computers

MDPI

Article

AI Gem: Context-Aware Transformer Agents as Digital Twin Tutors for Adaptive Learning

Attila Kovari [1,2,3,4]

[1] Institute of Digital Technology, Faculty of Informatics, Eszterházy Károly Catholic University, 3300 Eger, Hungary; kovari.attila@uni-eszterhazy.hu
[2] Institute of Computer Science, University of Dunaujvaros, 2400 Dunaujvaros, Hungary
[3] Institute of Electronics and Communication Systems, Kandó Kálmán Faculty of Electrical Engineering, Óbuda University, 1034 Budapest, Hungary
[4] GAMF Faculty of Engineering and Computer Science, John von Neumann University, 6000 Kecskemet, Hungary

Abstract

Recent developments in large language models allow for real time, context-aware tutoring. AI Gem, presented in this article, is a layered architecture that integrates personalization, adaptive feedback, and curricular alignment into transformer based tutoring agents. The architecture combines retrieval augmented generation, Bayesian learner model, and policy-based dialog in a verifiable and deployable software stack. The opportunities are scalable tutoring, multimodal interaction, and augmentation of teachers through content tools and analytics. Risks are factual errors, bias, over reliance, latency, cost, and privacy. The paper positions AI Gem as a design framework with testable hypotheses. A scenario-based walkthrough and new diagrams assign each learner step to the ten layers. Governance guidance covers data privacy across jurisdictions and operation in resource constrained environments.

Keywords: adaptive learning; personalized learning; digital twin tutor; context-aware AI agent; AI Gem; transformer architecture; retrieval augmented generation; Bayesian Knowledge Tracing; learner modeling; learning analytics; hallucination and bias mitigation; data privacy

check for updates

Academic Editor: Ananda Maiti

Received: 1 August 2025
Revised: 18 August 2025
Accepted: 23 August 2025
Published: 2 September 2025

Citation: Kovari, A. AI Gem: Context-Aware Transformer Agents as Digital Twin Tutors for Adaptive Learning. *Computers* **2025**, *14*, 367. https://doi.org/10.3390/computers14090367

1. Introduction

Personalized adaptive learning has long been a goal in education, adapting instruction, pacing, and feedback at the individual learner's requirements for maximum mastery. Traditional adaptive learning systems, usually epitomized as intelligent tutoring systems (ITSs), have realized enhanced learning results through the simulation of one-to-one tutoring with rule-based algorithms [1]. However, traditional ITSs can be brittle, offering very little flexibility in responding to the variety of real-world learners and learning environments. Such systems can rely on predefined content and rules, making it challenging to respond appropriately with unexpected student input or shifting curricula. The very recent advent of transformer-based AI architecture, led by large language models (LLMs) like OpenAI's GPT series (e.g., GPT-4 as the brain behind ChatGPT), is revolutionizing this situation. These models employ the transformer architecture's self-attention mechanism in order to provide hitherto unmatched natural language understanding and generation capabilities [2,3]. These can engage in human-like dialogue, spontaneously generate text, and adapt in real time, bringing entirely new prospects of personalized adaptive learning systems [1]. In practical implementations, one can have a transformer-based AI tutor that can comprehend a student's question written in a free text format and respond with personalized hints or spontaneous explanation, beyond the ability of earlier systems utilizing scriptable responses. Such tutors can be envisioned as a digital twin tutor, an AI-based instructional agent that mirrors an educator's presence, responsiveness, and pedagogical intent in real-time.

This evolution from rule-based ITS to transformer-based generative tutors marks a significant paradigm shift in how adaptivity and personalization are realized. Traditional systems operate on predefined content trees and rule sets, limiting their flexibility in real-world learning contexts. In contrast, transformer models dynamically generate responses, scaffold learning in context, and adjust to free-form learner input in real time. A high-level comparison of these two architectures is illustrated in Figure 1.

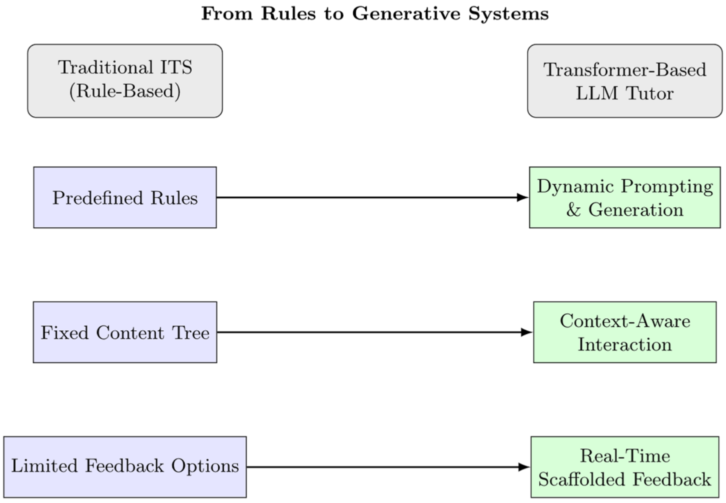
From Rules to Generative Systems
Traditional ITS (Rule-Based)
Transformer-Based LLM Tutor
Predefined Rules
Dynamic Prompting & Generation
Fixed Content Tree
Context-Aware Interaction
Limited Feedback Options
Real-Time Scaffolded Feedback

Figure 1. Conceptual comparison between rule-based ITS and transformer-based LLM tutors.

Rule-based systems rely on preorganized content maps and production rules that select from various hints, while transformer-based tutors generate responses conditioned on real-time context, retrieval over an established knowledge base, and an explicit learner model. The ensuing differences seem to appear on four axes: knowledge source, control logic (handcrafted rules versus probabilistic, policy constrained reasoning), interaction (menu driven feedback versus open ended dialog and Socratic scaffolding), and verification and governance (limited logging vs. post-generation verification and auditable policies).

This paper investigates the way in which transformer-based artificial intelligence is redrawing the face of personalized adaptive learning via a design-centered method of finding important pedagogical principles and technical considerations needed for implementation, including architecture, deployment, and integration. It details methods in which large language models could be incorporated to give real-time, individualized support and offers a critical examination of the promise and pitfalls of their adoption throughout educational settings. Emphasizing transformer-based generative models, such as ChatGPT, the report describes a future thinking but evidence-based approach to educational technology design. The term digital twin tutor denotes a transformer-focused teaching agent that provides context-aligned guidance within a larger personalized learning platform and operates as a persistent, context-aware instructor that executes autonomously in scalable, modular, and data-centric environments. These notions take shape in the AI Gem architecture, conceptualizing digital twin tutors as multi layered agents subject to pedagogical logic, runtime context, and verifiable system design to specify a transformative but practical path to deployment. The innovation extends rule-based intelligent tutor systems and recent large language model shells in formalizing cross layer context flow around a context fusion layer that conditions pedagogical alignment, personalization, generative reasoning, and interaction flow per turn by binding generative models to verifiable educational artifacts via the combination of retrieval-augmented generation (RAG), an explicit Bayesian knowledge tracing (BKT)-focused learner model, and policy-constrained dialog generating testable signals such as mastery probabilities, retrieval rationales, and verification results, and by managing trust, privacy, and deployment as first class pedagogical constraints via embedded governance, auditability, and coordinated edge and cloud approaches within the same agent architecture instead of as external addons.

