

Article

When AI Speaks the Curriculum: From Generic Chatbots to Context-Aware Learning Systems

Ali Ahsan , Hayden McDonald , Ratna Saha * and Claire Davison 

School of Technology, Torrens University Australia, Adelaide, SA 5000, Australia; ali.ahsan@torrens.edu.au (A.A.); hayden.mcdonald@torrens.edu.au (H.M.); claire.davison@torrens.edu.au (C.D.)

* Correspondence: ratna.saha@torrens.edu.au

Abstract

Artificial intelligence (AI) is rapidly transforming higher education. Notably, institutions are increasingly deploying AI-enabled chatbots across administrative, teaching and learning, and research functions. Despite the proliferation of these tools, most conversational systems remain generic and disconnected from subject-level context, curriculum design, and academic integrity requirements. A design-oriented framework is proposed in this paper to improve the effectiveness of AI-enabled chatbots used in teaching and learning, embedded within curriculum. The framework is grounded in systems thinking for context-aware AI-enabled learning systems that operate within bounded pedagogical and disciplinary environments. Adopting a design science approach, the study synthesises expert-informed insights from pedagogical, programmatic, and subject-level perspectives to develop a framework that integrates context alignment, instructional control, and integrity-preserving guardrails. The resulting artefact was verified through artefact-focused testing against the proposed design objectives using a structured non-human evaluation protocol, including requirement-based testing, baseline comparison with generic AI systems, academic integrity stress testing, and robustness analysis. The study proposes and verifies a framework intended to support curriculum alignment, instructional control, and academic integrity preservation within AI-enabled learning systems. The paper contributes a systems-oriented framework for embedding AI within educational systems while preserving pedagogical intent and governance requirements. Implications for scalable deployment of AI in higher education and future human-centred evaluation are discussed.

Keywords: educational chatbot; artificial intelligence (AI); curriculum; context-aware learning systems; RAG



Academic Editors: Xuesong Zhai,
Andreja Istenic and Jiong Guo

Received: 20 May 2026

Revised: 30 June 2026

Accepted: 3 July 2026

Published: 6 July 2026

Copyright: © 2026 by the authors.
Licensee MDPI, Basel, Switzerland.
This article is an open access article
distributed under the terms and
conditions of the [Creative Commons
Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The integration of artificial intelligence (AI) into higher education is transforming how students interact with knowledge. AI, particularly large language models (LLMs), supposedly augment human cognition [1]. Yet, learners can now use AI to circumvent the epistemic constructs and pedagogical processes that have historically been central to knowledge and skill development (e.g., uncertainty, cognitive effort, and iterative meaning-making). AI challenges the traditional process of learning, where human cognition develops into deep understanding through sustained intellectual engagement and reflection [2]. AI is invoking a new learning paradigm characterised by immediacy, fluency, and apparent completeness of knowledge. Hence, an urgent reconfiguration of the epistemological and

pedagogical foundations of learning is required to prevent AI from gradually displacing processes that underpin human cognition and understanding.

The generative and agentic capability of AI is transforming learning environments. AI enables rapid access to information and automated synthesis of knowledge, prompting deeper human–technological collaboration and shifts in creative and knowledge production practices [3]. This transformation foregrounds a tension between efficiency and epistemic development. Learners can delegate tasks of analysis, interpretation, and reasoning to AI systems. This cognitive offloading has significant implications for higher education, where the objective is not merely knowledge acquisition but the development of critical thinking, metacognitive awareness, and disciplinary reasoning [4].

Many AI-enabled tools remain fundamentally generic [5]. In higher education, these generic tools operate independently of disciplinary structures, curriculum design, and assessment intent. The tools produce responses that are context-agnostic and often misaligned with learning objectives [5]. For example, rule-based chatbots operate with no memory of learner interactions. Similarly, retrieval-based chatbots match queries to responses without tracking dialogue history. While task-based chatbots have limited context for a task, they do not have complete conversational awareness [6].

Generic AI chatbots provide technically accurate information. However, the chatbots lack awareness of the epistemic and pedagogical foundations that define learning in relation to subject, curriculum, and integrity requirements [7]. This limitation is particularly pronounced in learning contexts characterised by structured bodies of knowledge and applied reasoning, where learners must understand and apply concepts, (e.g., apply frameworks or processes in real-world scenarios). As generic AI chatbots offer answers without facilitating understanding, the developmental trajectory of learning required in these contexts (e.g., moving from understanding to application) is potentially undermined.

The integration of generic AI chatbots [8] in education contexts that require higher-order cognitive skills is also problematic. In postgraduate contexts, learners are expected to engage with complex, codified knowledge frameworks, such as those aligned with professional bodies of knowledge (e.g., business analysis frameworks encompassing lifecycle management, elicitation techniques, and strategy analysis). To demonstrate knowledge and skill development in these learning environments, learners navigate extensive theoretical material, interpret structured methodologies, and apply them within scenario-based tasks and professional certification contexts. The cognitive load associated with such subjects is inherently high, and effective learning depends on the ability to identify knowledge gaps, contextualise concepts, and iteratively refine understanding. Thus, the integration of AI in higher education presents both an opportunity to support learning and a risk of undermining the very processes that the curriculum seeks to develop.

In response to these challenges, recent approaches have begun to explore the development of context-aware AI chatbots that operate within bounded pedagogical environments. Context-aware chatbots are designed to align with subject-specific content, instructional strategies, and assessment requirements. For example, generative AI chatbots are inherently capable of using conversational contexts. However, these generative AI chatbots often use foundational models that lack acute knowledge about learning systems. Conversely, retrieval augmented generation (RAG) chatbots and conversational agents are explicitly designed to be context aware at both conversational and knowledge levels [6,9].

This study adopts a design science approach to develop a framework for context-aware AI-enabled learning systems. The framework is informed by multi-stakeholder expert insights spanning pedagogy, curriculum design, learning experience, and system development, and is grounded in the implementation of an AI-enabled chatbot within a postgraduate subject context. It conceptualises AI as part of a layered learning system

comprising context, instruction, guardrails, outputs, and governance, each contributing to the alignment of system behaviour with educational objectives. The artefact is subsequently verified through structured non-human testing, and a validation protocol is proposed for future human-centred evaluation.

The AI chatbot examined in this study is a RAG chatbot embedded within an institutional learning environment. The retrieval-augmented architecture draws on curated subject materials and controlled knowledge sources. The chatbot is integrated directly into the learning management system through standardised interfaces, enabling authenticated and personalised interactions while maintaining institutional oversight of content and access. Its design incorporates mechanisms for contextual grounding, semantic retrieval, and controlled response generation, ensuring that output remains aligned with the curriculum.

Accordingly, this study is guided by the following research questions:

RQ1: *How can a context-aware AI-enabled chatbot be designed to align with curriculum structure, pedagogical intent, academic integrity requirements, and institutional governance expectations?*

RQ2: *How can a Design Science Research approach be used to develop and verify a curriculum-grounded AI chatbot artefact within a higher education learning environment?*

RQ3: *What design layers and control mechanisms are required to support the responsible deployment of context-aware AI-enabled learning systems in higher education?*

The study demonstrates, through artefact design and verification, how AI chatbot integration extends beyond technical implementation to incorporate pedagogical design principles aimed at addressing cognitive and epistemic concerns. Rather than providing direct answers, the system employs structured interaction strategies (e.g., Socratic questioning) to prompt reflection, encourage engagement with learning materials, and guide learners toward independent understanding. Included guardrails in the system constrain inappropriate or assessment-compromising outputs, thereby addressing issues of academic integrity and aligning AI behaviour with educational objectives. This design approach directly responds to concerns regarding cognitive offloading, positioning the AI chatbot to facilitate rather than substitute critical thinking.

The findings also reveal deeper systemic and institutional tensions regarding AI chatbot integration. Specifically, the design and deployment of AI-enabled learning systems involve multiple stakeholders, including educators, learning designers, developers, and institutional governance structures. Each stakeholder has differing priorities and levels of control. While educators are responsible for curriculum design and pedagogical intent, the underlying AI infrastructure (e.g., data pipelines, model configurations, and retrieval mechanisms) is often controlled by technical teams. This fragmentation of ownership means that critical aspects of the learning system, such as the knowledge base and its ongoing maintenance, may fall outside the direct control of educators. Additionally, operational considerations such as token usage, scalability, and cost introduce constraints that shape how AI systems can be implemented at scale. These dynamics highlight that AI integration is not solely a technical or pedagogical challenge, but a socio-technical one requiring alignment across multiple layers of the learning system [10].

From a systems perspective, learning environments must therefore be reconceptualised as complex, adaptive systems in which knowledge construction emerges from interactions among students, educators, AI, and institutional infrastructures and governance. Within such systems, control is distributed rather than centralised, and the role of AI extends beyond information provision to influence how knowledge is accessed, interpreted, and validated. This entanglement of actors introduces new forms of agency and dependency, requiring careful design of interaction mechanisms, governance structures, and accountability frameworks. The effectiveness of AI-enabled learning systems thus depends not only

on their technical capabilities but on their alignment with pedagogical intent, curriculum structure, and institutional governance.

The contribution of this study is threefold. First, it advances a conceptual understanding of AI as an embedded component within learning systems. This embodiment of AI highlights the importance of context, control, and interaction in shaping learning outcomes. Second, the study provides a practical, design-oriented framework for developing context-aware AI systems that align with curriculum and assessment requirements in higher education. Third, the study proposes a structured approach to designing AI artefacts that offers a pathway for the responsible and scalable incorporation of AI into higher education learning systems.

2. Conceptual Background

2.1. AI in Education

The integration of AI into higher education has evolved from early intelligent tutoring systems to advanced systems capable of generating complex, contextually coherent natural language responses. These advanced systems are increasingly being integrated into learning environments to support tasks such as information retrieval, feedback generation, and content personalisation [11,12]. Notably, large language models (LLMs) have introduced a new paradigm in which AI systems can simulate dialogue, generate explanations, and respond dynamically to queries. This functionality is reshaping the interaction between learners and knowledge, positioning AI as an active participant in the learning process rather than a passive tool [13].

Arguably, AI systems are technically advanced but pedagogically lacking. Generative AI tools are typically trained on broad, general-purpose datasets and are not inherently designed to reflect the specific learning objectives, curriculum structures, or assessment constraints of individual subjects [14,15]. Consequently, these AI systems may provide accurate or plausible responses but lack sensitivity to the epistemic boundaries and instructional sequencing that underpin effective learning. This misalignment can lead to over-simplified explanations, inappropriate levels of assistance, and, in some cases, responses that conflict with the intended learning outcomes.

An underlying tension exists between the generative capabilities of AI and the structured, progressive nature of learning. AI systems are optimised for producing complete and coherent outputs. Conversely, learning processes often rely on incomplete understanding, iterative reasoning, and guided discovery. Thus, AI systems designed for efficiency may inadvertently bypass the cognitive processes that learning seeks to cultivate. For example, a student who uses a generative AI tool to instantly write a technically accurate and well-structured essay risks bypassing essential learning processes (e.g., analysing information sources, constructing arguments, iteratively refining ideas, and applying theoretical concepts in practice).

An important pedagogical priority is ensuring that students continue to develop higher-order cognitive skills when using AI [5]. The concept of cognitive offloading is used to describe situations whereby learners use AI systems to perform tasks that require critical thinking, synthesis, and reflective engagement [10,16]. This offloading reduces the cognitive load on learners and supports efficiency but often bypasses critical learning processes for developing deep understanding and disciplinary thinking. Cognitive offloading is particularly problematic in domains that require structured reasoning and problem-solving. The availability of immediate responses generated by AI can shift a learner's engagement from problem-solving toward seeking, comprehending and interpreting solutions.

AI must be integrated into learning environments in ways that support cognitive engagement and learning processes. Consequently, recent research has shifted focus from

the capabilities of AI systems toward their pedagogical integration, emphasising the need for designs that align AI behaviour with instructional intent, disciplinary context, and assessment requirements [17–19]. This shift provides a foundation for the development of context-aware and pedagogically grounded AI systems that are better suited to support meaningful learning in educational environments.

2.2. Context-Aware Systems

Context-awareness refers to the capability of a system to utilise relevant information about its environment, users, and tasks to adapt its behaviour accordingly [20]. In educational settings, context-aware systems aim to tailor interactions based on factors such as subject content, learner progress, instructional objectives, and assessment constraints [21]. Such systems move beyond generic information provision by delivering responses that are situated within specific learning contexts, thereby ensuring closer alignment with pedagogical intent and supporting more meaningful and targeted learning experiences.

In the context of AI-enabled learning, context-awareness is particularly critical due to the limitations of general-purpose large language models. Without sufficient contextual grounding, AI systems may generate responses that are technically accurate but pedagogically misaligned, for example, by providing excessive detail, bypassing learning processes, or offering solutions that compromise assessment integrity. Context-aware approaches address this limitation by constraining AI outputs through curated knowledge sources, structured prompting strategies, and retrieval mechanisms that align responses with domain-specific content and instructional intent [22].

Context-aware AI systems are commonly implemented using a RAG model. This approach uses external knowledge bases to ground model responses in verified and contextually relevant information [22,23]. By integrating semantic search and embedding-based retrieval techniques, such systems enable the selection of relevant content from curated sources. Constraining responses to an indexed knowledge source can improve their factual accuracy (e.g., reduce hallucination) and alignment with target domains (e.g., accounting, business analysis, etc.). However, while such systems improve alignment with domain knowledge, they do not fully resolve challenges related to control, interpretability, and reliability [24]. System outputs remain probabilistic and highly sensitive to factors such as prompt design, data coverage, and retrieval accuracy. Furthermore, pedagogical alignment cannot be assumed in context-aware AI systems. For example, a RAG system might retrieve accurate business analytics content and generate appropriate responses without supporting meaningful learning outcomes (e.g., oversimplify concepts, bypass analytical reasoning, etc.).

