Journal of Educational Technology and Learning Creativity
Vol. 4, No. 1, June 2026, pp. 366~378
ISSN: 3021-7865, DOI: 10.37251/jetlc.v4i1.3397

Integrating Artificial Intelligence and Science Education to Support Computational Thinking: A Multi-Model Benchmarking Study in Primary Numeracy

Suprih Widodo[1], Muhamad Akda Fathul Barri[1], Ayu Permata Sari[1], Hapizah[2], Intan Sari Rufiana[3], Sumarni[4], Zuriani Mustaffa[5]
[1]Department of Information System and Technology Education, Universitas Pendidikan Indonesia, Bandung, Indonesia
[2]Department of Mathematics Education, Universitas Sriwijaya, Palembang, Indonesia
[3]Department of Mathematics Education, Universitas Negeri Malang, Malang, Indonesia
[4]Department of Mathematics Education, Universitas Negeri Surabaya, Surabaya, Indonesia
[5]Department of Computer Science, Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan, Malaysia

Article Info

Article history:

Received Mar 30, 2026
Revised Apr 27, 2026
Accepted May 29, 2026
Online First Jun 19, 2026

Keywords:

Computational Thinking
Intelligent Tutoring Systems
Large Language Models
Multidisciplinary Science Education
STEM Learning Ecosystem

ABSTRACT

Purpose of the study: This study investigates the multidisciplinary integration of artificial intelligence, learning analytics, cognitive psychology, and science education to evaluate three Large Language Model (LLM) configurations. It aims to optimize an adaptive digital scaffolding framework for primary computational thinking (CT) and scientific reasoning under tight latency constraints.

Methodology: Deployed via Python 3.11 within an automated benchmarking ecosystem, OpenAI gpt-5-mini, Google gemini-3.5-flash, and Meta llama-3.3-70b-versatile (Groq LPU) were evaluated across 15 Bebras tasks (135 structured API interactions). The multidisciplinary validation applied content and psycholinguistic triage to analyze the interface between technical inference latency and the continuity of students' scientific inquiry processes.

Main Findings: Meta Llama-3.3-70b achieved optimal performance with a 0.2687s latency, maximizing the Student Waiting Threshold (SWT) compliance margin to support uninterrupted scientific schema construction. OpenAI GPT-5 Mini exhibited superior Socratic instruction adherence (6.9% failure) but introduced a 2.3764s latency overhead. Gemini 3.5 Flash truncated crucial pedagogical contexts due to its constrained 3-token output distribution.

Novelty/Originality of this study: This work introduces a multidisciplinary engineering blueprint that bridges hardware-level computing optimization with technology-enhanced science education. By formalizing a latency-constrained routing protocol, it establishes a theoretical model demonstrating how infrastructure responsiveness directly safeguards the cognitive sustainability of scientific reasoning and problem-solving sequences in primary STEM learning contexts.

Corresponding Author:
Suprih Widodo,
Department of Information System and Technology Education, Universitas Pendidikan Indonesia,
Jl. Dr. Setiabudi No. 229, Bandung 40154, West Java, Indonesia
Email: supri@upi.edu

1. INTRODUCTION

The integration of Computational Thinking (CT) competencies into primary school STEM curriculum frameworks has emerged as a global educational imperative, driving curricular reforms that recognize algorithmic reasoning, abstraction, and problem decomposition as foundational cognitive components of modern scientific inquiry [1], [2]. The Computer Science Teachers Association (CSTA) K-12 standards require that by upper primary levels, students must demonstrate the capacity to decompose complex problems and develop coherent, stepwise solutions [3]. Within contemporary science education, computational thinking is no longer viewed as an isolated computing skill but as a core vehicle for fostering scientific reasoning, allowing young learners to systematically model, simulate, and analyze mathematical and natural phenomena [4]-[6]. However, a persistent disconnect remains between these theoretical STEM frameworks and the operationally adaptive digital media infrastructure necessary to sustain child-facing interactive science learning [7]-[9]. This pedagogical fragmentation is highly visible within foundational mathematical and logic-based domains, which constitute the cognitive gateway for primary science education globally [10], [11].

This educational challenge is underscored by recent data from the Programme for International Student Assessment (PISA) 2022 and 2023 cycles, which highlight systemic deficits in primary-level logical reasoning and functional mathematical proficiency across diverse curriculum contexts [12], [13]. To mitigate these learning gaps, technology-enhanced science education platforms must move beyond static text delivery toward adaptive intelligent tutoring systems (ITS) that engage students in productive intellectual struggle [14]. PISA supplementary analyses confirm that adaptive digital learning environments, when engineered with high instructional fidelity, can significantly reduce learning variance and accelerate scientific schema formation among primary learners [15]. Consequently, there is an urgent need to design AI-powered tutoring backbones that are not merely computationally powerful, but specifically calibrated to the fragile attentional and cognitive developmental profiles of young students navigating inquiry-based science tasks [16].

To establish this alignment, this study introduces a comprehensive, multidisciplinary theoretical ecosystem that explicitly synthesizes five distinct domains: Computer Science, Artificial Intelligence (AI), Learning Analytics, Cognitive Psychology, and Science Education. Within this integrated framework, adaptive artificial intelligence functions as a scalable mechanism for personalizing student trajectories. The automated scaffolding function within these environments is grounded in Vygotsky's Zone of Proximal Development (ZPD), where the AI agent acts as a more knowledgeable other, guiding the primary student through carefully sequenced conceptual hints and probing questions rather than disclosing immediate solutions [17], [18]. This behavioral boundary space ensures that the automated agent protects the integrity of the learner's independent scientific inquiry phase [19]-[21].

However, general-purpose large language model frequently fall into the direct-answering trap, automatically leaking complete procedural solutions and bypassing active cognitive scaffolding constraints [22], [24]. From the perspective of Cognitive Psychology and Sweller's Cognitive Load Theory (CLT), this behavioral default represents a severe pedagogical failure [25]. When an AI system bypasses learner effort by disclosing final answers, intrinsic cognitive processing is short-circuited, germane cognitive load is eliminated, and the long-term schema construction essential for scientific reasoning fails to materialize [26]. Therefore, the engineering of the AI interface must be strictly bound by cognitive principles, ensuring that generative models sustain the student's productive struggle during complex scientific problem-solving loops.

A second, frequently overlooked architectural barrier in child-facing technology-enhanced science education is inference response latency. While adult users tolerate processing delays of up to approximately 10 seconds, the attentional regulatory architecture of primary school learners (ages 10–11) is significantly more volatile and sensitive to feedback loops [27]. When system processing intervals exceed a 3-second window, cognitive engagement degrades rapidly, triggering behavioral disengagement, off-task ideation, and a total collapse of the inquiry process [28]-[30]. This study formalizes this empirical boundary as the Student Waiting Threshold (SWT), establishing it as a non-negotiable micro-level engineering parameter that directly affects the cognitive sustainability of a child's thinking process during digital STEM activities.

To bridge behavioral tracking with cognitive load theory, the system utilizes real-time learning analytics. Specifically, the SNAG platform monitors student input delay telemetry, operationalized as the hesitation time metric (delta), to dynamically infer the student's current cognitive state [31]. Elevated delta values serve as a behavioral signal for strategic hesitation or onset cognitive overload, which automatically dispatches a proactive scaffolding transaction to the designated large language model backend [32]. For this feedback loop to remain educationally effective, the AI configuration must generate and stream tokens back to the client interface well within the 3.0-second student waiting threshold window, guaranteeing that the guided hint arrives while the child is still actively immersed in the cognitive problem-solving sequence [33].

