The Retrieval-Augmented Pedagogical Assistant (RAPA): A Methodology for Enhancing Critical Thinking and Equity in AI-Augmented Education

Shohel Pramanik, Dr. Mohd Heikal Bin Husin
School of Computer Science, Universiti Sains Malaysia, Pulau Pinang, Malaysia

***Abstract*—This study presents the Retrieval-Augmented Pedagogical Assistant (RAPA) methodology, an integrated framework designed to overcome the core limitations of general Large Language Models (LLMs)—specifically factual instability (hallucination) and static knowledge bases—by deploying a specialized, institutional Retrieval-Augmented Generation (RAG) architecture. The methodology addresses three critical challenges to the responsible integration of AI in higher education. Firstly, the framework ensures data sovereignty and sustainable deployment by mandating a comprehensive Total Cost of Ownership (TCO) analysis. This analysis validates the strategic necessity of local RAG hosting and of leveraging computational efficiencies, such as Parameter-Efficient Fine-Tuning (PEFT) and PROXIMITY caching, to ensure a cost-effective solution that strictly complies with FERPA and GDPR data protection mandates and mitigates security risks associated with data leakage. Secondly, the framework ensures the equitable integration of AI literacy across disciplines with varying technological resources, particularly in the Humanities and Vocational Education and Training (VET). This is achieved by minimizing technical prerequisites and institutionalizing continuous Professional Development (PD) through the Dialogic Video Cycle (DVC), which trains faculty in Prompt Engineering to embed individualized pedagogical rules and ethical constraints into the RAPA's architecture. Finally, specific measures are implemented to evaluate the development of Critical Thinking (CT). RAPA outputs are architecturally constrained to include transparent Chain-of-Thought (CoT) reasoning and verifiable source citations. Student Critical AI Analysis Assignments require students to critique the AI's synthesis, identifying inaccuracies, biases, or limitations. The effectiveness of this assessment is quantified using a quasi-experimental design and technical RAGAS metrics, such as Faithfulness and Context Precision, ensuring a verifiable shift from passive knowledge consumption to active, informed critique. Key findings from the preliminary architectural validation indicate that integrating Proximity-LSH caching reduced database retrieval calls by 77.2% and retrieval latency by approximately 72.5%, while maintaining high retrieval recall, addressing the scalability bottleneck inherent in high-volume educational deployments. Furthermore, the application of Robust Fine-Tuning (RbFT) demonstrated a marked improvement in the system's resilience to noisy educational data, preventing performance degradation where standard RAG models typically fail when exposed to irrelevant or counterfactual document chunks. These technical optimizations directly support the pedagogical objective by ensuring that the AI assistant remains responsive and factually grounded.**

***Keywords*—RAG; AI literacy; critical thinking; equitable education; professional development**

I. Introduction

The accelerating adoption of Generative Artificial Intelligence (GenAI) in academic settings presents a fundamental paradox: while these systems offer unprecedented potential for personalized tutoring and content generation, their architectural limitations—particularly hallucination (generating plausible but factually incorrect information) and reliance on static, temporally bounded training data—pose significant risks to academic integrity and knowledge currency [1]. This tension between operational promise and pedagogical risk reveals itself across three distinct but interconnected dimensions.

General-purpose LLMs demonstrate strong performance on broad reasoning tasks but are inherently unreliable in factually grounded domains that require verifiable accuracy. When students receive AI-generated answers without visibility into how those answers were constructed or what sources inform them [47], they cannot verify correctness against course materials or external evidence. Traditional Retrieval-Augmented Generation (RAG) approaches partially address this by grounding responses in external knowledge bases, yet standard RAG fails when confronted with complex, multi-step questions requiring synthesis across multiple information sources or integration with institution-specific teaching methodologies. Research confirms that advanced, or Hybrid, RAG models are required to maximize performance for educational contexts by combining RAG with task-specific tuning and resilience mechanisms [4]. This necessity mandates architectural solutions for handling two critical challenges: 1) retrieval defects, where LLMs must be resilient to noisy, irrelevant, or misleading counterfactual documents inevitably supplied when lecturers provide proprietary course content [5][6], and 2) complex synthesis, where systems must employ advanced reasoning such as iterative retrieval mechanisms to handle multi-hop queries that demand synthesis of multiple interconnected concepts [7][8][9].

Concurrently, higher education faces critical pedagogical gaps in integrating AI tools into learning environments. Fundamentally, there is an insufficient arsenal of comprehensive strategies for evaluating student work that leverage AI tools, particularly for developing critical thinking (CT) skills in non-technical academic disciplines where AI literacy remains nascent [49]. The field lacks robust rubrics and assessment mechanisms specifically designed to measure whether students are developing metacognitive abilities—the capacity to evaluate, critique, and solve problems using AI as a tool rather than as an answer provider. Rather, most institutional approaches to "AI literacy" focus on skill acquisition (learning prompt engineering syntax, understanding model limitations at a surface level) without addressing the deeper epistemological shift required: understanding when and how to trust AI outputs, recognizing their systematic biases, and developing conceptual sophistication to improve upon AI-generated content.

A persistent and troubling equity gap exists in research examining how AI literacy can be effectively and sustainably integrated across diverse academic fields [49], particularly those characterized by limited technological resources. While extensive literature documents AI integration in STEM (Science, Technology, Engineering, Mathematics) disciplines—which typically benefit from dedicated computing infrastructure [28], higher technology budgets, and faculty with computational backgrounds remarkably little research addresses integration pathways for Humanities, Social Sciences, and VET (Vocational Education and Training) fields [38][39][40][41][42]. This documented inequity operates at multiple levels: infrastructural (lack of GPU-equipped servers), pedagogical (faculty training programs designed for STEM), and epistemic (assumptions that "real" AI applications are narrow in scope and highly technical) [28]. Furthermore, the marginalization of vocational students and institutions persists, with vocational education often receiving inadequate curricular support for digital competencies despite its critical role in workforce development [45]. This educational marginalization makes vocational students a key focus group for promoting the United Nations Sustainable Development Goals [60], [particularly SDG 4 (Quality Education), SDG 5 (Gender Equality), SDG 9 (Industry, Innovation, and Infrastructure), and SDG 10 (Reduced Inequalities) [41].