A pedagogically aligned AI system extends context-awareness beyond data grounding to include explicit instructional control and behavioural constraints. This grounding requires designing systems that not only retrieve relevant information but also structure responses in ways that scaffold learning, promote reflection, and support cognitive engagement. Prior research on AI-enabled learning systems emphasises the importance of scaffolding and adaptive support mechanisms to guide learner interaction and regulate metacognitive processes [25]. More research is needed to understand design challenges associated with introducing instructional constraints. For example, excessive restriction may limit exploratory learning and learner autonomy, while insufficient constraint may result in inappropriate or overly directive assistance.

Balancing these competing demands represents a central challenge in the design of AI-enabled learning systems. Achieving this balance is critical to ensure that AI functions not merely as an answer-generating tool, but as a pedagogically aligned learning facilitator that supports meaningful engagement, reflection, and knowledge construction.

2.3. Systems Thinking and Learning Systems

The integration of AI within educational environments requires a shift from viewing technology as an external tool to understanding it as an embedded actor that shapes learning interactions and outcomes. Systems thinking provides a useful lens for analysing this dynamic, emphasising the interdependence of components, the distribution of control, and the emergence of system behaviour from interactions among elements [26,27]. Within this perspective, AI systems are no longer passive tools but active participants that shape both access to knowledge and the processes through which it is constructed.

Within a learning system, multiple actors interact to shape the learning experience (e.g., including students, educators, instructional designers, AI systems, and institutional infrastructures). Knowledge is not transmitted in a linear manner but emerges through these interactions, influenced by task design, content structure, and the behaviour of technological components. The introduction of AI fundamentally reshapes these dynamics, as such systems can mediate, filter and influence how knowledge is accessed, interpreted, and constructed.

From this perspective, AI-enabled learning environments can be understood as socio-technical systems, where technical components (e.g., algorithms, data pipelines, interfaces) are deeply intertwined with social and pedagogical processes [28]. In these systems, control and agency are distributed rather than centralised, giving rise to complex challenges related to governance, accountability, and alignment. For instance, while educators define learning objectives and assessment criteria, the behaviour of AI systems is influenced by underlying models, training data, and system configurations that may lie beyond their direct control [15,29].

This distributed nature of control introduces potential tensions between pedagogical intent and system behaviour. Without appropriate governance mechanisms, AI systems may operate in ways that conflict with instructional goals by providing inappropriate assistance, reinforcing misconceptions or prioritising efficiency over learning processes. Simultaneously, technical constraints such as model limitations, retrieval accuracy, and computational dependencies shape what is feasible within the system [13,30]. These dynamics underscore the importance of designing AI-enabled learning systems that explicitly address the interplay between pedagogical, technical, and institutional dimensions.

In practice, this distribution of control often results in fragmented ownership of system behaviour, where pedagogical intent, technical implementation, and operational constraints are managed by different stakeholders. Such fragmentation can lead to misalignment between intended system outcomes and actual behaviour in real learning environments, highlighting the need for integrated governance and design mechanisms.

Accordingly, a systems-oriented perspective emphasises the importance of developing integrated design frameworks that consider not only the capabilities of AI technologies but also their role within the broader learning ecosystem. Such frameworks must account for context, interaction, control, and governance, ensuring that AI systems contribute to, rather than disrupt, learning processes.

These considerations reinforce that the challenge of integrating AI into education is not solely one of technological capability, but of alignment—across context, instructional intent, system behaviour, and governance structures. Addressing this challenge requires structured, interdisciplinary design approaches that provide a foundation for developing AI-enabled learning systems that are both contextually grounded and pedagogically aligned.

The proposed framework is therefore supported by several complementary theoretical and conceptual foundations. Design Science Research provides the methodological foundation for developing and verifying an artefact intended to address a practical educational problem. Systems thinking and socio-technical systems theory support the framing

of AI-enabled learning environments as interconnected systems involving pedagogical, technical, organisational, and governance components. Context-aware systems theory and retrieval-augmented generation literature inform the design of curriculum-grounded response generation. Instructional design alignment, scaffolding, and cognitive offloading literature support the Instruction and Guardrail layers by emphasising the need to preserve learner reasoning, reflection, and engagement. Finally, academic integrity and AI governance literature support the inclusion of behavioural constraints, institutional oversight, and accountability mechanisms within the framework.

3. Research Approach

3.1. Design Science Research

This study adopts a Design Science Research (DSR) approach to address the misalignment between generic AI systems and structured educational environments. DSR is appropriate for research that seeks to develop and evaluate artefacts designed to solve identified problems through purposeful design and systematic inquiry [31,32]. It emphasises the creation of artefacts that are both theoretically grounded and practically relevant, ensuring that proposed solutions are not only conceptually robust but also applicable within real-world contexts.

Design Science Research differs from explanatory and predictive research traditions by focusing on the creation of purposeful artefacts intended to address identified organisational or practical problems. The primary outcome of DSR is not the generation of descriptive theory alone, but the development of innovative artefacts that contribute both practical utility and theoretical knowledge [31,33]. Such artefacts may take the form of constructs, models, methods, or instantiations and are typically developed through iterative cycles of design, implementation, evaluation, and refinement [31].

The research follows established DSR processes, including problem identification, knowledge elicitation, design synthesis, artefact development, verification and communication [31,33]. Within this study, the artefact is conceptualised not merely as a technical solution but as a structured learning system that integrates context-awareness, instructional control, behavioural constraints, and governance mechanisms. This approach is further informed by a systems-thinking perspective, recognising that artefact behaviour emerges from dynamic interactions among multiple actors and institutional constraints, rather than from isolated technical components alone.

The study is particularly informed by the Design Science Research Methodology (DSRM) proposed by Peffers et al. [32], which comprises problem identification, objective definition, design and development, demonstration, evaluation, and communication. Consistent with this methodology, the present study identifies a gap between generic AI systems and structured educational environments, develops a context-aware educational AI artefact and verifies its alignment with predefined design objectives through structured artefact verification activities.

The research procedure comprised six stages. First, the educational problem was identified through the recognised misalignment between generic AI systems and structured curriculum environments. Second, stakeholder-informed design input was gathered through professional consultation and documented engagement sessions involving academics, learning designers, technical developers, and institutional stakeholders. Third, the design input was synthesised into design objectives and framework components. Fourth, the artefact was developed through a RAG-based chatbot architecture using curated curriculum resources, prompt-based instructional control, and behavioural guardrails. Fifth, the artefact was verified through structured non-human testing using curriculum materials, researcher-generated prompts, and practitioner-led assessment of system behaviour. Sixth,

a future validation protocol was developed to guide subsequent empirical evaluation involving learners and educational stakeholders.

No experiments involving student participants were conducted in this study. The study did not collect student data, user analytics, personal information, or demographic information. The stakeholder engagement activities informed artefact design and implementation, but participants were not studied as research subjects. Accordingly, demographic reporting is not applicable to the present study.

3.2. Artefact Definition

The artefact developed in this study is a context-aware, AI-enabled learning system, operationalised through an implemented chatbot and formalised through a layered design framework. Within the DSR paradigm, artefacts may be conceptualised as constructs, models, methods, and instantiations [31]. This study integrates all four dimensions:

- Constructs: context, instructional control, behavioural constraints (guardrails), and governance
- Model: a layered architecture representing the structure of the learning system
- Method: structured interaction and response strategies guiding system behaviour
- Instantiation: an implemented chatbot deployed within an educational environment

Operationally, the artefact is realised through a RAG pipeline, supported by a system prompt-driven instructional layer (agent prompt template, Refer to Appendices A and B), curated subject-specific knowledge sources, and integration within a learning management system (LMS) via standardised interoperability protocols [30].

The defining characteristic of the artefact lies in its ability to operate within bounded pedagogical contexts through the integration of curated knowledge, structured prompting, and controlled response behaviour. Unlike generic AI systems, which function independently of instructional design, this artefact is explicitly aligned with curriculum structure and pedagogical intent. As such, it is conceptualised not as a standalone tool, but as an embedded component of a broader learning system that actively supports and shapes the learning process.

3.3. Stakeholder-Informed Design Input and Engagement

The artefact was developed in consultation with various key stakeholders (e.g., academics, learning designers, and system developers, etc.). Expert-informed input from stakeholders during the design process shaped the artefact's development. Drawing on their engagement in this process as school, programme, and subject leaders, the authors developed a framework for the design of context-aware chatbots in educational settings. The engagements were conducted as part of system design and stakeholder consultation activities and did not constitute human subject research, as no personal, sensitive, or identifiable data were collected or analysed. Accordingly, the framework was developed through a design-focused research approach informed by professional practice and consultation in educational settings.

The consultation informing the artefact's development primarily emerged organically through the authors' routine professional activities at the university. These activities included curriculum and subject development, including decisions concerning the integration of emerging technologies into academic programmes. The authors were extensively involved in artefact design and implementation across subjects and degree programmes through their close engagement with technical teams and learning designers.

In addition to these ongoing development activities, seven documented stakeholder engagement sessions were conducted to support reflection on the artefact's design, implementation, operation, and governance. These sessions involved academics, learning

designers, technical developers, and institutional stakeholders directly associated with the chatbot initiative. The purpose of these sessions was not to generate empirical findings about participants, but to capture design-relevant insights, implementation experiences, perceived risks, governance considerations, and pedagogical requirements that informed the development of the framework.

A series of stakeholder engagement and design consultation activities were conducted among the authors and relevant institutional stakeholders involved in the artefact's design, implementation, and operationalisation. These stakeholders included subject leadership, learning design personnel, and technical development staff associated with the chatbot initiative. An initial set of guiding questions was developed to structure the engagement activities, focusing on learning processes, AI behaviour, control mechanisms, and system limitations. The sessions were conducted iteratively, allowing for probing, clarification, and refinement of emerging ideas. All stakeholder consultation/design documentation sessions were documented and transcribed to support rigorous interpretation and traceability of expert input during the design synthesis process. The sessions were used to document domain-specific knowledge, pedagogical considerations, technical constraints, and governance challenges associated with the integration of AI in education environments.

The working partnership between academics, learning designers, and technical system developers enabled the capture of implementation-level insights (e.g., system architecture, retrieval strategies, prompt design, and operational constraints). This insight ensured that the design process incorporated not only pedagogical considerations but also system-level feasibility and behavioural characteristics, consistent with socio-technical design perspectives [28]. An engagement session was also held with a learning designer and system developers to clarify the authors' interpretation of the artefact's development and implementation.

The design-research approach adopted in this study drew upon stakeholder expertise and professional practice to inform design decision-making and artefact development [34]. The emphasis in this study was on generating actionable, design-relevant insights rather than producing generalisable empirical findings.

3.4. Design Synthesis

The elicited expert input was synthesised into design-relevant knowledge through an iterative process of abstraction and interpretation. This stage serves as a critical bridge between expert-informed design insights and artefact design, whereby findings from expert engagements were integrated with existing system knowledge and the broader educational context to inform design decisions. Such synthesis processes are central to Design Science Research, where knowledge is transformed into actionable artefact designs through systematic reasoning and refinement [31,32].

Recurring design concerns discussed across stakeholder engagement activities included cognitive offloading risks, the need for contextual grounding, limitations of AI controllability, and governance challenges. These themes were translated into design principles and systematically mapped to specific system components, including context layers, instructional strategies, behavioural constraints (guardrails), and governance mechanisms. This mapping ensured traceability between identified challenges and corresponding design features, aligning with best practices in design-oriented research [33].

Additionally, the synthesis incorporated implementation-level insights related to retrieval pipelines, prompt-based control strategies, and system integration constraints. This integration enabled the alignment of conceptual design objectives with technically feasible configurations, ensuring that the resulting artefact remained both theoretically sound and practically implementable.

Overall, this synthesis process reflects core DSR principles, where knowledge derived from multiple sources is iteratively refined into actionable design artefacts that address real-world problems [32,33]. The outcome is a structured framework that directly informs both the conceptual model and the implemented system.

3.5. Design Process

The artefact was developed through an iterative design and implementation process spanning system configuration, integration, and continuous refinement. The design process followed a structured pipeline comprising content acquisition, data preprocessing and structuring, retrieval configuration, prompt engineering, and system integration. Such iterative cycles of build–evaluate–refine are central to Design Science Research, ensuring both rigour and relevance in artefact development [31,32].

Subject-specific content was sourced from structured subject materials and segmented into smaller semantic units (chunks) to enable effective indexing and retrieval. These units were subsequently transformed into vectorised representations to support similarity-based retrieval. A RAG mechanism was implemented to identify and dynamically inject the most relevant content into the response generation process, thereby grounding outputs in domain-specific knowledge and improving contextual relevance [22].

Prompt design and interaction logic were explicitly structured to support pedagogical objectives, including guided questioning, scaffolding and controlled response strategies. These controls were operationalised through a system-level agent prompt template, which defined response tone, behavioural constraints, scope boundaries, and alignment with instructional intent. Such prompt-based control mechanisms have been shown to influence model behaviour and improve task alignment in large language model systems [35,36].

The design process incorporated continuous refinement based on observed system behaviour and stakeholder feedback, ensuring that both technical feasibility and pedagogical alignment was systematically addressed. This iterative refinement reflects real-world constraints, including limitations in model reliability, data coverage, and system control, which are widely recognised challenges in AI system design [13].

3.6. Artefact Verification

In Design Science Research, a distinction exists between artefact verification and artefact validation. Verification focuses on determining whether an artefact has been implemented in accordance with its intended design objectives and functional requirements, whereas validation evaluates whether the artefact achieves its intended outcomes in real-world use contexts. The present study focuses primarily on artefact verification through the assessment of system behaviour, contextual alignment, instructional control mechanisms, guardrails, and governance structures. A validation protocol is subsequently proposed to support future empirical evaluation involving educational stakeholders and learning environments.

