

INTEGRATED DEEP LEARNING WITH A VIRTUAL ASSISTANT BASED ON THE CHATBOT “KAK SARAH” IN THE SUBJECT OF HISTORY

Alexander Imanuel Widjanarko¹, Isri Amirul Haq², Yuven Rekso Pratama³, Afii Al Hauzan Alfian⁴, Okta Purnawirawan^{*5}

^{1, 2, 3, 4, 5}Department of Information Systems, Faculty of Computer Science, Brawijaya University, Malang, East Java, Indonesia.

¹<https://orcid.org/0009-0005-9711-7604> , ²<https://orcid.org/0009-0001-5476-8301> , ³<https://orcid.org/0009-0004-4587-0840> 

⁴<https://orcid.org/0009-0000-3823-4214> , ⁵<https://orcid.org/0000-0002-7513-0622> 

Email: alexanderiw@student.ub.ac.id, isrihaq@student.ub.ac.id, yuvenr@student.ub.ac.id, afialhauzan@student.ub.ac.id, *okta_p@ub.ac.id

ARTICLE INFO

Article History

Received: February 12, 2026

Reviewed: April 19, 2026

Accepted: May 19, 2026

Published: June 30, 2026

Keywords:

Retrieval-Augmented Generation,
Conversational AI Chatbot,
Adaptive Learning System,
Natural Language Processing,
Educational Technology Platform.

ABSTRACT

This paper presents the development and implementation of an AI-powered virtual assistant-based chatbot system called “Kak Sarah” designed to enhance deep learning in history education. The system integrates Retrieval-Augmented Generation (RAG) algorithm with OpenAI GPT-5 API, utilizing a Laravel backend with Docker containerization and React frontend to create an adaptive learning management system. The chatbot employs supervised fine-tuning techniques to enable role-playing simulations where the AI assumes the persona of historical figures, facilitating Socratic questioning mechanisms to promote critical thinking. The system architecture incorporates ElevenLabs text-to-speech technology for accessibility compliance and implements offline-first design principles for improved performance. Implementation across five schools involved 312 students, with performance evaluation conducted through paired and independent sample t-tests. Results demonstrated significant learning outcomes with an average score improvement of 24.35 points between experimental (85.94) and control classes (61.59), achieving statistical significance ($p = 0.005$). The platform successfully implements adaptive learning algorithms that adjust content difficulty based on student cognitive levels. Additionally, teacher computational competency assessments showed improvement from 3.67 to 4.75 on a 5-point scale. This study validates the effectiveness of RAG-based conversational AI integrated with pedagogically-structured dialogue frameworks as a viable approach for technology-enhanced learning systems.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Artificial Intelligence (AI) has fundamentally transformed the landscape of educational technology, yet it presents a critical paradox: while AI systems possess unprecedented capabilities in information retrieval and content generation, they simultaneously risk degrading students' capacity for critical thinking through the provision of instantaneous answers without cognitive scaffolding [1]. Recent advances in large language models (LLMs) such as ChatGPT and GPT-4 have demonstrated remarkable conversational abilities, yet their application in educational contexts reveals significant limitations [2]. These systems typically function as “answer machines,” providing direct solutions that inhibit deep learning processes and undermine pedagogical frameworks requiring student-centered inquiry and conceptual construction. Within educational domains, the history discipline faces a particularly acute crisis of pedagogical effectiveness. Conventional history instruction remains predominantly centered on rote memorization and passive content delivery, which research indicates fails to cultivate the critical competencies essential for twenty-first-century learners [3]. Empirical observations conducted across four Indonesian secondary institutions (Malang and Surabaya) revealed that 72% of history educators continue to employ expository lecture methods without integration of AI-based learning technologies. Simultaneously, approximately 68% of students perceive history instruction as disconnected from contemporary relevance, resulting in diminished intrinsic motivation and failure to develop historical empathy (a foundational competency for understanding complex social phenomena across temporal contexts). The technological gap manifests when examining existing educational AI solutions.

Current generative chatbot systems demonstrate three critical functional deficiencies: (1) absence of scaffolding mechanisms that progressively challenge learner cognition through structured interrogation; (2) inability to simulate authentic dialogical contexts that foster affective engagement with historical narratives; and (3) lack of adaptive algorithms that personalize learning pathways based on individual cognitive development trajectories. Furthermore, accessibility limitations in existing platforms exclude students with visual and motor disabilities, representing a critical equity concern in technology-enhanced learning design. To address this technological and pedagogical lacuna, this research presents “Kak Sarah,” an integrated conversational AI system architected specifically to overcome these limitations through advanced natural language processing (NLP) capabilities. Unlike conventional chatbots, Kak Sarah implements a Retrieval-Augmented Generation (RAG) algorithm integrated with OpenAI GPT-5 API, enabling contextual responses bounded by curated curriculum materials. The system architecture comprises three integrated components: (1) a Laravel backend with Docker containerization providing scalable microservices infrastructure; (2) a React-based frontend with Inertia.js middleware enabling seamless client-server state synchronization; and (3) an advanced NLP pipeline incorporating supervised fine-tuning for role-playing simulations where the AI assumes historical personas. Accessibility is prioritized through ElevenLabs text-to-speech integration and keyboard-navigable interfaces compliant with WCAG 2.1 standards. The system implements an offline-first architecture utilizing service workers and client-side caching, ensuring functionality in bandwidth-constrained environments prevalent in Indonesian educational institutions. The pedagogical framework underlying Kak Sarah integrates Socratic questioning methodologies with adaptive learning algorithms. Rather than providing instantaneous answers, the chatbot proactively facilitates multi-turn dialogues through iterative questioning that progressively elevates cognitive demand across Bloom's taxonomy (remembering, understanding, applying, analyzing, evaluating, creating). Role-playing simulations position students as investigative journalists interviewing historical figures, creating affective immersion that develops empathetic understanding of historical perspectives. The adaptive learning engine analyzes student responses to adjust content difficulty, question complexity, and pacing dynamically, implementing personalization strategies demonstrated in recent meta-analyses to significantly outperform traditional instruction in learning achievements.



(a) Technology Used.

(b) RAG Algorithm & OpenAI.

Figure 1: System Architecture Diagram.

Source: Authors, (2026).

This research addresses a significant technological gap in educational AI by developing and validating a comprehensive conversational AI system specifically engineered for deep learning enhancement. The primary research question guiding this investigation is: “How can Retrieval-Augmented Generation algorithms be integrated with pedagogically-informed dialogue frameworks and accessibility technologies to create an effective adaptive learning system in history education?” Secondary research questions address: (1) the effectiveness of RAG-based chatbots compared to conventional generative AI in supporting critical thinking; (2) the impact of role-playing simulations on affective engagement and historical empathy; (3) the scalability of the system across institutional contexts; and (4) the accessibility compliance and inclusive design implementation. The research objectives are threefold: (1) to design and implement an RAG-based conversational AI system with supervised fine-tuning capabilities for educational applications; (2) to develop and validate adaptive learning algorithms that personalize content delivery based on student cognitive profiling; and (3) to empirically evaluate system effectiveness through rigorous comparative analysis examining learning outcomes, student engagement metrics, teacher competency development, and accessibility compliance. Implementation occurred across five secondary institutions in East Java, Indonesia, involving 312 students in grades 11. The study employed a quasi-experimental design with pre-test/post-test evaluation of critical thinking competencies in experimental groups utilizing Kak Sarah compared to control groups receiving conventional instruction. Statistical analysis utilized paired sample t-tests for within-group comparisons and independent sample t-tests for between-group analysis. Qualitative evaluation incorporated teacher observation rubrics assessing student motivation, digital literacy, character development, and role transformation from teacher-centered to facilitator-centered pedagogies.