2. Transformer-Based Digital Twin Tutors in Adaptive Learning: A New Paradigm

Transformer-based language models have rapidly become central to next-generation adaptive learning innovations [3,4]. LLMs can now spontaneously generate contextualized explanations, examples, and questions. Examples include GPT-4 and other models that allow real-time interactivity and dialog in the voice of a human tutor [2]. Such systems can now be envisioned as digital twin tutors, AI-driven software agents that reproduce the instructional behavior, responsiveness, and adaptation of human educators in real time. Here is the new paradigm: Generative intelligent tutoring systems that utilize LLMs to spontaneously generate personalized interactions and content. A high-level schematic of such a transformer-based digital twin tutoring system is depicted in Figure 2, showing the integration of learner input, domain knowledge, and student modeling within a core language model architecture.

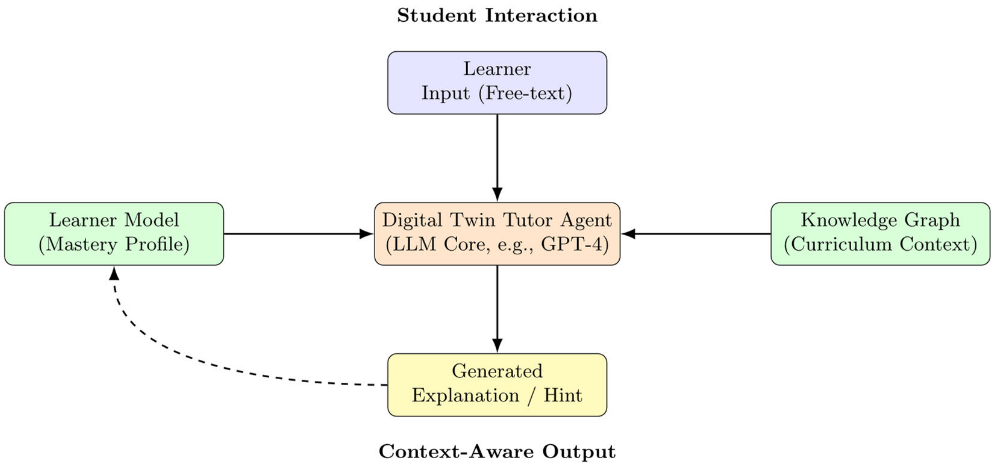
Student Interaction
Learner Input (Free-text)
Learner Model (Mastery Profile)
Digital Twin Tutor Agent (LLM Core, e.g., GPT-4)
Knowledge Graph (Curriculum Context)
Generated Explanation / Hint
Context-Aware Output

Figure 2. Architecture of a transformer-based digital twin tutoring system.

Transformer models have access to immense knowledge and advanced language capabilities and can be used to help students across numerous subject matter fields. As computerized instruction agents, they can provide real-time feedback on short-answer student input and can assess free-text response in natural language with contextual hints or follow-up questions [5]. Real-time scaffolding enables students to work at their own rate with skills that can be spontaneously adjusted to meet the learner's present understanding [6]. For instance, if a student is having difficulty with a math problem, an LLM can decompose the solution step-by-step or re-express its rationale at progressively simpler levels until it is at the learner's current understanding [7]. Adaptive scaffolding combines prerequisite selection, brief self-explanation prompts, and worked examples grounded by retrieval, which together reduce cognitive load while preserving productive struggle. Experiments combining LLMs with techniques of student modeling strengthen the benefit. The LLM provides progressively advanced help, spontaneous summaries, and prerequisite explanations [8]; however, human review remains necessary to prevent errors [9]. The tiered assistance leads to greater comprehension and superior task performance on pilot tests.

Another feature of transformer-based digital twin tutors is that they can engage learners in natural, open-ended conversation. Constructivist educational ideals strongly identify with such a feature: learning is usually most effective when learners build knowledge through conversation and discovery [10]. ChatGPT-5, for instance, can engage in a conversational mode encouraging learners to question without inhibition and explain their thinking. The research suggests that, with an AI tutor like ChatGPT, higher conversation initiation and construction of knowledge behaviors occur as students work together with dialog in negotiating meaning [11]. In language learning environments, ChatGPT's conversational interaction can enable greater language exercise and production through being an always-accessible conversational partner/coach [12]. AI's human-like response, e.g., through using an encouraging tone of voice or expressing empathetic sentiments, can enable user motivation and self-belief [13]. From the socio-constructivist perspective, the digital twin tutors are dialog partners collaborating with learning scaffolding while responding with learner's ideas and questions on an individual basis. Students can productively "think aloud" with the AI and be replied-to with immediate, personalized responses, aiding understanding refinement through iterative conversation. The overarching pedagogical affordances emerging under such transformer-based tutoring interactions are encapsulated within Figure 3, including context-aware dialog, personalized scaffolding, and real-time response systems rooted within learner modeling.

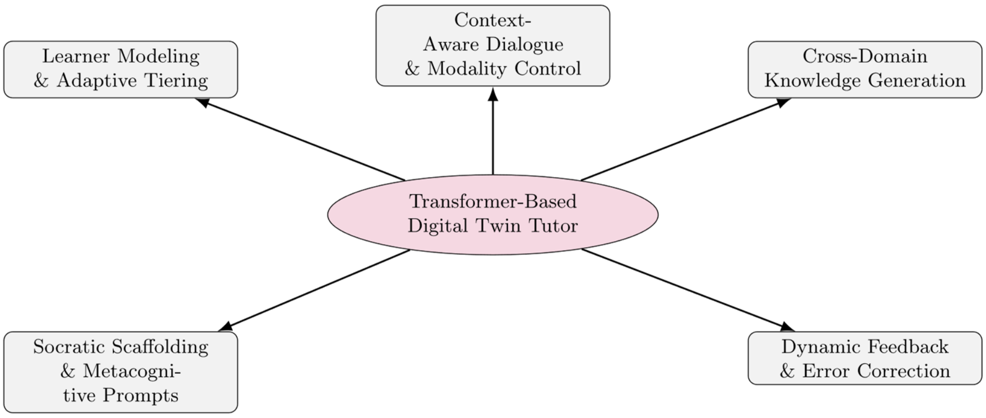

Figure 3. Educational affordances of transformer-based digital twin tutors.

Scalability and coverage allow an always-available tutor that can assist across subjects. This supports inclusion in settings where human tutoring is limited. They can instantly summon extensive knowledge on nearly all subjects and might perhaps give aid on any topic a learner would like to know more about [14]. Such broad coverage would enable an AI tutor to be issued as generalist coach-at-all-times, available 24/7, unlike human tutors or traditional ITSs that have typically had only narrow-domain coverage. Researchers have noted that generative AI tutors can potentially democratize high-quality individualized instruction, notably in cases where human tutoring or the creation of extensive content is not feasible [15]. Such prospects enable more inclusive and equitable learning. For example, under-resourced districts or rural territories might access an AI-driven learning companion with counsel on par with that of an individual tutor on subject matter for which their native school is unable to properly instruct. The patience and reliability of a digital twin tutor can be of assistance to those learners who need more time or alternative reconsideration and are without available resources within reach of their school. In effect, transformer-based digital twin tutors in themselves realize and enact the AI Gem paradigm framework of LLM-driven agents as modular, pedagogically focused, and context-aware instruction systems. Under such a framework, the digital twin tutor serves not only as response engine but as coordinated learning agent embedded within scalable and morally governed instruction systems. This is, indeed, a great leap forward for educational technology.

3. Design Principles for Transformer-Based Digital Twin Tutors

Transformer-based digital twin tutors require clear pedagogy and a simple, robust system design [16]. We treat the tutor as a context-aware digital twin that adapts to the learner's context. AI Gem provides a modular design that integrates learner modeling, context fusion, generative reasoning, orchestration, and governance, with clear interfaces for deployment.

This section outlines a set of pedagogical and technical design principles to guide the implementation of such systems (Tables 1 and 2).

Table 1. Design principles for transformer-based digital twin tutors.

