectification, and the Purposes and Politics of Teacher Education and Its Scholarship. *Asia-Pac. J. Teach. Educ.* **2023**, *51*, 105–112. [CrossRef]
48. 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]
49. Baker, B.; Mills, K.A.; McDonald, P.; Wang, L. AI, Concepts of Intelligence, and Chatbots: The "Figure of Man", the Rise of Emotion, and Future Visions of Education. *Teach. Coll. Rec.* **2023**, *125*, 60–84. [CrossRef]
50. Beerepoot, M. Formative and Summative Automated Assessment with Multiple-Choice Question Banks. *J. Chem. Educ.* **2023**, *100*, 2947–2955. [CrossRef]
51. Greiner, C.; Peisl, T.; Höpfl, F.; Beese, O. Acceptance of AI in Semi-Structured Decision-Making Situations Applying the Four-Sides Model of Communication—An Empirical Analysis Focused on Higher Education. *Educ. Sci.* **2023**, *13*, 865. [CrossRef]
52. Khan, R.A.; Jawaid, M.; Khan, A.R.; Sajjad, M. ChatGPT—Reshaping Medical Education and Clinical Management. *Pak. J. Med. Sci.* **2023**, *39*, 605–607. [CrossRef] [PubMed]
53. Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learn. Individ. Differ.* **2023**, *103*, 102274. [CrossRef]
54. Lacey, M.; Smith, D. Teaching and Assessment of the Future Today: Higher Education and AI. *Microbiol. Aust.* **2023**, *44*, 124–126. [CrossRef]
55. Milton, C.L. ChatGPT and Forms of Deception. *Nurs. Sci. Q.* **2023**, *36*, 232–233. [CrossRef] [PubMed]
56. Zhang, B. ChatGPT, an Opportunity to Understand More About Language Models. *Med. Ref. Serv. Q.* **2023**, *42*, 194–201. [CrossRef] [PubMed]
57. Castro, C.A. de A Discussion about the Impact of ChatGPT in Education: Benefits and Concerns. *J. Bus. Theory Pract.* **2023**, *11*, 28. [CrossRef]
58. Alasadi, E.A.; Baiz, C.R. Generative AI in Education and Research: Opportunities, Concerns, and Solutions. *J. Chem. Educ.* **2023**, *100*, 2965–2971. [CrossRef]
59. Barrot, J.S. Using ChatGPT for Second Language Writing: Pitfalls and Potentials. *Assess. Writ.* **2023**, *57*, 100745. [CrossRef]
60. Clark, T. Investigating the Use of an Artificial Intelligence Chatbot with General Chemistry Exam Questions. *J. Chem. Educ.* **2023**, *100*, 1905–1916. [CrossRef]
61. Cloesmeijer, M.; Janssen, A.; Koopman, S.; Cnossen, M.; Mathôt, R. ChatGPT in Pharmacometrics? Potential Opportunities and Limitations. *Br. J. Clin. Pharmacol.* **2023**, *90*, 360–365. [CrossRef]
62. Currie, G.; Singh, C.; Nelson, T.; Nabasenja, C.; Al-Hayek, Y.; Spuur, K. ChatGPT in Medical Imaging Higher Education. *Radiography* **2023**, *29*, 792–799. [CrossRef] [PubMed]
63. Dalalah, D.; Dalalah, O. The False Positives and False Negatives of Generative AI Detection Tools in Education and Academic Research: The Case of ChatGPT. *Int. J. Manag. Educ.* **2023**, *21*, 100822. [CrossRef]
64. Ibrahim, H.; Asim, R.; Zaffar, F.; Rahwan, T.; Zaki, Y. Rethinking Homework in the Age of Artificial Intelligence. *IEEE Intell. Syst.* **2023**, *38*, 24–27. [CrossRef]
65. Kooli, C. Chatbots in Education and Research: A Critical Examination of Ethical Implications and Solutions. *Sustainability* **2023**, *15*, 5614. [CrossRef]
66. Masters, K. Ethical use of Artificial Intelligence in Health Professions Education: AMEE Guide No. 158. *Med. Teach.* **2023**, *45*, 574–584. [CrossRef] [PubMed]
67. Michel-Villarreal, R.; Vilalta-Perdomo, E.; Salinas-Navarro, D.E.; Thierry-Aguilera, R.; Gerardou, F.S. Challenges and Opportunities of Generative AI for Higher Education as Explained by ChatGPT. *Educ. Sci.* **2023**, *13*, 856. [CrossRef]