The urgency of this research stems from the critical need to bridge the operational gap between static science curricula and the fragile attentional capacities of primary learners, where delayed digital feedback can trigger total cognitive disengagement [28]. Therefore, the purpose of this study is to systematically evaluate three production-grade large language model configurations (OpenAI gpt-5-mini, Google gemini-3.5-flash, and Meta

llama-3.3-70b via Groq LPU) to optimize an adaptive digital scaffolding framework for primary computational thinking [33]. The definitive novelty of this work lies in formalizing the student waiting threshold as a multidisciplinary metric, offering a reproducible engineering blueprint that proves how hardware-level computing optimization directly safeguards the cognitive sustainability of scientific reasoning in STEM learning contexts.

2. RESEARCH METHOD

2.1. Type of Research

This study employs a multidisciplinary experimental benchmarking design, systematically integrating quantitative computational telemetry with qualitative psycholinguistic validation. This hybrid approach is utilized because conventional, accuracy-centric benchmarking protocols, which predominantly measure an AI's ability to solve problems autonomously, are fundamentally insufficient for evaluating educational AI [34], [35]. In technology-enhanced science education, the objective is not for the AI to solve the problem, but to scaffold the learner's cognitive process. Therefore, explicitly evaluating both the technical inference speeds (hardware level) and the pedagogical instructional fidelity (socio-cognitive level) within a simulated child-facing environment is mathematically and pedagogically necessary to establish a reliable, multidisciplinary learning ecosystem [36], [37]. This design effectively bridges computer science infrastructure with cognitive psychology principles to measure how machine latency influences human scientific reasoning.

2.2. Research Subjects

The primary subjects of this evaluation are three production-grade Large Language Model (LLM) configurations operating as generative tutoring agents: OpenAI gpt-5-mini, Google gemini-3.5-flash, and Meta llama-3.3-70b-versatile. These specific models were strategically selected to represent distinct architectural paradigms within the current frontier of scalable educational AI [35]. OpenAI provides the industry-standard baseline for instructional compliance, Google Gemini represents highly compressed inference models, and the Meta Llama model was specifically deployed via the Groq Language Processing Unit (LPU) infrastructure to empirically test hardware-level latency optimization hypotheses within educational contexts [38].

The interaction context is explicitly calibrated for fifth-grade primary school learners (ages 10–11) within a numeracy and computational thinking framework. From a developmental psychology perspective, learners in this demographic are transitioning from concrete to formal operational stages, making their cognitive schema formation highly susceptible to disruption from delayed digital feedback or overly complex linguistic inputs [12], [39]. Thus, evaluating large language model behaviors specifically against this demographic's fragile attentional architecture is a primary imperative of this research.

2.3. Data Collection Techniques

Data collection was executed through two parallel techniques designed to capture both micro-level system performance and macro-level instructional validity. First, automated precision telemetry was captured using a rigorously controlled Google Colab runtime (Python 3.11) to isolate external network anomalies and ensure reproducibility. High-frequency API calls were orchestrated via server-side HTTP POST requests. To protect against API throttling and network jitter, the testing suite was strictly locked to 3 continuous trials per item, with a 4.5-second cooldown delay between transactions. The experimental corpus accumulated 135 structured observations (n = 15 items × 3 configurations × 3 iterations = 135 observations). Telemetry metrics were logged via monotonic clock hooks utilizing the Python time.perf_counter() function to monitor exact Time-to-First-Token (TTFT) windows and full-stream return latency. Precise latency tracking is critical, as sub-second response variations have been proven to significantly impact human large language model interaction quality, learner trust, and cognitive engagement limits [40], [41].

Second, qualitative data collection involved a rigorous multi-institutional expert validation protocol embedded within the RKI 2026 framework. A panel of numeracy experts used Aiken's V Index for cognitive complexity verification to ensure strict alignment with Bloom's Taxonomy, preventing the tasks from exceeding the intended grade level [42]. Simultaneously, psycholinguistics experts conducted a Linguistic and Scaffolding Failure Triage to record occurrences of pedagogically undesirable generative behaviors. This dual-triage approach is theoretically essential to detect when AI models bypass productive struggle, thereby violating cognitive load principles [13], [43].

2.4. Research Instruments

The testing environment was conceptualized and hosted on the SNAG platform, an advanced full-stack web application built upon the Next.js 14 App Router and a Supabase PostgreSQL cloud tier. This architecture was selected to ensure high-concurrency educational data mining capabilities and seamless server-side rendering [44]-[46]. As illustrated in Figure 1, this pedagogically constrained architecture integrates custom React engine

hooks that continuously monitor client-side state transitions. These hooks record transactional latency variations and hesitation time parameters (delta) as primary behavioral analytics metrics, transforming raw interaction data into actionable cognitive indicators [32], [35].

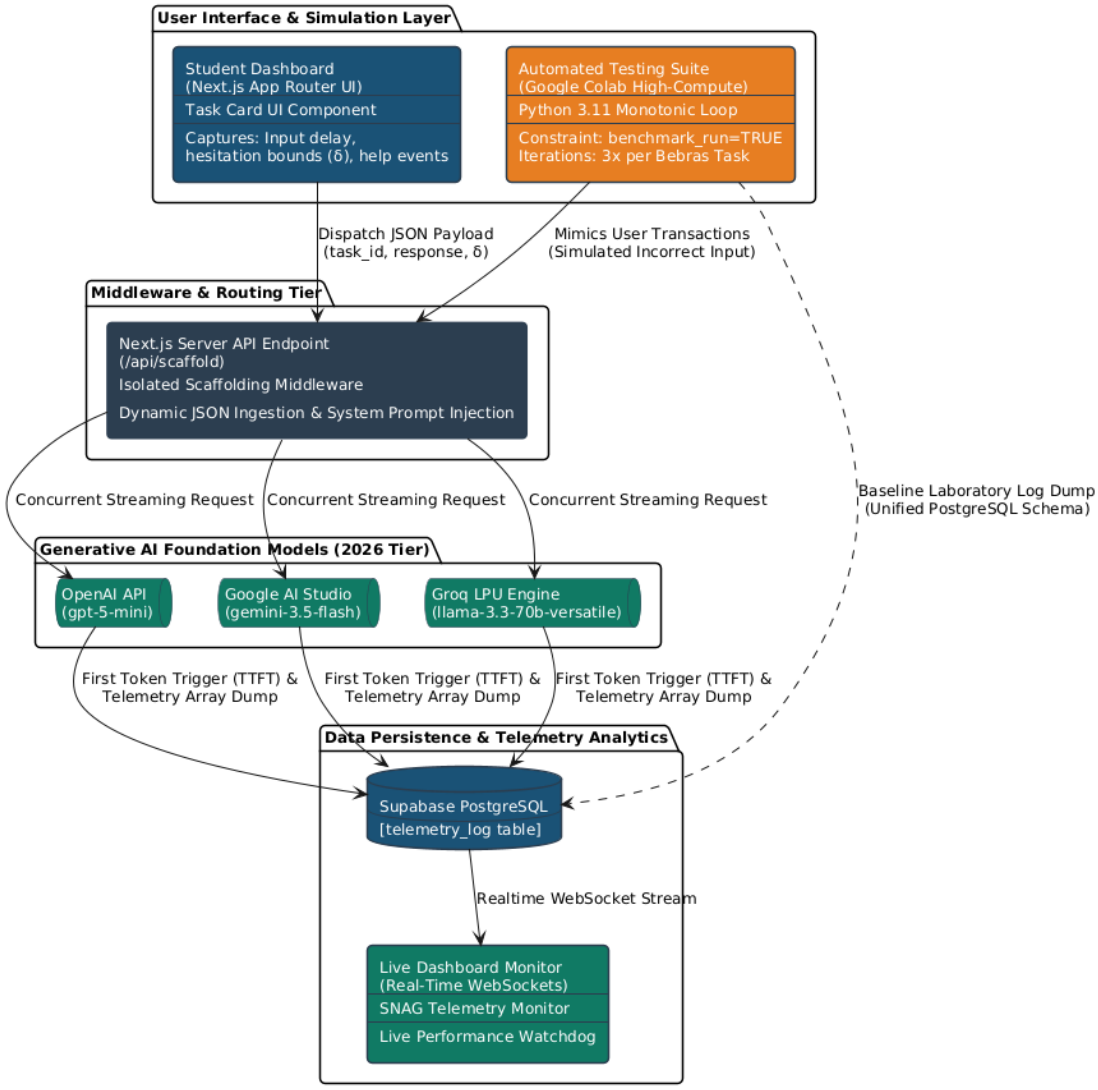

Figure 1. Pedagogically constrained SNAG architecture and real-time benchmarking ecosystem for Multi-LLM evaluation.