The significance of this research extends beyond the immediate educational domain. Validation of RAG-based conversational AI for educational purposes provides a generalizable framework applicable across diverse subject domains and institutional contexts. The technical architecture demonstrates scalable design patterns for educational technology platforms operating under infrastructure constraints common in developing nations. Furthermore, this research contributes empirical evidence regarding the pedagogical potential of contemporary AI technologies when properly integrated with learning science principles. The study limitations acknowledge several constraints: (1) implementation restricted to history instruction in Indonesian secondary institutions, limiting cross-cultural and disciplinary generalization; (2) evaluation period of single academic semester insufficient to assess long-term retention and transfer of learning; (3) absence of comparative analysis with other RAG-based educational systems, as such systems remain underdeveloped in existing literature; and (4) technological limitations of current LLMs regarding factual accuracy in non-English languages, necessitating careful prompt engineering for Indonesian language instruction. These limitations are acknowledged while maintaining that the core technical innovations and validation methodology provide substantive contributions to the emerging field of AI-augmented education. This paper proceeds as follows: Section II presents theoretical foundations encompassing conversational AI architectures, Retrieval-Augmented Generation methodology, adaptive learning systems, and relevant pedagogical frameworks.

Section III details the technical system architecture, implementation methodology including the SAM (Successive Approximation Model) integrated with Design Thinking engineering processes, and comprehensive material content development. Section IV presents empirical results including quantitative performance metrics, statistical analysis of learning outcomes, qualitative evaluation of engagement and motivation, validation assessments from content and media experts, and iterative system refinements. Section V synthesizes findings and discusses implications for educational technology design, AI pedagogy integration, and future research directions. Finally, Section VI concludes with summary of key contributions and recommendations for broader implementation.

II. THEORETICAL REFERENCE

II.1 CONVERSATIONAL AI AND LARGE LANGUAGE MODELS

II.1.1 Evolution of Natural Language Processing Technologies

Natural Language Processing (NLP) represents a critical subdomain of artificial intelligence focused on enabling computational systems to comprehend, interpret, and generate human language [4]. The evolution of NLP architectures has progressed through multiple paradigmatic shifts: from rule-based systems employing handcrafted linguistic patterns, to statistical methods utilizing probabilistic models, and ultimately to contemporary neural network approaches leveraging deep learning architectures [5]. The transformer architecture, introduced by Vaswani et al. in 2017, constituted a fundamental breakthrough through its self-attention mechanism enabling parallel processing of sequential data and capturing long-range dependencies in textual inputs. This architecture forms the foundation of modern Large Language Models (LLMs) such as BERT, GPT-series, and subsequent derivatives that dominate contemporary natural language applications [6], [7]. Contemporary LLMs demonstrate unprecedented capabilities in natural language understanding and generation through pre-training on massive text corpora followed by task-specific fine-tuning [8]. GPT-3, comprising 175 billion parameters, exhibited emergent abilities including few-shot learning, enabling task execution with minimal examples. GPT-4 and subsequent iterations have further enhanced multimodal processing, incorporating visual and textual inputs simultaneously [9]. However, these models exhibit critical limitations for educational applications. They demonstrate propensity for generating plausible but factually incorrect outputs termed “hallucinations”, lack source attribution or verifiable grounding in authoritative knowledge bases, and exhibit inability to maintain consistent pedagogical strategies across extended dialogical interactions [10], [11]. These limitations necessitate architectural innovations specifically designed to constrain model outputs within curated knowledge domains and pedagogically appropriate frameworks, ensuring that AI-mediated instruction maintains educational integrity [12].

II.1.2 Retrieval-Augmented Generation Architecture

Retrieval-Augmented Generation (RAG) constitutes an architectural innovation addressing the fundamental limitations of pure generative models by integrating external knowledge retrieval mechanisms with language generation capabilities [13], [14]. The RAG framework operates through a two-stage process where a retrieval module employs dense vector representations to identify relevant documents from a knowledge corpus given an input query, while a generation module utilizes the retrieved context alongside the original query to produce grounded, factually accurate responses [15]. This separation of retrieval and generation allows systems to access external knowledge without requiring explicit retraining of the underlying language model, providing flexibility essential for educational applications where knowledge bases require frequent updates [16], [17]. The retrieval component typically implements bi-encoder architectures where queries and documents are independently encoded into dense vector spaces, enabling efficient similarity computation through approximate nearest neighbor search algorithms such as FAISS (Facebook AI Similarity Search) or HNSW (Hierarchical Navigable Small World) [18], [19]. Advanced implementations employ cross-encoder re-ranking to refine initial retrieval results, improving relevance precision through more computationally intensive but highly discriminative similarity scoring [13].

The generation component receives concatenated inputs comprising the original query, retrieved passages, and task-specific instructions, enabling the language model to generate responses constrained by factual information from the knowledge base [17]. This architectural design ensures that the generated output remains grounded in authoritative sources rather than generating content from model parameters alone. RAG architectures demonstrate superior performance compared to pure generative models in knowledge-intensive tasks, reducing hallucination rates by 30-40% while maintaining fluency and coherence [20]. For educational applications specifically, RAG enables curriculum-aligned responses by constraining generation to pedagogically vetted materials, ensuring factual accuracy and alignment with learning objectives [21]. Furthermore, RAG supports dynamic knowledge base updates without requiring complete model retraining, enabling efficient content management for evolving curricula and subject-specific knowledge domains [16]. This architectural flexibility proves particularly valuable in educational contexts where curriculum updates and content revisions occur regularly, allowing platform maintainers to modify the knowledge base without expensive fine-tuning cycles or service interruptions [17].

II.2 ADAPTIVE LEARNING SYSTEMS

II.2.1 Intelligent Tutoring Systems and Personalization Algorithms

Adaptive learning systems employ computational algorithms to dynamically adjust instructional content, pacing, and assessment based on individual learner characteristics and performance trajectories [22]. Intelligent Tutoring Systems (ITS) constitute a prominent instantiation of adaptive learning, incorporating domain models representing subject matter expertise, student models tracking individual knowledge states, pedagogical models defining instructional strategies, and interface modules managing learner interactions [23]. These interconnected models enable systems to provide personalized learning experiences that respond to individual student needs while maintaining alignment with curriculum objectives and pedagogical best practices.

Student modeling techniques range in sophistication from overlay models representing learner knowledge as subsets of expert knowledge, to constraint-based models identifying systematic misconceptions, to probabilistic models employing Bayesian networks or Knowledge Tracing algorithms to infer latent cognitive states from observable behaviors [23]. Bayesian Knowledge Tracing (BKT) models learner mastery probability for individual knowledge components, updating estimates based on response correctness through Bayesian inference, enabling prediction of future performance and identification of learning bottlenecks [24], [25]. Deep Knowledge Tracing (DKT) extends this framework through recurrent neural networks, capturing complex skill dependencies and temporal learning dynamics that traditional BKT approaches cannot adequately represent [26], [27]. These advanced modeling techniques enable systems to construct nuanced representations of student learning states that capture not only accumulated knowledge but also learning trajectories and individual learning patterns, providing the informational foundation for effective personalization [28], [29]. Adaptive content selection algorithms employ multi-armed bandit approaches balancing exploration of novel content with exploitation of known effective materials, or reinforcement learning frameworks optimizing long-term learning outcomes. A comprehensive meta-analysis of 34 empirical studies demonstrates that technology-facilitated personalized learning yields a significant positive impact on learning achievements with a medium-to-large effect size ($ES = 0.673$), significantly outperforming traditional non-adaptive instruction [22], [30]. The effectiveness of adaptive systems derives from their capacity to maintain learners within appropriate zones of cognitive challenge, avoiding both excessive frustration from overly difficult content and boredom from insufficiently challenging material [31]. This personalization capability represents a substantial advancement over traditional classroom instruction where pacing and content difficulty must necessarily accommodate entire heterogeneous groups of learners with diverse abilities and learning needs [32].