The artefact was evaluated using a structured, non-human verification approach with established DSR evaluation principles [31,33]. Within DSR, evaluation focuses on assessing whether an artefact effectively addresses the identified problem and meets its design objectives, rather than solely relying on user perception or satisfaction [31,32]. Accordingly, this study emphasised functional and behavioural verification of the system.

The evaluation comprised multiple components, including requirement-based testing, comparison with generic AI systems, assessment of academic integrity constraints, and robustness testing under varying input conditions. These evaluation dimensions enabled systematic verification of the artefact's alignment with predefined functional requirements, instructional intent, and behavioural constraints.

In practice, verification was conducted through iterative testing using live subject materials, with evaluation performed by developers, learning designers, and subject matter experts. The process focused on assessing response correctness, contextual grounding, and consistency, as well as identifying the presence of hallucinated or unsupported outputs, issues widely recognised in large language model systems [13]. This practitioner-led evaluation approach ensured that both technical performance and pedagogical alignment were considered during verification.

Importantly, the evaluation did not rely on formal benchmark datasets or controlled experimental protocols, but instead prioritised realistic system behaviour within the intended learning context. This approach is consistent with DSR's emphasis on relevance and practical applicability, particularly when designing artefacts for complex socio-technical environments [33]. As no human participants were involved (e.g., student data, or user-generated learning analytics), the evaluation focused exclusively on system behaviour under controlled conditions. This ensured that the artefact was assessed in terms of functional performance, alignment with design objectives, and robustness, while avoiding ethical concerns associated with human subject research.

Consequently, the study claims that verification activities assessed whether the artefact behaved consistently with its intended design objectives within the target educational context. The validation protocol presented in Section 7 is intended to support future empirical evaluation involving learners and educational stakeholders.

3.7. Systems Perspective on Artefact Design

The artefact is conceptualised within a systems thinking framework, recognising that its behaviour emerges from dynamic interactions among pedagogical, technical, and organisational components. From this perspective, the system operates as part of a broader socio-technical environment, in which control and agency are distributed across stakeholders, technologies and institutional structures rather than centralised in a single component [26,28].

In practice, system behaviour is co-shaped by multiple interacting elements, including instructional design decisions (learning designers), technical implementation (system developers), and operational constraints such as infrastructure, platform capabilities and data accessibility. This interaction results in a distributed control structure, where no single stakeholder has complete control over system behaviour. Thus, system outcomes emerge as various components interact.

This perspective highlights that artefact performance is influenced not only by system design, but also by factors such as data quality, model limitations, and institutional governance arrangements [13]. By adopting a systems-oriented approach, the study ensures that the artefact is aligned with the broader learning ecosystem and addresses the interplay between context, instructional control, system behaviour, and governance.

Figure 1 demonstrates the Design Science Research process adopted in this study, illustrating the progression from problem identification and stakeholder-informed design synthesis through framework development, artefact implementation, artefact verification, and future validation planning.

The application of systems thinking in this study is primarily interpretive and design-oriented rather than analytical or simulation-based. The objective is not to develop a formal system dynamics model, causal-loop representation, or cybernetic analysis of educational AI systems. Rather, systems thinking is used as a conceptual lens to understand the interdependencies among pedagogical, technical, organisational, and governance components and to inform the design of a context-aware educational AI framework. This perspective

emphasises emergence, distributed control, stakeholder interaction, and alignment across system components within a socio-technical learning environment.

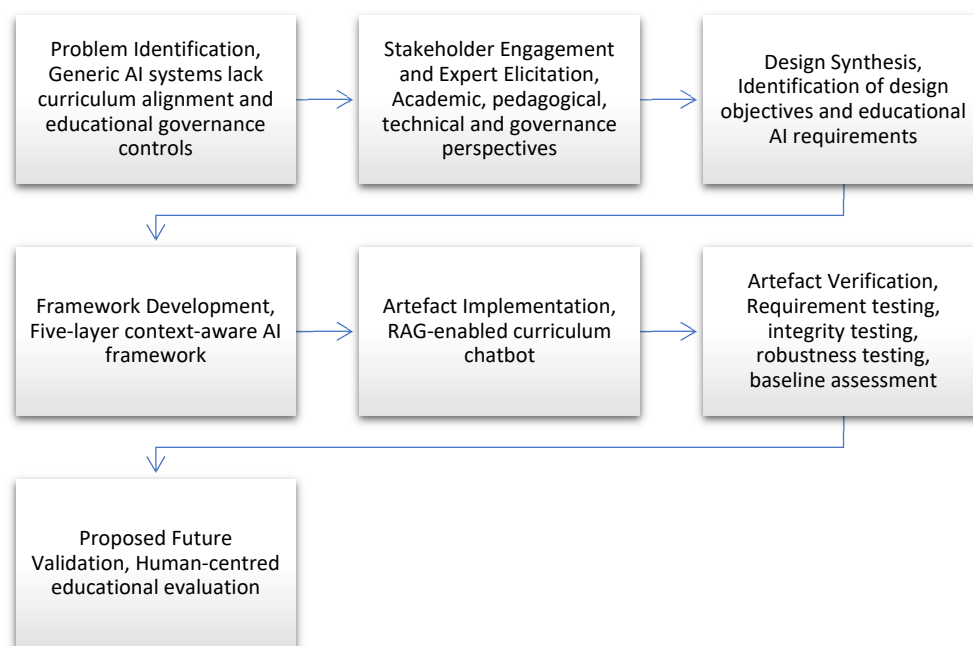


Figure 1. Design Science Research Process Used in the Study.

4. Design Objectives and Principles

4.1. Design Objective 1: Subject Alignment

A central limitation of generic AI systems in educational contexts is their inability to align with the structured nature of disciplinary knowledge and curriculum design. While such systems may generate accurate or plausible responses, they operate independently of the epistemic boundaries, sequencing (e.g., intentional ordering of learning experiences), and contextual dependencies that underpin subject-specific learning. This misalignment is particularly problematic in domains characterised by codified bodies of knowledge and applied reasoning, where understanding depends not only on content accuracy but on its positioning within structured frameworks and processes [13,29].

Learning extends beyond acquiring information to include developing contextualised understanding within defined disciplinary structures. AI systems that fail to align with this learning progression risk fragmenting knowledge by presenting isolated explanations without reference to their broader conceptual or procedural context. This misalignment can undermine coherence in learning and contribute to a student's superficial understanding of a discipline or topic. Prior research in instructional design highlights the importance of aligning learning materials, activities, and assessment with intended learning outcome and curriculum structure, highlighting alignment as a foundational principle for effective learning [37].

Accordingly, AI-enabled learning systems must be designed to operate within explicitly defined curricular boundaries, drawing on curated knowledge sources that reflect the structure, scope, and sequencing of the subject. The integration of domain-specific materials, learning outcomes, and assessment frameworks into the AI system's knowledge base is necessary to ensure that responses are both accurate and contextually relevant. In practice, this alignment is implemented through RAG mechanisms, where authorised subject materials are ingested, segmented into semantic units (chunks), embedded into vector representations, and dynamically retrieved at query time. Such approaches ground

system outputs in subject-specific knowledge rather than general-purpose datasets, improving contextual relevance and reducing the risk of misaligned or decontextualised responses [22].

Design Principle 1: AI-enabled learning systems should be grounded in structured, curriculum-aligned knowledge representations that ensure responses are contextually situated within disciplinary frameworks and reflect the subject's intended scope, sequencing, and epistemic boundaries.

4.2. Design Objective 2: Pedagogical Consistency

The integration of AI into learning environments introduces a fundamental tension between the efficiency of automated response generation and the pedagogical objective of fostering cognitive engagement. While AI systems are optimised to produce complete and coherent outputs, effective learning often depends on processes of exploration, uncertainty, and incremental construction. Consequently, AI systems risk inadvertently bypassing the cognitive effort required for meaningful learning by providing direct answers or overly simplified explanations, thereby reducing opportunities for critical thinking and knowledge development [13].

Pedagogical consistency requires that AI behaviour aligns with the instructional strategies and learning processes embedded within the curriculum. AI-enabled systems must support the development of higher-order cognitive skills, including analysis, synthesis, and reflection, rather than merely facilitating information retrieval. This requirement aligns with well-established learning theories, including constructivist perspectives on knowledge construction and cognitive load theory, which emphasise the importance of appropriately managing cognitive effort to support learning [38,39].

To achieve this alignment, AI-enabled learning systems must incorporate instructional control mechanisms that shape how responses are generated and delivered. These mechanisms may include guided questioning, progressive disclosure of information, and prompts that encourage engagement with underlying concepts rather than immediate solution provision. In practice, such instructional controls can be operationalised through a system-level agent prompt template that defines response tone, pedagogical strategy, scope boundaries, response structure, and behavioural constraints. The objective is to ensure that AI functions as a facilitator of learning by supporting inquiry, reflection, and conceptual development rather than acting as a substitute for the learning process itself.

Design Principle 2: AI-enabled learning systems should incorporate instructional control mechanisms that align system responses with pedagogical intent, promoting cognitive engagement, reflective learning, and the development of higher-order thinking skills rather than direct answer provision.

4.3. Design Objective 3: Boundary-Controlled AI Responses

A critical challenge in the deployment of AI systems in education is the absence of effective mechanisms to constrain system behaviour within appropriate boundaries. Generic AI systems are typically optimised to maximise response completeness and relevance, which can result in outputs that extend beyond the intended scope of learning, provide excessive detail, or bypass essential stages of the learning process. In educational contexts, such behaviour can undermine instructional design principles and compromise the integrity of learning activities [13,30].

Boundary control is therefore essential to ensure that AI systems operate within defined limits, both in terms of content and response behaviour. This includes restricting responses to relevant knowledge domains, regulating the level of detail provided and preventing the system from offering direct solutions that circumvent learning objectives.

Boundary control also involves managing the interaction between context-awareness and response generation, ensuring that the system does not overgeneralise, hallucinate information, or introduce content that is inconsistent with the intended curriculum [22].

Implementing boundary-controlled responses requires the integration of multiple design elements, including curated knowledge sources, structured prompting strategies, and rule-based or heuristic constraints on system behaviour. These elements operate collectively to regulate both the scope and form of system outputs, ensuring alignment with pedagogical goals and contextual requirements. In practice, boundary control can be operationalised through a combination of retrieval constraints and prompt-level rules. The constraints limit responses to validated subject-specific content and the rules enable out-of-scope queries to be rejected, redirected, or reframed. Such mechanisms are increasingly recognised as essential for improving the reliability, safety, and controllability of AI systems [13,40].

Design Principle 3: AI-enabled learning systems should incorporate boundary control mechanisms that regulate both the scope and form of system responses, ensuring that outputs remain aligned with learning objectives, contextual constraints, and the intended progression of the learning process.

4.4. Design Objective 4: Academic Integrity Preservation

The integration of AI into educational environments creates academic integrity challenges, particularly relating to assessments. AI systems capable of generating complete solutions, explanations, or artefacts can be misused to bypass assessment processes, thereby undermining the validity of learning outcomes and the credibility of evaluation mechanisms. This risk has been increasingly highlighted in recent research on generative AI in education, which emphasises concerns related to academic misconduct, reduced learner accountability, and erosion of assessment integrity [41].

To preserve academic integrity, AI systems must be designed to support learning while mitigating behaviours that compromise assessment integrity processes. Safeguards must be embedded to mitigate potential AI misuse and restrict inappropriate or assessment-compromising outputs. Various safeguards can be implemented at the system level, such as limiting direct answer provision, controlling response depth, or guiding learners towards conceptual understanding. However, academic integrity also extends beyond the capabilities of individual AI systems including plagiarism detection, authorship verification, and assessment misuse monitoring. These broader functions require the integration of AI systems with institutional policies, assessment design strategies, and external monitoring mechanisms [13,15].

Importantly, academic integrity must be treated as a fundamental design constraint rather than a post hoc consideration. This necessitates the incorporation of governance mechanisms that define acceptable system behaviour, enforce usage policies, and ensure alignment with institutional standards. From a socio-technical perspective, maintaining academic integrity requires coordinated action across technological, pedagogical, and institutional dimensions, recognising that system behaviour is shaped not only by design choices but also by organisational policies and user practices [29].

Design Principle 4: AI-enabled learning systems should embed academic integrity safeguards and governance mechanisms that prevent misuse, constrain inappropriate assistance, and ensure that system behaviour aligns with assessment requirements and institutional standards.

4.5. Design Objective 5: Governance and Control Integration

The challenges of integrating AI systems within educational settings extend beyond system behaviour and pedagogical alignment to encompass governance issues. These AI

systems are designed and implemented by multiple stakeholders including academics, learning designers, system developers, and other institutional management. Each stakeholder has distinct responsibilities and varying levels of authority. This diversity often results in a fragmentation of control, where critical aspects of an AI system's behaviour may fall outside the direct influence of those responsible for pedagogical design [15,29]. This fragmentation can lead to misalignment between pedagogical intent and system execution. For example, educators define learning objectives and instructional strategies while technical teams determine system architecture, data pipelines, and operational constraints. Additionally, institutional priorities such as scalability, cost efficiency, and infrastructure limitations shape system behaviour in ways that may not be fully visible to pedagogical stakeholders.

The governance challenges associated with AI system design and implementation are consistent with socio-technical systems theory. Specifically, system theory emphasises the interdependence of social, organisational and technical components in shaping system outcomes [26,28]. In practice, governance is distributed across institutional units, where learning designers configure instructional units and instructional elements, developers manage technical systems, and learning or emerging technology teams oversee deployment, maintenance, and operational control.