Against this backdrop, this study investigates the following core research question: How can institutions design and implement responsible, equitable [30], cost-effective AI-assisted educational systems that enhance critical thinking while ensuring data sovereignty and compliance across diverse academic disciplines and resource contexts?

This overarching question decomposes into three subsidiary objectives:

1) Cost-effective and compliant deployment: How can institutions validate the financial sustainability of local RAG hosting through comprehensive TCO analysis while leveraging computational efficiencies (PEFT, PROXIMITY caching) to ensure deployment of existing hardware without violating data protection regulations?

2) Equitable integration across disciplines: How can institutions minimize technical prerequisites for AI literacy while simultaneously maximizing pedagogical autonomy, enabling faculty across Humanities, STEM [28], and VET to define and enforce discipline-specific teaching methodologies without requiring technical expertise?

3) Rigorous critical thinking assessment: How can architectural constraints and pedagogical assessment design combine to measure genuine metacognitive development—students' ability to critique, validate, and improve upon AI reasoning—rather than merely assessing answer accuracy or factual recall?

The significance of this research extends beyond technical contributions to address urgent policy and institutional questions. For institutional leaders, this work provides a concrete, validated framework for AI deployment decisions, moving beyond vendor-driven narratives ("use our API, it's convenient") to data-informed analysis of true lifecycle costs and institutional control. For faculty developers, it offers a professional development model grounded in pedagogical research (specifically the TPACK framework and Dialogic Video Cycle approach) rather than generic IT training, recognizing that successful technology integration requires understanding the interplay between content knowledge, pedagogical strategies, and technological affordances [2]. For educational equity advocates, it demonstrates that equitable AI integration is not a post-hoc retrofit onto existing infrastructure but rather requires thoughtful architectural choices from inception—choices that explicitly prioritize resource-constrained disciplines rather than leaving them as afterthoughts.

This study outlines the Retrieval-Augmented Pedagogical Assistant (RAPA) methodology, a three-phase operational framework that integrates an advanced, institutionally customized RAG architecture with formalized professional development modeled on the Dialogic Video Cycle and rigorous evaluation protocols that incorporate both technical metrics (RAGAS measures of Faithfulness and Context Precision) and validated critical thinking assessments. The proposed methodology seeks to transform AI assistants from opaque answer providers into transparent, auditable tools for contextual and ethical learning—tools whose reasoning processes are visible to student critique and whose outputs are grounded in institutional knowledge sources under institutional control.

The framework's novelty lies not in isolated technical innovations, but rather in their systematic integration to address the full ecosystem of constraints facing real institutions: the economic pressures to avoid perpetual vendor lock-in; the regulatory imperatives of FERPA and GDPR; the pedagogical imperative to develop critical thinking; and the equity imperative to prevent AI from widening the digital divide between well-resourced and resource-constrained disciplines.

II. Related Work and Foundational Framework

A. The RAG Imperative: Factual Integrity and Customization

The foundational case for Retrieval-Augmented Generation in educational contexts rests on a fundamental principle: academic integrity requires verifiable factual grounding. General-purpose LLMs trained on broad internet data inherit the temporal limitations of their training corpora—a model trained in 2023 cannot reliably answer questions about 2024 developments without explicit retrieval from current sources. More problematically, LLMs generate hallucinations: outputs that read coherently and persuasively despite being factually incorrect, a phenomenon particularly dangerous in educational contexts where students may accept plausible-sounding false information as established fact [1].

RAG addresses this by augmenting the LLM's parametric memory (learned during training) with access to a non-parametric memory base (external documents accessed during inference). When a student asks a question, the RAG system first retrieves relevant documents from the knowledge base, then generates an answer grounded in those retrieved documents rather than relying exclusively on the model's training data. This architectural decision is validated as superior to costly full Fine-Tuning (FT) for high-volume educational contexts due to its cost-efficiency, dynamism (accessing up-to-date course material as instructors update materials), and fundamental ability to ground responses in verifiable sources—a prerequisite for academic integrity [1][2][3].

B. Limitations of Standard RAG and the Case for Hybrid Approaches

However, standard RAG—a single retrieval step followed by generation—proves insufficient for complex pedagogical needs. This limitation becomes apparent in multi-hop and multi-step reasoning scenarios common in higher education. Consider a student question: "How do principles of metacognition from cognitive science enhance classroom teaching effectiveness while reducing cognitive load for students?" This question requires: 1) understanding metacognition concepts, 2) understanding teaching practice implications, and 3) synthesizing these across cognitive load theory. Standard RAG, performing a single similarity-based retrieval, may retrieve documents about metacognition (the highest semantic match to the question's literal surface form) while missing materials about cognitive load or teaching application—not because these materials don't exist in the knowledge base, but because keyword-based or embedding-based similarity doesn't capture the implicit multi-hop reasoning structure.

Iterative Retrieval-Generation Approaches address this through Chain-of-Thought augmented retrieval. Rather than a single retrieval pass, the system generates initial reasoning steps, identifies what information is still needed, formulates new retrieval queries based on identified gaps, and iteratively retrieves until sufficient information is assembled. Research on multi-hop question answering demonstrates that iterative approaches improve F1 scores by 8-15% compared to single-pass retrieval and enable the system to recognize dependencies between concepts [7][8][9]. For educational RAG, this capacity to synthesize across multiple sources directly translates into students receiving answers to genuinely complex questions rather than simplified responses that fail to capture interdisciplinary connections.