Design Principle	Description
Learner-centered personalization	The digital twin tutor applies real-time learning analytics (such as quiz scores, task latency, and error type) to customize feedback and progression for every learner.
Adaptive feedback and scaffolding	Instead of offering solutions, the tutor provides Socratic suggestions, clarification-seeking questions, and graduated prompts according to learner state.
Content generation with pedagogical alignment	Instructional material is generated spontaneously through LLMs but is based on curricular aims and pedagogically confirmed templates.
Multi-modal and flexible interaction	Tutor customizes delivery modalities (text, audio, visuals, simulations) on the basis of device capacity and learner preference inferred using context signals.
Human–AI collaboration and transparency	The system offers explainable reasoning, conveys the degree of confidence, and gives teachers review/override of tutor choices.
Ethical and responsible AI integration	The tutor incorporates bias detection, age-appropriate safeguards, and transparent consent protocols to support ethical AI use.

Table 2. Technical design principles for LLM-based digital twin tutor systems.

Technical Principle	Description
Modular architecture and pipeline design	Separate the system into independent modules (e.g., learner model, context merging, LLM engine) to facilitate updating, maintenance, and testing.
Deployment strategy: Cloud, edge, or hybrid	Determine the deployment mode based on latency, data sovereignty, and bandwidth: edge for real-time responsiveness; cloud for scale; hybrid for resilience.
Scalability and performance optimization	Apply model compression (distillation, quantization), batching on the GPU, and caching to support rapid, parallel interaction for numerous concurrent users.

Table 2. *Cont.*

Technical Principle	Description
Integration with educational platforms and standards	Enable LMS tools (e.g., Moodle, Canvas) compatibility with protocols such as LTI, SCORM, or xAPI and allow for seamless adoption and interoperability of data.
Data privacy, security, and compliance	Reduce and obscure learner information, implement encryption, honor legal protocols (e.g., GDPR, FERPA), and record AI choices for traceability.

Compliance with these principles not only ensures that the digital twin tutor reacts appropriately to learner performance but also honors pedagogical integrity and institutional trust. Context-aware decision-making enables the AI to differentiate support, pace flexibly, and tailor content without rigid pre-programming, which are essential characteristics of fair personalization at scale. As part of the AI Gem framework, the principles underlie the pedagogical layer of an agent-based architecture for tutoring that aligns LLM affordances with learner modeling, interaction design, and instructional governance.

Alongside learning objectives, system constructors must also guarantee that the AI tutor is technologically sound, interoperable, and regulatory compliant. Table 2 describes the key IT and infrastructure design principles underpinning successful rollout.

While Tables 1 and 2 set out the pedagogical and infrastructural requirements, the concrete integration of RAG, BKT, and bias mitigation are given in the operational pipeline (Figures 4 and 7); the step-to-layer map appears in Figure 6.

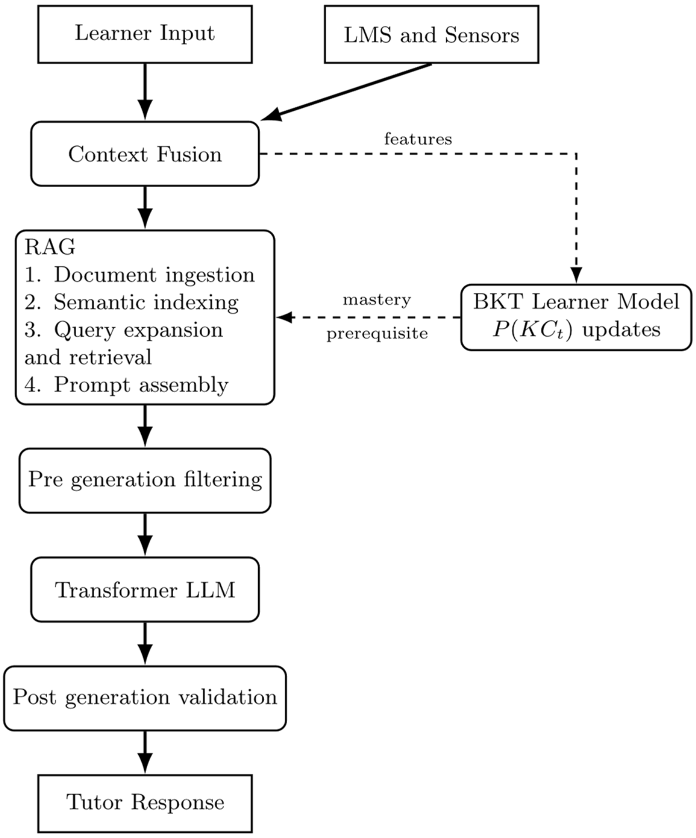

Figure 4. System diagram of the AI Gem pipeline integrating the RAG process, the BKT learner model, and bias-mitigation filters (dashed lines indicate feature signals and mastery prerequisite links for learner modeling and adaptive retrieval).

Together, these pedagogical and technical principles operationalize the digital twin tutor not just as an LLM frontend, but as an orchestrated multi-agent system with clear educational intent and regulatory resilience. This holistic approach, exemplified by the AI Gem model, ensures that digital twin tutors are not isolated components but integral actors in a socio-technical learning ecosystem, one that harmonizes generative AI power with human oversight and institutional trust. Building such systems requires close collaboration between AI engineers, instructional designers, school IT administrators, and educational policy stakeholders.

Section 4 applies these principles across the ten layers and uses Figure 6 to map learner steps and Figure 7 to show the runtime pipeline.

4. AI Gem in Practice: A Layered Framework for Context-Aware Digital Twin Tutoring Agents

Educational AI agents follow a simple loop. They observe context, update a learner model, plan the next step, generate a response, then reflect. In AI Gem this agent acts as a digital twin tutor that senses affect and progress, maintains a compact learner model, and adapts turns in real time.

This multi-layered architecture reflects a socio-technical design: it embeds pedagogical theory, dialogic strategies, and system-level requirements into a unified agent structure. The result is a scalable, responsive, and ethically aligned tutoring system. In the spirit of AI Gem, each layer is not merely a software component, but a pedagogical affordance with measurable contribution to learning outcomes, ethical compliance, and system transparency.

The aspiration of this section is to give a single, cohesive explanation of the ten layers of the AI Gem architecture and demonstrate how these layers collaborate within the agent cycle to provide a product that is pedagogy aligned, transparent, and which runs at scale. Figure 5 shows the complete architecture in a single diagram, and Table 3 provides a summary of the whole system per layer of the primary artifact or service, its overriding educational advantage, and principal threat that needs to be controlled.