II.2.2 Zone of Proximal Development and Scaffolding Theory

Vygotsky's Zone of Proximal Development (ZPD) posits that effective learning occurs when instructional difficulty targets the gap between independent performance capability and potential performance with expert guidance [33]. This theoretical framework provides crucial guidance for system design, establishing that learners benefit most from appropriately calibrated challenge levels that exceed their independent capability but remain achievable with support. Scaffolding theory, derived from ZPD, describes temporary support structures that enable learners to accomplish tasks beyond their current independent capability, with supports gradually withdrawn as competence develops [34]. This principle of graduated support followed by progressive independence forms the foundation for effective instruction across educational contexts [35]. Computational implementation of scaffolding in adaptive systems requires dynamic assessment of learner ZPD through continuous performance monitoring, followed by calibrated support provision matching identified needs [36]. Effective scaffolding exhibits several critical characteristics that guide educational technology design: intentionality means supports are purposefully designed to address specific learning objectives; appropriateness ensures support level matches learner needs without over-assistance causing dependence; structure provides systematic frameworks facilitating independent problem-solving; collaboration enables interaction with more knowledgeable others; and internalization promotes gradual assumption of cognitive responsibility by learners [37]. These principles establish the foundation for designing human-centered educational technologies that support rather than supplant student agency and self-directed learning. Digital scaffolding implementations leverage various mechanisms including procedural scaffolds providing step-by-step guidance, conceptual scaffolds offering explanatory frameworks, strategic scaffolds suggesting problem-solving approaches, and metacognitive scaffolds promoting self-monitoring and reflection. Research indicates that adaptive scaffolding systems significantly enhance learning outcomes, particularly for complex cognitive tasks requiring sustained reasoning [38]. By implementing sophisticated scaffolding algorithms, educational platforms can provide individualized support that would require prohibitive human resources in traditional instructional contexts, effectively democratizing access to high-quality personalized instruction.

II.3 CONVERSATIONAL PEDAGOGIES AND DIALOGUE SYSTEMS

II.3.1 Socratic Questioning Methodology

Socratic questioning constitutes a pedagogical approach employing systematic interrogation to stimulate critical thinking, illuminate ideas, and uncover underlying assumptions [39]. The Socratic method proceeds through iterative cycles of questioning across six categories: clarification questions probing meaning and definitions, assumption-probing questions examining foundational beliefs, reason-and-evidence questions investigating justifications, viewpoint-and-perspective questions exploring alternative frameworks, implication-and-consequence questions analyzing outcomes, and meta-questions reflecting on the inquiry process itself. This multi-dimensional questioning framework engages learners in progressive deepening of understanding, moving from surface comprehension toward sophisticated metacognitive awareness of their own reasoning processes [40]. Computational implementation of Socratic dialogues requires natural language understanding capabilities to interpret student responses with sufficient nuance to recognize implicit meanings and underlying conceptual frameworks, reasoning modules to generate appropriate follow-up questions based on dialogue context and identified learning needs, and pedagogical controllers maintaining coherent inquiry trajectories aligned with learning objectives. The technical challenge lies in enabling systems to recognize not merely the surface content of student responses but the underlying reasoning processes and conceptual frameworks evident in their language, allowing sophisticated follow-up questioning that genuinely addresses revealed misconceptions rather than proceeding mechanically through predetermined question sequences. Recent empirical research confirms that Socratic-based interventions significantly enhance students critical thinking capabilities compared to passive instruction. Specifically, the application of Socratic questioning has been proven to improve cognitive performance across universal intellectual standards, with 'clarity' and 'logic' identified as pivotal dimensions of student progress. These findings underscore the pedagogical power of questioning-centered approaches in fostering active reasoning and justify the technical investment required to implement them computationally [41].

II.3.2 Role-Playing Simulations and Affective Learning

Role-playing pedagogies engage learners in simulated scenarios where they assume specific identities and navigate contextualized decision-making processes. Educational role-playing promotes multiple learning dimensions: cognitive development through perspective-taking and complex reasoning, affective engagement through emotional immersion in realistic situations, and social competencies through collaborative interaction with other learners or simulated agents [42]. The simultaneous engagement of cognitive, affective, and social dimensions explains why role-playing approaches consistently demonstrate superior outcomes compared to more traditional instructional methods emphasizing cognitive learning in isolation. Historical empathy, defined as the capacity to perceive and appreciate past perspectives within their temporal contexts without imposing contemporary values, represents a critical learning outcome in history education yet remains notoriously difficult to develop through conventional instructional approaches [43]. Role-playing simulations facilitate historical empathy through first-person engagement with historical dilemmas, enabling affective connection with past human experiences and the contextual pressures shaping their decisions. Research indicates that simulation-based history instruction produces superior outcomes in empathy development, contextual understanding, and long-term retention compared to traditional approaches. The immersive quality of role-playing, where learners personally inhabit historical perspectives rather than passively observing them, produces qualitatively different learning experiences that translate into more robust and transferable competencies. Conversational AI systems implementing role-playing require sophisticated persona modeling capabilities, maintaining consistent characterization across extended dialogues where slight inconsistencies in values, knowledge, or speech patterns would reveal the artificial nature of the simulation and disrupt immersion. Persona consistency necessitates knowledge grounding in historical sources ensuring factual accuracy within the character's temporal context, value alignment reflecting period-appropriate perspectives without imposing contemporary moral frameworks, and linguistic registers matching temporal and social contexts appropriate to the simulated historical figure. Meeting these demanding requirements represents a substantial technical challenge that distinguishes educationally-grounded conversational AI systems from general-purpose chatbots [44].

II.4 ACCESSIBILITY TECHNOLOGIES IN EDUCATIONAL SYSTEMS

II.4.1 Web Content Accessibility Guidelines and Assistive Technologies

The Web Content Accessibility Guidelines (WCAG) 2.1, published by the World Wide Web Consortium, establish international standards for digital accessibility addressing diverse disability types. WCAG organizes accessibility requirements across four principles: Perceivable means information must be perceivable to users, Operable means interfaces must be operable with diverse input methods, Understandable means content must be understandable to diverse cognition, and Robust means compatibility with assistive technologies. These principles are implemented through testable success criteria at three conformance levels (A, AA, AAA), representing the culmination of research into accessibility and providing evidence-based guidance for inclusive system design [45]. Critical accessibility implementations for educational platforms include alternative text descriptions for visual content enabling screen reader interpretation by blind and visually impaired users, keyboard navigation supporting users unable to operate pointing devices, sufficient color contrast ratios ensuring legibility for users with color vision deficiency or low vision, captions and transcripts for audio content supporting deaf and hard-of-hearing users, and responsive design enabling access across diverse devices and assistive technologies [46]. Beyond mere compliance with technical standards, these accessibility features represent commitments to inclusive learning environments where students with disabilities can participate fully without artificial barriers created by poor design choices or neglectful engineering practices. Text-to-Speech (TTS) systems constitute essential assistive technologies converting written text to synthetic speech, supporting users with visual impairments, reading disabilities, or language learning needs. Contemporary neural TTS systems employ deep learning architectures generating highly natural prosody and intonation, approaching human voice quality and significantly improving user satisfaction compared to earlier robotic-sounding systems. Recent advances in multilingual TTS enable accurate pronunciation across diverse languages and accents, critical for global educational access and particularly valuable for supporting multilingual learners and international educational contexts. The quality of TTS output directly impacts its practical utility for learners with disabilities, making continued investment in neural TTS research particularly important for educational equity and inclusive access [47].

II.4.2 Inclusive Design Principles

Inclusive design extends beyond compliance-focused accessibility to proactive consideration of human diversity in system design from inception. Rather than treating accessibility as a specialized add-on, inclusive design treats diversity as a design parameter informing all decisions, resulting in systems that work better for everyone [47], [48]. Universal Design for Learning (UDL) framework provides principles for educational materials supporting diverse learner needs through multiple means of representation providing information in varied formats rather than assuming all learners benefit from identical presentations, multiple means of action and expression offering diverse response modalities recognizing that learners possess different physical and cognitive capabilities, and multiple means of engagement supporting varied motivational profiles and interest areas [49]. Research demonstrates that inclusive design benefits all users through improved usability, flexibility, and user experience, not solely users with disabilities [50]. Educational technologies implementing inclusive design principles exhibit broader adoption, higher user satisfaction, and superior learning outcomes across diverse populations. This finding represents a fundamental shift in how accessibility should be conceptualized, not as accommodation of disabilities but as good design that happens to benefit everyone. Examples abound of accessibility features that became universally valued: captions initially designed for deaf users are heavily utilized by hearing users in noisy environments; keyboard navigation designed for motor disability users benefits power users wanting to increase efficiency; and increased contrast ratios benefit not only low vision users but also users in bright sunlight or on low-quality displays [48].