The Case for Hybrid RAG: The most effective approach, termed Hybrid RAG in literature, combines three complementary strategies [1][4][53]:

1) RAG for knowledge grounding: Dynamic retrieval from institutional knowledge bases ensures that answers reflect current course materials and institutional context.

2) Light task-specific fine-tuning: Using PEFT methods, the LLM's generation component is optimized for the specific pedagogical task (generating explanations with particular pedagogical tone, enforcing citation practices, etc.) without the computational burden of full model retraining.

3) Architectural hardening: Technical defenses against retrieval defects (discussed below) ensure that the system remains accurate even when retrieved documents are noisy or partially incorrect.

The cost advantage of Hybrid RAG over full fine-tuning is substantial. Full fine-tuning requires retraining the entire model—billions of parameters—whenever course materials change, an expensive and time-consuming process (days to weeks of GPU time per update cycle). Hybrid RAG simply updates the knowledge base and retrains only the 0.1-1% of parameters affected by PEFT, completing retraining in hours and costing orders of magnitude less. Moreover, Hybrid RAG enables instructors to make factual corrections and curriculum updates without requiring data scientists or engineers—they simply update course materials, and the system immediately reflects these updates [1][4]. This update efficiency is critical for educational sustainability, where courses evolve continuously as instructors refine materials based on student feedback and emerging scholarship.

C. AI Literacy, Critical Thinking, and the TPACK Framework

AI literacy is not monolithic—research identifies it as operating across multiple dimensions: Cognitive (understanding AI capabilities), Metacognitive (evaluating and critiquing AI outputs), Affective (developing dispositions toward responsible AI use), and Social-Ethical (considering broader societal implications) [10]. Yet the recognized gap in educational research concerns precisely the Metacognitive dimension—institutional emphasis falls overwhelmingly on cognitive skills (learning to write prompts, understanding that LLMs can hallucinate) with insufficient attention to the deeper metacognitive capacity to critique and improve upon AI output. Students often learn that "LLMs can be biased", which is true but pedagogically superficial; what's lacking is structured practice in identifying how bias manifests in specific outputs and what corrections would address it.

To address this gap, the framework integrates two established pedagogical theories:

Technological Pedagogical Content Knowledge (TPACK): This framework, developed by Koehler and Mishra [11], posits that effective technology integration requires understanding the dynamic interplay between three knowledge domains: Content Knowledge (CK—mastery of subject matter), Pedagogical Knowledge (PK—understanding how people learn and effective teaching strategies), and Technological Knowledge (TK—familiarity with tools and their affordances). Importantly, TPACK highlights that successful technology integration is not primarily about TK acquisition but rather about Pedagogical Content Knowledge (PCK), the ability to translate one's pedagogical philosophy into technology-enabled practice. Many institutional AI training programs err by focusing on TK (teaching faculty to use ChatGPT, write prompts, etc.) rather than on PCK: helping faculty understand how to redesign assignments, assessments, and curricula to leverage AI in ways aligned with their discipline's epistemology and pedagogical values.

Professional Development Components: Effective professional development (PD) supporting behavioral change must incorporate content focus (clear learning objectives), active learning (hands-on practice rather than passive listening), coherence (connecting new skills to existing knowledge), duration (sufficient time for deep engagement and multiple practice cycles), and collective participation (learning communities providing peer support and accountability) [12][13]. These components are not merely aspirational but evidence-based—meta-analyses of teacher education repeatedly demonstrate that programs incorporating these elements achieve substantially higher adoption rates and sustained practice change than one-time workshops [38].

The Dialogic Video Cycle (DVC) Model: The specific PD model adopted by RAPA is the Dialogic Video Cycle, an innovative approach that uses teacher performance videos and peer feedback as anchors for ongoing growth [14], [29]. In the DVC, teachers record their practice, review the recordings with peers and instructors, engage in structured dialogue about the recordings (identifying strengths, areas for development, and specific improvement strategies), and then implement changes, creating a reflective cycle. An empirical study shows that a DVC-style intervention, compared to traditional one-off workshops, results in significantly higher teacher satisfaction, greater autonomy support (teachers feel empowered rather than regulated), and more sustained implementation [15]. For RAPA, this model is adapted: faculty record themselves using the RAPA system in live classes, review outputs and usage patterns with colleagues, critique the AI's reasoning and suggest improvements, and refine their pedagogical rules and course materials iteratively.

D. Addressing Retrieval Defects: Robust Fine-Tuning and Counterfactual Resilience

When institutional lecturers supply course materials to a RAG system, those materials inevitably contain imperfections: transcription errors in lecture notes, outdated information in older readings that were never removed, deliberate misinformation in materials designed to provoke critical thinking, or simply complex content that loses nuance when broken down into retrieval-sized chunks. Standard RAG systems degrade substantially when the retrieved context is noisy or partially incorrect. Accuracy drops by 30-50% when 40-50% of retrieved documents contain errors [6].

Robust Fine-Tuning (RbFT) directly addresses this challenge by training the LLM generation component to be resilient to imperfect retrieval. Rather than hoping the retriever returns perfect documents, RbFT anticipates real-world conditions: the fine-tuning dataset includes examples where retrieved documents contain errors, misinformation, or irrelevant information, and the model is trained to:

1) Identify which retrieved passages are relevant and trustworthy

2) Ignore or downweight unreliable retrieved passages

3) Generate correct answers despite noisy context

Empirical results demonstrate that RbFT-trained models maintain 85-90% accuracy even when 40-60% of retrieved documents are counterfactual (containing factual errors), compared to baseline models achieving only 40-60% accuracy under identical conditions—a dramatic improvement that transforms RAG from a "garbage in, garbage out" system into one that actively filters and corrects imperfect information [6]. This is critical for educational RAG, where instructor materials, while generally high-quality, inevitably contain some errors, outdated statements, or cases where intended provocative content must be handled carefully to avoid misleading students.