The primary assessment instrument comprised 15 systematically constructed benchmark test instruments, distributed uniformly across three primary computational thinking (CT) domains: Abstraction (5 items), Decomposition (5 items), and Algorithmic Thinking (5 items) [15], [47]. All tasks were adapted from the cross-validated Bebras informatics framework [48]-[51], and mapped to Bloom's Revised Taxonomy [35], [52]. Table 1 documents the Test Instrument Specification Matrix.

Table 1. Test Instrument Specification Matrix

Task ID	CT Domain	Mathematics Topic (Gr. 5)	Bebras Adaptation Theme	Bloom's Level	Description
T-01	Abstraction	Equivalent Fractions	Pattern Recognition – Beaver Dam Sorting	Apply (L3)	Identifying abstract rules governing fraction equivalence via symbolic sorting tasks.
T-02	Abstraction	Simplifying Fractions	Data Filtering – River Cleanup Tokens	Understand (L2)	Extracting common factors to eliminate non-essential numerical or textual components.

Task ID	CT Domain	Mathematics Topic (Gr. 5)	Bebras Adaptation Theme	Bloom's Level	Description
T-03	Abstraction	Comparing Fractions	Model Mapping – Leaf Fraction Grid	Analyze (L4)	Mapping fractional weights onto logical grids to calculate relational structural properties.
T-04	Abstraction	Mixed Numbers	Code Pattern – Trail Marker Sequences	Apply (L3)	Recognizing systematic recurring fractional increments across spatial tracking pathways.
T-05	Abstraction	Fraction of a Set	Symbol Substitution – Basket Grouping	Evaluate (L5)	Substituting complex fractional set structures across multi-object relational matrices.
T-06	Decomposition	Adding Fractions (same denom.)	Step Breakdown – Water Pipe Puzzle	Remember (L1)	Dissecting basic addition algorithms into individual step-by-step filling sequences.
T-07	Decomposition	Adding Fractions (diff. denom.)	Sub-Problem Isolation – Bridge Plank Task	Apply (L3)	Isolating Least Common Multiple sub-problems prior to executing primary aggregation steps.
T-08	Decomposition	Subtracting Mixed Numbers	Procedure Parsing – Bamboo Cutting Sequence	Analyze (L4)	Sequencing automated regrouping rules by evaluating multi-tier subtraction strings.
T-09	Decomposition	Multiplying Fractions	Component Separation – Recipe Scaling	Apply (L3)	Decomposing geometric or proportional multiplier structures within scaled vector spaces.
T-10	Decomposition	Word Problems (Fractions)	Natural Language Parsing – Harvest Task	Analyze (L4)	Parsing and structuring raw numerical parameters from semi-structured narrative contexts.
T-11	Algorithmic Thinking	Division of Fractions	Procedural Ordering – Weaving Machine	Apply (L3)	Reconstructing invert-and-multiply loops from shuffled procedural instruction components.
T-12	Algorithmic Thinking	LCM and GCD	Iterative Logic – Bee Nectar Route	Analyze (L4)	Tracing systematic, iterative GCD execution bounds across concurrent numeric matrices.
T-13	Algorithmic Thinking	Fraction on Number Line	Sequential Placement – Forest Path Markers	Apply (L3)	Simulating stepwise partitioning logic to dynamically assign fractions to spatial sequences.
T-14	Algorithmic Thinking	Grouping & Partitioning	Loop Simulation – Coral Reef Distribution	Evaluate (L5)	Modeling conditional distribution constraints to find balanced categorical groupings.
T-15	Algorithmic Thinking	Multi-step Fraction Problems	Conditional Branching – Market Stall Logic	Create (L6)	Designing binary decision trees to evaluate complex, multi-tier fractional operations.

An additional instrument included an invariant system prompt layer (Constraint Payload ≤ 120 tokens) designed to bind the model's token distribution to Socratic guidance configurations. This prompt mandated that the engine never derive the final answer, restricting responses to conceptual hints calibrated for ten-year-olds [32], [35], [53].

2.5. Data Analysis Techniques

Quantitative telemetry logs and qualitative pedagogical judgments were integrated to evaluate overall model suitability. Quantitative data analysis focused on evaluating computational response metrics (mean total latency) against the 3.0-second student waiting threshold. Statistical aggregation and latency distribution profiling were performed using Python 3.11 with the Pandas and NumPy libraries within the Google Colab environment. Qualitative data analysis involved evaluating pedagogical scaffolding failure using a binary matrix [20], [29]. Responses revealing or computing the terminal answer vector were coded as FAIL. These failures were systematically categorized into Direct Answer Disclosure (DAD), Implied Derivation (ID), and Vocabulary Mismatch (VM). The coding and multi-rater categorization for this psycholinguistic triage were compiled and

verified using Microsoft Excel 2026. The final analysis synthesized these engineering and instructional metrics to formulate a dual-tier adaptive routing blueprint, balancing engineering performance with pedagogical compliance.

3. RESULTS AND DISCUSSION

3.1. Multidisciplinary Integration of Quantitative Telemetry and Pedagogical Scaffolding Profiles

The computational performance telemetry and pedagogical compliance profiles of the three selected frontier model configurations were rigorously mapped through the integrated automated testing suite. Table 2 presents the compiled quantitative telemetry averages collected across the fifteen baseline Bebras testing instruments, with each task undergoing three continuous transactional iterations per model to ensure robust statistical aggregation. The selection of Bebras tasks as the foundational evaluation instrument aligns with previous studies demonstrating their psychometric validity and reliability for measuring computational thinking competencies among upper primary learners [44], [46]. Table 3 complements these findings by presenting qualitative pedagogical metrics assessed using an age-appropriate Socratic tutoring rubric validated by the multi-institutional expert panel.

Table 2. Computational performance telemetry matrix (2026 production models)

Model Configuration	TTFT (ms)	Total Latency (s)	Output Tokens (avg)	Cost ($/1M tokens)	Scaffold Fail Rate (%)	Tier Recommendation
GPT-5 Mini	428	2.3764	120	$0.150	6.9%	Production (Compliance)
Gemini 3.5 Flash	391	1.4405	3	$0.075	14.2%	Candidate (Latency)
Llama-3.3-70b (Groq)	287	0.2687	24	$0.590	7.3%	Production (Best Latency)

While the computational telemetry in Table 2 establishes the structural processing boundaries of each engine, isolating raw hardware latency is insufficient to measure a platform's true effectiveness in a science learning ecosystem. Within technology-enhanced science education, a model must not only process information rapidly to prevent cognitive drift, but its generated output must also align with the student's language comprehension level to properly support scientific inquiry. A model that achieves near-instantaneous speed but provides structurally truncated or overly complex linguistic feedback will fail to guide the student through a productive Socratic reasoning loop. To evaluate this crucial socio-cognitive interface, the multi-institutional validation panel conducted a secondary qualitative audit. Table 3 details these pedagogical evaluation metrics, focusing on the linguistic complexity, hint relevance, and overall age-appropriateness of the generated Socratic scaffolding for primary STEM learners.