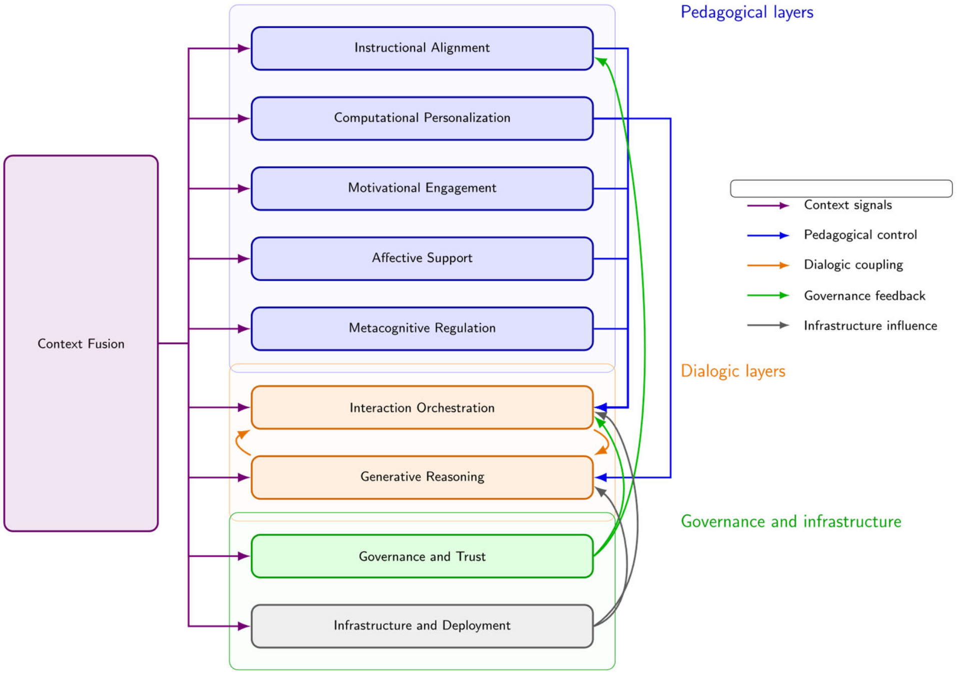

Figure 5. Layered Design Framework of the AI Gem: Context-Aware Adaptive AI-Agent Tutoring with Transformer-Based Generative Models.

Table 3. AI Gem Layered Framework: Context-Aware Adaptive AI-Agent Tutoring Layers.

Layer	Core Artifact/Service	Key Benefits	Principal Risk
Context fusion	Multimodal context vector (sensors + LMS inputs)	Situational adaptation across all layers	Missing or noisy signals
Instructional alignment	Curriculum map aligned to the learner's zone of proximal development	Optimal challenge calibration	Over/under-scaffolding
Computational personalization	Dynamic learner model with BKT	Granular, learner-specific targeting	Stale or inaccurate learner state
Motivational engagement	Strategy selector grounded in the attention, relevance, confidence and satisfaction motivational model	Sustained effort and engagement	Learner disengagement
Affective support	Emotion classifier and empathy response generator	Faster recovery from frustration	Misclassification of affect
Metacognitive regulation	Prompting engine that is sensitive to self-regulated learning	Improved self-monitoring and reflection	Prompt fatigue or cognitive overload
Generative reasoning	Prompting engine that uses RAG with a factuality verification module	Rich, contextualized instructional content	Hallucinated or misleading output
Interaction orchestration	Dialogue manager with adaptive turn policies	Deep, personalized discourse flow	Shallow, repetitive interaction loops
Governance and trust	Teacher-facing dashboard and policy filters	Transparency, auditability, fairness	Latent bias or opaque system behavior
Infrastructure and deployment	Edge and cloud orchestration and autoscaling services	Low-latency, scalable delivery	Latency spikes or resource constraints

The instructional agent is referred to as a digital twin tutor, and when the agent's architectural role is emphasized the term context-aware agent is used. Figure 5 visualizes how the tutor handles context, pedagogy, dialog, governance, and deployment in one view. A typical example is frustration detection: when Context Fusion identifies rising frustration, the Affective Support layer offers an empathetic rephrasing, Interaction Orchestration shortens the next prompt and defers a hint until a brief self-explanation, Generative Reasoning selects a worked example drawn from trusted sources, and Governance & Trust records the intervention for teacher review.

Table 3 encapsulates the essential artifacts, pedagogical advantages, and primary pitfalls of each layer. These layers together establish not only the system's functional capabilities but also fault lines along which breakdowns can occur if input is absent, logic is incoherent, or constraints are violated. Such architecture provides the basis of the AI Gem framework, whereby digital twin tutoring is made operable through context-driven reasoning across modular, interpretable artifacts.

This cross-layer propagation capability distinguishes context-aware digital twin tutors from traditional modular ITSs or even contemporary LLM wrappers. It enables truly adaptive, real-time educational interaction that is grounded in learner state and operationalized across the software stack. In addition, the framework facilitates the formation of layer-specific hypotheses.

4.1. Integrated Architecture and the Role of Context Fusion

The system is structured around a context fusion layer that provides situated intelligence to each phase of the agent loop. Sensor signals, activity traces, and learning management system data are combined and normalized into a multimodal representation robust to noise and missing data. The top levels represent pedagogical intention and decide what to adapt for the student. The middle levels implement reasoning and dialog faithfully to pedagogy. The bottom levels ensure trust, privacy, compliance, and performance. This division of concerns adheres to microservice style but stays rooted in learning science, such as socio-cognitive scaffolding, self-regulated learning, and dialogic instruction.

More concretely, the ten layers can be interpreted across three macro-level functional categories:

- Pedagogical layers (instructional alignment to metacognitive regulation) convert context into directed learning experiences. The layers continuously adapt with differential difficulty of subject matter, motivational strategy, and metacognitive support based on the learner’s changing ZPD, affective cues, and goals [1].
- Dialogic layers (generative reasoning and interaction orchestration) regulate the tutor’s communication behavior. RAG facilitates pedagogically bounded content construction, while policies of orchestration decide the timing and form of the prompts, hints, and reflective turns [17,18].
- Governance and infrastructure layers provide for trust, equity, and efficiency. Data flows that preserve privacy, explainable dashboards, and edge and cloud deployment all facilitate compliance with ethical, legal, and performance requirements [19].

By incorporating the AI Gem design lens, each layer can be proactively audited for learning alignment, context responsiveness, and generativity, thus permitting the digital twin tutor to become an adaptive pedagogical agent rather than as a passive automaton device. Context is inherent across the architecture here. Should, for example, the context fusion layer detects building frustration through multimodal input, the system can at once undertake the following:

- simplify content (instructional alignment),
- trigger empathetic feedback (affective support),
- shorten LLM prompts to reduce load (generative reasoning),
- surface an alert to the teacher (governance).

Such cross-layer propagation enables truly adaptive, real-time educational interaction.

4.2. Mapping the Agent Loop to the Ten Layers

The agent loop has five stages: Perceive, Update, Plan, Generate and Reflect. Agent loop overview:

- Perceive: uses Context Fusion to gather signals from the device and the learning platform.
- Updates: refresh the learner model and the current curriculum target.
- Plan: selects the next support using the motivation, affect, metacognition, and dialogic layers.
- Generate: composes the tutor message using retrieval constrained generation plus a factuality check.
- Reflect: records the decision, applies policy filters, and adjusts resources.

Each phase corresponds to one or more layers, the layer map: Perceive → Context Fusion; Update → Instructional Alignment, Computational Personalization; Plan → Motivation, Affect, Metacognition, Orchestration; Generate → Generative Reasoning; Reflect → Governance, Infrastructure.

This loop enables the tutor to operate as an independent, reflective digital pedagogical twin, with each step being informed by context as well as being of theoretical foundation. Fitting with the AI Gem model, this loop is not only an execution engine but is equally a pedagogical monitoring circuit enabling traceability, instructional quality, and modular evolvability across layers.

4.3. Pedagogical Layers

Instructional Alignment aligns context with concept and prerequisite. It retains challenge within learner's proximal zone. Computerized Personalization has a Bayesian estimate of proficiency and revises it on each turn. Motivational Engagement implements the Attention, Relevance, Confidence, Satisfaction model of maintaining effort. Affective Support recognizes effect and supplies brief empathetic reframes. Metacognitive Regulation elicits short self-explanations and reflection, with limited intensity to prevent fatigue.

4.4. Dialogic Layers

Interaction Orchestration makes decisions about when to ask, whether to prompt, and when to wait and observe for self-explanation. Generative Reasoning uses the Section 3 pipeline (retrieval → prompt assembly → generation → verification); here the focus on its dialogic role in the turn (see Figure 7).