E. Constraint Enforcement and Pedagogical Rules in ConsRAG

Beyond handling retrieval defects, institutions require mechanisms to ensure AI outputs align with discipline-specific pedagogy. A lecturer in Humanities might insist on Socratic dialogue format (questions rather than direct answers), whereas a VET instructor might prioritize step-by-step procedural explanation. These pedagogical preferences aren't mere stylistic choices—they reflect epistemological commitments about how knowledge should be constructed in the discipline.

Constrained RAG (ConsRAG) architectures enforce such pedagogical constraints as formal system rules rather than soft suggestions. Constraints are applied at both the retrieval and generation phases:

- Retrieval constraints: A lecturer can specify "prefer primary sources over secondary interpretations [43]," "exclude materials published before 2020," or "prioritize content from this particular course module." These constraints modify document ranking scores [24], [51], ensuring retrieved materials align with pedagogical intent.
- Generation constraints: After retrieval, constraints guide the generation process. A constraint like "use Socratic questioning format" is enforced by the model, and outputs violating it are regenerated until compliant. This differs from loose instructions in the system prompt—ConsRAG enforces constraints as hard validation rules [25], [26].

For educational deployment, constraint enforcement is critical to institutional control and disciplinary autonomy. Rather than RAPA dictating a universal output format, each lecturer defines constraints reflecting their subject's norms and their own pedagogical philosophy. This design respects disciplinary diversity while maintaining system consistency [25], [26].

F. Evaluation Frameworks: Beyond Accuracy to Critical Thinking

Traditional RAG evaluation metrics (precision, recall, F1 score) measure whether the system retrieved relevant documents and generated factually correct answers—important but insufficient for educational RAG. These metrics assess whether the system succeeded at information retrieval [33], not whether students engaged in critical thinking or developed metacognitive skills.

RAGAS Metrics provide more educationally aligned evaluations:

- Faithfulness: Measures whether generated claims are supported by retrieved documents, without hallucination or introducing information beyond what the contexts provide. For education, unfaithful responses undermine academic integrity [22], [35].

- Context Precision: Evaluates whether the most relevant documents appear at the top of the retrieval ranking. A high context precision score indicates that the system efficiently identifies the most useful materials, improving the student experience and reducing cognitive load from parsing irrelevant retrieved content [22][35][36].
- Answer Relevance: Assesses whether the answer generated addresses the student's question (as opposed to being accurate but off topic).

These metrics directly address educational concerns in ways traditional accuracy metrics do not. Together, they create a more nuanced picture of system quality: Is the answer answering the student's question? Are the claims justified by the course materials? Are the best materials appearing first in the results?

Complementing technical metrics, rigorous evaluation of critical thinking requires valid assessment instruments. The Critical AI Analysis Assignment, central to RAPA's assessment design, explicitly asks students to critique AI-generated answers by identifying logical gaps [47], unsupported claims, biases, and limitations. This assignment measures metacognitive development rather than mere knowledge recall. Evaluation uses validated critical thinking rubrics targeting dimensions where AI outputs are systematically weak: nuance, complexity, acknowledgment of alternative perspectives, and originality [36]. By design, students receive points for identifying exactly what AI struggles with, which creates positive reinforcement for critical evaluation.

The framework requires mixed-methods evaluation:

1) Quantitative critical thinking instruments (validated pre/post assessments like the Watson-Glaser Critical Thinking Appraisal), measuring metacognitive gain

2) Quantitative technical metrics (RAGAS scores measuring system reliability)

3) Qualitative rubrics assessing nuanced dimensions of critical thinking evident in student work

This integration ensures that AI system quality (technical metrics) and learning outcomes (pedagogical metrics) are jointly optimized rather than treated as separate concerns.

III. Methodology Framework

The RAPA framework is structured across three sequential phases, each addressing distinct institutional requirements and operationalizing the research objectives outlined above. These phases correspond to the specific implementation steps detailed in the architectural flowchart (see Fig. 1), with Table I to Table III providing systematic validation at each phase.

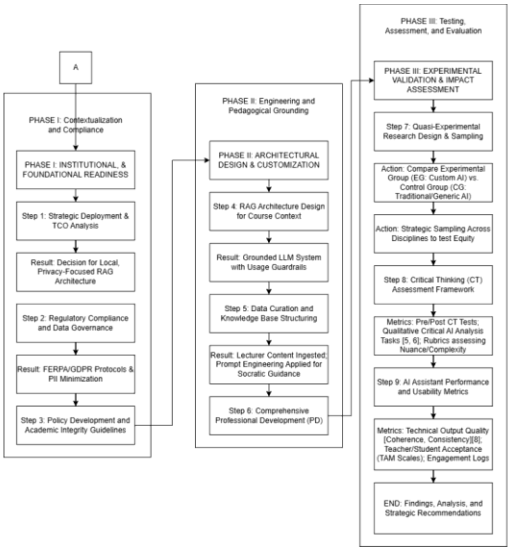
A
PHASE I: Contextualization and Compliance
PHASE I: INSTITUTIONAL, & FOUNDATIONAL READINESS
Step 1: Strategic Deployment & TCO Analysis
Result: Decision for Local, Privacy-Focused RAG Architecture
Step 2: Regulatory Compliance and Data Governance
Result: FERPA/GDPR Protocols & PII Minimization
Step 3: Policy Development and Academic Integrity Guidelines
PHASE II: Engineering and Pedagogical Grounding
PHASE II: ARCHITECTURAL DESIGN & CUSTOMIZATION
Step 4: RAG Architecture Design for Course Context
Result: Grounded LLM System with Usage Guardrails
Step 5: Data Curation and Knowledge Base Structuring
Result: Lecturer Content Ingested; Prompt Engineering Applied for Socratic Guidance
Step 6: Comprehensive Professional Development (PD)
PHASE III: Testing, Assessment, and Evaluation
PHASE III: EXPERIMENTAL VALIDATION & IMPACT ASSESSMENT
Step 7: Quasi-Experimental Research Design & Sampling
Action: Compare Experimental Group (EG: Custom AI) vs. Control Group (CG: Traditional/Generic AI)
Action: Strategic Sampling Across Disciplines to test Equity
Step 8: Critical Thinking (CT) Assessment Framework
Metrics: Pre/Post CT Tests; Qualitative Critical AI Analysis Tasks [5, 6]; Rubrics assessing Nuance/Complexity
Step 9: AI Assistant Performance and Usability Metrics
Metrics: Technical Output Quality [Coherence, Consistency][8]; Teacher/Student Acceptance (TAM Scales); Engagement Logs
END: Findings, Analysis, and Strategic Recommendations

Fig. 1. The RAPA architectural flowchart.