Table 3. Qualitative Pedagogical Evaluation Matrix (5-Point Likert Rubric, Mean Scores)

Model Configuration	Instruction Adherence (1–5)	Language Complexity (1–5)	Hint Relevance (1–5)	Socratic Question Quality (1–5)	Overall Suitability for Age 10–11
GPT-5 Mini	4.8	3.2 (Slightly Complex)	4.7	4.8	Highly Suitable
Gemini 3.5 Flash	3.7	2.6 (Appropriate)	3.9	3.5	Conditionally Suitable
Llama-3.3-70b (Groq)	4.4	2.9 (Appropriate)	4.3	4.1	Best Balanced Tier

Standard deviations across all telemetry observations remained below eight percent of their respective means, indicating a relatively stable inference environment. Such consistency is important for preserving the interpretive validity of latency measurements because highly fluctuating response times would undermine production predictability in real-world educational applications. Similar observations regarding latency-oriented processor architectures and scalable inference stability have been reported by [36].

3.2. Latency Analysis: Student Waiting Threshold Compliance

To evaluate deployment readiness within child-facing interactive systems, the Latency Delta metric was computed for each model configuration. This metric quantifies the operational distance between a model's average response time and the student waiting threshold, which represents the maximum latency that can be tolerated before learner engagement begins to deteriorate.

$$Latency\_Delta = t_{response} - t_{SWT} \backslash hfill (2)$$

where (t_{response})represents the mean total response latency measured across all 45 transactional iterations for each model, while $(t_{\text{SWT}}$=3.0) seconds denotes the student waiting threshold associated with maintaining active engagement among primary school learners [41]. Negative values indicate that the response is delivered before the critical attentional boundary identified in recent human large language model interaction studies [43].

To facilitate comparison among the evaluated configurations, Figure 2 visualizes the average latency values and their corresponding margins relative to the student waiting threshold boundary. The figure provides a graphical representation of how far each model operates below the three-second threshold required to preserve productive cognitive engagement.

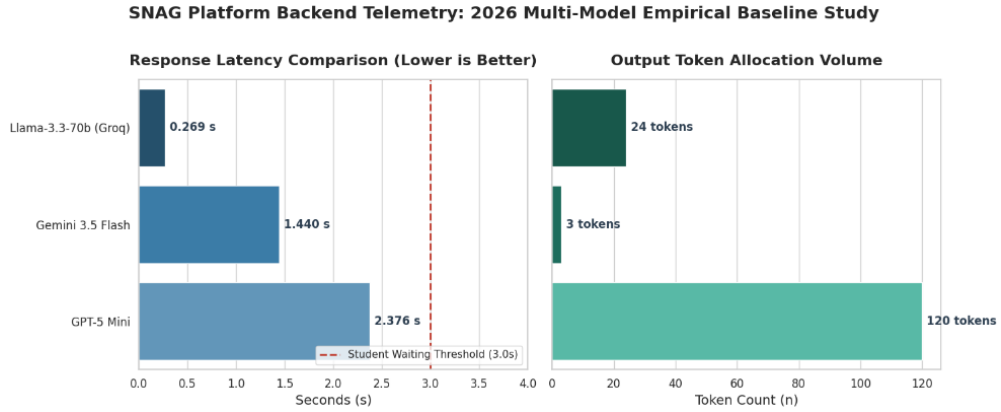

Figure 2. Telemetry bar chart: mean latency and student waiting threshold compliance margin by model configuration

As illustrated in Figure 2, all evaluated models successfully remained below the established Student Waiting Threshold. However, substantial differences were observed in the magnitude of their compliance margins. Among the tested configurations, Llama-3.3-70B deployed through the Groq infrastructure exhibited the most favorable latency profile, achieving a mean response time of only 0.2687 seconds and yielding a Latency Delta of −2.7313 seconds. Such near-instantaneous responses indicate that scaffolding cues can be delivered before measurable attentional disruption occurs.

In contrast, GPT-5 Mini demonstrated the highest latency among the evaluated models, averaging 2.3764 seconds and producing a Latency Delta of −0.6236 seconds. Although this value remains within the acceptable range, the comparatively narrow safety margin highlights the importance of incorporating fallback mechanisms to preserve robustness under real-world load fluctuations.

These findings extend previous human computer interaction research by explicitly incorporating inference latency into the design of child-facing AI systems. Recent comparative studies have shown that conventional chatbot architectures frequently experience delays and inconsistent instructional quality [52]. Meanwhile, hardware-oriented optimization approaches demonstrate that specialized inference processors substantially improve response efficiency and scalability [36]. Taken together, these results suggest that hardware acceleration and pedagogical design should be viewed as complementary dimensions in educational AI deployment rather than independent considerations.

3.2.1. Multidisciplinary Interpretation: Latency and the Continuity of Scientific Reasoning

From a science education and cognitive psychology perspective, the exceptional sub-second latency achieved by Meta Llama-3.3-70b via the Groq language processing unit architecture (0.2687s) is not merely a technical performance victory; it is a critical requirement for maintaining the student's active inquiry process. Inquiry-based science learning and scientific reasoning depend heavily on the working memory's capacity to hold complex, multi-tiered problem components, such as fraction sets, partitioning rules, and algorithmic variables, in an active mental workspace [[42]. When a child experiences a system latency spike that approaches or exceeds the 3.0-second student waiting threshold, the fragile cognitive chain of scientific reasoning is disrupted. The delay causes an abrupt shift in attention, leading to cognitive drift and forcing the child to expend additional executive function resources to rebuild their mental schema state once the prompt finally loads [54].

By establishing a Latency Delta of minus 2.7313 seconds, the language processing unit inference engine ensures that the AI's Socratic feedback arrives almost instantaneously at the exact moment of strategic hesitation (detected via delta telemetry). This immediate response ensures the absolute continuity of the student's problem-

solving process, effectively anchoring their cognitive focus within the Zone of Proximal Development without entering a state of learned helplessness [23]. Conversely, while OpenAI GPT-5 Mini delivers superior pedagogical alignment, its 2.3764-second processing delay pushes the system close to the attentional boundary window. This proximity increases the risk of cognitive disengagement if minor network jitter occurs. Therefore, the integration of these technical telemetry profiles with cognitive load metrics confirms that optimizing hardware-level inference throughput is a foundational prerequisite for sustaining productive scientific inquiry and computational thinking development in technology-enhanced primary STEM classrooms [38].

3.3. Scaffolding Compliance Analysis

While latency determines whether assistance can be delivered in a timely manner, response quality determines whether that assistance remains pedagogically beneficial. Therefore, a complementary analysis was conducted to investigate the extent to which each model preserved Socratic tutoring principles without degenerating into direct-answer behaviour [41]. Examination of the complete 135-response corpus revealed three recurring linguistic failure modes.

Direct Answer Disclosure (DAD)

Direct Answer Disclosure occurs when a model explicitly provides the final solution immediately after receiving a query. Such behavior violates the productive struggle principle and interrupts the foundational schema construction processes described by Cognitive Load Theory [43]. By preemptively resolving the cognitive conflict, the system prevents the learner from engaging in the deep analytical thinking required for genuine conceptual mastery.

Implied Derivation (ID)

Implied Derivation refers to situations in which excessive procedural hints make the target solution trivially recoverable. Although the final answer remains hidden, learners may bypass genuine reasoning processes by simply following the exposed algorithmic steps. Consequently, this excessive guidance limits opportunities for durable schema development and reduces the overall efficacy of the scaffolding intervention [43].

Vocabulary Mismatch (VM)

Vocabulary Mismatch arises when generated explanations employ linguistic structures that exceed the developmental reading capabilities of upper primary students. Excessive terminology and syntactic complexity increase extraneous cognitive load and reduce the accessibility of instructional scaffolding [39]. When learners expend cognitive energy decoding complex vocabulary rather than solving the underlying scientific problem, the instructional intervention inherently fails.