4.5. Governance and Infrastructure Layers

Governance and trust reveal teacher-facing explanations, policy controls, and audit records. They track information used, why a prompt or hint was selected, and with what certainty the answer was provided.

Human oversight operates through a simple workflow with named roles. Low-confidence or policy triggers issue a neutral hold and teacher review; verification and logging steps are detailed in the Section 3 pipeline. Bias mitigation executes pre- and post-generation within the same pipeline, and Infrastructure & Deployment executes it across edge and cloud for latency and privacy. Figure 7 shows the low-confidence hold, the teacher and admin dashboards, and the logging and audit store that implement these workflows, including an on-device fallback for constrained settings.

The paper rank data as high (raw learner text/audio/video/physiology), medium (per-turn mastery, affect tags), and low sensitivity (de-identified timing, cache keys). Retention: raw text/media retained on edge $\leq$30 days then removed; derived features/mastery $\leq$12 months; audit logs (roles/decisions) $\leq$24 months. Consent: age-appropriate enrollment consent, right to erasure against raw and derived data, machine-readable manifest of holdings. Jurisdictions: GDPR, legitimate interests or explicit consent, data minimization, SCCs across borders; FERPA, institution retains education records, inspection/correction rights, vendor DPAs. Restricted settings: offer text-only/offline/on-device modes and switch high-sensitivity collection off by default.

4.6. Cross-Layer Propagation and Worked Examples

Context signals spread through layers so that a single detection may induce coordinated modification. Context signals propagate across layers:

- Frustration ↑ → simplify target or add prerequisite (Instructional Alignment)
- Add brief empathy (Affective Support);
- Choose a worked example (Generative Reasoning);
- Alert the teacher (Governance).

If mastery of a prerequisite falls below threshold, retrieval is adjusted, and the next hint waits for a one-sentence self-explanation (Orchestration). Because these changes use explicit artifacts, they are inspectable and testable.

4.7. Scenario-Based Proof-of-Concept Sketch

This sub-section tracks one specific learner interaction across the ten layers in Figure 5 and then summarizes a minimal proof of concept that enacts the same flow in practice. Figure 6 shows a learner journey mapped to layers using a swim lane layout. Each step is numbered and appears in exactly one lane.

For example, consider a Grade 8 learner working on adding fractions with unlike denominators. Context fusion aggregates three signals over two minutes: firstly, the novice learner's changed context, with response latency rising from 8 to 25 s; then, two consecutive errors share a common-denominator mistake; finally, the learner pauses before requesting an additional hint. Instructional alignment maps the task to the prerequisite concept of finding a common denominator. Computational personalization updates the learner's mastery estimate of that concept from 0.62 to 0.41 based upon the error type, the time costs incurred in making the errors, and the nature of the hint being requested. Motivational engagement selects a new shorter learning prompt that reframes the step; affective support prepares an empathetic lead-in that acknowledges their work and pressure; metacognitive regulation prompts the learner to submit a self-explanation about the 'next step' in one sentence. The interaction orchestration layer schedules a short clarifying question before the hint display. Generative reasoning retrieves two validated passages, and a worked example in least common multiples from the knowledge base and constructs a grounded prompt constraining the transformer model, which is then verified against the retrieved passages. Governance and trust record the rationale, the retrieved sources and the confidence signal for teacher review. Infrastructure and deployment route retrieval to a nearby cache and manage total round-trip latency below 600 milliseconds. The learner then completes the item with a single guided step, and the performance evidence goes to computational personalization, which updates the mastery estimate to 0.58, and then to governance for audit.

Figures 6 and 7 make these mechanisms observable: context signals propagate to target selection, dialogic policy, and retrieval; the low-confidence gate routes responses to teacher review.

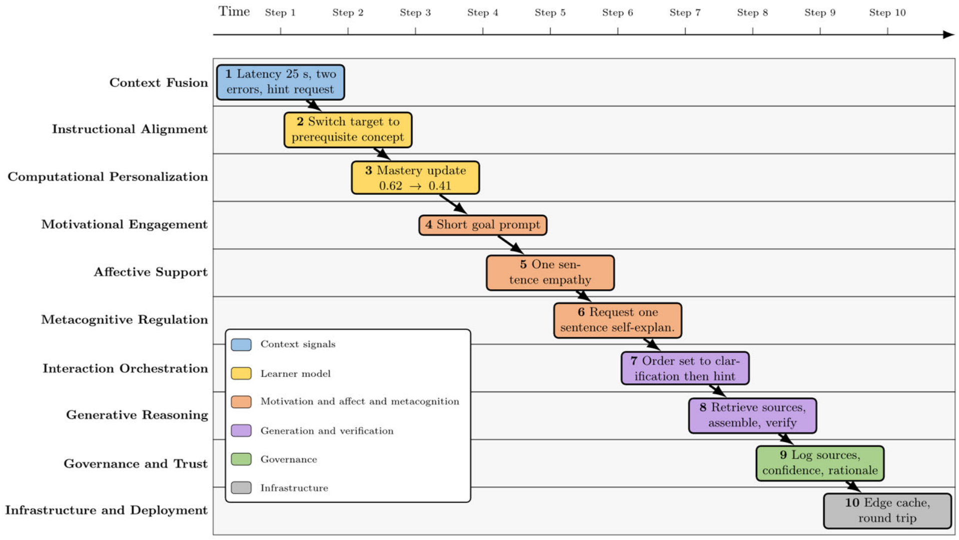
Time
Step 1
Step 2
Step 3
Step 4
Step 5
Step 6
Step 7
Step 8
Step 9
Step 10
Context Fusion
Instructional Alignment
Computational Personalization
Motivational Engagement
Affective Support
Metacognitive Regulation
Interaction Orchestration
Generative Reasoning
Governance and Trust
Infrastructure and Deployment
1 Latency 25 s, two errors, hint request
2 Switch target to prerequisite concept
3 Mastery update 0.62 → 0.41
4 Short goal prompt
5 One sentence empathy
6 Request one sentence self-explan.
7 Order set to clarification then hint
8 Retrieve sources, assemble, verify
9 Log sources, confidence, rationale
10 Edge cache, round trip
Context signals
Learner model
Motivation and affect and metacognition
Generation and verification
Governance
Infrastructure

Figure 6. Learner journey to layer mapping for Grade 8 fractions. Lanes list layers. Steps show what artifact or decision each layer produces.

Figure 7 summarizes a minimal proof of concept that enacts the same flow. The figure shows the operational pipeline with edge and cloud areas and category colors for clarity. Device and LMS signals come into Context Fusion, whose outputs feed two edge functions. One stream updates the learner state in the BKT update and local store, and the other exports de-identified features to the edge cache such that they can be retrieved fast. On the cloud side an ingestion and semantic index underlies Retrieval with query expansion. Retrieval takes input from the BKT state and sends the selected sources on to Prompt assembly where the curriculum rules and safety policies are also used and receives constraints from the Teacher dashboard and the Admin dashboard. Assembled prompt drives Generation, and the draft response is verified in Verification against retrieved text. The result goes on to the Low confidence gate that sends the tutor turn on if confidence exceeds it or forwards the case on to the Teacher dashboard to be reviewed if it falls below it, while all events are recorded in the Logging and audit store and admin actions are logged. In resource limited environments Device and LMS signals can also be sent to an On device light model that sends the tutor turn if and when the central services are down. This figure thus relates context acquisition, learner modeling, retrieval, prompt assembly, generation, verification, governance, and delivery together in one auditable flow.