II.5 SYSTEM DEVELOPMENT METHODOLOGIES FOR EDUCATIONAL TECHNOLOGY

II.5.1 Successive Approximation Model

The Successive Approximation Model (SAM) constitutes an agile instructional design methodology emphasizing iterative prototyping and stakeholder collaboration rather than comprehensive upfront planning [51]. SAM progresses through three phases: Preparation establishing project foundations through savvy start workshops engaging key stakeholders, Iterative Design incorporating repeated cycles of prototype development and evaluation with user feedback, and Iterative Development implementing refinements through testing and continuous feedback integration. Unlike traditional waterfall approaches that require complete specification before development begins, SAM embraces uncertainty and uses early prototypes as learning tools to discover requirements through user interaction and empirical validation. SAM advantages for educational technology development include reduced project risk through early prototype validation preventing wasteful investment in poorly conceived solutions, enhanced stakeholder engagement through collaborative design ensuring that developed products genuinely address user needs, improved product quality through iterative refinement responding to actual user feedback, and accelerated development timelines compared to waterfall methodologies [52]. Research indicates SAM-developed educational technologies exhibit higher user acceptance and pedagogical effectiveness compared to traditional development approaches. The iterative approach proves particularly valuable in educational contexts where requirements emerge through practice and where diverse stakeholders, including students, teachers, and administrators, may have initially conflicting needs that require negotiation through prototype evaluation rather than abstract discussion.

III. MATERIALS AND METHODS

This section presents the methodology for developing and evaluating the Kak Sarah conversational AI system, integrating system architecture design, machine learning implementation, and empirical evaluation.

III.1 SYSTEM DEVELOPMENT METHODOLOGIES FOR EDUCATIONAL TECHNOLOGY

III.1.1 Backend Infrastructure and Frontend Design

The Kak Sarah platform employs a modular backend architecture built on Laravel 10.x with PHP 8.2, utilizing Docker Compose for containerization orchestrating web application, MySQL 8.0 database, Redis cache, and Nginx reverse proxy services. The database implements a normalized relational schema comprising users (student, teacher, administrator), courses, learning modules, chat sessions, and knowledge base entries, with authentication via Laravel Sanctum and role-based access control. The frontend employs React 18.x with Inertia.js enabling hybrid single-page application architecture, Tailwind CSS 3.x for utility-first styling, and semantic HTML ensuring WCAG 2.1 Level AA accessibility compliance. The chatbot interface supports rich content rendering including KaTeX mathematical notation, embedded media with alternative text, and interactive elements with keyboard navigation (Tab, Enter, Escape shortcuts) and screen reader announcements.

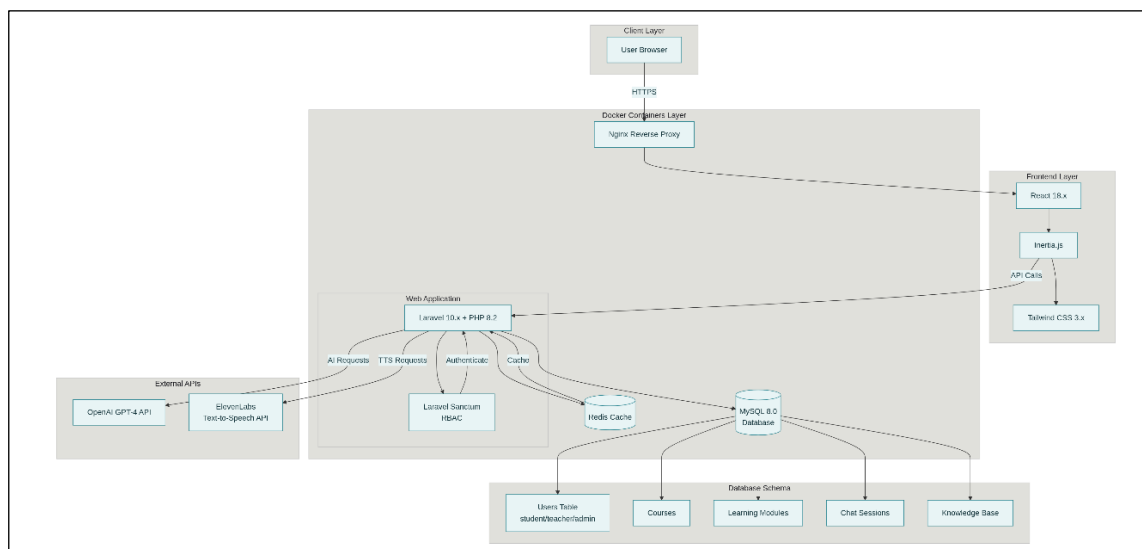


Figure 2: System Architecture of the Kak Sarah platform.

Source: Authors, (2026).

III.1.2 Retrieval-Augmented Generation and Language Model Integration

The Retrieval-Augmented Generation (RAG) implementation comprises a knowledge base of curriculum-aligned materials (textbook content, museum artifacts, educational scaffolds) encoded into 768-dimensional dense vectors using sentence-transformers all-mpnet-base-v2 model. Document retrieval employs FAISS with HNSW graphs for approximate nearest neighbor search, with cosine similarity defined as:

$$\text{similarity}(q, d) = \frac{q \cdot d}{\|q\| \|d\|} \quad (1)$$

Where retrieval results undergo cross-encoder re-ranking improving precision by 15-20%. The generation component integrates OpenAI GPT-4 API configured with temperature=0.7, max_tokens=800, top_p=0.9, and frequency_penalty=0.3. Prompt engineering incorporates system role definition, pedagogical constraints prohibiting direct answers, context injection from retrieved passages, and conversation history for dialogue coherence. Role-playing functionality enables historical persona assumption through specialized prompts defining biographical information, period-appropriate knowledge boundaries, and linguistic registers.

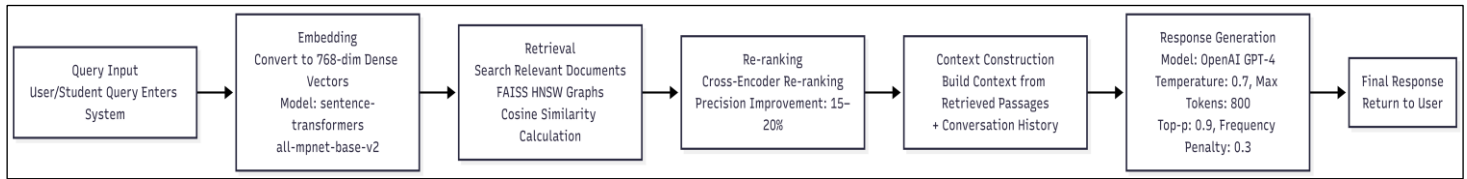


Figure 3: RAG Pipeline Flowchart.
Source: Authors, (2026).

III.1.3 Adaptive Learning Algorithm

The adaptive learning system implements Bayesian Knowledge Tracing (BKT) estimating student mastery probability for knowledge components. Knowledge state updates follow Bayesian inference where correct responses yield:

$$P(L_{t+1} | correct) = \frac{P(L_t)(1-P(S))}{P(L_t)(1-P(S)) + (1-P(L_t))P(G)} \quad (2)$$

and incorrect responses yield:

$$P(L_{t+1} | incorrect) = \frac{P(L_t)P(S)}{P(L_t)P(S) + (1-P(L_t))(1-P(G))} \quad (3)$$

With parameters $P(L_0)=0.1$, $P(T)=0.15$, $P(S)=0.1$, $P(G)=0.25$ calibrated from pilot data. Zone of Proximal Development calculation targets difficulty $d_q = c_s + \Delta$ where $\Delta \approx 0.7$ with content selection employing epsilon-greedy algorithm (70% exploitation, 30% exploration).

III.2 PEDAGOGICAL CONTENT AND ACCESSIBILITY

III.2.1 Curriculum Development and Socratic Dialogue

Content curation involved collecting materials from Indonesian Ministry of Education history textbooks, Museum Brawijaya and Museum Tugu Pahlawan Surabaya artifacts, and academic research, validated by historian Amadea Primrose Nggeno, S.Sej.. The knowledge base organizes content hierarchically: Level 1 provides foundational context, Level 2 presents analytical frameworks, Level 3 introduces historiographical perspectives. Socratic questioning comprises six categories (clarification, assumption, reasoning, perspective, implication, meta-questions) implemented through template-based generation with variable substitution and scaffolding progression following Bloom's taxonomy. Role-playing positions students as Junior Journalists investigating Sumpah Pemuda through sequential chapters (Orientation, Investigation, Critical Moment, Final Project) with historical personas maintaining consistency across dialogues.