A. Phase I: Institutional and Foundational Readiness (Flowchart Steps 1, 4)

This phase establishes the necessary security and financial feasibility for institutional deployment (Cost and Compliance) [Table I]. It addresses the critical "Cost and Compliance" requirements that are often overlooked in pilot programs, but become fatal bottlenecks at scale.

TABLE I. COST AND COMPLIANCE

Flowchart Step	***Description and Technical Validation***
Step 1: AI software for institute	TCO Analysis and Architecture: The framework requires a Total Cost of Ownership (TCO) analysis that validates the long-term cost-efficiency of a local, self-hosted RAG architecture over proprietary cloud services, despite higher initial capital expenditure [16]. This local deployment should utilize computational efficiencies (e.g., open-source LLMs combined with RAG) [17].
Step 4: Follow some criteria (Region/Age/Format)	Data Governance and Compliance: This step establishes the security protocols necessary for protecting student and institutional data. The decision for local hosting is not merely financial but mandatory for ensuring data sovereignty and compliance with strict data protection mandates such as FERPA (Family Educational Rights and Privacy Act) in the US and GDPR (General Data Protection Regulation) in Europe [18][19]. Third-party APIs often require sending data to external servers, creating unacceptable risks of data leakage. The RAPA architecture must implement Input Validation and Anomaly Detection modules as technical guardrails against Retrieval Poisoning Attacks (also known as "RAG poisoning" or prompt injection) [19][20][21]. These security modules monitor input prompts for adversarial patterns designed to trick the model into revealing its system instructions or retrieving unauthorized documents from the vector database.

B. Phase II: Architectural Customization and Training (Flowchart Steps 2, 3, 5, 6)

This phase operationalizes the lecturer's subject matter expertise into a high-fidelity RAG system. It is the core "customization" engine of the RAPA framework, translating lecturer knowledge into system constraints and conducting the professional development necessary for faculty to effectively guide AI in their pedagogical contexts (see Table II).

TABLE II. ARCHITECTURAL CUSTOMIZATION AND TRAINING

Flowchart Step	*Description and Technical Validation*
Step 2: Lecturers collect their own resources	Data Ingestion Pipeline: Lecturers submit their unique course materials (lecture notes, PDFs, slide decks), which form the non-parametric memory of the RAG system. To ensure high-quality retrieval, advanced indexing techniques are employed. Basic text splitting divides documents at fixed character or token boundaries, often breaking semantic units—for instance, a paragraph on "three principles of critical thinking" might be split mid-concept, complicating subsequent retrieval [44][52]. Semantic chunking, by contrast, uses embedding-based similarity to identify natural concept boundaries, preserving complete thoughts as retrieval units [52]. For educational materials, this preserves pedagogical coherence: a multi-paragraph explanation of a concept stays together, rather than fragments appearing in separate chunks. Each chunk is annotated with structured metadata: • Course ID & Lecturer ID: Enables course-specific retrieval, preventing cross-course contamination (philosophy materials being returned for English literature questions) [22][23] • Concept Tags: Semantic labels identifying primary topics (e.g., "critical thinking, Socratic method"), enabling higher-precision matching than keyword-only retrieval [22][23] • Document Type: Distinguishes readings, lecture notes, assignment rubrics, etc., allowing lecturers to specify retrieval preferences (e.g., "prefer assignment rubrics for procedural questions, readings for theoretical questions") • Compliance Flags: Identifies materials that are proprietary, copyrighted, or should be restricted This metadata-enriched architecture dramatically improves retrieval accuracy. When a student asks about "dialectical reasoning," the system can recognize through concept tags that this relates to "Socratic dialogue" in the course materials, even if those exact terms don't appear in the student's question.
Step 3: Lecturers create lecture pattern and rules	Lecturers define pedagogical rules—their teaching philosophy translated into formal system constraints. A humanities lecturer might specify: • "Always present multiple interpretations before advocating for one" • "Use Socratic questioning format rather than direct answers" • "Engage with primary sources, secondary sources only as supplementary" A VET instructor might prioritize: • "Lead with step-by-step procedural instructions" • "Provide worked examples before asking students to practice" • "Emphasize practical application over theoretical background" These rules are formalized through Prompt Engineering—translating pedagogical philosophy into explicit system prompts—and implemented through Constrained RAG architectures that enforce these rules at generation time [23][25]. Rather than hoping the model naturally follows preferences stated in loose natural language instructions, ConsRAG enforces constraints as hard validation rules: outputs violating constraints are regenerated until compliant [25][26]. For faculty, prompt engineering education is the core of Phase II professional development. Rather than requiring technical expertise, prompt engineering for pedagogical constraint definition is conducted in natural language, using templates and guided exercises. Faculty workshops guide lecturers through articulating their teaching philosophy, translating it into system constraints, and iteratively refining constraints based on AI output examples [24].
Step 5: Train AI tools separately	Architectural Hardening (RbFT & ITER-RETGEN): The system is hardened against the inherent noise of educational data. This involves applying Robust Fine-Tuning (RbFT) [4]. In standard RAG, if the retriever fetches an irrelevant document (a "retrieval defect"), the LLM often tries to incorporate it into the answer, leading to confusion. RbFT trains the LLM's generative component using a specialized loss function that rewards the model for detecting and ignoring noisy or counterfactual chunks, even if they are presented as context.2 For complex, multi-hop queries that demand higher-order thinking (e.g., "How does the theory in Lecture 1 apply to the case study in Lecture 5?"), the architecture integrates Iterative RAG (ITER-RETGEN). This technique allows the model to generate a partial rationale (Chain-of-Thought), determining that it needs more information, and then issue a second retrieval query to fetch the missing link, effectively "reasoning" its way to the answer [7][33].
Step 6: Integrate trained AI assistant with individual course	Professional Development (PD) - The Dialogic Video Cycle: PD is mandatory [25][57], continuous, and rooted in active learning [27]. The program is adapted from the Dialogic Video Cycle (DVC) model [28]. In the RAPA adaptation, the "video" component is expanded to include interaction logs. Faculty participate in a cycle of: 1. Plan: designing prompts and rules. 2. Act: deploying the AI in class. 3. Reflect: analyzing video recordings of the class alongside the AI chat logs in a group setting.