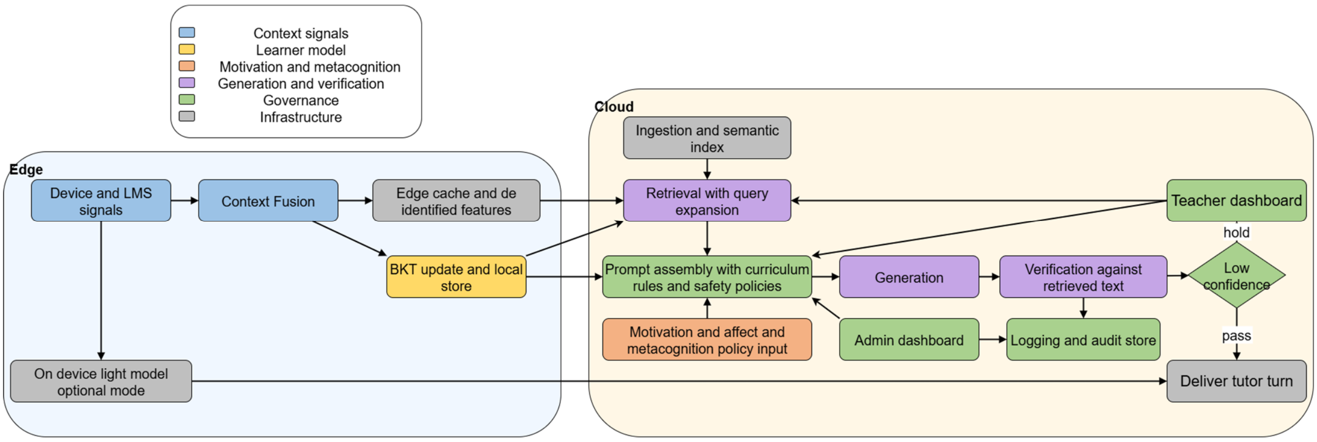
Context signals
Learner model
Motivation and metacognition
Generation and verification
Governance
Infrastructure
Edge
Device and LMS signals
Context Fusion
Edge cache and de identified features
BKT update and local store
On device light model optional mode
Cloud
Ingestion and semantic index
Retrieval with query expansion
Teacher dashboard
hold
Prompt assembly with curriculum rules and safety policies
Generation
Verification against retrieved text
Low confidence
Motivation and affect and metacognition policy input
Admin dashboard
Logging and audit store
pass
Deliver tutor turn

Figure 7. Operational pipeline of AI Gem with edge and cloud separation.

A proof-of-concept can be constructed with a small knowledge base of teacher-approved snippets indexed by embeddings, a retrieval function returning the top candidates by cosine similarity, a transformer model that is constrained by the retrieved text and policy instructions, a BKT store indexed by knowledge components, and two lightweight dashboards exposing the retrieved sources, mastery trajectory, and basic time signals for teachers and administrators. The architecture described by Figure 5 works as anticipated, even at this limited scale: context becomes fused into features, pedagogy specifies the target and scaffolding, dialog regulates the turn, RAG is utilized, and governance captures decisions for reporting and monitor oversight.

5. Opportunities, Benefits, and Challenges of Context-Aware AI-Agent Tutoring Within the AI Gem Framework

Based on the layered architecture, in this chapter we assess the tangible benefits that context-aware digital twin tutors already provide within learning environments and identify the overt vulnerabilities that need to be managed prior to such systems being deemed educationally reliable. The discussion is integrated with the synthesis of recent empirical results with the design principles outlined previously, framing them within contemporary debates on equity, efficacy, and trustworthiness within AI-augmented learning environments. The evaluation is framed within the AI Gem framework, where it situates the digital twin tutors as context-aware pedagogical agents acting through a compositional architecture of interoperable layers.

5.1. Opportunities and Benefits

The most common reported advantage of AI-controlled digital twin tutors is individualization of learning at scales that would be impossible for human teachers alone. Such transformer-activated agents replicate expert tutor instruction on a turn-by-turn basis as they adapt to the learner's state. A recent randomized controlled field experiment matching an LLM-controlled tutor with best practice active learning showed considerably greater learning gains in shorter time with higher pupil motivation [20]. These results concur with earlier laboratory work relating the moment-by-moment adaptation of feedback with greater mastery learning.

Scalability constitutes a second benefit. Once fine-tuned, a transformer model can serve thousands of simultaneous learners without degradation in response latency, enabling resource-constrained schools to offer subject support that would otherwise require specialist staff. The recent integration of generative AI into commercial platforms such as Canvas demonstrates how digital twin tutors can operate in mainstream learning management workflows, providing twenty-four hour conversational assistance while keeping teachers in control of assessment decisions [21].

Context awareness further enriches learner engagement. Agents that fuse device telemetry, affective cues, and curricular state can modulate the difficulty of tasks, switch explanatory styles, or inject empathetic prompts when frustration is detected. Physiological indicators such as real-time heart-rate variability have likewise proved sensitive markers of moment-by-moment arousal in live classroom settings, providing actionable context for adaptive interventions [22]. Early evidence suggests that such multi-signal perception, as implemented in the agent loop, explains a substantial portion of the observed learning gains in vocabulary acquisition studies using GPT-4-based tutoring agents [20]. Because digital twin tutors do not fatigue, students obtain timely feedback during evenings and weekends, a pattern that has been shown to reinforce self-regulated learning habits and reduce achievement gaps between high and low resource schools.

Instructors also benefit. Automated generation of practice items, first pass grading, and dashboard analytics frees teacher time for conceptual clarification and mentoring. Surveys conducted after semester long deployments consistently report improvements in perceived instructional efficacy and reduced administrative workload when generative assistants are integrated through well designed teacher dashboards [21].

In the AI Gem framework, these functions are not framed as discrete affordances but rather as emergent behaviors of correlated layer activity, such that feedback, personalization, and engagement are all optimized at once via cross-layer propagation.

5.2. Challenges and Mitigation Strategies

Though the encouraging possibilities of digital twin tutor agents hold much promise, some limitations remain unresolved and need remediation prior to their deployments at scale. Factual reliability is the first among them. Benchmark research like that of MathTutorbench identifies that the present LLMs of the state of the art continue to yield pedagogically incorrect feedback [23]. The adoption of RAG pipelines, ensemble prompting, and post-hoc verification modules decreased but did not eradicate those inaccuracies. Because digital twin tutors provide expert-like guidance, designers of systems need to adopt confidence-aware response mechanisms and permit the fallback to human review in high-stakes instructional scenarios.

Bias and safety are other issues. Big scale corpus pre training of models subjects to cultural stereotypes that can appear at a delicate level as the tutor dialogue proceeds. An extensive 2025 survey of intelligent tutors of all kinds ranked bias amplification and occasional culturally insensitive examples as common failures, particularly under conditions of gendered or socio-economic reference [24]. Responsible deployments thus include bias sensitivity tests across diverse learner profiles, filtering of contents with pedagogical harm-tuned protection layers, and clear reporting of mitigation outcomes to teachers and parents.

Data protection and privacy are of foremost importance when digital twin agents hold persistent learner models. The latter typically encompass fine-grained information regarding a pupil's misconceptions, emotional condition, and problem-solving strategies. To align with protocols such as FERPA and GDPR, builders need to deploy edge and cloud hybrid systems that keep sensitive computation localized while centrally storing only anonymized embeddings. Data erasure procedures and informed consent must be incorporated within the platform's governing layer along with audit logs.

Overdependency on guidance by AI poses pedagogical danger. Longitudinal research suggests that, with unrestricted hinting on offer, some of the students offload problem solving to the agent instead of acquiring persistence and metacognitive strategies. Adaptive withholding procedures, where the agent successively reduces scaffolding or asks for pupil self-explanations prior to exposing the next step, have been successful in re-establishing productive struggle without cutting down on learner engagement.