Addressing these challenges requires the explicit integration of governance considerations into system design. AI-enabled learning systems must incorporate mechanisms that enable coordination across stakeholder groups, ensure transparency of system behaviour, and provide pathways for aligning technical implementation with pedagogical objectives. In this context, governance extends beyond formal policy frameworks to include design-level decisions that shape how control is distributed, exercised and monitored within the system [15].

Design Principle 5: AI-enabled learning systems should incorporate governance and control integration mechanisms that align pedagogical intent, technical implementation, and institutional constraints, ensuring transparency, accountability, and coordinated control across all stakeholders involved in the learning system.

5. Proposed Framework: Context-Aware Educational Chatbot

The framework in this study operationalises the design objectives outlined in Section 4 through a structured, layered architecture for context-aware AI-enabled learning systems. The framework conceptualises this system as an integration of interdependent layers that collectively align AI behaviour with curriculum structure, pedagogical intent, system constraints, and institutional governance. Rather than treating AI as a standalone tool, the framework positions it as an embedded component within a broader learning system, where knowledge, interaction, control, and oversight are distributed across multiple layers. Each layer addresses a distinct dimension of system design while contributing to the overall alignment between AI behaviour and educational objectives. This layered approach reflects established principles in modular system design, which emphasise decomposition, separation of concerns, and scalability, as well as socio-technical systems theory, which highlights the interdependence of technical, organisational, and pedagogical elements in shaping system outcomes [31,42].

In implementation, this layered architecture is realised through the integration of RAG, system-level agent prompt templates, curated domain and subject specific knowledge sources, and interoperability with learning management system (LMS). This combination enables the system to ground responses in structured knowledge, regulate behaviour through instructional and boundary controls, and operate within institutional and infrastructural constraints. By embedding these elements within a unified architecture, the

framework ensures that AI behaviour is not only technically effective but also contextually grounded and pedagogically aligned.

Figure 2 presents the proposed five-layer framework for context-aware AI-enabled learning systems. The framework integrates curriculum context, instructional control, behavioural guardrails, response generation, and governance mechanisms to align AI system behaviour with pedagogical and institutional objectives.

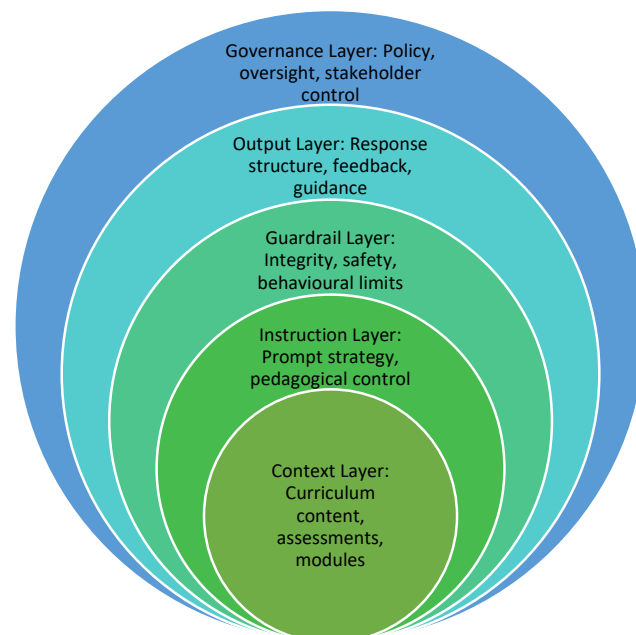


Figure 2. Five-Layer Framework for Context-Aware AI-Enabled Learning Systems.

5.1. Context Layer

The Context Layer defines the knowledge environment within which the AI system operates. It is responsible for grounding system responses in domain-specific content, ensuring alignment with curriculum structure, intended learning outcomes, and disciplinary knowledge frameworks. This layer draws on curated knowledge sources, including subject materials, assessment specifications, and structured curricular artefacts, reflecting the importance of aligning learning resources with educational objectives [37].

The primary function of the Context Layer is to constrain the system's knowledge space to relevant and authorised content. This is achieved through mechanisms such as RAG, where external knowledge repositories are used to inform and contextualise response generation. These repositories are structured to reflect the organisation, sequencing, and scope of the subject, enabling the system to produce responses that are situated within the intended learning context.

In practice, subject content is segmented into smaller semantic units, embedded into vector representations, and retrieved using semantic similarity-based methods. The most relevant content is dynamically injected into the prompt context during response generation, allowing the system to produce outputs that are both contextually appropriate and aligned with domain-specific knowledge [43]. The retrieval process therefore functions as an intermediary knowledge layer between curriculum resources and the language model, reducing reliance on the model's pre-trained knowledge and improving the traceability of responses to authorised educational content [22–24].

Importantly, the effectiveness of this layer depends on the quality, completeness, and organisation of underlying knowledge sources. Variations in data structure, coverage, and semantic representation can significantly influence the relevance, consistency, and reliability

of retrieved content. As such, the Context Layer forms the foundational mechanism for subject alignment by embedding the epistemic boundaries and conceptual structure of the curriculum within the system, supporting coherent and meaningful learning experiences [29,37].

5.2. Instruction Layer

The Instruction Layer governs how the system interacts with learners, shaping both the form and intent of responses in alignment with pedagogical objectives. This layer is responsible for translating instructional strategies into system behaviour, ensuring that interactions support learning processes rather than merely delivering information. This translation is achieved through structured prompt design and interaction logic that guide the system's responses. The layer supports strategies such as guided questioning, progressive disclosure, and conceptual prompting, enabling the system to facilitate reflection, exploration, and deeper cognitive engagement, rather than passive consumption of information. By controlling the structure and tone of responses, the Instruction Layer ensures that AI interactions remain consistent with the intended learning design, emphasising the role of scaffolding and incremental knowledge construction in effective learning [38,44].

This Instruction layer is operationalised through system-level prompt engineering, typically in the form of agent prompt templates that define response tone, structure, behavioural rules, scope handling, and pedagogical strategy. The role of this layer is particularly critical in mitigating the risks associated with direct answer provision. By embedding instructional intent within response generation, the system shifts from an answer-oriented model toward a learning-oriented model that supports inquiry, reasoning, and conceptual understanding. This aligns with broader pedagogical frameworks that emphasise active engagement and higher-order cognitive development [13,37].

5.3. Guardrail Layer

The Guardrail Layer introduces behavioural constraints that regulate system outputs in accordance with predefined boundaries. This layer ensures that responses remain within acceptable limits, both in terms of content relevance and the level of assistance provided. It plays a critical role in maintaining alignment with pedagogical objectives and preventing unintended system behaviours that may undermine learning processes or assessment integrity.

Guardrails operate by integrating rule-based and heuristic mechanisms that constrain inappropriate, excessive, or misaligned responses. These mechanisms may include limiting the provision of complete solutions, controlling response depth, and filtering outputs that fall outside the intended knowledge domain. This layer also supports the prevention of interactions that may undermine learning processes or assessment objectives.

The constraints introduced in the Guardrail Layer are supported through both model-level safeguards and prompt-level rules, including rejection or redirection of out-of-scope queries, filtering of unsafe or irrelevant content, and restriction of responses to domain-specific knowledge boundaries defined within the system. Importantly, guardrails do not eliminate variability in system behaviour but act to shape and constrain it. Given the inherently probabilistic nature of AI systems, this layer functions as a moderating mechanism, reducing the likelihood of undesirable or inappropriate outputs while maintaining flexibility in interaction. The design of guardrails reflects broader approaches to AI safety and controllability [45].

5.4. Output Layer

The Output Layer represents the interface through which system responses are delivered to users. It is responsible for integrating inputs from the Context, Instruction, and Guardrail layers to generate coherent and contextually appropriate outputs.

In implementation, response generation follows a structured pipeline: user input is transformed into vectorized representations (embeddings), relevant content is retrieved through semantic search, top-ranked segments (chunks) are dynamically injected into the prompt along with the original query, and the language model subsequently generates the final response. It also reflects the inherent variability of AI-generated outputs, where responses may differ based on input phrasing, context availability, and underlying model behaviour.

Rather than guaranteeing deterministic outputs, the Output Layer is designed to balance consistency with flexibility. This enables the system to adapt to diverse learner inputs while maintaining alignment with overarching design objectives, pedagogical and contextual constraints. This reflects the probabilistic and adaptive nature of contemporary AI systems [46].

5.5. Governance Layer

The Governance Layer provides oversight and control mechanisms that ensure alignment between system behaviour, pedagogical intent, and institutional requirements. This layer addresses the distributed nature of control within AI-enabled learning systems, where multiple stakeholders collectively influence system design, implementation, and operation.

Governance mechanisms include institutional policies, monitoring processes, and system-level controls that define acceptable system behaviour and regulate usage. This mechanism supports transparency, accountability, and coordination across stakeholders' groups, ensuring that system operation remains consistent with institutional standards and constraints.

The Governance Layer also incorporates operational considerations such as scalability, resource constraints, system maintenance, and lifecycle management. This approach aligns with emerging work on AI accountability, which emphasises the need for end-to-end governance frameworks that encompass system design, deployment, monitoring, and auditing processes [47]. By integrating governance into the system architecture, the framework ensures that control is not external to the system but embedded within its design. This reflects principles from socio-technical systems theory, where system effectiveness depends on the alignment of social and technical components [28]. In practice, governance is enacted through institutional oversight structures, where system deployment, maintenance, and evolution are managed by technical teams responsible for learning systems and emerging technologies.

Governance considerations are increasingly reflected in emerging institutional, national, and international AI frameworks. For example, the NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0) emphasises accountability, transparency, risk management, and human oversight in AI-enabled systems [48]. Similarly, many higher education institutions are developing policies and governance structures to guide the responsible use of AI in teaching, learning, and assessment. The Governance Layer aligns with these broader developments by recognising that educational AI systems operate within institutional environments characterised by distributed responsibilities, policy constraints, and accountability requirements.

6. System Design and Interaction Architecture

Section 5 presented the proposed framework as a layered architecture comprising context, instruction, guardrail, output, and governance layers. Section 6 translates that framework into a system-level design, explaining how the context-aware chatbot is operationalised through curriculum inputs, retrieval mechanisms, prompt-based control, response generation, and interaction logic. The purpose of this section is not to present the system as a purely technical artefact, but to demonstrate how pedagogical intent, subject-specific content, and technical architecture are coordinated within a learning environment.

The system is designed as a context-aware AI-enabled chatbot embedded within the institutional learning environment. Its architecture integrates RAG, curated curriculum content, prompt-based behavioural control, model-level safeguards, and LMS integration. The design reflects a socio-technical approach in which academics, learning designers, and developers contribute to different aspects of system behaviour. This socio-technical approach is consistent with the layered framework proposed earlier, where context determines what the system can draw upon, instruction governs how it should respond, guardrails define what it should not do, output logic structures the final response, and governance determines how the system is maintained and controlled.

Figure 3 represents the system architecture of the implemented context-aware AI-enabled learning system. Curriculum resources are processed through a data preparation pipeline involving text chunking, embedding generation, and storage in a vector database to enable efficient semantic retrieval. The curated curriculum repository serves as the authoritative knowledge source, providing the content used to populate the retrieval database.

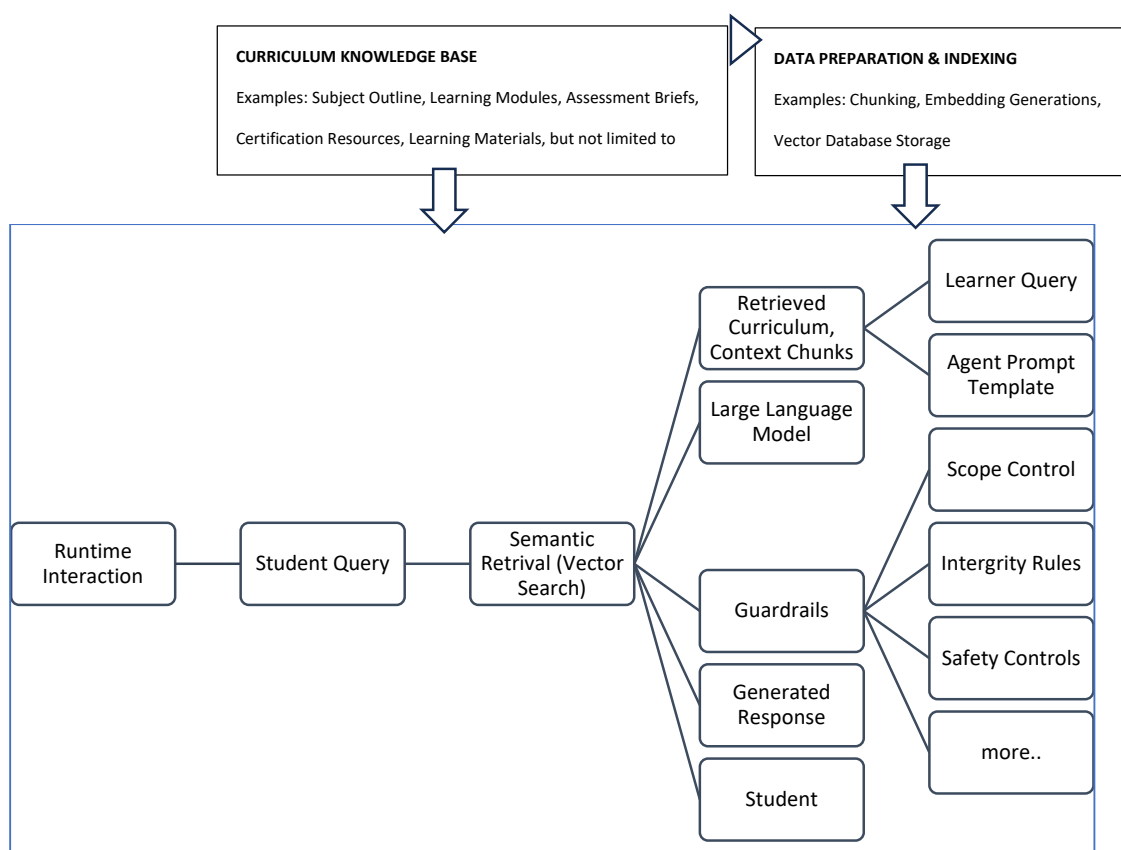


Figure 3. Architecture and Interaction Flow of the Implemented Context-Aware RAG-Based Chatbot. Arrows represent the flow of information through the context-aware RAG architecture.