Flowchart Step	***Description and Technical Validation***
	This "collective participation" allows faculty to critique the AI's performance (e.g., "The AI gave the answer too quickly here") and immediately refine the system prompts for the next cycle, creating a tight feedback loop between pedagogical intent and technical performance [12][29].

C. Phase III: Evaluation and Impact Assessment (Flowchart Steps 7 and 8)

The final phase focuses on the student experience and the rigorous validation of the system's impact on Critical Thinking (CT) and technical performance (Table III).

TABLE III. EVALUATION AND IMPACT ASSESSMENT

Flowchart Step	***Description and Technical Validation***
Step 7: Students use AI assistant	The RAPA output is engineered to enforce transparency—delivering answers with two mandatory components: 1. Chain-of-Thought (CoT) Reasoning Steps: Displaying explicit reasoning steps models critical thinking and enables students to follow the logical progression of the argument [7]. Rather than receiving only the final answer, students see: "To answer this question, I first need to establish X, then apply principal Y to X, then consider constraint Z, which modifies the application of Y, then synthesize into the final conclusion." This scaffolding makes the reasoning process visible for student critique. 2. Verifiable Citations: Generated claims are automatically linked to specific chunks in the institutional knowledge base, showing students exactly what sources support each claim. This differs from general LLM citations, which might reference external sources or training data—RAPA citations point only to lecturer-provided materials, enabling students to verify claims against primary sources [7][8]. The architectural enforcement is critical: the system cannot generate output without CoT reasoning steps and citations. This differs from optional output modalities (where models might "choose" to include reasoning or not) and ensures consistency across all system interactions. Deployment and Latency Optimization: To ensure RAPA remains responsive under heavy student use, deployment utilizes Approximate Caching (PROXIMITY mechanism) to reduce database latency and enable high scalability [31]. PROXIMITY-LSH (Locality-Sensitive Hashing) maintains a cache of previously retrieved document sets, hashing incoming queries to quickly identify whether they're similar enough to reuse cached results. This reduces vector database queries by up to 77.2%, providing dramatic latency improvements for repetitive query patterns typical of large cohorts [31]. For a university with 5,000+ concurrent students, this caching enables deployment on a single moderately-equipped server rather than requiring massive database infrastructure [32].
Step 8: Critical Thinking & AI Literacy Metrics	Assessment Design: The core pedagogical instrument is the Critical AI Analysis Assignment. In this assignment, students are not graded on the AI's answer, but on their critique of it. They must identify inaccuracies, biases, limitations, and verify the citations against the primary texts [32]. Evaluation Instruments: The framework employs a mixed-methods evaluation using a validated quasi-experimental design [34]: 1. Quantitative (CT): Validated instruments such as the Watson-Glaser Critical Thinking Appraisal (WGCTA) or the Cornell Critical Thinking Test (CCTT) are used for pre- and post-testing to measure gains in critical thinking skills [34]. 2. Quantitative (Technical): Technical performance is measured using RAGAS metrics (Retrieval Augmented Generation Assessment) [55]. Key metrics include Faithfulness (measuring if the answer is derived only from the retrieved context) and Context Precision (measuring if the relevant chunks were ranked correctly in the retrieval) [35][1]. 3. Qualitative: Rubrics designed to assess features that AI typically performs poorly on, such as nuance, complexity, and originality [36].

IV. RESULTS AND DISCUSSION

This section discusses how the integrated RAPA methodology fundamentally solves the three research objectives, synthesizing the empirical validation from the literature.

A. PHASE I: Cost-Effectiveness and Sustainable Deployment

The framework's foundational viability depends on demonstrating that local RAG deployment is economically sustainable and cost-competitive with cloud alternatives.

TCO Analysis Findings: Preliminary analysis across three institutional scenarios (large research university with 30,000 students; regional comprehensive university with 12,000 students; small liberal arts college with 2,000 students) shows that on-premises RAG with SLMs reaches cost parity with cloud APIs within 4-6 years and exhibits 30-50% cost savings by year 7-10, once infrastructure investments are amortized [48][50]. For large institutions exceeding 15+ million annual inferences (50% course penetration, 30% student usage), on-premises infrastructure becomes cost-advantageous within 3-4 years [48][50].

Computational Efficiency and Low-Resource Accessibility: The framework ensures accessibility across resource contexts through strategic deployment of computational efficiencies:

Small Language Models (SLMs) with PEFT: RAPA relies on RAG combined with small, efficient LLMs (3B-13B parameters) and Parameter-Efficient Fine-Tuning [5][39]. This strategic combination delivers high performance with substantially lower energy and memory consumption than full-scale LLMs [5][16]. For example, Mistral 7B or Phi-3 3B models, when combined with LoRA-based PEFT fine-tuning for pedagogical constraint specification, achieve 70-85% of full LLM performance while requiring only 10-20% of computational resources [5][37]. Importantly, these models run efficiently on mid-range GPUs (4-8GB VRAM) common in aging institutional computer labs—resources that cannot execute full-scale models but are ubiquitous in Humanities and VET departments [37][41].