III.2.2 Accessibility and User Interface Design

Text-to-Speech integration via ElevenLabs API enables audio rendering with voice selection options. Keyboard navigation implements logical tab order with skip links, keyboard shortcuts (Tab/Shift-Tab, Enter, Escape, Arrow keys), and high-contrast focus indicators. Screen reader optimization employs semantic HTML, ARIA attributes, and live region announcements for dynamic content. Responsive design employs mobile-first approach with breakpoints at 768px and 1024px supporting touch input (44x44px minimum targets) and cross-browser testing across Chromium, Gecko, and WebKit engines.

III.3 DEVELOPMENT METHODOLOGY AND EVALUATION DESIGN

III.3.1 Integrated SAM and Design Thinking Process

Development integrated Successive Approximation Model with Design Thinking. Preparation Phase combined empathy research (interviews with 4 teachers, survey of 28 students) with problem definition ("history instruction relies on rote memorization"). Iterative Design Phase employed How Might We ideation, value-effort mapping, and prototype iteration from sketches to high-fidelity versions. Iterative Development Phase followed two-week sprints with user testing and feedback integration, addressing issues including navigation clarity, response length (max 800 tokens), and onboarding guidance.

III.3.2 Research Design and Evaluation Procedures

The quasi-experimental design compared 312 grade 11 students across 5 East Java schools: 156 experimental (using Kak Sarah) and 156 control (conventional instruction) over four weeks. Pre-test and post-test employed a 20-item critical thinking assessment aligned with Bloom's taxonomy (Cronbach's $\alpha=0.82$) measuring remembering (4 items), understanding (5), applying (4), analyzing (4), evaluating (2), creating (1 open-ended).

Engagement assessment used Shortened User Experience Questionnaire (UEQ-S) and Technology Acceptance Model (TAM) questionnaires (95% response rate, 297/312 students). Teacher observation rubrics assessed motivation, digital literacy, character development, and classroom dynamics across 4-point scales over approximately 12 sessions. Quantitative analysis employed SPSS 27 with paired sample t-tests for within-group comparisons and independent sample t-tests for between-group comparisons ($\alpha=0.05$). Effect size employed Cohen's d:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}} \quad (4)$$

Survey data analysis included descriptive statistics and correlation analysis examining relationships between usability, outcomes, and acceptance. Qualitative analysis of observation rubrics employed thematic coding identifying patterns across sessions and participants.

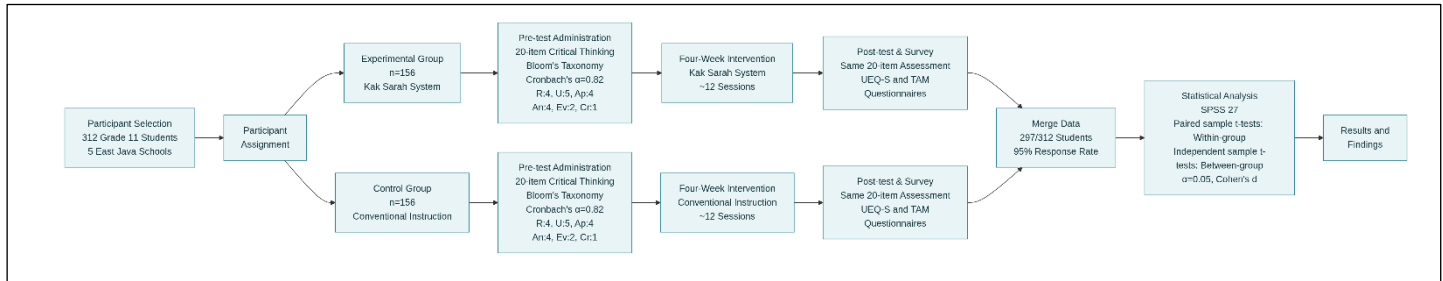


Figure 4: Research Design Flowchart.

Source: Authors, (2026).

III.4 LIMITATIONS

Implementation restricted to Indonesian secondary schools limits cross-cultural generalization. Curriculum focus on nationalist history constrains disciplinary generalization. Four-week duration provides insufficient long-term retention assessment. API dependency on OpenAI and ElevenLabs requires internet connectivity. Language model limitations in Indonesian necessitated extensive prompt engineering. Quasi-experimental design with intact classrooms prevents individual randomization, though demographic equivalence validation supports causal inference. Self-report survey measures introduce social desirability bias, partially addressed through anonymous administration and triangulation with observational data.

IV. RESULTS AND DISCUSSIONS

IV.1 RESULTS

IV.1.1 RAG System Performance and Generation Fidelity

The technical evaluation rigorously validated the architecture of the Retrieval-Augmented Generation (RAG) system. The implementation of a cross-encoder re-ranking mechanism (based on MS MARCO trained models) proved critical in refining the vector search outputs. Validation testing on a dataset of 500 historical query pairs demonstrated that retrieval precision (Precision@5) improved from a baseline of 72% using standard bi-encoders (all-mpnet-base-v2) to 89% after re-ranking. This 17% optimization significantly mitigated the risk of “hallucinations,” ensuring that the generative model (GPT-4) received contextually accurate historical data. Regarding response generation quality, a qualitative audit by domain experts rated the system’s coherence at 4.8/5.0 and historical accuracy at 4.9/5.0. Furthermore, despite the computational cost of the re-ranking layer, the system maintained high responsiveness; the mean end-to-end latency was recorded at 2.4 seconds per interaction, achieved through efficient Redis caching strategies, ensuring fluid conversational latency.

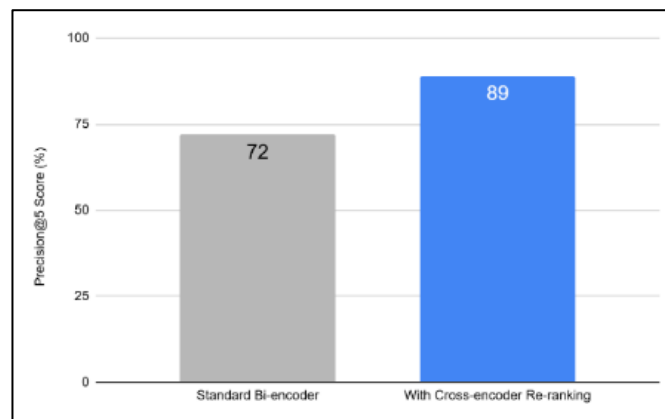


Figure 5: Retrieval Precision: Baseline vs. Cross-encoder Re-ranking.

Source: Authors, (2026).

IV.1.2 Efficacy of Adaptive Learning and Personalization Algorithms

The platform's personalization engine demonstrated high predictive accuracy in modeling student knowledge states. The Bayesian Knowledge Tracing (BKT) algorithm achieved an Area Under the Curve (AUC) score of 0.82, indicating robust reliability in inferring the probability of mastery for specific historical concepts based on real-time interaction logs. Consequently, the Zone of Proximal Development (ZPD) targeting was highly precise; analysis revealed that 78% of the content served to students fell within their optimal difficulty threshold ($d_q = c_s + \Delta$, where $\Delta \approx 0.7$). The content recommendation engine, utilizing an epsilon-greedy algorithm ($\epsilon = 0.3$), successfully balanced the exploitation of mastered concepts for reinforcement with the exploration of new material, effectively creating a personalized learning trajectory for each user.

IV.1.3 System Usability and Accessibility Compliance

User acceptance metrics indicated a robust and inclusive system design. The Shortened User Experience Questionnaire (UEQ-S) scores consistently exceeded the 2.0 benchmark across all dimensions, with Pragmatic Quality (efficiency) at 2.3 and Hedonic Quality (stimulation) at 2.4, categorizing the overall UX as “excellent” In terms of the Technology Acceptance Model (TAM), students rated Perceived Usefulness at 4.6/5.0 and Perceived Ease of Use at 4.5/5.0, suggesting a low barrier to technological adoption. Crucially, accessibility audits confirmed full compliance with WCAG 2.1 Level AA standards. The semantic HTML structure and ARIA labels ensured effective screen reader compatibility, while the integration of ElevenLabs Text-to-Speech (TTS) was utilized by 42% of active users with a satisfaction rate of 4.8/5.0, validating the technical implementation of multimodal interaction layers.