Lastly, expense and infrastructure continue to be realistic hurdles. Usage of high-end LLM endpoints continues to be beyond the technology budget of numerous public institutions. Open-source distilled models provide partial mitigation but often underperform in applications involving fine-grained reasoning. Cooperative procurement and federated fine tuning initiative gain credence as realistic solutions, but empirical cost–benefit analyses remain scant.

Equity and access need explicit focus. Schools with inadequate resources will be under bandwidth limits, use older devices, have inconsistently available power, and offer limited local IT capacity, all of which can impede use and increase gaps. Therefore, in all platform deployments, text-only and low-bandwidth versions, optional suppression of images and audio, offline caches of core content and logs with delayed syncing, and on-device light-weight models should be used to provide basic tutoring when connectivity is unavailable. To achieve greater equity and access, it will also be important to consider localization of language and examples, keyboard-only navigation, and screen-reader-compatible interfaces. Procurement should offer transparent education-tier pricing with flat caps and enable national or consortium-managed edge deployments.

In AI Gem, limitations work through certain risk mitigation paths attendant to every tier: bias mitigation using governance filters, latency management using infrastructure modules, and learner freedom using orchestration policies. Such collaboratively integrated safeguards forestall educational benefits being lost through systems fragilities.

5.3. Research and Development Agenda

To sustainably realize the benefits of context-aware digital twin tutors, ongoing interdisciplinary research is needed across three key domains. First, benchmark suites such as Math TutorBench and the CODE framework provide reproducible, context-aware task sets for measuring both factual accuracy and pedagogical quality, but they must be expanded to cover multimodal and affective tutoring behaviors [25]. Second, cross layer interventions should be evaluated experimentally; for instance, whether more granular emotion detectors in the context fusion layer improve dialogue coherence and learning outcomes through the orchestration layer. Third, teacher-in-the-loop governance models require systematic study to establish workflows that balance rapid AI feedback with human pedagogical judgment, especially in formative assessment.

Digital twin tutors (context-aware agents), as instantiated in the AI Gem framework, especially when implemented as digital twin tutors that emulate key instructional roles of human educators, demonstrate measurable improvements in learning gains, engagement, and instructional efficiency while extending personalized support to previously underserved populations. These digital twin agents operate continuously within a structured pedagogical loop, observing learner inputs, planning appropriate interventions, generating tailored outputs, and reflecting on their own instructional performance.

Yet the very capacity for generativity animating those benefits carries with it the risks of hallucination, bias, transgression of privacy, and overly strong dependence on the learner. Affecting lasting impact therefore depends on marrying transformer capability with severe verification pipelines, clear governance, and cost-sensitivity in infrastructure. Where such conditions obtain, context-aware digital twin tutors can be trustworthy partners augmenting human expertise, bringing personal adaptation learning down from aspirational ideal to working reality. The AI Gem architecture provides an architecture and evaluation framework for starting such transformation under controlled conditions. Table 4 encapsulates the key opportunities, benefits, and challenges of context-aware AI-agent tutoring.

Table 4. Overview of opportunities, benefits, and challenges of context-aware digital twin tutors within the AI Gem framework.

Category	Opportunities/Benefits	Key Challenges/Risks
Personalized learning outcomes	The context-aware AI tutor adapts explanations and activities to the learner's immediate knowledge in real time, enabling much faster progression than traditional instruction.	Impact depends strongly on model accuracy; incorrect feedback may mislead learners.
Accessibility and scalability	A fine-tuned LLM can serve thousands of learners simultaneously, providing subject support around the clock in under-resourced regions.	Peak-time latency and limited device access may restrict practical adoption.
Teacher augmentation	Automatically generated exercises, first-pass grading, and learner analytics free instruction time for mentoring and creative planning.	Verifying AI-generated content can be time-consuming, offsetting part of the time savings.
Context-aware engagement	Through the integration of affective cues, pacing, and task context, the system changes difficulty and explanation style, increasing motivation and self-regulated learning ability.	Noisy or missing context data can trigger inappropriate adaptation, reducing learning quality.
Reliability of explanations	RAG and post verification decrease the incidence of wrong statements.	Partially incorrect steps still arise in open-ended exercises at roughly twenty percent frequency, especially in longer calculations.
Bias and cultural fairness	Targeted fine tuning, bias audits, and inclusive libraries of examples counteract inherited stereotypes and enhance cultural relevance.	Performance and sensitivity gaps continue for underrepresented populations and low-resource languages.
Data privacy and security	An edge and cloud hybrid keeps personal data local while sending only de-identified features to the cloud for analysis.	Compliance with FERPA and GDPR requires robust consent, delete, and encryption workflows.

Table 4. *Cont.*

Category	Opportunities/Benefits	Key Challenges/Risks
Learner over-reliance	Gradual hint withdrawal and self-explanation prompts foster independent problem solving.	Surveys have reported that some students experience quality loss and help dependency with unrestricted AI support.
Cost and infrastructure	Distilled open-source models and education tier licenses assist in cutting down entry fees.	Token-based pricing for frontier models can exceed institutional budgets; hosting open models shifts hardware and maintenance burdens locally.

5.4. Design Feature to Outcome Mapping and Ablation Plan

To close the mechanism–evidence gap, the paper derives four causal hypotheses directly from AI Gem's artifacts in Figures 6 and 7. Each hypothesis relates one unique mechanism to observable happenings and outcomes (signals, gates, logs). The hypotheses (H1–H4) are presented below with mechanism, prediction, and measures.

H1 Cross-layer propagation (Figure 6).

- Mechanism: frustration and error signals from Context Fusion concurrently adjust target concept, dialogic policy, and retrieval.
- Prediction: faster recovery.
- Measures: median time to next correct step; repeated-error rate.

H2 Governance low confidence hold (Figure 7).

- Mechanism: verifier sends low-confidence drafts to teacher review.
- Prediction: fewer high-severity factual errors in delivered turns.
- Measures: proportion of turns with high-severity factual error; teacher override rate.

H3 RAG with verification (Figure 7).

- Mechanism: constrain generation to retrieved text and verify claims.
- Prediction: higher factual precision with limited loss of coverage.
- Measures: precision/recall of factual statements vs. reference key.

H4 BKT-informed scaffolding (Figures 6 and 7).

- Mechanism: hints/prerequisites conditioned on mastery.
- Prediction: fewer off-target hints; faster recovery.
- Measures: hint-relevance ratings; recovery time.

Ablation. Toggle one mechanism off at a time while others stay on; randomize at the turn level; report effect sizes with 95% CIs; stratify by bandwidth tier and device class.

6. Conclusions

AI Gem introduces a modular, context-aware agent architecture for transformer-based digital twin tutors that integrates pedagogical alignment, generative reasoning, interaction orchestration, and governance in a deployable stack. Coupled with verification, bias mitigation, and human oversight, the approach can enhance effective, engaging and scalable instruction while being auditable and compliant. The primary risks are factual errors, bias, privacy and consent responsibilities, operational expense and latency, and potential learner over-reliance, mitigations and safeguards. Prototype assessments will provide a succinct suite of indicators: latency distributions across deployment configurations, hallucination detection rates post-generation, pre and post learner interest, correctness and mastery growth; governance and equity signals, including teacher override rates, precision and recall of safety flags, time to review, per learner compute and bandwidth by deployment tier; and parity of outcomes across bandwidth, device, and language groups. Constraints remain in the form of low resource contexts, no large scale in situ bias audits, and the ongoing challenge of privacy and data protection across jurisdictions. The research agenda centers on the development of multimodal and affect aware benchmarks for factual accuracy, pedagogical quality, and equity; running longitudinal field studies across grades and domains; and piloting an extensible governance model that involves teacher dashboards, auditable logs, data minimization, consent management, and transparent cost control. In sum, AI Gem offers an actionable and testable pathway from idea to equitable classroom practice.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