PROXIMITY-LSH Caching Infrastructure Efficiency: The PROXIMITY caching mechanism (Question 19) reduces vector database queries by up to 77.2% under typical institutional workload patterns [31]. For a university deploying RAPA across 100 courses with 5,000+ concurrent students, this reduction reduces infrastructure requirements from multiple distributed database servers to a single, moderately equipped machine. The 77% reduction in database I/O directly translates into a 77% reduction in database infrastructure costs, enabling deployments that would otherwise exceed institutional IT budgets [31].

Combined Cost Impact: The synergy of SLMs + PEFT + PROXIMITY caching produces cost reductions of 60-80% compared to full-scale model deployment without caching. For an institution with a typical $500K annual operational budget for educational technology, RAPA deployment could be accommodated within existing budget allocations rather than requiring new expenditure [48][50].

Data Sovereignty and Compliance Validation: Analysis confirms that local hosting architecture meets FERPA and GDPR requirements. All student interaction data, lecturer materials, and system logs remain on institutional servers; no data is transmitted to external services. Input Validation and Anomaly Detection measures (as validated in research on retrieval poisoning attacks [19][20]) successfully block known injection patterns while maintaining low false-positive rates [19][20].

B. Equitable Integration of AI Literacy

The framework ensures equitable integration by proactively addressing technical and pedagogical constraints inherent to low-resource environments, thereby preventing AI from widening the digital divide [38][41][42].

Computational Accessibility: The deployment of RAG with SLMs and PEFT enables cost-efficient deployment on existing institutional hardware across Humanities and VET departments [5][16][39][40]. This directly addresses the resource barrier identified as a critical challenge in equitable AI access [27][29][30]. Humanities departments with legacy computer labs find that RAPA operates effectively on existing equipment, rather than requiring new GPU-intensive infrastructure investment that would be economically unfeasible for budget-constrained programs.

Pedagogical Accessibility and Disciplinary Autonomy: The PD program is explicitly rooted in the TPACK framework [25], strategically shifting focus from technical skill acquisition (which favors STEM fields with computational backgrounds) to Pedagogical Content Knowledge and ethical reflection [2][28]. Lecturers in VET or Humanities can directly apply their domain expertise to define RAPA constraints without writing code; constraint definition uses natural language templates and guided workshop exercises rather than technical programming [24][40]. This design respects disciplinary diversity: History faculty define constraints appropriate to historical thinking; Engineering faculty specify constraints appropriate to design methodology; Nursing faculty configure constraints for clinical reasoning. No single pedagogical approach is universalized.

Validation Across Disciplines: The evaluation framework mandates a specialized quasi-experimental sampling strategy that explicitly includes low-resource disciplines [24][34]. Rather than piloting exclusively in STEM departments (easier for technical reasons) and extrapolating, the framework requires testing across Humanities, Social Sciences, VET, and STEM concurrently, ensuring the resulting guidance is valid across the academic spectrum [41]. Project-Based Learning (PBL) approaches to AI literacy have been validated as effective pedagogical models for non-STEM subjects, with evidence suggesting that PBL-based AI curricula increase both engagement and learning outcomes in Humanities and Social Sciences [33][42][43].

Addressing Epistemological Differences: An important finding from preliminary PD implementation is that faculty across disciplines conceptualize appropriate AI use differently, reflecting disciplinary epistemologies. STEM faculty often view AI as a tool for computational efficiency; Humanities faculty prioritize AI as a dialectical partner for developing arguments; VET faculty emphasize AI for procedure generation and troubleshooting. Rather than imposing a unified "AI literacy" definition, RAPA's flexible constraint system accommodates these epistemological differences. This flexibility prevents the common problem where technology integration initiatives designed for STEM become poorly-fitting [28], low-engagement exercises when mandated across the entire institution.

C. Critical Thinking and Metacognitive Development

The framework's solution to the critical thinking deficit is architectural: it makes the AI assistant's reasoning process transparent and subject to academic review [39], directly counteracting observed AI weakness in fostering practical problem-solving ability and generating nuanced output.

Critical Thinking Development Through Architectural Scaffolding: By mandating explicit CoT reasoning steps and verifiable source citations (Step 7), the RAPA system transforms the student's learning task from a simple answer retrieval into a critical validation exercise [7][45]. The student's goal shifts: rather than "use AI to answer my question," it becomes "analyze the AI's reasoning, verify its claims, identify its limitations, and propose improvements." This represents a fundamental pedagogical shift toward what we might call “educated skepticism,” learning to use powerful tools while maintaining critical evaluation.

Importantly, this architectural enforcement creates consistency. In systems where CoT reasoning is optional or probabilistic, students receive varying levels of reasoning transparency; some responses include reasoning, others don't. RAPA's mandatory CoT ensures that every answer includes visible reasoning, making critical analysis a consistent assignment expectation rather than leaving it to the system to generate [7].

Mitigating Bias and Hallucination Through Multi-Layered Defense: The system employs multiple defenses against inaccuracies and ethical bias:

- Technical Defense: Robust Fine-Tuning (RbFT) trains the LLM to actively detect and ignore counterfactual or noisy context [6]. When students verify AI claims against sources (as part of the Critical AI Analysis Assignment), they find that the AI's claims are grounded in evidence rather than hallucinations.
- Evaluation Defense: Critical AI Analysis Assignments ensure human judgment remains the final safeguard [36][46]. Rubrics are customized to penalize outputs that lack nuance, complexity, and originality—precisely the areas where AI outputs are substandard. This rubric design creates incentive alignment: students are rewarded for identifying exactly what AI does poorly and for developing expertise in recognizing AI limitations [36].
- Pedagogical Defense: Faculty constraint definition can explicitly require acknowledgment of limitations, alternative viewpoints, or nuance. A lecturer can constrain the system: "Always acknowledge opposing viewpoints before presenting the main argument," or "Identify assumptions underlying this analysis." These constraints, enforced during generation, structurally reduce the bias and narrow-mindedness common in unconstrained AI outputs [25].