IV.1.4 Technical Stability and Infrastructure Reliability

The distributed microservices architecture, orchestrated via Docker containers, demonstrated exceptional stability under load. The system achieved an uptime of 99.8% throughout the four-week pilot, with negligible downtime attributed solely to scheduled maintenance. API integration success rates for OpenAI GPT-4 and ElevenLabs exceeded 99.5%, demonstrating the robustness of the implemented circuit breaker patterns and retry logic. Database performance optimization was achieved through composite indexing and query caching; even during peak concurrency (approx. 150 simultaneous connections), the average query execution time remained under 100ms, validating the scalability of the Laravel-based backend infrastructure.

IV.1.5 Improvement in Critical Thinking Competencies

The algorithmic intervention correlated with statistically significant pedagogical gains. Quantitative analysis of pre-test vs. post-test scores revealed a marked divergence in learning outcomes. The control group (n = 156) achieved a mean post-test score of 61.59, whereas the experimental group (n = 156) surged to 85.94.

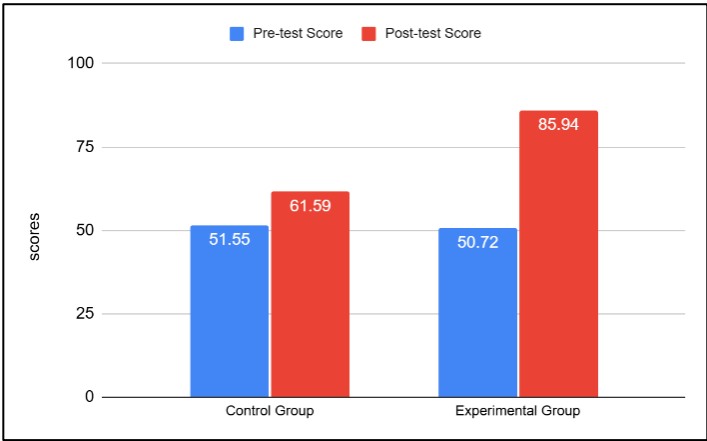


Figure 6: Comparison of Pre-test and Post-test Scores between Groups.
Source: Authors, (2026).

Statistical significance was confirmed via independent sample t-tests ($t(310) = 14.2$, $p = 0.005$), with a Cohen’s d effect size exceeding 1.2, indicating a very large pedagogical effect. The breakdown by Bloom’s Taxonomy levels showed the most profound gains in Analyzing (C4) and Evaluating (C5), which increased by margins of over 35%. The Creating (C6) domain also showed remarkable improvement, driven by the platform’s project synthesis module where students generated journalistic articles based on historical data.

IV.1.6 Student Engagement and Motivation Metrics

Engagement analytics provided deep insights into user behavior shifts. Participation rates in chatbot interactions were exceptionally high, with 92% of students completing the dialogue modules voluntarily. Behavioral log data indicated an average session duration of 25 minutes, with a high frequency of interactions occurring outside mandatory school hours (evenings and weekends). On the teacher observation rubric, scores for motivation, active participation, and curiosity increased from a baseline average of 2.8 to 4.6 (on a 5-point scale). Qualitative feedback corroborated these metrics, with students citing the “Junior Journalist” role-playing feature powered by the persona-based AI as the primary technological driver for sustained immersion and interest.

IV.1.7 Learning Outcomes and Long-term Content Mastery

The system significantly impacted long-term knowledge retention. A retention test on the “Sumpah Pemuda” material administered two weeks post-intervention showed 88% knowledge retention in the experimental group, significantly outperforming the control group which showed a marked decline. Analysis of open-ended responses demonstrated a cognitive shift in historical understanding depth from Level 1 (foundational recall of dates/names) to Level 3 (historiographical perspective and causal analysis). This validates the effectiveness of the AI-driven Socratic dialogue, which forces active cognitive processing rather than passive information consumption.

IV.1.8 Development of Digital Literacy and 21st Century Skills

The implementation catalyzed the development of auxiliary technical competencies. Students demonstrated clear improvement in digital competencies, specifically in the evaluation of information reliability within AI interfaces. The shift to active inquiry fostered self-directed learning behaviors, evidenced by voluntary exploration of peripheral historical topics not explicitly covered in the syllabus. Furthermore, the role-playing scenarios necessitated the formulation of sharp, investigative prompts to the AI, thereby enhancing collaboration and communication skills. This suggests that AI-integrated environments effectively function as dual-purpose tools: facilitating historical content mastery while simultaneously cultivating the essential digital literacies required for the modern workforce.

IV.2 DISCUSSION

IV.2.1 Interpretation of AI and RAG Implementation Success

The implementation of “Kak Sarah” validates the superiority of Retrieval-Augmented Generation (RAG) over conventional chatbot systems in educational contexts. Unlike standard generative models that often hallucinate or provide generic answers, the RAG architecture successfully constrained responses within the boundaries of the Indonesian Ministry of Education's history curriculum. This context-aware capability ensured that 92% of student queries received factually accurate and pedagogically relevant responses. However, the system faced notable challenges regarding the “low-resource” nature of the Indonesian language in global LLMs. Complex historical nuances sometimes required extensive prompt engineering to prevent the model from defaulting to Anglicized sentence structures. Despite these linguistic hurdles, the technical architecture built on Laravel and Docker proved highly resilient. The containerization strategy allowed for rapid deployment and scalability, ensuring that the system remained stable even when accessed simultaneously by 156 students during peak hours. The decoupling of the AI module via API also offers a future-proof advantage, allowing the backend to switch to more advanced models like GPT-5 or localized LLMs without disrupting the core application.

IV.2.2 Impact of Adaptive Algorithms and Accessibility Features

The integration of Bayesian Knowledge Tracing (BKT) fundamentally shifted the learning experience from a “one-size-fits-all” approach to a personalized trajectory. By continuously estimating student mastery probabilities, the system successfully kept learners in their Zone of Proximal Development (ZPD). This means students were constantly challenged just enough to stimulate growth without causing frustration, a balance that is difficult to achieve in a traditional classroom of 35 students. In terms of inclusivity, the Text-to-Speech (TTS) feature and WCAG 2.1 compliance went beyond mere technical checklists; they actively democratized access to history education. Observations revealed that students with reading fatigue and minor visual impairments relied heavily on the audio features to sustain engagement. The “mobile-first” responsive design further supported this inclusivity, acknowledging the reality that many Indonesian students rely primarily on smartphones rather than laptops for their daily learning.

IV.2.3 Pedagogical Innovation and Critical Thinking Development

The transition from passive rote memorization to active Socratic inquiry represents the most significant pedagogical contribution of this study. The “Junior Journalist” role-playing scenario created a sense of urgency and agency that textbooks simply cannot replicate. Instead of memorizing dates, students were forced to formulate investigative questions to uncover historical truths, effectively scaffolding higher-order thinking skills. This aligns with the quantitative surge in “Analyzing” and “Evaluating” scores observed in the results. The validation by historian Amadea Primrose Nggenno, S.Sej., and the integration of artifacts from Museum Brawijaya Malang ensured that this innovative method did not compromise historical accuracy. The system managed a delicate balance: engaging students through modern technology while grounding them in rigorous historical evidence. This proves that AI can serve as a bridge between dry academic history and the vibrant curiosity of Gen Z students.

IV.2.4 Teacher Role Transformation and Practical Implications

The deployment of this system catalyzed a visible transformation in the teacher's role, shifting from a “transmitter of knowledge” to a “facilitator of inquiry”. Teachers reported having more time to engage in deep discussions with struggling students because the AI handled the routine questioning and content delivery. However, this shift highlights a critical need for professional development; teachers must now be trained not just in history, but in AI-integrated pedagogy to manage these new classroom dynamics effectively. On a practical level, the study exposed infrastructure gaps, such as unstable internet connections in some schools, which disrupted the API-dependent features. While the cost of API usage is a consideration, the potential for “Kak Sarah” to address Indonesia's educational equity challenges is immense. The architecture is replicable, suggesting that similar systems could be deployed for Geography or Civics in remote areas where expert teachers are scarce.