68. Oh, N.; Choi, G.-S.; Lee, W.Y. ChatGPT Goes to the Operating Room: Evaluating GPT-4 Performance and its Potential in Surgical Education and Training in the Era of Large Language Models. *Ann. Surg. Treat. Res.* **2023**, *104*, 269. [CrossRef]
69. Pretorius, L. Fostering AI Literacy: A Teaching Practice Reflection. *J. Acad. Lang. Learn.* **2023**, *17*, T1–T8.
70. Rahman, M.; Watanobe, Y. ChatGPT for Education and Research: Opportunities, Threats, and Strategies. *Appl. Sci.* **2023**, *13*, 5783. [CrossRef]
71. Sanchez Ruiz, L.M.; Moll-López, S.; Nuñez-Pérez, A.; Moraño, J.; Vega, E. ChatGPT Challenges Blended Learning Methodologies in Engineering Education: A Case Study in Mathematics. *Appl. Sci.* **2023**, *13*, 6039. [CrossRef]
72. Shaikh, S.; Yildirim Yayilgan, S.; Klimova, B.; Pikhart, M. Assessing the Usability of ChatGPT for Formal English Language Learning. *Eur. J. Investig. Health Psychol. Educ.* **2023**, *13*, 1937–1960. [CrossRef] [PubMed]
73. Thomas, S.P. Grappling with the Implications of ChatGPT for Researchers, Clinicians, and Educators. *Issues Ment. Health Nurs.* **2023**, *44*, 141–142. [CrossRef] [PubMed]
74. Tsang, R. Practical Applications of ChatGPT in Undergraduate Medical Education. *J. Med. Educ. Curric. Dev.* **2023**, *10*, 23821205231178449. [CrossRef] [PubMed]
75. Weng, T.-L.; Wang, Y.-M.; Chang, S.; Chen, T.-J.; Hwang, S.-J. ChatGPT Failed Taiwan's Family Medicine Board Exam. *J. Chin. Med. Assoc.* **2023**, *86*, 762–766. [CrossRef] [PubMed]
76. Bhattacharya, K.; Bhattacharya, N.; Bhattacharya, A.; Yagnik, V.; Garg, P. ChatGPT in Surgical Practice-a New Kid on the Block. *Indian J. Surg.* **2023**, *85*, 1346–1349. [CrossRef]
77. Eager, B.; Brunton, R. Prompting Higher Education Towards AI-Augmented Teaching and Learning Practice. *J. Univ. Teach. Learn. Pract.* **2023**, *20*, 2. [CrossRef]
78. French, F.; Levi, D.; Maczo, C.; Simonaityte, A.; Triantafyllidis, S.; Varda, G. Creative Use of OpenAI in Education: Case Studies from Game Development. *Multimodal Technol. Interact.* **2023**, *7*, 81. [CrossRef]
79. Fütterer, T.; Fischer, C.; Alekseeva, A.; Chen, X.; Tate, T.; Warschauer, M.; Gerjets, P. ChatGPT in Education: Global Reactions to AI Innovations. *Sci. Rep.* **2023**, *13*, 15310. [CrossRef] [PubMed]
80. Glaser, N. Exploring the Potential of ChatGPT as an Educational Technology: An Emerging Technology Report. *Technol. Knowl. Learn.* **2023**, *28*, 1945–1952. [CrossRef]
81. Livberber, T.; Ayvaz, S. The Impact of Artificial Intelligence in Academia: Views of Turkish Academics on ChatGPT. *Heliyon* **2023**, *9*, e19688. [CrossRef]
82. Marzuki; Widiati, U.; Rusdin, D.; Darwin; Indrawati, I. The Impact of AI Writing Tools on the Content and Organization of Students' Writing: EFL Teachers' Perspective. *Cogent Educ.* **2023**, *10*, 2236469. [CrossRef]