Among the evaluated configurations, GPT-5 Mini demonstrated the strongest instructional adherence, achieving the lowest failure rate (6.9%) and producing no instances of Direct Answer Disclosure. This performance can be attributed to the model's robust alignment mechanisms, which consistently preserve Socratic questioning patterns. Gemini 3.5 Flash exhibited the highest failure rate (14.2%). Its highly concise outputs, averaging only three tokens per interaction, frequently generated generic questions lacking sufficient contextual support. Similar observations have been reported in studies examining the effects of abbreviated AI feedback on learner persistence and instructional quality [36]. Llama-3.3-70B achieved a balanced performance profile with a failure rate of 7.3%, a language complexity score of 2.9, and a Socratic question quality score of 4.1. Most violations occurred within higher-order algorithmic thinking tasks, where distinctions between guidance and procedural disclosure become increasingly subtle. Because these errors primarily originated from prompt interpretation rather than architectural limitations, they are considered amenable to further prompt refinement.

The psycholinguistic triage conducted at Universitas Negeri Malang emphasizes that the prevention of these scaffolding failures (DAD, ID, and VM) directly impacts the quality of students' scientific reasoning during inquiry-based learning tasks. When a model configuration successfully avoids the direct-answering trap, it preserves the student's active inquiry phase, forcing them to engage in critical hypothesis testing and problem decomposition. Conversely, when a model fails through direct answer disclosure or implied derivation, it short-circuits the cognitive feedback loop necessary for mathematical and scientific schema synthesis. In the context of technology-enhanced science education, maintaining strict scaffolding compliance ensures that the AI framework operates as a true pedagogical bridge within the Zone of Proximal Development, converting potential frustration into structured, autonomous scientific thinking patterns.

3.4. Engineering Trade-Off and Deployment Blueprint

The benchmark outcomes further suggest that effective child-facing generative tutoring systems should be conceptualized as reliability-oriented educational infrastructures rather than as static single-model agents. Recent surveys in AI in education emphasize that large language models are most beneficial when embedded within adaptive orchestration frameworks capable of balancing responsiveness, personalization, and pedagogical safeguards rather than maximizing isolated benchmark scores alone [52], [55]. Similarly, whole-learner support architectures advocate treating model selection as part of a broader socio-technical ecosystem in which cognitive, motivational, and behavioral dimensions jointly influence learning effectiveness [55]. These perspectives imply

that model heterogeneity should not be interpreted as a system limitation but rather as an engineering mechanism that enables educational agents to satisfy multiple instructional objectives simultaneously. Consequently, the present findings support a transition from monolithic large language model deployment toward dynamically orchestrated tutoring ecosystems that prioritize both developmental appropriateness and operational reliability [17], [35].

From an infrastructure perspective, latency management emerges not merely as a computational optimization problem but as a pedagogically consequential design parameter. Advances in latency-optimized inference architectures demonstrate that response time constitutes a critical determinant of user experience in large language model systems [36]. Moreover, recent human large language model interaction studies indicate that delayed responses negatively influence user trust, perceived intelligence, and interaction quality [43]. Within educational settings, these temporal effects become even more pronounced because scaffolding effectiveness depends on delivering guidance at the moment when cognitive disequilibrium occurs. Educational studies on generative AI-based scaffolding further emphasize the importance of immediate feedback loops in sustaining productive engagement and maintaining learners within their zone of proximal development [18], [41], [56]. Therefore, maintaining a balance between latency performance and instructional fidelity represents a fundamental requirement for the practical deployment of child-centered tutoring architectures.

Beyond identifying performance differences among candidate models, the present benchmarking procedure serves a prescriptive role by transforming empirical measurements into actionable architectural decisions. Recent literature argues that evaluations of educational large language models should extend beyond accuracy-centric comparisons and instead inform the design of adaptive instructional ecosystems capable of supporting diverse learner needs [52], [56]. In this perspective, benchmark telemetry becomes a decision-support mechanism through which latency characteristics, pedagogical reliability, and interaction quality are translated into operational policies. Such an approach aligns with contemporary views of AI-enhanced learning systems as socio-technical infrastructures in which educational effectiveness emerges from the interaction between cognitive theories, behavioral analytics, and engineering constraints rather than from model capability alone [15], [55]. Consequently, the findings of the present study are not interpreted merely as comparative statistics, but as empirical evidence guiding the formulation of a robust deployment strategy for child-centered generative tutoring environments.

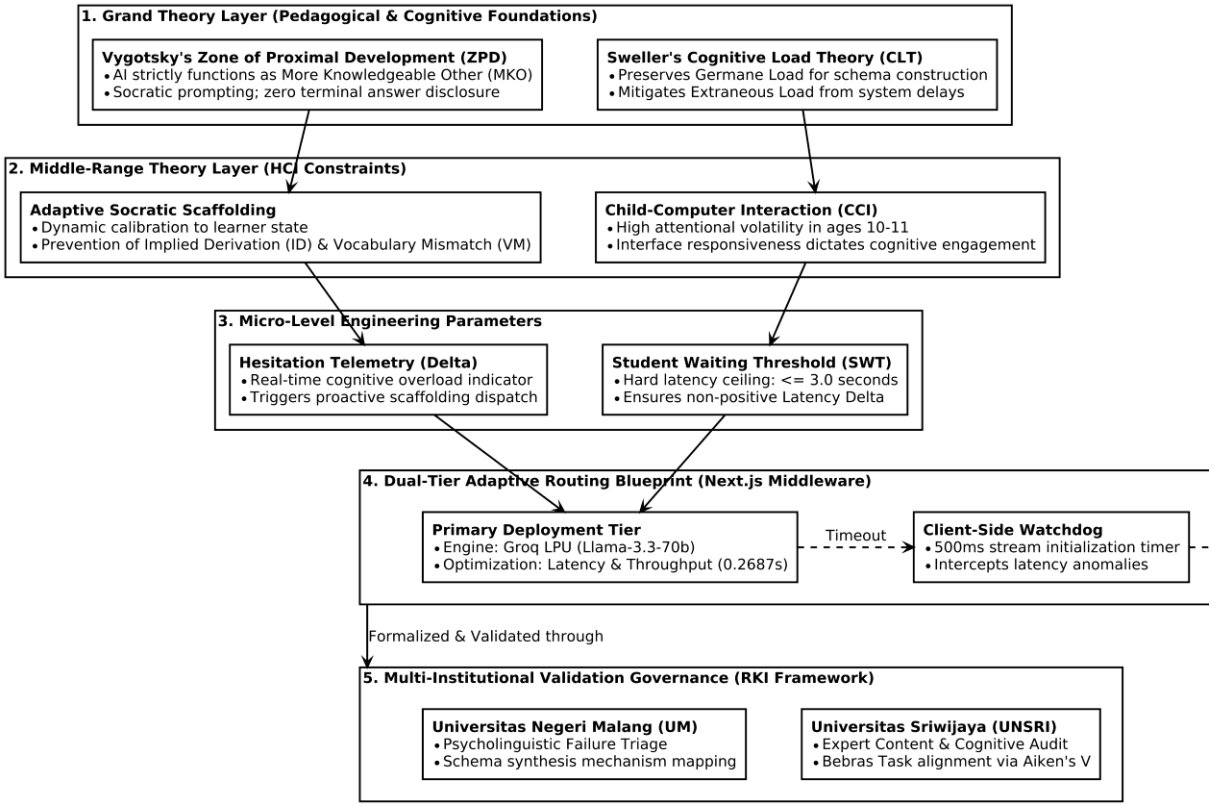

Figure 3. Grand theory and deployment roadmap diagram

As shown in Figure 3, the adaptive workflow begins with continuous monitoring of learner interaction behaviour. Elevated hesitation intervals are interpreted as indicators of cognitive overload or strategic uncertainty and subsequently trigger the scaffolding engine. Under normal conditions, the primary Groq-hosted Llama-3.3-

70B model is activated to maximize response speed. When latency anomalies are detected, the watchdog mechanism automatically redirects the request to GPT-5 Mini, thereby prioritizing instructional fidelity over computational efficiency.The resulting architecture establishes a dynamic coupling between hesitation telemetry and adaptive scaffolding dispatch. By ensuring that guidance is delivered within the three-second Student Waiting Threshold, the system maintains learners within a state of productive cognitive struggle and prevents disengagement or learned helplessness [41]. This real-time integration of behavioral telemetry and multi-model routing distinguishes SNAG from conventional static-hint intelligent tutoring systems and constitutes the principal infrastructure-level contribution validated in the present study.