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

References

1. Liu, S.; Guo, X.; Hu, X.; Zhao, X. Advancing generative intelligent tutoring systems with GPT-4: Design, evaluation, and a modular framework for future learning platforms. *Electronics* **2024**, *13*, 4876. [CrossRef]
2. Bernal, M.E. Revolutionizing eLearning Assessments: The Role of GPT in Crafting Dynamic Content and Feedback. *J. Artif. Intell. Technol.* **2024**, *4*, 188–199. [CrossRef]
3. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
4. Mirzababaei, B.; Pammer-Schindler, V. Facilitating the Learning Engineering Process for Educational Conversational Modules Using Transformer-Based Language Models. *IEEE Trans. Learn. Technol.* **2024**, *17*, 1210–1223. [CrossRef]
5. Burke, H.B.; Hoang, A.; Lopreiato, J.O.; King, H.; Hemmer, P.A.; Montgomery, M.; Gagarin, V. Assessing the Ability of a Large Language Model to Score Free-Text Medical Student Clinical Notes: Quantitative Study. *JMIR Med. Educ.* **2024**, *10*, e56342. [CrossRef] [PubMed]
6. Zhang, M.; Dilling, A.; Gondelman, L.; Lyngdorf, N.; Lindsay, E.; Bjerva, J. SEFL: Harnessing Large Language Model Agents to Improve Educational Feedback Systems. *arXiv* **2025**, arXiv:2502.12927. [CrossRef]
7. Ye, B.; Xi, Y.; Zhao, Q. Optimizing Mathematical Problem-Solving Reasoning Chains and Personalized Explanations Using Large Language Models: A Study in Applied Mathematics Education. *J. AI-Powered Med. Innov.* **2024**, *3*, 67–83. [CrossRef]
8. Keerthichandra, M.; Vihidun, T.; Lakshan, S.; Perera, I. Large Language Model-Based Student Intent Classification for Intelligent Tutoring Systems. In Proceedings of the 2024 9th International Conference on Information Technology Research (ICITR), Colombo, Sri Lanka, 5–6 December 2024; pp. 1–6. [CrossRef]
9. Albuquerque da Silva, D.; de Mello, C.E.; Garcia, A. Analysis of the Effectiveness of Large Language Models in Assessing Argumentative Writing and Generating Feedback. In Proceedings of the 16th International Conference on Agents and Artificial Intelligence–Volume 2: ICAART, Rome, Italy, 24–26 February 2024; SciTePress: Lisbon, Portugal, 2024; pp. 573–582. [CrossRef]
10. Nouzri, S.; EL Fatimi, M.; Guerin, T.; Othmane, M.; Najjar, A. Beyond Chatbots: Enhancing Luxembourgish Language Learning Through Multi-agent Systems and Large Language Model. In *PRIMA 2024: Principles and Practice of Multi-Agent Systems*; Lecture Notes in Computer Science; Arisaka, R., Sanchez-Anguix, V., Stein, S., Aydoğan, R., van der Torre, L., Ito, T., Eds.; Springer: Cham, Switzerland, 2025; Volume 15395. [CrossRef]
11. Liu, H. Applicability of ChatGPT in Online Collaborative Learning: Evidence Based on Learning Outcomes. In *Proceedings of the International Academic Conference on Education*; Diamond Scientific Publishing: Vilnius, Lithuania, 2024; pp. 22–43. [CrossRef]
12. Wang, F.; Cheung, A.C.K.; Neitzel, A.J.; Chai, C.S. Does Chatting with Chatbots Improve Language Learning Performance? A Meta-Analysis of Chatbot-Assisted Language Learning. *Rev. Educ. Res.* **2024**, *95*, 623–660. [CrossRef]
13. Anjum, F.; Raheem, B.R.; Ghafar, Z.N. The Impact of ChatGPT on Enhancing Students' Motivation and Learning Engagement in Second Language Acquisition: Insights from Students. *J. E-Learn. Res.* **2025**, *3*, 1–11. [CrossRef]
14. Chang, S.-W.; Kim, D.-S. Scalable Transformer Accelerator with Variable Systolic Array for Multiple Models in Voice Assistant Applications. *Electronics* **2024**, *13*, 4683. [CrossRef]
15. Gabriel, S. Generative AI and Educational (In)Equity. In *Proceedings of the International Conference on AI Research*; ICAIR: Chicago, IL, USA, 2024; Volume 4, pp. 133–142. [CrossRef]

16. Molnar, G.; Nagy, E. Current Issues in Effective Learning: Methodological and Technological Challenges and Opportunities Based on Modern ICT and Artificial Intelligence. In *International Scientific Conference on Distance Learning in Applied Informatics*; DiVAI 2024; Innovations in Communication and Computing; Springer Nature: Cham, Switzerland, 2024; Chapter 1; pp. 1–11. [CrossRef]
17. Naznin, K.; Mahmud, A.A.; Nguyen, M.T.; Chua, C. ChatGPT Integration in Higher Education for Personalized Learning, Academic Writing, and Coding Tasks: A Systematic Review. *Computers* **2025**, *14*, 53. [CrossRef]
18. Salih, S.; Husain, O.; Hamdan, M.; Abdelsalam, S.; Elshafie, H.; Motwakel, A. Transforming Education with AI: A Systematic Review of ChatGPT's Role in Learning, Academic Practices, and Institutional Adoption. *Results Eng.* **2025**, *25*, 103837. [CrossRef]
19. Rincón, Y.R.; Munárriz, A.; Campión Arrastia, M.J.; Goicoechea López-Vailo, M.I. Instructional Design for Tutoring on Interactive Platforms: Creating Educational Interventions Overcoming the Digital Gap. In *Educational Technology Research and Development*; Springer: Berlin/Heidelberg, Germany, 2025. [CrossRef]
20. Kestin, G.; Miller, K.; Klales, A.; Milbourne, T.; Ponti, G. AI Tutoring Outperforms In-Class Active Learning: An RCT Introducing a Novel Research-Based Design in an Authentic Educational Setting. *Sci. Rep.* **2025**, *15*, 17458. [CrossRef] [PubMed]
21. Francisti, J.; Balogh, Z.; Reichel, J.; Benko, Ľ.; Fodor, K.; Turčáni, M. Identification of heart rate change during the teaching process. *Sci. Rep.* **2023**, *13*, 16674. [CrossRef] [PubMed]
22. Zandvakili, R.; Liu, D.; Li, A.T.; Santhanam, R.; Schanke, S. Design and Evaluation of a Gamified Generative AI Chatbot for Canvas LMS Courses. In *HCI International 2024 Posters*; Communications in Computer and Information Science; Stephanidis, C., Antona, M., Ntoa, S., Salvendy, G., Eds.; Springer: Cham, Switzerland, 2024; Volume 2117, pp. 259–264. [CrossRef]
23. Macina, J.; Daheim, N.; Hakimi, I.; Kapur, M.; Gurevych, I.; Sachan, M. MathTutorBench: A Benchmark for Measuring Open-ended Pedagogical Capabilities of LLM Tutors. *arXiv* **2025**, arXiv:2502.18940. [CrossRef]
24. Sumanasekara, S.; Deckker, D. ChatGPT and the Evolution of AI-Powered Tutoring Systems. *EPRA Int. J. Environ. Econ. Commer. Educ. Manag.* **2025**, *12*, 71–87. [CrossRef]
25. Pit, H. Henry at BEA 2025 Shared Task: Improving AI Tutor's Guidance Evaluation Through Context-Aware Distillation. In Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, Vienna, Austria, 31 July–1 August 2025; pp. 1164–1172.

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.