83. Abu hammour, K.; Alhamad, H.; Al-Ashwal, F.; Halboup, A.; Abu Farha, R. ChatGPT in Pharmacy Practice: A Cross-Sectional Exploration of Jordanian Pharmacists' Perception, Practice, and Concerns. *J. Pharm. Policy Pract.* **2023**, *16*, 115. [CrossRef]
84. Borger, J.; Ng, A.; Anderton, H.; Ashdown, G.; Auld, M.; Blewitt, M.; Brown, D.; Call, M.; Collins, P.; Freytag, S.; et al. Artificial Intelligence Takes Center Stage: Exploring the Capabilities and Implications of ChatGPT and Other AI-Assisted Technologies in Scientific Research and Education. *Immunol. Cell Biol.* **2023**, *101*, 923–935. [CrossRef]
85. Chang, D.; Lin, M.; Hajian, S.; Wang, Q. Educational Design Principles of Using AI Chatbot That Supports Self-Regulated Learning in Education: Goal Setting, Feedback, and Personalization. *Sustainability* **2023**, *15*, 12921. [CrossRef]
86. Ellis, A.; Slade, E. A New Era of Learning: Considerations for ChatGPT as a Tool to Enhance Statistics and Data Science Education. *J. Stat. Data Sci. Educ.* **2023**, *31*, 128–133. [CrossRef]
87. Hosseini, M.; Gao, C.; Liebovitz, D.; Carvalho, A.; Ahmad, F.; Luo, Y.; MacDonald, N.; Holmes, K.; Kho, A. An Exploratory Survey about Using ChatGPT in Education, Healthcare, and Research. *PLoS ONE* **2023**, *18*, e0292216. [CrossRef]
88. Ibrahim, H.; Liu, F.; Asim, R.; Battu, B.; Benabderrahmane, S.; Alhafni, B.; Adnan, W.; Alhanai, T.; AlShebli, B.; Baghdadi, R.; et al. Perception, Performance, and Detectability of Conversational Artificial Intelligence across 32 University Courses. *Sci. Rep.* **2023**, *13*, 12187. [CrossRef]
89. Kohnke, L.; Moorhouse, B.; Zou, D. ChatGPT for Language Teaching and Learning. *Relc J.* **2023**, *54*, 537–550. [CrossRef]
90. Lower, K.; Seth, I.; Lim, B.; Seth, N. ChatGPT-4: Transforming Medical Education and Addressing Clinical Exposure Challenges in the Post-pandemic Era. *Indian J. Orthop.* **2023**, *57*, 1527–1544. [CrossRef] [PubMed]
91. Mohapatra, D.; MT, F.; Tripathy, S.; Rajan, S.; Vathulya, M.; Lakshmi, P.; Singh, V.; Haq, A. Leveraging Large Language Models (LLM) for the Plastic Surgery Resident Training: Do They Have a Role? *Indian J. Plast. Surg.* **2023**, *56*, 413–420. [CrossRef] [PubMed]
92. Moshirfar, M.; Altaf, A.W.; Stoakes, I.M.; Tuttle, J.J.; Hoopes, P.C. Artificial Intelligence in Ophthalmology: A Comparative Analysis of GPT-3.5, GPT-4, and Human Expertise in Answering StatPearls Questions. *Cureus* **2023**, *15*, e40822. [CrossRef] [PubMed]
93. Passmore, J.; Woodward, W. Coaching Education: Wake up to the New Digital and AI Coaching Revolution! *Int. Coach. Psychol. Rev.* **2023**, *18*, 58–72. [CrossRef]
94. Su, Y.; Lin, Y.; Lai, C. Collaborating with ChatGPT in Argumentative Writing Classrooms. *Assess. Writ.* **2023**, *57*, 100752. [CrossRef]
95. Walters, W.H.; Wilder, E.I. Fabrication and Errors in the Bibliographic Citations Generated by ChatGPT. *Sci. Rep.* **2023**, *13*, 14045. [CrossRef] [PubMed]