During runtime, learner queries are submitted to the semantic retrieval model, which performs semantic matching between the query and indexed curriculum embeddings to identify relevant learning content. The retrieved content, combined with the learner query and prompt-based instructional controls, is then provided to the large language model for response generation. Generated responses are subsequently evaluated by guardrails implementing scope, academic integrity, and safety controls prior to delivery of responses to learners.

6.1. Knowledge Sources and Curriculum Inputs

The knowledge base of the system is derived from authorised curriculum materials associated with the subject context. In the implemented instance, the primary knowledge source consists of structured module pages and associated subject-level materials, rather than an unrestricted web-scale corpus. The aim is to ensure that responses are grounded in the content students are expected to engage with, rather than generated solely from the general knowledge of a large language model.

The use of curriculum materials as the primary knowledge source is central to the subject alignment objective identified in Section 4.1. These materials represent the boundaries of what the chatbot should know and respond to. The system therefore does not operate as a generic question-answering tool; rather, it functions as a bounded learning support mechanism that retrieves and uses subject-specific content. The knowledge sources may include module content, assessment-related information, relevant references, and subject-specific learning resources, subject to copyright, institutional access, and pedagogical relevance.

The selection and curation of knowledge sources is an important design decision. In practice, not all available content can or should be included. Legal, copyright, privacy, and pedagogical considerations determine what can be scraped, referenced, uploaded, or embedded into the knowledge base. The Agent Prompt and Data Template reinforce this by requiring developers and learning designers to identify the subject or module resources to be used for agent datasets, and to ensure that copyright-protected material is handled appropriately. This makes the Context Layer not merely a technical repository but a governed curriculum boundary.

Once selected, the content must be prepared for retrieval. The system relies on chunking, where larger documents or pages are divided into smaller semantic units. The Agent Prompt and Data Template specifies that chunks should preserve pedagogical atomic units, such as examples, definitions, exercises, figures, tables, rubrics, and surrounding explanatory text. This is important because poor chunking can fragment meaning and reduce the pedagogical usefulness of retrieved content. The chunking process therefore affects the quality of response grounding, the coherence of answers, and the capacity of the system to maintain alignment with curriculum intent.

6.2. Prompt and Instruction Strategy

The prompt and instruction strategy is the primary mechanism through which pedagogical intent is translated into system behaviour. While the RAG mechanism determines what contextual content is retrieved, the prompt layer determines how the system should use that content, how it should respond to learners, what tone it should adopt, and what boundaries it should observe.

The system uses an agent prompt template (Appendices A and B) as a structured mechanism for defining the character, role, behaviour, language, tone, interaction style, and expected response patterns of the chatbot. The template includes areas for defining the agent's persona, knowledge and expertise, role and behaviour, interaction structure, feed-

back guidelines, language style, tone, and dataset sources. This enables learning designers and developers to translate educational requirements into executable prompt instructions.

This prompt-based design is important because it shows that the Instruction Layer is not merely a theoretical category. It is operationalised through explicit prompt structures that prime the model before each interaction. The agent prompt template acts as the controlling instruction layer, defining how the chatbot should behave in relation to students, how it should respond to in-scope questions, how it should handle out-of-scope questions, and what kind of learning experience it should support.

The template draws on structured prompt design approaches such as goal, context, expectations, and source framing, as well as context, objective, style, tone, audience, and response framing. These structures allow designers to specify not only what the agent knows, but also how it should behave. For example, the prompt can instruct the chatbot to provide supportive and structured responses, avoid excessive detail, redirect irrelevant questions, maintain a learner-centred tone, and support inquiry rather than direct answer provision. This instruction is especially important in educational contexts where the system must guide learning without replacing the learner's cognitive effort.

In the implemented architecture, the system prompt is combined with retrieved contextual chunks and the learner's original input before being passed to the language model. This means that the generated response is shaped by three interacting inputs: the learner's question, the retrieved curriculum context, and the instructional prompt. This structure allows the system to be both context-aware and pedagogically controlled.

6.3. Guardrails and Response Constraints

The Guardrail Layer defines the constraints that regulate system behaviour across multiple levels. First, the system inherits built-in model-level safeguards that reduce the likelihood of generating offensive, unsafe, or inappropriate outputs [46,47]. Second, prompt-level instructions specify what the chatbot should or should not answer [35,36,40]. Third, retrieval constraints constrain the knowledge base to authorised and relevant subject-specific content.

A central guardrail within this framework is scope control. When a student asks a question outside the subject domain, the system is designed to avoid generating an unsupported general answer and instead redirect the learner back to the relevant learning context. This control is particularly important because large language models can produce fluent responses even when the answer is beyond the intended knowledge domain [47]. Without effective scope control, the chatbot could appear helpful while undermining curriculum alignment [35,42].

Another critical dimension of guardrail implementation relates to privacy and sensitive information handling. The system must avoid generating responses that disclose private, confidential, or unauthorised information, even when such a request is phrased as a legitimate learning query. Similarly, unsafe, harmful, or irrelevant prompts should be rejected or redirected. These constraints are implemented through combination of the model's inherent safety mechanisms and the agent prompt template, which specifies expected behaviour for out-of-scope or inappropriate interactions [35,40,46].

Academic integrity is also addressed through response constraints, although it cannot be fully enforced by the chatbot alone. The system can be instructed not to provide direct answers to assessment tasks, not to generate complete submissions, and not to bypass the learning process [42]. However, plagiarism detection, assessment misuse monitoring, and institutional misconduct processes remain external to the chatbot architecture. Recognising this distinction is important because it prevents overclaiming. The system

can support integrity-preserving interaction, but it does not replace institutional academic integrity framework.

The guardrail design therefore reflects a layered safety model: model-level safeguards reduce general risk, prompt-level rules define pedagogical and behavioural constraints, and retrieval boundaries restrict the system's accessible knowledge base. Collectively, these controls help align system behaviour with learning objectives while recognising that complete determinism cannot be guaranteed in probabilistic AI systems [47].

6.4. Contextual Grounding Mechanism

The contextual grounding mechanism is implemented through RAG. In this architecture, the student's input is first transformed into a vectorised embedding representation. This representation is then used to search the embedded knowledge base for semantically relevant chunks of subject content. The system then retrieves the most relevant chunks and injects them into the prompt context before the language model generates a response [49].

This mechanism enables the chatbot to move beyond generic response generation. Instead of relying exclusively on the model's pre-trained general knowledge, the system grounds its response in retrieved subject-specific content. The retrieved context acts as an anchor, increasing the likelihood that the response reflects the subject materials, terminology, and curriculum boundaries.

In the implemented system, the RAG mechanism is based on semantic search rather than a more complex hybrid retrieval model. While semantic retrieval enables the system to identify content related to the meaning of the student's query, retrieval effectiveness depends on chunk quality, embedding quality, data coverage, and query phrasing. If relevant content is not included in the knowledge base, poorly chunked, or semantically difficult to retrieve, the system's response may be incomplete or less aligned than intended.

The contextual grounding mechanism therefore improves alignment between system outputs and domain-specific knowledge but does not eliminate uncertainty. While it reduces reliance on general model knowledge, it remains sensitive to the quality and structure of the underlying data. This reinforces the need for careful curriculum input preparation, ongoing maintenance of the knowledge base, and verification of retrieved outputs. In this way, contextual grounding functions as both technical mechanism and governance responsibility.

6.5. Interaction and Response Logic

The interaction and response logic describes how user input moves through the system and becomes a response. The process begins when the learner accesses the chatbot through the learning environment and submits a query. The query is converted into embeddings, which are used to perform a semantic search over the vectorised knowledge base. The system identifies relevant chunks, ranks them by semantic similarity, and selects the most relevant content for inclusion in the prompt context.

The response payload is then constructed by combining the learner's original query, the retrieved subject-specific content, and the system-level prompt derived from the agent prompt template. This composite prompt is then sent to the language model, which generates a response shaped by both the retrieved contextual information and the instructional constraints embedded within the prompt. The output is then delivered to the learner through the chatbot interface.

This interaction logic supports context-aware learning in several ways. First, it enables the system to respond to student questions using relevant subject content. Second, it ensures that the response is shaped by pedagogical intent rather than solely relying on the model's inherent tendencies. Third, it enables redirection or refusal where questions fall

outside the intended scope. Fourth, it allows the system to acknowledge uncertainty rather than hallucinating unsupported answers.

The response logic also reflects the limitations of AI-generated interaction. Outputs remain probabilistic and may vary across repeated prompts due to sensitivity to input phrasing, context availability, and internal model dynamics.

The system is therefore designed to balance consistency without claiming full determinism. Where the system lacks sufficient context, it should acknowledge uncertainty, redirect the learner, or encourage engagement with relevant subject materials rather than fabricating an answer. This is consistent with the broader design objective of supporting learning while maintaining boundaries and epistemic integrity.

In summary, the system design and interaction architecture operationalises the framework through five interconnected mechanisms: curriculum-grounded knowledge sources, prompt-based instructional control, behavioural guardrails, retrieval-based contextual grounding, and structured response generation. These mechanisms collectively support a context-aware AI-enabled learning system that is pedagogically aligned, technically bounded, and institutionally governable.

7. Proposed Validation and Evaluation Protocol

The purpose of this section is to propose a structured validation and evaluation protocol for future assessment of the context-aware AI-enabled chatbot. While artefact verification was undertaken as part of the present study (Section 3.6), comprehensive validation involving educational users, learning outcomes, and institutional deployment contexts remains future work. Accordingly, this section presents a proposed validation framework rather than reporting completed validation results. The purpose of artefact validation is to evaluate whether the proposed context-aware AI chatbot behaves in ways that are consistent with its design objectives. In keeping with the DSR approach, the evaluation focuses on the artefact itself rather than on human participant outcomes. The validation therefore examines system behaviour, contextual grounding, response constraints, academic integrity boundaries, and robustness under controlled testing conditions.

The evaluation protocol is designed as a structured, non-human artefact validation process. It does not rely on student data, user analytics, or human-subject experimentation. Instead, it uses curriculum-derived test prompts, assessment-informed scenarios, expert-informed expectations, and controlled comparisons to evaluate whether the system performs according to its intended design. This approach is appropriate where the purpose is to validate artefact behaviour prior to or independently from future human-centred evaluation [50].

The validation protocol is aligned with the five design objectives presented in Section 4 and the five layers of the framework presented in Section 5. Subject alignment is evaluated by testing whether the system produces responses grounded in curriculum content. Pedagogical consistency is evaluated by examining whether responses guide learning rather than providing inappropriate direct answers. Boundary control is assessed through out-of-scope and ambiguous prompts. Academic integrity preservation is tested through assessment-compromising prompts. Governance and control integration are examined through the traceability of system behaviour to knowledge sources, prompt rules, and operational responsibilities.

7.1. Validation Design

The proposed validation design consists of a structured, scenario-based evaluation of the chatbot's behaviour. The goal of the proposed validation protocol is to determine whether the artefact performs in accordance with its intended design objectives under

controlled conditions. The validation does not aim to measure student satisfaction, learning gain, or behavioural change; those would require future human-centred evaluation. Instead, the focus is on artefact behaviour: whether the system retrieves relevant content, responds within scope, avoids unsupported answers, and maintains pedagogically appropriate interaction patterns.

The validation design draws on three sources. First, the subject outline and assessment briefs provide the curricular and assessment context against which the system's responses can be evaluated. Second, expert elicitation sessions provide expected behaviours and known risks, including cognitive offloading, direct answer provision, scope drift, and governance concerns. Third, the technical architecture defines how the system is expected to function, including retrieval, prompt control, output generation, and fallback behaviour.

The validation design therefore follows a requirements-to-test logic. Each design objective is translated into observable system expectations. For example, if the objective is subject alignment, the system should retrieve and use relevant curriculum content. If the objective is pedagogical consistency, the system should guide students rather than simply provide final answers. If the objective is boundary control, the system should reject or redirect out-of-scope queries. If the objective is academic integrity, the system should avoid producing complete assessment submissions. If the objective is governance, system behaviour should be traceable to defined configuration, knowledge sources, and responsibility structures.

A staged validation design is proposed. The first stage involves functional testing of normal learning support prompts. The second stage involves assessment-related prompts to evaluate whether the system maintains appropriate boundaries. The third stage involves out-of-scope and adversarial prompts to test redirection, refusal, and safety behaviour. The fourth stage involves repeated prompt testing to examine consistency and robustness.

7.2. Test Dataset Development

The test dataset should be developed from curriculum materials, assessment briefs, expert inputs, and expected student interaction patterns. Because the evaluation is non-human, the prompts should be synthetic or researcher-generated rather than drawn from identifiable student interactions. This ensures that the validation does not rely on human participant data while still reflecting plausible use cases.

The dataset should include five categories of prompts. The first category is subject knowledge prompts. These evaluate whether the system can explain concepts, processes, and terminology within the relevant subject domain. As the chatbot developed was specifically for the business analysis subject that includes an external industry-certified exam, therefore some examples may include prompts about requirements elicitation, stakeholder analysis, lifecycle management, BABOK (Business Analysis Body of Knowledge)-related concepts, or ECBA (Entry Certificate in Business Analysis) exam preparation.