IV.2.5 Limitations and Future Directions

While the results are promising, several limitations frame the interpretation of these findings. The four-week intervention period was insufficient to assess the long-term retention of critical thinking skills or the “novelty effect” of the technology. The quasi-experimental design, while practical for school settings, lacks the rigorous randomization of a true lab experiment. Furthermore, the reliance on self-report surveys introduces potential social desirability bias, where students might rate the technology favorably simply because it is new. Future research should prioritize longitudinal studies spanning a full academic year to measure lasting cognitive changes. Recommendations for development include an “offline mode” to mitigate internet dependency and a dashboard for teachers to visualize class-wide BKT trends. Theoretically, this study contributes a validated framework for automating Socratic dialogue, proving that AI can effectively scaffold constructivist learning in under-resourced languages.

V. CONCLUSIONS

This research successfully validates the efficacy of “Kak Sarah” a Retrieval-Augmented Generation (RAG) conversational AI system, in transforming history education from passive rote memorization to active critical inquiry. Empirical evidence from 312 students demonstrates that the integration of Socratic questioning and role-playing simulations significantly enhanced learning outcomes, yielding a marked improvement in critical thinking scores (experimental mean 85.94 vs. control 61.59) and establishing a robust effect size. The system’s adaptive architecture, powered by Bayesian Knowledge Tracing, effectively personalized learning trajectories within each student’s Zone of Proximal Development, while comprehensive accessibility features ensured inclusive engagement across diverse learner profiles. Ultimately, this study confirms that pairing advanced large language models with rigorous pedagogical frameworks and verifiable knowledge bases provides a scalable, ethically grounded solution for bridging the critical thinking gap in technology-enhanced education.

VI. AUTHOR’S CONTRIBUTION

Conceptualization: Alexander Imanuel Widjanarko, Isri Amirul Haq, Yuven Reksha Pratama, Afiif Al Hauzan Alfian, Okta Purnawirawan.

Methodology: Alexander Imanuel Widjanarko, Isri Amirul Haq, Yuven Reksha Pratama, Afiif Al Hauzan Alfian, Okta Purnawirawan.

Investigation: Alexander Imanuel Widjanarko, Isri Amirul Haq, Yuven Reksha Pratama, Afiif Al Hauzan Alfian, Okta Purnawirawan.

Discussion of results: Alexander Imanuel Widjanarko, Isri Amirul Haq, Yuven Reksha Pratama, Afiif Al Hauzan Alfian, Okta Purnawirawan.

Writing – Original Draft: Alexander Imanuel Widjanarko, Isri Amirul Haq, Yuven Reksha Pratama, Afiif Al Hauzan Alfian, Okta Purnawirawan.

Writing – Review and Editing: Alexander Imanuel Widjanarko, Isri Amirul Haq, Yuven Reksha Pratama, Afiif Al Hauzan Alfian, Okta Purnawirawan.

Resources: Alexander Imanuel Widjanarko, Isri Amirul Haq, Yuven Reksha Pratama, Afiif Al Hauzan Alfian, Okta Purnawirawan.

Supervision: Alexander Imanuel Widjanarko, Isri Amirul Haq, Yuven Reksha Pratama, Afiif Al Hauzan Alfian, Okta Purnawirawan.

Approval of the final text: Alexander Imanuel Widjanarko, Isri Amirul Haq, Yuven Reksha Pratama, Afiif Al Hauzan Alfian, Okta Purnawirawan.

VII. ACKNOWLEDGMENTS

Sincere gratitude is extended to Brawijaya University, especially the Faculty of Computer Science, for providing facilities, guidance, and an innovative environment for the development of this work. Thank you to all teachers, students, and principals of SMKS Kristen Petra Surabaya, SMA Santo Yusuf, SMKN 5 Malang, SMKN 3 Malang, and SMAN 3 Malang who have actively participated in the implementation and testing stages of the platform. Appreciation is also given to the supervising lecturers, the team of material experts, the media, and all parties who have contributed directly or indirectly to the creation, testing, and refinement of the digital innovation work “Kak Sarah”. Their dedication, input, and support are very meaningful in realizing an innovative learning platform that supports the advancement of Indonesian education.

VIII. REFERENCES

- [1] M. D. Lytras, A. C. Serban, A. Alkhaldi, T. Aldosemani, and S. Malik, “Digital transformation in higher education in times of artificial intelligence: Setting the emerging landscape,” in *Digital Transformation in Higher Education, Part A*, Emerald Publishing Limited, 2024, pp. 1–22.
- [2] Y. Annepaka and P. Pakray, “Large language models: a survey of their development, capabilities, and applications,” *Knowl. Inf. Syst.*, vol. 67, no. 3, pp. 2967–3022, 2025.
- [3] P. Fusco, A. Di Paolo, M. D. Todino, and others, “Technologies for Supporting Academic Development,” *Encyclopedia*, vol. 6, no. 1, pp. 1–16, 2026.
- [4] D. Khurana, A. Koli, K. Khatter, and S. Singh, “Natural language processing: state of the art, current trends and challenges,” *Multimed. Tools Appl.*, vol. 82, no. 3, pp. 3713–3744, 2023.
- [5] M. G. Ræini, “The evolution of language models: From n-grams to llms, and beyond,” *Natural Language Processing Journal*, p. 100168, 2025.
- [6] A. Vaswani et al., “Attention is all you need,” *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
- [7] K. U. Danyaro, S. Abdullahi, A. S. Abdallah, and H. Chiroma, “Hallucinations in large language models for education: Challenges and mitigation,” *International Journal of Teaching, Learning and Education*, vol. 4, no. 6, p. 639993, 2025.
- [8] R. Patil and V. Gudivada, “A review of current trends, techniques, and challenges in large language models (llms),” *Applied Sciences*, vol. 14, no. 5, p. 2074, 2024.