These findings align with and extend recent literature emphasizing that latency and pedagogical fidelity are intrinsically linked in AI-assisted education [57]. For instance, previous studies on chatbot learning assistants [34] have highlighted that delayed feedback negatively impacts user trust and cognitive engagement. Furthermore, recent research on AI integration supports our observation that hyper-concise AI responses, such as those generated by Gemini 3.5 Flash, fail to provide the necessary conceptual anchoring for younger learners, reinforcing the need for contextually rich scaffolding [13], [58].

The explicit novelty of this research lies in the multidisciplinary formalization and operationalization of the student waiting threshold as a socio-technical metric [28]. Unlike conventional large language model benchmarks that focus strictly on technical accuracy or standalone throughput metrics [27], [28], this work establishes a direct, empirically validated link between hardware-level cloud infrastructure performance and the real-time cognitive sustainability of primary school learners [32]. By intersecting the domains of computer science, learning analytics, and cognitive load theory, this study provides a reproducible architectural blueprint showing how sub-second machine latency directly prevents working memory decay and strategic disengagement in young students during complex computational thinking sequences [30], [35].

The practical and theoretical implications of these findings suggest that educational developers and system architects cannot treat AI modeling and hardware deployment as independent considerations [28]. To build effective child-facing intelligent tutoring ecosystems, platforms must prioritize specialized, latency-optimized inference processors, such as Language Processing Units, alongside highly constrained Socratic prompt structures to guarantee immediate feedback delivery [30], [35]. Furthermore, the successful implementation of the SNAG dual-tier routing framework implies that adaptive educational software must incorporate real-time learning analytics, like strategic hesitation tracking, to dynamically protect the student's independent inquiry phase [36], [39]. Cultivating this equilibrium between technical engineering and pedagogical fidelity is essential for developing trustworthy digital tutoring infrastructure within primary science education [28], [49].

However, a primary limitation of this study is its reliance on controlled laboratory benchmarking simulations that isolate network jitter and programmatic variables [27]. Consequently, the observed telemetry thresholds, algorithmic routing switches, and child scaffolding dynamics must be further validated within active, highly unpredictable classroom environments where varied internet bandwidth and diverse environmental distractions occur [18], [27]. Additionally, the evaluation corpus remains bound to fifth-grade primary school numeracy contexts using Bebras task variants [35], [36]. Future long-term implementations will need to examine whether these specific latency parameters and adaptive Socratic interaction models remain stable across different age demographics and non-mathematical scientific domains [23], [33].

4. CONCLUSION

This study successfully evaluated three large language model configurations within a multidisciplinary framework, revealing that Meta Llama-3.3-70b via the Groq language processing unit provides the optimal infrastructure for primary STEM education by satisfying the 3.0-second Student Waiting Threshold while maintaining high pedagogical fidelity. Conversely, while OpenAI GPT-5 Mini excels in Socratic adherence to serve as a robust fallback model, Google Gemini 3.5 Flash's hyper-concise outputs lead to severe contextual truncation, proving that the effectiveness of educational AI depends on a delicate equilibrium between computational efficiency and cognitive appropriateness. For further research, it is recommended that subsequent studies deploy this dual-tier middleware architecture in longitudinal, real-world classroom settings to analyze live hesitation-time telemetry against student learning outcomes. Additionally, future investigations should focus on developing adaptive, taxonomy-aligned prompt libraries to entirely mitigate vocabulary mismatch anomalies across diverse cognitive tasks.

ACKNOWLEDGEMENTS

This work was supported by the Riset Kolaborasi Indonesia (RKI) 2026 Scheme A program through a multi-institutional collaboration involving Universitas Pendidikan Indonesia, Universitas Sriwijaya, and Universitas Negeri Malang.

AUTHOR CONTRIBUTIONS

Conceptualization, S.W. and M.A.F.B.; Methodology, M.A.F.B., H., and A.P.S.; Software, M.A.F.B.; Validation, H., I.S.R., and A.P.S.; Formal Analysis, M.A.F.B., I.S.R., and A.P.S.; Investigation, M.A.F.B. and A.P.S.; Resources, S.W. and H.; Data Curation, M.A.F.B.; Visualization, M.A.F.B.; Writing – Original Draft Preparation, M.A.F.B.; Writing – Review and Editing, S.W., H., I.S.R., S., Z.M., and A.P.S.; Supervision, S.W.; Project Administration, M.A.F.B. and S.W.; Funding Acquisition, S.W. and H. All authors have read and agreed to the published version of the manuscript. Contributions are reported according to the CRediT taxonomy.

INFORMED CONSENT STATEMENT

This study did not involve human participants, human biological materials, or identifiable personal data. The research consisted exclusively of automated benchmarking experiments using large language model configurations and secondary educational task datasets. Therefore, informed consent was not required.

CONFLICTS OF INTEREST

The authors declare no conflict of interest. The sponsoring institutions and internal DPPM management of the three participating universities have no direct operational intervention in script data compilation, API algorithm parameter determination, statistical inference analysis, or journal submission target selection.

USE OF ARTIFICIAL INTELLIGENCE (AI)-ASSISTED TECHNOLOGY

During the preparation of this work, the authors used Gemini to align raw lab benchmarking datasets into structured Markdown tables and format mathematical equations into clean text arrays, as well as Claude 3.5 Sonnet (Anthropic) for light grammatical proofreading of select paragraphs in the final revision stage. The scientific content, data generation methodology, empirical tables, and all interpretive analyses were conducted entirely by the human authors. AI assistance was not used for literature synthesis, result interpretation, or conclusion formulation. After using these tools, the authors reviewed and edited all content as needed and take full responsibility for the content of the publication.

REFERENCES

[1] S. K. Nordby, L. Mifsud, and A. H. Bjerke, "Computational thinking in primary mathematics classroom activities," *Front. Educ.*, vol. 9, 2024, doi: 10.3389/feduc.2024.1414081.

[2] S. K. Nordby, A. H. Bjerke, and L. Mifsud, "Computational thinking in the primary mathematics classroom: A systematic review," *Digit. Exp. Math. Educ.*, vol. 8, no. 1, pp. 27–49, Apr. 2022, doi: 10.1007/s40751-022-00102-5.

[3] V. Barr and C. Stephenson, "Bringing computational thinking to K–12: What is involved and what is the role of the computer science education community?," *ACM Trans. Comput. Educ.*, vol. 11, no. 1, pp. 1–11, Mar. 2011, doi: 10.1145/1929887.1929905.

[4] K. Bati, "A systematic literature review regarding computational thinking and programming in early childhood education," *Educ. Inf. Technol.*, vol. 27, no. 2, pp. 2059–2082, Mar. 2022, doi: 10.1007/s10639-021-10700-2.

[5] W. Welyta and M. G. Vega, "Discovery learning and scientific literacy: Integrating PISA indicators in high school science," *J. Acad. Biol. Biol. Educ.*, vol. 2, no. 1, pp. 79–87, Jun. 2025, doi: 10.37251/jouabe.v2i1.1941.

[6] J. B. Khamali and H. O. Mondoh, "The effect of flash-based learning media on students' achievement in learning atomic structure in Kenyan senior high schools," *J. Chem. Learn. Innov.*, vol. 2, no. 2, pp. 186–192, Dec. 2025, doi: 10.37251/jocli.v2i2.2584.