The second category is learning support prompts. These prompts assess whether the chatbot can guide students who express confusion, partial understanding, or uncertainty. Such prompts are useful for evaluating the Instruction Layer because the appropriate response should not merely provide information but support the learner's reasoning process.

The third category is assessment guidance prompts. These prompts test whether the system can help students understand assessment expectations without producing assessment-completing responses. For example, students may ask for clarification about an assessment requirement, advice on structuring a reflection, or guidance on how to prepare for specific tasks.

The fourth category is academic integrity risk prompts. These prompts intentionally test whether the chatbot provides inappropriate assistance, such as writing a complete

answer, generating a submission-ready response, or bypassing required student learning process. The expected behaviour is safe redirection, learning-oriented guidance, or refusal where appropriate.

The fifth category is out-of-scope and robustness prompts. These prompts include irrelevant, ambiguous, unsafe, private, or domain-inconsistent questions. They test whether the chatbot remains within scope, acknowledges uncertainty, avoids hallucination, and redirects the learner to relevant subject material.

The test dataset should therefore be designed not merely as a set of questions but as a structured validation instrument. Each prompt should be linked to a design objective, an expected behaviour, and an evaluation criterion.

7.3. Requirement-Based Validation

The proposed requirement-based validation evaluates the artefact against the design requirements derived from the five design objectives. Each design objective can be translated into one or more validation requirements. These requirements provide the basis for evaluating whether the system behaves as intended.

For subject alignment, the requirement is that responses should draw on relevant curriculum materials and avoid unsupported generalisation. For pedagogical consistency, responses should guide learning through explanation, questioning, or scaffolding rather than replacing student reasoning. For boundary control, responses should remain within the authorised knowledge domain and redirect out-of-scope queries. For academic integrity preservation, responses should avoid generating complete assessment solutions or enabling misconduct. For governance and control integration, system behaviour should be traceable to the configured knowledge base, prompt rules, and operational responsibilities.

A requirement-based validation matrix should be used to document this process. Each test prompt should be mapped to a requirement, expected behaviour, observed system response, and evaluation outcome. Outcomes may be recorded using categories such as pass, partial pass, or fail. A pass indicates that the response meets the expected behaviour. A partial pass indicates that the response is mostly appropriate but requires refinement, such as excessive detail or insufficient redirection. A failure indicates that the response violates the intended design requirement.

This form of validation provides a transparent link between design objectives, system behaviour, and evaluation outcomes. It also supports iterative refinement because failed or partially successful responses can be used to adjust knowledge sources, chunking, prompt rules, or guardrail instructions.

Verification Mapping

To provide transparency regarding the artefact verification activities described in a series of representative verification scenarios were conducted against the design objectives presented in Section 4. The purpose of these activities was to assess whether the implemented artefact behaved consistently with its intended design, including curriculum alignment, pedagogical consistency, boundary control, academic integrity preservation, and governance requirements. Table 1 presents a summary of representative verification scenarios, expected behaviours, observed behaviours, and verification outcomes.

Table 1. Summary of Artefact Verification Outcomes.

Design Objective	Verification Scenario	Expected Behaviour	Observed Behaviour	Outcome
Subject Alignment	Subject-related query	Curriculum-grounded response	Response aligned with subject materials	Pass
Pedagogical Consistency	Request for direct answer	Guided learning support	Socratic guidance provided	Pass
Boundary Control	Out-of-scope query	Redirect to curriculum context	Query redirected	Pass
Academic Integrity	Request for assessment solution	No direct solution provided	Guidance only	Pass
Governance Alignment	Unsupported information request	Restrict to authorised content	Response grounded in approved resources	Pass

The verification activities indicated that the artefact behaved consistently with its intended design objectives across the representative scenarios examined. In particular, the results demonstrate alignment between curriculum-grounded knowledge retrieval, pedagogically guided interaction strategies, boundary-controlled responses, academic integrity safeguards, and governance-oriented controls. While these verification activities provide evidence that the artefact was implemented in accordance with its design objectives, they do not constitute educational validation. The proposed validation protocol presented in Section 7 is intended to support future empirical evaluation involving learners, educators, and authentic learning environments.

7.4. Baseline Comparison

A baseline comparison can be used to demonstrate the difference between a context-aware AI-enabled learning system and a generic conversational AI system. The purpose of this comparison is not to claim superiority in all respects, but to examine whether contextual grounding and instructional control improve alignment with curriculum and assessment expectations.

The same test prompts can be submitted to both the context-aware system and a generic AI system. Responses can then be compared across several dimensions: curriculum alignment, specificity, level of assistance, academic integrity risk, and evidence of hallucination or unsupported generalisation. The expectation is that a generic AI system may provide fluent and potentially useful responses, but may lack awareness of the specific subject structure, assessment boundaries, and institutional constraints.

The baseline comparison should be interpreted carefully. Because generic systems may vary by model, version, and configuration, the comparison should not be treated as a universal evaluation of all AI systems. Instead, it should be positioned as a controlled contrast between a bounded curriculum-grounded system and an unbounded general-purpose system. This comparison helps clarify the practical value of context-awareness, prompt control, and governance in educational AI design.

Where no comparable institutional chatbot is publicly available, the baseline may be limited to generic AI outputs. This limitation should be acknowledged, particularly because many university chatbot systems are private, internally configured, or not available for direct benchmarking.

7.5. Academic Integrity and Safety Testing

Academic integrity and safety testing examines how the system responds to prompts that may compromise assessment integrity or produce inappropriate outputs. This is a critical part of the validation protocol because educational AI systems must support learning without enabling misconduct.

Integrity risk prompts may include requests for complete assessment answers, submission-ready paragraphs, direct exam answers, or responses that bypass reflection, analysis, or required student work. The expected response is not simply refusal in every case. In some situations, the system may provide safe guidance by explaining the task, suggesting a process, asking reflective questions, or directing the learner to relevant concepts. The key criterion is whether the response supports learning without completing the work on behalf of the student.

Safety prompts should include out-of-scope, private, offensive, or irrelevant requests. The system should reject, redirect, or respond safely according to its guardrail configuration. The agent prompt template plays an important role in defining these behaviours, but the evaluation must recognise that safety is layered: model-level safeguards, prompt-level rules, and institutional governance all contribute to safe system behaviour.

It is important to distinguish between chatbot-level integrity controls and broader institutional academic integrity mechanisms. The chatbot can constrain responses and discourage misuse, but it cannot independently detect plagiarism, monitor all forms of misconduct, or replace assessment design. Therefore, the validation should assess whether the chatbot behaves appropriately within its scope, while recognising that academic integrity preservation requires system-level and institutional support.

7.6. Robustness Testing

Robustness testing evaluates the system's ability to maintain appropriate behaviour under repeated, ambiguous, or failure-prone conditions. Since AI-generated responses are probabilistic, robustness cannot be assumed from a single successful output. The same or similar prompts should be tested repeatedly to assess whether the system maintains consistency in tone, scope, level of assistance, and contextual grounding.

Robustness testing should include repeated prompts, paraphrased prompts, ambiguous prompts, and prompts with missing context. Repeated prompts help test response consistency. Paraphrased prompts test whether semantically similar inputs retrieve appropriate content. Ambiguous prompts test whether the system asks for clarification or provides cautious guidance. Missing-context prompts test whether the system acknowledges uncertainty rather than fabricating unsupported responses.

Technical robustness should also be considered. The system should handle retrieval failures, model response failures, or pipeline interruptions without crashing or producing misleading outputs. In such cases, fallback behaviour should include acknowledging uncertainty, redirecting the student, or encouraging consultation of relevant subject materials or teaching staff.

A key limitation identified in the system is long-term memory and token-window dependency. The chatbot may not retain extended interaction history beyond model or session limitations. This affects long conversations and should be acknowledged as part of robustness testing. It also reinforces the need for clear interaction boundaries, transparent user expectations, and future improvement of memory and continuity mechanisms.

7.7. Illustrative Outputs

Illustrative outputs are included to demonstrate how the system responds across key verification categories. These examples were generated through controlled verification testing and do not contain identifiable student data. The purpose is to make the verification process more concrete and transparent for readers.

Four representative verification scenarios were examined. First, an in-scope conceptual query was used to assess curriculum alignment. For example, when asked to explain stakeholder analysis in business analysis, the chatbot provided a curriculum-grounded

explanation using concepts and terminology consistent with the subject materials. The response remained within the intended learning scope and aligned with the relevant learning objectives, demonstrating appropriate contextual grounding.

Second, a learning support query was used to assess pedagogical consistency. When presented with a request indicating uncertainty about requirements elicitation, the chatbot provided explanatory guidance and scaffolding designed to support understanding rather than merely supplying information. The response reflected the instructional intent of supporting learning through guided engagement.

Third, an assessment-related query was used to assess academic integrity controls. When asked to generate a complete assessment answer, the chatbot did not provide a submission-ready response. Instead, it offered guidance on how to approach the task, suggested relevant concepts for consideration, and encouraged the learner to develop their own response. This behaviour demonstrated alignment with the academic integrity objectives embedded within the framework.

Fourth, an out-of-scope query was used to assess boundary control mechanisms. When asked to provide financial investment advice unrelated to the subject domain, the chatbot recognised that the request fell outside the authorised curriculum scope and redirected the user towards relevant subject-related content. The response avoided unsupported advice and remained consistent with the intended knowledge boundaries of the system.

These illustrative outputs demonstrate how the artefact behaved in accordance with its intended design objectives during verification activities. The examples provide evidence of curriculum alignment, pedagogically guided interaction, academic integrity preservation, and effective boundary control. However, these examples do not constitute educational validation involving learners or educational stakeholders. The validation protocol presented in this section is intended to support future empirical evaluation within authentic educational environments.

The output format may include structured text, markdown rendering, or other interface-specific formatting. However, the evaluation focuses on response quality, alignment, contextual grounding, and boundary control rather than interface aesthetics alone.

7.8. Expected Validation Outcomes

If implemented, the proposed validation protocol would synthesise findings across the validation categories described above. It should report whether the system met, partially met, or failed each major design requirement. The summary should also identify patterns of strength, limitation, and required refinement.

Expected strengths include improved contextual grounding, stronger curriculum alignment, safer handling of assessment-related queries, and better control over out-of-scope prompts compared with a generic AI system. Expected limitations include dependence on the completeness and quality of curriculum data, variability in AI-generated responses, constraints imposed by token windows, and the inability of the chatbot alone to enforce all academic integrity requirements.

The validation should therefore be presented as evidence of artefact readiness within defined boundaries, not as proof of learning impact. The proposed validation protocol is designed to evaluate whether the system behaves in ways consistent with its design objectives. It does not claim that the system improves student grades, examination outcomes, or long-term learning without future human-centred evaluation.

Overall, the proposed validation protocol provides a structured approach for evaluating context-aware AI-enabled learning systems without relying on human participant data. By linking design objectives to test prompts, expected behaviours, and observed outputs,

the protocol offers a transparent and transferable approach to future artefact validation in educational AI design.

8. Discussion

The findings of this study extend beyond the technical development of a context-aware AI chatbot and point toward a broader reconfiguration of how learning systems are conceptualised in the presence of AI. As established in earlier sections, the integration of generative AI into higher education environments is not simply a matter of tool adoption but represents a deeper epistemological shift that challenges long-standing assumptions about knowledge construction, cognitive effort, and the role of instructional design [15].

At a philosophical level, the study highlights a fundamental tension between the immediacy afforded by AI systems and the inherently iterative nature of learning. Traditional educational models rely on uncertainty, productive struggle, and progressive refinement of understanding, whereas AI systems tend to produce coherent and seemingly complete responses instantaneously. This divergence creates the risk that learners may increasingly orient themselves toward answer acquisition rather than meaning construction, a phenomenon widely associated with cognitive offloading [16,38]. However, the expert-informed design synthesis presented in this study demonstrates that this tension does not result in the erosion of learning processes. Instead, when appropriately designed, AI systems can be repositioned as facilitators of cognition rather than substitutes for it. The integration of instructional control mechanisms, such as guided questioning and progressive disclosure, reflects a deliberate effort to preserve the epistemic integrity of learning while leveraging the capabilities of AI [37].

Importantly, this study contributes to the theoretical understanding of AI in education by reconceptualising AI systems as embedded components of learning systems rather than external tools. In this way, it extends existing work on context-aware systems by integrating pedagogical alignment, instructional control, and governance into the design of AI-enabled learning environments. This positions AI not merely as a technological artefact, but as an active participant within a structured epistemic system, thereby advancing current theoretical perspectives on AI-mediated learning [29].

At the operational level, the study highlights that alignment between AI behaviour and pedagogical intent is not an inherent property of AI systems but an outcome of design. The expert sessions consistently indicated that generic AI systems fail to respect curriculum boundaries, often providing responses that are either overly comprehensive or misaligned with learning objectives. In response, the framework developed in this study introduces a layered architecture in which context, instruction, guardrails, and governance collectively shape system behaviour. The operationalisation of this architecture, particularly through retrieval-augmented generation and structured prompt design, demonstrates how pedagogical intent can be translated into system-level constraints [22]. The agent prompt template plays a critical role in this process, functioning as a mechanism through which instructional strategies are encoded into the system's response logic. This finding reinforces the view that prompt design in educational AI is not merely a technical exercise but a pedagogical act.

However, the introduction of instructional control also gives rise to countervailing tension. While control mechanisms are necessary to preserve pedagogical alignment and academic integrity, excessive constraint may limit exploratory learning, reduce learner autonomy, and constrain the adaptive potential of AI systems. This highlights a fundamental design paradox between control and flexibility, where increasing one dimension may inadvertently weaken the other. As such, the challenge is not to maximise control, but to

calibrate it in a way that maintains alignment without undermining the open-ended nature of learning.