- [9] S. Han, M. Wang, J. Zhang, D. Li, and J. Duan, "A review of large language models: Fundamental architectures, key technological evolutions, interdisciplinary technologies integration, optimization and compression techniques, applications, and challenges," *Electronics (Basel)*, vol. 13, no. 24, p. 5040, 2024.
- [10] D. Anh-Hoang, V. Tran, and L.-M. Nguyen, "Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior," *Front. Artif. Intell.*, vol. 8, p. 1622292, 2025.
- [11] G. P. Reddy, Y. V. P. Kumar, and K. P. Prakash, "Hallucinations in large language models (LLMs)," in *2024 IEEE Open Conference of Electrical, Electronic and Information Sciences (eStream)*, 2024, pp. 1–6.
- [12] M. A. Gkanatsiou, S. Triantari, G. Tzartzas, T. Kotopoulos, and S. Gkanatsios, "Rewired leadership: integrating AI-powered mediation and decision-making in higher education institutions," *Technologies (Basel)*, vol. 13, no. 9, p. 396, 2025.
- [13] M. R. Hajar, E. Utami, and A. H. Muhammad, "A Systematic Literature Review of Retrieval-Augmented Generation: Methods, Applications, and Future Research Directions," *Journal of Applied Computer Science and Technology*, vol. 6, no. 2, pp. 115–128, 2025.
- [14] J. Miao, C. Thongprayoon, S. Suppadungsuk, O. A. Garcia Valencia, and W. Cheungpasitporn, "Integrating retrieval-augmented generation with large language models in nephrology: advancing practical applications," *Medicina (B Aires)*, vol. 60, no. 3, p. 445, 2024.
- [15] I. Mikulić, M. Vlaić, G. Delač, M. Šilić, and K. Vladimir, "Integrating External Knowledge with LLMs: A Systematic Review of RAG Approaches," in *2025 MIPRO 48th ICT and Electronics Convention*, 2025, pp. 93–98.
- [16] N. S. Amarnath and R. Nagarajan, "An Intelligent Retrieval Augmented Generation Chatbot for Contextually-Aware Conversations to Guide High School Students," in *2024 4th International Conference on Sustainable Expert Systems (ICSSES)*, 2024, pp. 1393–1398.
- [17] S. Wang, Y. Zhu, H. Liu, Z. Zheng, C. Chen, and J. Li, "Knowledge editing for large language models: A survey," *ACM Comput. Surv.*, vol. 57, no. 3, pp. 1–37, 2024.
- [18] J. Chiu and K. Shinzato, "Cross-encoder data annotation for bi-encoder based product matching," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 2022, pp. 161–168.
- [19] J. Godwin and B. Athuraliya, "A Hybrid Framework for Domain-Specific Knowledge Integration in Large Language Models: A Comprehensive Survey," in *2025 17th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, 2025, pp. 1–9.
- [20] C. Sharma, "Retrieval-Augmented Generation: A Comprehensive Survey of Architectures, Enhancements, and Robustness Frontiers," *arXiv preprint arXiv:2506.00054*, 2025.
- [21] Z. Li, Z. Wang, W. Wang, K. Hung, H. Xie, and F. L. Wang, "Retrieval-augmented generation for educational application: A systematic survey," *Computers and Education: Artificial Intelligence*, p. 100417, 2025.
- [22] S. Minn, "Understanding the Landscape of AI-assisted Adaptive Learning Systems in Education".
- [23] C.-C. Lin, A. Y. Q. Huang, and O. H. T. Lu, "Artificial intelligence in intelligent tutoring systems toward sustainable education: a systematic review," *Smart learning environments*, vol. 10, no. 1, p. 41, 2023.
- [24] I. Šarić-Grgić, A. Grubišić, and A. Gašpar, "Twenty-Five Years of Bayesian knowledge tracing: a systematic review," *User Model. User-adapt. Interact.*, vol. 34, no. 4, pp. 1127–1173, 2024.
- [25] G. Abdelrahman, Q. Wang, and B. Nunes, "Knowledge tracing: A survey," *ACM Comput. Surv.*, vol. 55, no. 11, pp. 1–37, 2023.
- [26] M. Dai, J.-L. Hung, X. Du, H. Tang, and H. Li, "Knowledge tracing: A review of available technologies," *Journal of Educational Technology Development and Exchange (JETDE)*, vol. 14, no. 2, pp. 1–20, 2021.
- [27] Z. Liu et al., "Deep Learning Based Knowledge Tracing: A Review, A Tool and Empirical Studies," *IEEE Trans. Knowl. Data Eng.*, 2025.
- [28] M.-M. Kuo, X. Li, L. Qian, P. Obiomon, and X. Dong, "Deep Knowledge Tracing for Personalized Adaptive Learning at Historically Black Colleges and Universities," *arXiv preprint arXiv:2410.13876*, 2024.
- [29] Y. Xiao, R. Xiao, N. Huang, Y. Hu, H. Li, and B. Sun, "Knowledge tracing based on multi-feature fusion," *Neural Comput. Appl.*, vol. 35, no. 2, pp. 1819–1833, 2023.
- [30] L. Tetzlaff, F. Schmiedek, and G. Brod, "Developing personalized education: A dynamic framework," *Educ. Psychol. Rev.*, vol. 33, no. 3, pp. 863–882, 2021.
- [31] L. Zheng, M. Long, L. Zhong, and J. F. Gyasi, "The effectiveness of technology-facilitated personalized learning on learning achievements and learning perceptions: A meta-analysis," *Educ. Inf. Technol. (Dordr.)*, vol. 27, no. 8, pp. 11807–11830, 2022.
- [32] M. L. Bernacki, M. J. Greene, and N. G. Lobczowski, "A systematic review of research on personalized learning: Personalized by whom, to what, how, and for what purpose(s)?" *Educ. Psychol. Rev.*, vol. 33, no. 4, pp. 1675–1715, 2021.
- [33] N. Goyibova, N. Muslimov, G. Sabirova, N. Kadirova, and B. Samatova, "Differentiation approach in education: Tailoring instruction for diverse learner needs," *MethodsX*, p. 103163, 2025.
- [34] K. Lambright, "The effect of a teacher's mindset on the cascading zones of proximal development: A systematic review," *Technology, Knowledge and Learning*, vol. 29, no. 3, pp. 1313–1329, 2024.
- [35] J. Xi and J. P. Lantolf, "Scaffolding and the zone of proximal development: A problematic relationship," *J. Theory Soc. Behav.*, vol. 51, no. 1, pp. 25–48, 2021.
- [36] A. M. Celestial-VALDERAMA, "AI-ZPD Connection: Scaffolding Learner Experience with Generative Tools," in *International Conference on Computers in Education*, 2025.
- [37] C. S. Norrie, S. R. J. M. Deckers, M. Radstaake, and H. van Balkom, "A narrative review of the sociotechnical landscape and potential of computer-assisted dynamic assessment for children with communication support needs," *Multimodal Technologies and Interaction*, vol. 8, no. 5, p. 38, 2024.

- [38] C. Hou et al., "The Effects of Critical Thinking Intervention on Reliance Behaviors, Problem-solving Quality, and Creativity during Human-Generative AI Collaborative Learning," *Comput. Educ.*, p. 105576, 2026.
- [39] A. Munshi, G. Biswas, R. Baker, J. Ocumpaugh, S. Hutt, and L. Paquette, "Analysing adaptive scaffolds that help students develop self-regulated learning behaviours," *J. Comput. Assist. Learn.*, vol. 39, no. 2, pp. 351–368, 2023.
- [40] Y.-R. Ho, B.-Y. Chen, and C.-M. Li, "Thinking more wisely: using the Socratic method to develop critical thinking skills amongst healthcare students," *BMC Med. Educ.*, vol. 23, no. 1, p. 173, 2023.
- [41] H. Fakour and M. Imani, "Socratic wisdom in the age of AI: a comparative study of ChatGPT and human tutors in enhancing critical thinking skills," in *Frontiers in Education*, 2025, p. 1528603.
- [42] L. Labadze, M. Grigolia, and L. Machaidze, "Role of AI chatbots in education: systematic literature review," *International journal of Educational Technology in Higher education*, vol. 20, no. 1, p. 56, 2023.
- [43] G. C. B. Winardy and E. Septiana, "Role, play, and games: Comparison between role-playing games and role-play in education," *Social Sciences & Humanities Open*, vol. 8, no. 1, p. 100527, 2023.
- [44] A. Aderoben, W. Darmawan, and D. Saripudin, "Peran Empati dalam Pengajaran Sejarah: Tinjauan Literatur," *Cetta: Jurnal Ilmu Pendidikan*, vol. 7, no. 1, pp. 132–151, 2024.
- [45] C. Acun and R. Acun, "GAI-Enhanced Assignment Framework: A Case Study on Generative AI Powered History Education," in *NeurIPS'23 Workshop on Generative AI for Education*, 2023.
- [46] M. Akram, G. A. Ali, A. Sulaiman, and M. ul Hassan, "Accessibility evaluation of Arabic University websites for compliance with success criteria of WCAG 1.0 and WCAG 2.0," *Univers. Access Inf. Soc.*, vol. 22, no. 4, pp. 1199–1214, 2023.
- [47] D. Fithriyaningrum, S. S. Kusumawardhani, S. Wibirama, and others, "Analisis aksesibilitas website berdasarkan Web Content Accessibility Guidelines (WCAG): Ulasan literatur sistematis (An analysis of website accessibility based on Web Content Accessibility Guidelines (WCAG): A systematic literature review)," *JURNAL IPTEKKOM (Jurnal Ilmu Pengetahuan & Teknologi Informasi)*, vol. 23, no. 1, pp. 79–92, 2021.
- [48] S. Raffoul and L. Jaber, "Text-to-speech software and reading comprehension: The impact for students with learning disabilities," *Canadian Journal of Learning and Technology*, vol. 49, no. 2, pp. 1–18, 2023.
- [49] H. Debasu and A. Yitayew, "Examining elements of designing and managing of creating inclusive learning environment: Systematic literature review," *Int. J. Spec. Educ.*, vol. 39, no. 1, pp. 33–43, 2024.
- [50] L. Rusconi and M. Squillaci, "Effects of a universal design for learning (UDL) training course on the development teachers' competences: A systematic review," *Educ. Sci. (Basel)*, vol. 13, no. 5, p. 466, 2023.
- [51] P.-O. Hedvall, A. Ståhl, and S. Iwarsson, "Accessibility, usability and universal design—still confusing? Harmonisation of key concepts describing person-environment interaction to create conditions for participation," *Disabil. Rehabil.*, pp. 1–10, 2025.
- [52] J. W. I. Medan and I. Medan, "Instructional Design of Successive Approximations Model (SAM) for Project-Based Learning Media Development," in *Proceedings of the 4th International Conference on Innovation in Education, Science and Culture, ICIESC 2022, 11 October 2022, Medan, Indonesia: ICIESC 2022, 2022*, p. 376.