96. Wilby, R.; Esson, J. AI Literacy in Geographic Education and Research: Capabilities, Caveats, and Criticality. *Geogr. J.* **2023**, *190*, e12548. [CrossRef]
97. Bender, S.M. Coexistence and Creativity: Screen Media Education in the Age of Artificial Intelligence Content Generators. *Media Pract. Educ.* **2023**, *24*, 351–366. [CrossRef]
98. Chiu, T.K.F. The Impact of Generative AI (GenAI) on Practices, Policies and Research Direction in Education: A Case of ChatGPT and Midjourney. *Interact. Learn. Environ.* **2023**. [CrossRef]
99. Clark, T.; Anderson, E.; Dickson-Karn, N.; Soltanirad, C.; Tafini, N. Comparing the Performance of College Chemistry Students with ChatGPT for Calculations Involving Acids and Bases. *J. Chem. Educ.* **2023**, *100*, 3934–3944. [CrossRef]
100. Ruiz, L.; Acosta-Vargas, P.; De-Moreta-Llovet, J.; Gonzalez, M. Empowering Education with Generative Artificial Intelligence Tools: Approach with an Instructional Design Matrix. *Sustainability* **2023**, *15*, 11524. [CrossRef]
101. Sison, A.J.; Daza, M.; Gozalo-Brizuela, R.; Garrido-Merchán, E. ChatGPT: More Than a Weapon of Mass Deception, Ethical Challenges and Responses from the Human-Centered Artificial Intelligence (HCAI) perspective. *Int. J. Hum. Comput. Interact.* **2023**. [CrossRef]
102. Zhang, B.; Qian, M. ChatGPT Related Technology and Its Applications in the Medical Field. *Adv. Ultrasound Diagn. Ther.* **2023**, *7*, 158. [CrossRef]
103. Dobbs, T. ChatGPT: Do We Need to Write Anything Ever Again? *Bulletin* **2023**, *105*, 82–83. [CrossRef]
104. Johnson, W. How to Harness Generative AI to Accelerate Human Learning. *Int. J. Artif. Intell. Educ.* **2023**, 1–5. [CrossRef]
105. Kikerpill, K.; Siibak, A. App-Hazard Disruption: An Empirical Investigation of Media Discourses on ChatGPT in Educational Contexts. *Comput. Sch.* **2023**, *40*, 334–355. [CrossRef]
106. Marquez, R.; Barrios, N.; Vera, R.E.; Mendez, M.E.; Tolosa, L.; Zambrano, F.; Li, Y. A Perspective on the Synergistic Potential of Artificial Intelligence and Product-Based Learning Strategies in Biobased Materials Education. *Educ. Chem. Eng.* **2023**, *44*, 164–180. [CrossRef]
107. Ratten, V.; Jones, P. Generative Artificial Intelligence (ChatGPT): Implications for Management Educators. *Int. J. Manag. Educ.* **2023**, *21*, 100857. [CrossRef]
108. Kaneda, Y.; Takahashi, R.; Kaneda, U.; Akashima, S.; Okita, H.; Misaki, S.; Yamashiro, A.; Ozaki, A.; Tanimoto, T. Assessing the Performance of GPT-3.5 and GPT-4 on the 2023 Japanese Nursing Examination. *Cureus* **2023**, *15*, e42924. [CrossRef] [PubMed]
109. Luo, W.; He, H.; Liu, J.; Berson, I.; Berson, M.; Zhou, Y.; Li, H. Aladdin's Genie or Pandora's Box for Early Childhood Education? Experts Chat on the Roles, Challenges, and Developments of ChatGPT. *Early Educ. Dev.* **2023**, *2023*, 2214181. [CrossRef]
110. Orrù, G.; Piarulli, A.; Conversano, C.; Gemignani, A. Human-like Problem-Solving Abilities in Large Language Models Using ChatGPT. *Front. Artif. Intell.* **2023**, *6*, 1199350. [CrossRef]
111. Giannos, P. Evaluating the Limits of AI in Medical Specialisation: ChatGPT's Performance on the UK Neurology Specialty Certificate Examination. *BMJ Neurol. Open* **2023**, *5*, e000451. [CrossRef]
112. Romero-Rodríguez, J.-M.; Ramírez-Montoya, M.-S.; Buenestado-Fernández, M.; Lara Lara, F. Use of ChatGPT at University as a Tool for Complex Thinking: Students' Perceived Usefulness. *J. New Approaches Educ. Res.* **2023**, *12*, 323–339. [CrossRef]
113. Yan, D. Impact of ChatGPT on Learners in a L2 Writing Practicum: An Exploratory Investigation. *Educ. Inf. Technol.* **2023**, *28*, 13943–13967. [CrossRef]
114. Elkhatat, A.; Elsaid, K.; Almeer, S. Evaluating the Efficacy of AI Content Detection Tools in Differentiating between Human and AI-Generated Text. *Int. J. Educ. Integr.* **2023**, *19*, 17. [CrossRef]
115. Kortemeyer, G. Could an Artificial-Intelligence Agent Pass an Introductory Physics Course? *Phys. Rev. Phys. Educ. Res.* **2023**, *19*, 010132. [CrossRef]
116. Perkins, M.; Roe, J. Decoding Academic Integrity Policies: A Corpus Linguistics Investigation of AI and Other Technological Threats. *High Educ Policy* **2023**, 1–21. [CrossRef]
117. Chan, C. A Comprehensive AI Policy Education Framework for University Teaching and Learning. *Int. J. Educ. Technol. High. Educ.* **2023**, *20*, 38. [CrossRef]
118. Esplugas, M. The Use of Artificial Intelligence (AI) to Enhance Academic Communication, Education and Research: A Balanced Approach. *J. Hand Surg. Eur. Vol.* **2023**, *48*, 819–822. [CrossRef]
119. Shoufan, A. Exploring Students' Perceptions of ChatGPT: Thematic Analysis and Follow-Up Survey. *IEEE Access* **2023**, *11*, 38805–38818. [CrossRef]
120. Alshater, M.M. Exploring the Role of Artificial Intelligence in Enhancing Academic Performance: A Case Study of ChatGPT. *SSRN* **2022**. *pre-print*. [CrossRef]
121. Deshpande, S.; Szefer, J. Analyzing ChatGPT's Aptitude in an Introductory Computer Engineering Course. *arXiv* **2023**, arXiv:2304.06122. [CrossRef]
122. Chan, C.K.Y.; Hu, W. Students' Voices on Generative AI: Perceptions, Benefits, and Challenges in Higher Education. *arXiv* **2023**, arXiv:2305.00290. [CrossRef]

MDPI AG
Grosspeteranlage 5
4052 Basel
Switzerland
Tel.: +41 61 683 77 34

Education Sciences Editorial Office
E-mail: education@mdpi.com
www.mdpi.com/journal/education

Academic Open
Access Publishing

mdpi.com

ISBN 978-3-7258-6683-0