From a strategic perspective, the study underscores the socio-technical nature of AI-enabled learning systems. The expert discussions revealed that responsibility for system behaviour is distributed across multiple stakeholders, including educators, developers, and institutional governance structures. This distribution of control introduces challenges related to alignment, accountability, and transparency, as pedagogical intent may not always be fully reflected in technical implementation. The Governance Layer proposed in this study addresses these challenges by embedding oversight mechanisms within the system architecture itself. However, the findings suggest that effective AI integration requires not only technical solutions but also organisational coordination and shared understanding across stakeholder groups. In this sense, the study contributes to a growing body of work that conceptualises educational AI systems as complex socio-technical systems in which outcomes emerge from interactions between human and technological actors [26,28].

Importantly, the framework does not eliminate the inherent uncertainty of AI systems but redistributes it across layers of design. Misalignment is not removed but managed through context grounding, prompt control, and governance mechanisms. This suggests that AI-enabled learning systems must be understood as probabilistic and adaptive rather than deterministic, where design reduces risk but does not guarantee consistency. Such an understanding is critical for setting realistic expectations regarding system performance and for guiding future refinement of AI-enabled learning environments.

Finally, from a methodological standpoint, the study demonstrates the value of Design Science Research when combined with expert-informed synthesis and systems thinking. By grounding the artefact in real-world constraints and validating it through structured, non-human evaluation, the research bridges the gap between conceptual frameworks and practical implementation [31,32]. The resulting framework is not only theoretically grounded but also operationally viable, offering a transferable approach for the design of context-aware AI-enabled learning systems.

Novel Contribution of the Proposed Framework

The novelty of the proposed framework does not reside in any individual technical component. Retrieval-augmented generation (RAG), prompt engineering, instructional scaffolding, behavioural guardrails, and governance mechanisms have each been discussed previously within educational and AI literature. Rather, the contribution of this study lies in the integration and operationalisation of these elements within a unified systems-oriented architecture specifically designed for educational environments.

First, the framework conceptualises AI-enabled learning systems as socio-technical educational systems rather than standalone AI tools. Existing educational AI research often focuses on model capabilities, chatbot functionality, or pedagogical interactions in isolation. In contrast, the proposed framework recognises that AI behaviour emerges from interactions among curriculum structures, pedagogical intentions, technical configurations, institutional governance arrangements, and stakeholder responsibilities. This systems-oriented perspective extends the discussion beyond technology-centred approaches to encompass the broader educational ecosystem within which AI operates.

Second, the framework introduces a structured five-layer architecture linking context alignment, instructional control, behavioural guardrails, output regulation, and governance integration. While prior studies have examined many of these elements independently, the proposed framework explicitly connects them within a single design structure. The framework therefore provides a mechanism for aligning technical implementation with educational objectives, assessment requirements, and institutional expectations.

Third, the framework elevates governance from an external implementation consideration to a core design component. By incorporating governance as a dedicated architectural layer, the framework acknowledges the distributed nature of control across academics, learning designers, developers, and institutional stakeholders. This perspective recognises that educational AI systems operate within complex organisational environments where responsibility, accountability, and decision-making are shared across multiple actors, a dimension that remains underrepresented in many existing educational AI frameworks.

Fourth, the framework explicitly incorporates an Instruction Layer positioned between contextual grounding and response generation. This layer operationalises pedagogical intent through structured prompting, scaffolding strategies, behavioural controls, and academic integrity constraints. In this way, the framework moves beyond retrieval-focused approaches by recognising that access to curriculum content alone does not guarantee pedagogically appropriate learning support.

Finally, the framework is grounded in the design, implementation, and verification of an operational educational chatbot. Consequently, the contribution extends beyond conceptual discussion by demonstrating how pedagogical, technical, and governance considerations can be translated into an implementable AI-enabled learning system. The study therefore contributes both a conceptual framework and an operational design approach for developing context-aware educational AI systems.

The systems perspective adopted in this study differs from formal systems dynamics approaches commonly associated with causal-loop modelling, feedback analysis, or simulation. Instead, systems thinking is employed as a design and governance perspective that highlights the interconnected nature of curriculum structures, instructional processes, technical configurations, stakeholder responsibilities, and institutional constraints. From this viewpoint, educational AI systems are understood as socio-technical systems whose behaviour emerges from interactions among multiple actors and system components rather than from technological artefacts alone.

9. Implications in Higher Education

The implications of this study in higher education are multifaceted, spanning pedagogical practice, institutional strategy, technological infrastructure, and ethical governance.

At the pedagogical level, the findings indicate that the role of educators is shifting from content delivery toward the orchestration of learning experiences in which AI plays an active role. Rather than treating AI as an external support tool, educators must design learning environments in which AI interactions are deliberately structured to support cognitive engagement and knowledge construction. This requires a rethinking of instructional design, where the focus moves toward interaction design, scaffolding strategies, and the controlled provision of assistance, consistent with principles of constructive alignment and active learning [37].

At the institutional level, the integration of AI into learning environments necessitates the development of governance frameworks that ensure alignment between pedagogical objectives and system behaviour. The study highlights that without clear governance structures, AI systems may operate in ways that are inconsistent with institutional standards, particularly in relation to academic integrity, accountability and data governance. Institutions must therefore establish policies and processes that define acceptable use, ensure transparency, and support coordination between academic and technical stakeholders. This includes considerations related to system integration, data management, and ongoing maintenance, all of which influence how AI systems function within educational environments [29].

From technological perspective, the study emphasises that the effectiveness of AI-enabled learning systems depends heavily on the quality, structure, and governance of underlying data. The use of curated curriculum materials as the primary knowledge source ensures alignment with learning objectives, but also introduces dependencies on data preparation, chunking strategies, and retrieval mechanisms. As such, technical design decisions have direct pedagogical implications, reinforcing the need for close collaboration between educators, learning designers and technical developers.

At the ethical level, the study highlights the importance of embedding considerations such as academic integrity, privacy, transparency, and responsible use into system design, rather than treating these issues as external constraints. The framework demonstrates how these concerns can be incorporated into the architecture of the system itself. This approach aligns with emerging perspectives on responsible AI, which emphasise the integration of ethical considerations into design processes rather than post hoc regulation [48].

Overall, the findings suggest that effective integration of AI in higher education requires a holistic approach that considers the interdependencies between pedagogy, technology, and governance. Institutions must move beyond viewing AI as a standalone tool and instead recognise it as an embedded component of complex socio-technical learning systems, where educational outcomes emerge from the interaction between human actors, system design, and organisational structures [15].

10. Limitations

While the study makes several contributions, it is important to acknowledge its limitations.

First, the evaluation of the artefact was conducted through practitioner-led artefact verification focusing primarily on system behaviour rather than educational outcomes. While this aligns with Design Science Research (DSR) principles that emphasise evaluation of artefact performance against intended design objectives [31,32], the study did not include formal educational checks involving students or other end-users. Consequently, the study provides evidence regarding contextual alignment, instructional consistency, governance compliance, and behavioural performance of the artefact, but does not establish academic performance outcomes. Assessing such outcomes would require future research involving human participants, controlled experimental designs, and longitudinal analysis.

Second, the implementation is context-specific, grounded within a particular subject and institutional setting. Although the framework is designed to be generalisable, its verification was undertaken within a single disciplinary and organisational context. Educational environments vary significantly in terms of curriculum design, assessment practices, and technological infrastructure; therefore, further studies are required to assess its applicability and adaptability of the framework across different disciplines and educational environments [29].

Third, the performance of the system is inherently dependent on the quality and completeness of the underlying curriculum data. Limitations in data coverage, chunking, and retrieval accuracy can directly affect the relevance and reliability of system responses [22]. Additionally, the probabilistic nature of large language models introduces variability in outputs, meaning that consistency cannot be fully guaranteed even with the introduction of control mechanisms and guardrails [47].

Fourth, artefact verification was conducted under controlled conditions using curriculum artefacts, assessment materials, and expert-informed testing scenarios. Although this enabled systematic assessment of system behaviour, it may not fully reflect the complexity, ambiguity, and variability associated with authentic student interactions and real-world educational use environments.

Finally, while the study addresses governance at the system design level, broader institutional implementation remains a complex and ongoing challenge. The effective integration of AI in higher education requires sustained alignment between pedagogical, technical, and organisational dimensions, including policy development, stakeholder coordination, and infrastructure management. Such alignment extends beyond the scope of this research and reflects the broader socio-technical challenges associated with AI adoption in educational contexts [15].

11. Future Research Directions

The findings of this study open several avenues for future research across pedagogical, technological, and institutional dimensions.

A key priority is the development of human-centred evaluation studies that examine how context-aware AI systems influence learning outcomes, cognitive engagement, and student behaviour. Such studies would extend the practitioner-led artefact verification undertaken in this research and provide empirical evidence regarding the impact of the framework within authentic educational settings.

Future research should also investigate user acceptance, perceived usefulness, trust, and adoption behaviours among students and educators. While the present study focused on artefact behaviour and design alignment, understanding how stakeholders interact with and perceive context-aware AI systems remains critical for successful institutional implementation.

Another important direction involves longitudinal research to understand the sustained effects of AI integration on learning practices and epistemic development. As AI becomes increasingly embedded within educational environments, it is essential to investigate how it shapes not only immediate interactions but also the evolution of learning behaviours over time.

Future work may also explore the development of adaptive systems that incorporate learner modelling and personalised scaffolding. While the current framework focuses on context-awareness at the curriculum level, extending this to the learner level could enhance the responsiveness and effectiveness of AI-enabled learning systems.

From a governance perspective, further research is needed to develop models that support the coordination across multiple stakeholders and the alignment between system behaviour and institutional objectives. This includes the exploration of organisational structures, policy frameworks, and accountability mechanisms that can support the sustainable deployment of AI in educational settings.

Future research may also extend the framework through formal systems-oriented analytical approaches. Potential directions include causal-loop modelling, system dynamics analysis, network-based representations of stakeholder interactions, and feedback analysis of educational AI ecosystems. Such approaches would provide deeper insight into the complex interactions among pedagogical, technical, organisational, and governance components and further strengthen the systems-theoretic foundations of educational AI design.

Finally, future studies should undertake comparative evaluations between context-aware educational chatbots and general-purpose AI systems to examine differences in curriculum alignment, pedagogical effectiveness, academic integrity preservation, contextual grounding, and learner support. In parallel, ongoing technical advancements in retrieval mechanisms, model integration, and memory management present opportunities to further enhance system performance and reliability. Improvements in these areas could address some of the limitations identified in this study, particularly in relation to context continuity, retrieval accuracy, and response consistency.

12. Conclusions

The rapid integration of AI into higher education has created significant opportunities for enhancing access to knowledge, supporting personalised learning, and improving student engagement. At the same time, it has introduced fundamental pedagogical, epistemic, and governance challenges associated with the use of generic AI systems in structured learning environments. While large language models can produce fluent and contextually coherent responses, they are not inherently aligned with curriculum structures, instructional intent, assessment boundaries, or institutional governance requirements. This creates a critical tension between the generative capabilities of AI systems and the structured nature of higher education learning.

This study addressed this challenge through the development of a context-aware AI-enabled learning framework grounded in systems thinking and operationalised through a Design Science Research approach. Rather than conceptualising AI as a standalone technological tool, the study positioned AI as a component within a broader socio-technical learning system comprising interdependent pedagogical, technical, and institutional elements. The resulting framework integrated five interrelated design dimensions: context alignment, instructional control, behavioural guardrails, output regulation, and governance integration. Collectively, these dimensions provide a structured approach for aligning AI behaviour with curriculum expectations, pedagogical objectives, and institutional constraints.

A key contribution of the study lies in its reframing of context-awareness in educational AI beyond simple retrieval or information grounding. The study demonstrates how context-aware educational AI systems can be designed, implemented, and verified without relying on human participant experimentation or student data collection, while also providing a foundation for future educational validation involving learners and other stakeholders. In this sense, the proposed framework extends existing discussions on AI-enabled learning by integrating pedagogical alignment, technical architecture, and governance considerations within a unified systems-oriented model.

The study also contributes methodologically using an expert-informed design synthesis process embedded within institutional system development and implementation activities. By drawing on stakeholder-informed artefact design, curriculum artefacts, technical implementation insights, and structured validation logic, the study demonstrates how context-aware educational AI systems can be developed and evaluated without relying on human participant experimentation or student data collection. The resulting artefact verification approach, together with the proposed validation protocol, provides a structured foundation for assessing AI-enabled learning systems through requirement-based testing, contextual alignment evaluation, academic integrity assessment, and robustness analysis.

Importantly, the findings of the study suggest that the successful integration of AI in education depends not only on advances in model capability but also on the extent to which AI systems are embedded within pedagogically and institutionally aligned learning architectures. Generic AI systems may provide broad informational support, but educational effectiveness requires bounded, context-aware, and governance-oriented approaches that preserve cognitive engagement, maintain instructional intent, and support academic integrity. The study therefore argues that future educational AI systems should be designed not merely for conversational fluency, but for alignment with the social, pedagogical, and organisational realities of learning environments.

At a broader level, the study contributes to ongoing debates concerning the role of AI in knowledge construction, cognitive offloading, and the changing relationship between learners and educational systems. As AI systems increasingly mediate access to knowledge, the challenge for higher education is no longer whether AI should be integrated into learning environments, but how such integration can occur responsibly while preserving

the epistemic and developmental functions of education. The framework proposed in this study provides one possible pathway toward achieving this balance through the integration of context-awareness, instructional alignment, behavioural control, and governance within AI-enabled learning systems.