[7] J. Su and W. Yang, "A systematic review of integrating computational thinking in early childhood education," *Comput. Educ. Open*, vol. 4, p. 100122, Dec. 2023, doi: 10.1016/j.caeo.2023.100122.

[8] H. Riah, "Augmented reality-based interactive learning media: Enhancing understanding of chemical bonding concepts," *J. Chem. Learn. Innov.*, vol. 2, no. 1, pp. 55–63, Jun. 2025, doi: 10.37251/jocli.v2i1.1919.

[9] R. Rosana, A. Darda, and Z. Zaenudin, "The effect of using offline web-based interactive multimedia on students' biology learning outcomes," *J. Acad. Biol. Biol. Educ.*, vol. 1, no. 2, pp. 177–185, Dec. 2024, doi: 10.37251/jouabe.v1i2.3178.

[10] P. Chen, D. Yang, A. H. S. Metwally, J. Lavonen, and X. Wang, "Fostering computational thinking through unplugged activities: A systematic literature review and meta-analysis," *Int. J. STEM Educ.*, vol. 10, no. 1, 2023, doi: 10.1186/s40594-023-00434-7.

[11] Y. H. Ching and Y. C. Hsu, "Educational robotics for developing computational thinking in young learners: A systematic review," *TechTrends*, vol. 68, no. 3, pp. 423–434, May 2024, doi: 10.1007/s11528-023-00841-1.

[12] S. Tao, "Aligning technology with cognitive development: A five-tiered framework for generative AI in K–12 education," *AI Brain Child*, vol. 1, no. 1, 2025, doi: 10.1007/s44436-025-00024-0.

[13] M. Vendrell and S. K. Johnston, "Scaffolding critical thinking with generative AI: Design principles for integrating large language models in higher education," *Comput. Educ.: Artif. Intell.*, vol. 10, p. 100572, Jun. 2026, doi: 10.1016/j.caeai.2026.100572.

[14] T. K. F. Chiu, "Applying the self-determination theory (SDT) to explain student engagement in online learning during the COVID-19 pandemic," *J. Res. Technol. Educ.*, vol. 54, no. S1, pp. S14–S30, 2022, doi: 10.1080/15391523.2021.1891998.

[15] L. El-Hamamsy, A. Kilic, M. Seiter, F. Heitzmann, U. Bergner, S. M. Volpe, and T. Reuter, "The competent computational thinking test (cCTt): A valid, reliable and gender-fair test for longitudinal CT studies in grades 3–6," *Technol. Knowl. Learn.*, vol. 30, no. 3, pp. 1607–1661, Sep. 2025, doi: 10.1007/s10758-024-09777-8.

[16] H. S. Hsiao, Y. W. Lin, K. Y. Lin, C. Y. Lin, J. H. Chen, and J. C. Chen, "Using robot-based practices to develop an activity that incorporated the 6E model to improve elementary school students' learning performances," *Interact. Learn. Environ.*, vol. 30, no. 1, pp. 85–99, 2022, doi: 10.1080/10494820.2019.1636090.
[17] J. Fagerlund, P. Häkkinen, M. Vesisenaho, and J. Viiri, "Computational thinking in programming with Scratch in primary schools: A systematic review," *Comput. Appl. Eng. Educ.*, vol. 29, no. 1, pp. 12–28, Jan. 2021, doi: 10.1002/cae.22255.
[18] R. Tariq, B. M. Aponte Babines, J. Ramirez, I. Alvarez-Icaza, and F. Naseer, "Computational thinking in STEM education: Current state-of-the-art and future research directions," *Front. Comput. Sci.*, vol. 6, 2024, doi: 10.3389/fcomp.2024.1480404.
[19] E. A. Kyza, Y. Georgiou, A. Agesilaou, and M. Souropetsis, "A cross-sectional study investigating primary school children's coding practices and computational thinking using ScratchJr," *J. Educ. Comput. Res.*, vol. 60, no. 1, pp. 220–257, Mar. 2022, doi: 10.1177/07356331211027387.
[20] X. Liu, X. Wang, K. Xu, and X. Hu, "Effect of reverse engineering pedagogy on primary school students' computational thinking skills in STEM learning activities," *J. Intell.*, vol. 11, no. 2, Art. no. 36, Feb. 2023, doi: 10.3390/jintelligence11020036.
[21] U. H. Rusmin, A. Awaluddin, and M. T. Ajadi, "Development of Nusa Ra Island as a marine tourism object in increasing regional original income (PAD) in South Halmahera Regency," *Multidiscip. J. Tour. Hosp. Sport Phys. Educ.*, vol. 1, no. 2, pp. 50–59, 2024, doi: 10.37251/jthpe.v1i2.1188.
[22] S. Zhang, C. D. Jaldi, N. L. Schroeder, A. A. López, J. R. Gladstone, and S. Heidig, "Pedagogical agent design for K–12 education: A systematic review," *Comput. Educ.*, vol. 223, Dec. 2024, Art. no. 105165, doi: 10.1016/j.compedu.2024.105165.
[23] J. B. Bush, "Software-based intervention with digital manipulatives to support student conceptual understandings of fractions," *Br. J. Educ. Technol.*, vol. 52, no. 6, pp. 2299–2318, Nov. 2021, doi: 10.1111/bjet.13139.
[24] S. Sharman, U. Axunov, and M. A. Balushi, "A study of reflective practice within a multidisciplinary team in an elite football academy," *Multidiscip. J. Tour. Hosp. Sport Phys. Educ.*, vol. 2, no. 1, pp. 41–49, 2025, doi: 10.37251/jthpe.v2i1.1856.
[25] S. C. Shih, C. C. Chang, B. C. Kuo, and Y. H. Huang, "Mathematics intelligent tutoring system for learning multiplication and division of fractions based on diagnostic teaching," *Educ. Inf. Technol.*, vol. 28, no. 7, pp. 9189–9210, Jul. 2023, doi: 10.1007/s10639-022-11553-z.
[26] R. H. Huang, D. J. Liu, A. Tlili, Y. Yang, and J. F. Wang, *Handbook on Facilitating Flexible Learning During Educational Disruption: The Chinese Experience in Maintaining Undisrupted Learning in COVID-19 Outbreak*. Beijing, China: Smart Learning Institute of Beijing Normal University, Mar. 2020. [Online]. Available: http://creativecommons.org/licenses/by-sa/3.0/igo/
[27] D. Najmudin, L. Susanti, and I. Pebrian, "Digital transformation in education: Challenges and opportunities in the age of AI," *Pedagogia*, vol. 23, no. 1, Jun. 2025, doi: 10.17509/pdgia.v23i1.78405.
[28] Y. Copur-Gencturk, J. Li, and S. Atabas, "Improving teaching at scale: Can AI be incorporated into professional development to create interactive, personalized learning for teachers?," *Am. Educ. Res. J.*, vol. 61, no. 4, pp. 767–802, Aug. 2024, doi: 10.3102/00028312241248514.
[29] T. Son, "Intelligent tutoring systems in mathematics education: A systematic literature review using the substitution, augmentation, modification, redefinition model," *Computers*, vol. 13, no. 10, Oct. 2024, Art. no. 270, doi: 10.3390/computers13100270.
[30] Y. F. Lee, G. J. Hwang, and P. Y. Chen, "Technology-based interactive guidance to promote learning performance and self-regulation: A chatbot-assisted self-regulated learning approach," *Educ. Technol. Res. Dev.*, vol. 73, no. 4, pp. 2279–2304, Aug. 2025, doi: 10.1007/s11423-025-10478-x.
[31] T. K. F. Chiu, "A classification tool to foster self-regulated learning with generative artificial intelligence by applying self-determination theory: A case of ChatGPT," *Educ. Technol. Res. Dev.*, vol. 72, no. 4, pp. 2401–2416, Aug. 2024, doi: 10.1007/s11423-024-10366-w.
[32] C. C. Chien, H. Y. Chan, and H. T. Hou, "Learning by playing with generative AI: Design and evaluation of a role-playing educational game with generative AI as scaffolding for instant feedback interaction," *J. Res. Technol. Educ.*, vol. 57, no. 4, pp. 894–913, 2025, doi: 10.1080/15391523.2024.2338085.
[33] F. Wang, X. Zhou, K. Li, A. C. K. Cheung, and M. Tian, "The effects of artificial intelligence-based interactive scaffolding on secondary students' speaking performance, goal setting, self-evaluation, and motivation in informal digital learning of English," *Interact. Learn. Environ.*, vol. 33, no. 7, pp. 4633–4652, 2025, doi: 10.1080/10494820.2025.2470319.
[34] P. Smutny and P. Schreiberova, "From rules to language models: A comparative study of chatbot learning assistants," *Front. Educ.*, vol. 11, May 2026, doi: 10.3389/feduc.2026.1794807.
[35] S. Wang, et al., "Large language models for education: A survey and outlook," *IEEE Signal Process. Mag.*, vol. 42, no. 6, pp. 51–63, 2025, doi: 10.1109/MSP.2025.3594309.
[36] A. Mannekote, et al., "Large language models for whole-learner support: Opportunities and challenges," *Front. Artif. Intell.*, vol. 7, 2024, doi: 10.3389/frai.2024.1460364.
[37] A. Létourneau, M. Deslandes Martineau, P. Charland, J. A. Karran, J. Boasen, and P. M. Léger, "A systematic review of AI-driven intelligent tutoring systems (ITS) in K–12 education," *NPJ Sci. Learn.*, vol. 10, no. 1, 2025, doi: 10.1038/s41539-025-00320-7.
[38] S. Moon, et al., "A latency processing unit: A latency-optimized and highly scalable processor for large language model inference," *IEEE Micro*, vol. 44, no. 6, pp. 17–33, 2024, doi: 10.1109/MM.2024.3420728.
[39] S. Jiang and G. K. W. Wong, "Exploring age and gender differences of computational thinkers in primary school: A developmental perspective," *J. Comput. Assist. Learn.*, vol. 38, no. 1, pp. 60–75, Feb. 2022, doi: 10.1111/jcal.12591.