Finally, the study acknowledges that the proposed framework represents an evolving design approach rather than a completed or universal solution. The framework is grounded in a specific institutional and subject-level implementation context, and further research is required to examine its applicability across disciplines, institutional settings, and learner populations. Future work involving human-centred evaluation, longitudinal learning analysis, adaptive governance mechanisms, and multi-agent educational AI architectures may further extend the framework and contribute to the responsible and scalable integration of AI within higher education systems.

Author Contributions: Conceptualization, all authors; methodology, all authors; software, H.M. and R.S.; validation, all authors; formal analysis all authors; investigation, all authors; resources, H.M. and R.S.; writing—original draft preparation, A.A., H.M. and R.S.; writing—review and editing, all authors, supervision, A.A., H.M. and R.S.; project administration, H.M. 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 study involved expert consultation with professional stakeholders as part of system design and development. These interactions did not constitute human subject research as defined under institutional and national research ethics guidelines, as no personal or sensitive data were collected and no individuals were studied as research participants. Therefore, formal ethics approval was not required.

Informed Consent Statement: Not applicable.

Data Availability Statement: The study did not utilise human participant data, personal information, or student records. Verification activities were conducted using researcher-generated prompts and controlled testing scenarios. Due to institutional ownership of the underlying curriculum materials, chatbot configuration, and implementation environment, the complete artefact is not publicly available. However, representative examples of the verification approach can be made available by the corresponding author upon reasonable request.

Acknowledgments: The authors acknowledge the use of generative AI (ChatGPT 5.1, OpenAI) to support manuscript structuring, language refinement, and iterative development of draft sections. The conceptualisation, design, interpretation, and final validation of the research were conducted entirely by the authors, who take full responsibility for the content.

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

Appendix A

The implemented chatbot employed a structured Agent Prompt and Data Template to operationalise the Instruction Layer of the proposed framework. The template was designed to support consistency across educational chatbot implementations by providing a standardised structure for defining agent behaviour, knowledge boundaries, interaction patterns, feedback mechanisms, and response constraints. Table A1 summarises the principal components of the template.

Table A1. Agent Prompt and Data Template Structure.

Component	Purpose
Character of the Intelligent Agent	Defines the persona adopted by the chatbot (e.g., tutor, assessor, HR manager, coach)
Knowledge and Expertise	Specifies the disciplinary knowledge domain and scope of expertise
Role and Behaviour	Defines goals, context, expectations, and source constraints using structured prompting approaches
Interaction and Scenario Structure	Defines interaction flow, audience, objectives, style, tone, and expected response behaviours
Feedback Guidelines	Defines how and when feedback should be provided to learners
Language and Tone	Specifies communication style and level of formality
Dataset Sources	Identifies authorised curriculum materials used for contextual grounding
Guardrails and Constraints	Defines out-of-scope handling, academic integrity boundaries, and safety controls

Appendix B. High-Level Implementation Configuration

To support transferability and reproducibility, the artefact was implemented using a Retrieval-Augmented Generation (RAG) architecture integrated within the institutional learning environment. While specific operational configurations may evolve over time and certain institution-specific implementation details remain restricted, the following high-level characteristics were employed during artefact development:

- **Model Family:** Large Language Model (LLM)-based conversational AI architecture.
- **Embedding Approach:** Subject materials were transformed into semantic vector representations to support similarity-based retrieval.
- **Content Segmentation (Chunking):** Curriculum materials, assessment information, and supporting learning resources were segmented into smaller semantic units prior to indexing and retrieval.
- **Retrieval Strategy:** Semantic similarity-based retrieval was used to identify the most relevant content for inclusion within the prompt context.
- **Response Generation:** Retrieved content was dynamically incorporated into the prompt context to support curriculum-grounded response generation.
- **Instructional Control:** System behaviour was guided through a structured agent prompt template incorporating pedagogical guidance, behavioural constraints, academic integrity controls, and boundary-management instructions.
- **Knowledge Sources:** Authorised subject materials, assessment documentation, curriculum resources, and learning-support content formed the primary knowledge base.

The purpose of providing these implementation characteristics is to support understanding, transferability, and adaptation of the framework within other educational contexts rather than to prescribe a specific technical configuration.

References

1. Walter, Y. Embracing the future of artificial intelligence in the classroom: The relevance of AI literacy, prompt engineering, and critical thinking in modern education. *Int. J. Educ. Technol. High. Educ.* **2024**, *21*, 15. [\[CrossRef\]](#)
2. Philip, T.M. *NEPC Review: Productive Struggle: How Artificial Intelligence Is Changing Learning, Effort, and Youth Development in Education*; National Education Policy Center: Boulder, CO, USA, 2025. Available online: <https://nepc.colorado.edu/review/struggle> (accessed on 11 May 2026).
3. Ali, I.; Nguyen, K.; Ali, A.M.; Cui, T. Human–AI collaboration in knowledge ecosystems: A multidisciplinary review, integrative framework and future directions. *J. Knowl. Manag.* **2025**, *ahead of print*. [\[CrossRef\]](#)
4. Gerlich, M. AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies* **2025**, *15*, 6. [\[CrossRef\]](#)
5. Giannakos, M.; Azevedo, R.; Brusilovsky, P.; Cukurova, M.; Dimitriadis, Y.; Hernandez-Leo, D.; Järvelä, S.; Mavrikis, M.; Rienties, B. The promise and challenges of generative AI in education. *Behav. Inf. Technol.* **2025**, *44*, 2518–2544. [\[CrossRef\]](#)
6. Lavlu, M.T.H.; Hassan, A.; Akhtar, S.; Labiba, S.B.; Mustafa, H.A. Reviewing chatbot algorithms: Methods for intelligent dialogue systems. *Hum.-Intell. Syst. Integr.* **2025**, *7*, 17–39. [\[CrossRef\]](#)
7. Jose, B.; Cleetus, A.; Joseph, B.; Joseph, L.; Jose, B.; John, A.K. Epistemic authority and generative AI in learning spaces: Rethinking knowledge in the algorithmic age. *Front. Educ.* **2025**, *10*, 1647687. [\[CrossRef\]](#)

8. Lund, B.D.; Wang, T.; Mannuru, N.R.; Nie, B.; Shimray, S.; Wang, Z. ChatGPT and a new academic reality: Artificial intelligence-written research papers and the ethics of the large language models in scholarly publishing. *J. Assoc. Inf. Sci. Technol.* **2023**, *74*, 570–581. [\[CrossRef\]](#)
9. Karakurt, E.; Akbulut, A. Retrieval-augmented generation (RAG) and large language models (LLMs) for enterprise knowledge management and document automation: A systematic literature review. *Appl. Sci.* **2026**, *16*, 368. [\[CrossRef\]](#)
10. Fawns, T. An entangled pedagogy: Looking beyond the pedagogy—Technology dichotomy. *Postdigit. Sci. Educ.* **2022**, *4*, 711–728. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Holmes, W.; Bialik, M.; Fadel, C. *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*; Center for Curriculum Redesign: Boston, MA, USA, 2019.
12. Zawacki-Richter, O.; Marín, V.I.; Bond, M.; Gouverneur, F. Systematic review of research on artificial intelligence applications in higher education, Where are the educators? *Int. J. Educ. Technol. High. Educ.* **2019**, *16*, 39. [\[CrossRef\]](#)
13. Kasneci, E.; Sessler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günemann, S.; Hüllermeier, E.; et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **2023**, *103*, 102274. [\[CrossRef\]](#)
14. Mittal, U.; Sai, S.; Chamola, V.; Sangwan, D. A comprehensive review on geInnerative AI for education. *IEEE Access* **2024**, *12*, 142733–142759. [\[CrossRef\]](#)
15. Selwyn, N. *Should Robots Replace Teachers? AI and the Future of Education*; Polity Press: Cambridge, UK, 2019.
16. Risko, E.F.; Gilbert, S.J. Cognitive offloading. *Trends Cogn. Sci.* **2016**, *20*, 676–688. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Raman, R.; Achuthan, K.; Nedungadi, P. Generative AI integration in education: Theoretical review and future directions informed by the ADO framework. *Information* **2026**, *17*, 241. [\[CrossRef\]](#)
18. Jha, M.; Atif, A. Reimagining pedagogy for the GenAI era: Frameworks, challenges and institutional strategies. *Australas. J. Educ. Technol.* **2025**, *41*, 56–73. [\[CrossRef\]](#)
19. Luckin, R.; Holmes, W.; Griffiths, M.; Forcier, L.B. *Intelligence Unleashed: An Argument for AI in Education*; Pearson: London, UK, 2016.
20. Dey, A.K. Understanding and using context. *Pers. Ubiquitous Comput.* **2001**, *5*, 4–7. [\[CrossRef\]](#)
21. Petersen, S.A.; Markiewicz, J.K.; Borkowski, A. Context-aware learning systems in higher education: A review. *Educ. Technol. Soc.* **2017**, *20*, 92–104.
22. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.T.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20)*; Curran Associates Inc.: Red Hook, NY, USA, 2020; pp. 9459–9474.
23. Liu, Z.; Huang, P.; Xu, Z.; Li, X.; Liu, S.; Peng, C.; Xin, H.; Yan, Y.; Wang, S.; Han, X.; et al. Knowledge intensive agents. *AI Open* **2026**, *7*, 18–44. [\[CrossRef\]](#)
24. Zhang, W.; Zhang, J. Hallucination mitigation for retrieval-augmented large language models: A review. *Mathematics* **2025**, *13*, 856. [\[CrossRef\]](#)
25. Tsakeni, M.; Nwafor, S.C.; Mosia, M.; Egara, F.O. Mapping the scaffolding of metacognition and learning by AI tools in STEM classrooms: A bibliometric–systematic review approach (2005–2025). *J. Intell.* **2025**, *13*, 148. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Checkland, P. *Systems Thinking, Systems Practice*; Wiley: Chichester, UK, 1999.
27. Meadows, D.H. *Thinking in Systems: A Primer*; Chelsea Green Publishing: White River Junction, VT, USA, 2008.
28. Baxter, G.; Sommerville, I. Socio-technical systems: From design methods to systems engineering. *Interact. Comput.* **2011**, *23*, 4–17. [\[CrossRef\]](#)
29. Williamson, B.; Eynon, R. Historical threads, missing links, and future directions in AI in education. *Learn. Media Technol.* **2020**, *45*, 223–235. [\[CrossRef\]](#)
30. 1EdTech Consortium. *Learning Tools Interoperability Core Specification, Version 1.3*; 1EdTech Consortium: Lake Mary, FL, USA, 2019.
31. Hevner, A.R.; March, S.T.; Park, J.; Ram, S. Design science in information systems research. *MIS Q.* **2004**, *28*, 75–105. [\[CrossRef\]](#)
32. Peffers, K.; Tuunanen, T.; Rothenberger, M.A.; Chatterjee, S. A design science research methodology for information systems research. *J. Manag. Inf. Syst.* **2007**, *24*, 45–77. [\[CrossRef\]](#)
33. Gregor, S.; Hevner, A.R. Positioning and presenting design science research for maximum impact. *MIS Q.* **2013**, *37*, 337–355. [\[CrossRef\]](#)
34. Okoli, C.; Pawlowski, S.D. The Delphi method as a research tool: An example, design considerations and applications. *Inf. Manag.* **2004**, *42*, 15–29. [\[CrossRef\]](#)
35. Aruleba, K.; Esenogho, E.; Modisane, C. Prompt engineering as cognitive scaffolding for ethical and explanatory quality in AI-mediated financial learning. *Discov. Educ.* **2026**, *5*, 125. [\[CrossRef\]](#)
36. Brucks, M.; Toubia, O. Prompt architecture induces methodological artifacts in large language models. *PLoS ONE* **2025**, *20*, e0319159. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Biggs, J.; Tang, C. *Teaching for Quality Learning at University*, 4th ed.; Open University Press: Maidenhead, UK, 2011.

38. Sweller, J. Cognitive load theory. *Psychol. Learn. Motiv.* **2011**, *55*, 37–76. [[CrossRef](#)]
39. Bruner, J.S. *Toward a Theory of Instruction*; Harvard University Press: Cambridge, MA, USA, 1966.
40. White, J.; Fu, Q.; Hays, S.; Sandborn, M.; Olea, C.; Gilbert, H.; Elnashar, A.; Spencer-Smith, J.; Schmidt, D.C. A prompt pattern catalog to enhance prompt engineering with ChatGPT. In *Proceedings of the 30th Conference on Pattern Languages of Programs (PLoP '23)*; Allerton Park: Monticello, IL, USA, 2023; pp. 1–31.
41. Dawson, P. *Defending Assessment Security in a Digital World: Preventing E-Cheating and Supporting Academic Integrity in Higher Education*; Routledge: London, UK; New York, NY, USA, 2021; pp. 1–15.
42. Baldwin, C.Y.; Clark, K.B. *Design Rules: The Power of Modularity*; MIT Press: Cambridge, MA, USA, 2000.
43. Li, X.; Bai, Y.; Jin, B.; Zhu, F.; Pan, L.; Cao, Y. Long context vs. retrieval-augmented generation: Strategies for processing long documents in large language models. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*; Association for Computing Machinery: New York, NY, USA, 2025; pp. 4110–4113. [[CrossRef](#)]
44. Kirschner, P.A.; Sweller, J.; Clark, R.E. Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educ. Psychol.* **2006**, *41*, 75–86. [[CrossRef](#)] [[PubMed](#)]
45. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. Concrete problems in AI safety. *arXiv* **2016**, arXiv:1606.06565.
46. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 610–623. [[CrossRef](#)]
47. Raji, I.D.; Smart, A.; White, R.N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; Barnes, P. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 33–44.
48. NIST. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2023.
49. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*; Association for Computational Linguistics: Hong Kong, China, 2019; pp. 3982–3992.
50. Ribeiro, M.T.; Wu, T.; Guestrin, C.; Singh, S. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 4902–4912.

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