[40] F. Fang, Y. Tan, M. A. Messerschmidt, W. Yin, O. Nov, and A. Messerschmidt, "The impact of response latency and task type on human–LLM interaction and perception," in *Proc. CHI Conf. Hum. Factors Comput. Syst. (CHI '26)*, Barcelona, Spain, Apr. 2026, Art. no. 17, doi: 10.1145/3772318.

[41] T. K. F. Chiu, "Student engagement in K–12 online learning amid COVID-19: A qualitative approach from a self-determination theory perspective," *Interact. Learn. Environ.*, vol. 31, no. 6, pp. 3326–3339, 2023, doi: 10.1080/10494820.2021.1926289.

[42] E. H. S. Y. Elim, "Promoting cognitive skills in AI-supported learning environments: The integration of Bloom's taxonomy," *Educ. 3–13*, 2024, doi: 10.1080/03004279.2024.2332469.

[43] J. Sweller, "Cognitive load during problem solving: Effects on learning," *Cogn. Sci.*, vol. 12, no. 2, pp. 257–285, Apr. 1988, doi: 10.1207/s15516709cog1202_4.

[44] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 10, no. 3, p. e1355, May 2020, doi: 10.1002/widm.1355.

[45] L. Kohnke, D. Zou, and H. Xie, "Microlearning and generative AI for pre-service teacher education: A qualitative case study," *Educ. Inf. Technol.*, vol. 30, no. 15, pp. 21221–21248, 2025, doi: 10.1007/s10639-025-13606-5.

[46] A. Kumar, "Design and performance optimization of server-side rendering systems in modern web applications," *J. Inf. Syst. Eng. Manag.*, vol. 9, no. 3, Sep. 2024, Art. no. 76, doi: 10.52783/jisem.v9i3.76.

[47] D. Weintrop, et al., "Defining computational thinking for mathematics and science classrooms," *J. Sci. Educ. Technol.*, vol. 25, no. 1, pp. 127–147, Feb. 2016, doi: 10.1007/s10956-015-9581-5.

[48] M. Zapata-Cáceres, P. Marcelino, L. El-Hamamsy, and E. Martín-Barroso, "A Bebras computational thinking (ABC-Thinking) program for primary school: Evaluation using the competent computational thinking test," *Educ. Inf. Technol.*, vol. 29, no. 12, pp. 14969–14998, Aug. 2024, doi: 10.1007/s10639-023-12441-w.

[49] V. Dagienė and G. Futschek, "Bebras international contest on informatics and computer literacy: Criteria for good tasks," in *Informatics Education—Supporting Computational Thinking*, R. T. Mittermeir and M. M. Sysło, Eds., Lecture Notes in Computer Science, vol. 5090. Berlin, Germany: Springer, 2008, pp. 19–30, doi: 10.1007/978-3-540-69924-8_2.

[50] L. El-Hamamsy, M. Zapata-Cáceres, E. Martín-Barroso, F. Mondada, J. D. Zufferey, and B. Bruno, "The competent computational thinking test: Development and validation of an unplugged computational thinking test for upper primary school," *J. Educ. Comput. Res.*, vol. 60, no. 7, pp. 1818–1866, Dec. 2022, doi: 10.1177/07356331221081753.

[51] F. Heintz, L. Mannila, and T. Färnqvist, "A review of models for introducing computational thinking, computer science and computing in K–12 education," in *Proc. IEEE Front. Educ. Conf. (FIE)*, 2016, pp. 1–9, doi: 10.1109/FIE.2016.7757410.

[52] Y. M. Al-Yafaei, N. J. Jomaa, and R. A. Attamimi, Eds., *Human-Centered Approaches to AI-Enhanced English Language Learning and Teaching*. Hershey, PA, USA: IGI Global Scientific Publishing, 2026, doi: 10.4018/979-8-3373-3561-2..

[53] M. Tran, C. Balasooriya, C. Semmler, and J. Rhee, "Generative artificial intelligence: The 'more knowledgeable other' in a social constructivist framework of medical education," *NPJ Digit. Med.*, vol. 8, no. 1, p. 430, 2025, doi: 10.1038/s41746-025-01823-8.

[54] R. Moreno and B. Park, "Cognitive load theory: Historical development and relation to other theories," in *Cognitive Load Theory*, J. L. Plass, R. Moreno, and R. Brünken, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2010, pp. 9–28, doi: 10.1017/CBO9780511844744.003.

[55] L. Cai, M. M. Msafiri, and D. Kangwa, "Exploring the impact of integrating AI tools in higher education using the zone of proximal development," *Educ. Inf. Technol.*, vol. 30, no. 6, pp. 7191–7264, 2025, doi: 10.1007/s10639-024-13112-0.

[56] İ. Çetin, A. K. Erümit, V. Nabiyev, H. Karal, T. Kösa, and M. Kokoç, "The effect of gamified adaptive intelligent tutoring system Artibos on problem-solving skills," *Particip. Educ. Res.*, vol. 10, no. 1, pp. 344–374, 2023, doi: 10.17275/per.23.19.10.1.

[57] H. Bai, W. C. Lui, and P. V. Khiatani, "Promoting student engagement with GPTutor: An intelligent tutoring system powered by generative AI," *Int. J. Educ. Technol. High. Educ.*, vol. 22, no. 1, 2025, doi: 10.1186/s41239-025-00571-9.

[58] I. H. Y. Yim and J. Su, "Artificial intelligence (AI) learning tools in K–12 education: A scoping review," *J. Comput. Educ.*, vol. 12, no. 1, pp. 93–131, Mar. 2025, doi: 10.1007/s40692-023-00304-9.