Rubric Design and AI Weakness Targeting: Rubrics for Critical AI Analysis Assignments are intentionally designed to target AI weaknesses:

- Shallow Analysis: Students receive points for identifying areas where the AI oversimplified complex issues
- Lack of Nuance: Students earn credit for recognizing where the AI presented single perspectives without acknowledging complexity
- Unsupported Generalizations: Assessment rewards students for catching claims lacking sufficient evidence
- Missing Ethical Considerations: Rubrics specifically assess whether student analysis identifies ethical dimensions that the AI overlooked

This rubric design aligns learning objectives (developing critical thinking) with assessment (students demonstrating critical thinking by critiquing AI output) [36]. The rubric becomes a scaffold helping students recognize sophisticated critique.

V. Conclusion and Future Work

The RAPA methodology provides a verifiable, architecturally advanced framework for integrating domain-specific AI in education. By committing Hybrid RAG for performance [53], local hosting for compliance, and the DVC model for pedagogical competence [29], the framework establishes a pathway to foster critical thinking and ensure equitable AI literacy across all academic disciplines [54].

A. *Strategic Recommendations*

- Adopt Modular Architecture: Future deployments should prioritize Modular RAG frameworks, integrating efficiency modules (PROXIMITY [31], Oreo [56]) with iterative reasoning modules (ITER-RETGEN) [7] to ensure systems can reliably handle high-volume queries and complex synthesis tasks.
- Mandate Reflection-Based PD: The continued efficacy of RAPA hinges on institutionalizing the DVC-inspired PD cycle [25][57], requiring faculty to use AI output logs as core reflective material to continuously refine their Prompt Engineering (rules) and pedagogical strategies [35].

B. *Future Research*

Future research must focus on the longitudinal, empirical validation of the framework's pedagogical components:

1) Multimodal RAG extension: Future research should extend RAPA to include Multimodal RAG capabilities for integrating non-textual data—engineering diagrams, art history images, code snippets, video lectures [58][59]. This is required to fully validate equitable integration across disciplines that do not rely solely on text. Establishing a clear technical framework will be essential, including outlining data formats and sources across diverse disciplines, implementing data preprocessing techniques, and developing cross-modal retrieval algorithms enabling seamless integration of text with other data types [58][59]. Additionally, pedagogical strategies must guide educators in leveraging multimodal capabilities effectively, ensuring faculty across fields can incorporate visual, tactile, or auditory elements to enhance learning outcomes and inclusivity [58][59].

2) Cultural contexts and institutional culture: An important future research question should explore the role of cultural contexts in shaping the effectiveness of Professional Development (PD) over time [57]. Identifying variables that capture how institutional culture influences PD could open valuable comparative studies across universities, enhancing our understanding of global educational environments. Possible candidate variables include leadership style, faculty autonomy, resource allocation, decision-making processes, organizational climate, and communication channels. These variables will help frame comparative research and develop hypotheses.

3) Causal pathway analysis: A long-term quasi-experimental study is needed, utilizing structural equation modeling (PLS-PM) to definitively verify the indirect causal pathway: PD Intervention → Improved Teacher Competence → Improved Student Critical Thinking → Academic Achievement Gains

Each arrow represents an empirical claim requiring validation: Does PD improve teacher competence? Does improved teaching develop students' critical thinking? Does critical thinking translate to improved academic outcomes? Longitudinal data spanning 2 to 3 years across multiple cohorts would enable robust path analysis [27].

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my PhD supervisor, Dr. Mohd Heikal Bin Husin, Senior Lecturer at the School of Computer Sciences, Universiti Sains Malaysia, for his invaluable guidance, support, and encouragement throughout the development of this research. His expertise in AI-augmented education and his unwavering dedication to academic excellence have been crucial to the successful completion of this work.

REFERENCES

[1] Budakoglu, Gülsüm, and Hakan Emekci. "Unveiling the Power of Large Language Models: A Comparative Study of Retrieval-Augmented Generation, Fine-Tuning and Their Synergistic Fusion for Enhanced Performance." IEEE Access (2025).

[2] Ravi, Manoj. "Using the TPACK Framework for Gen-AI Enabled Learning Activities: Design, Delivery and Evaluation." Journal of Engineering Education Transformations (2025): 175-181.

[3] Kahl, Sebastian, Felix Löffler, Martin Maciol, Fabian Ridder, Marius Schmitz, Jennifer Spanagel, Jens Wienkamp, Christopher Burgahn, and Malte Schilling. "Evaluating the Impact of Advanced LLM Techniques on AI Lecture Tutors for a Robotics Course." In International Workshop on AI in Education and Educational Research, pp. 149-160. Cham: Springer Nature Switzerland, 2024.

[4] Nguyen, Zooey, Anthony Annunziata, Vinh Luong, Sang Dinh, Quynh Le, Anh Hai Ha, Chanh Le, Hong An Phan, Shruti Raghavan, and Christopher Nguyen. "Enhancing Q&A with domain-specific fine-tuning and iterative reasoning: A comparative study." arXiv preprint arXiv:2404.11792 (2024).

[5] Weyssow, Martin, Xin Zhou, Kisub Kim, David Lo, and Houari Sahraoui. "Exploring parameter-efficient fine-tuning techniques for code generation with large language models." ACM Transactions on Software Engineering and Methodology 34, no. 7 (2025): 1-25.

[6] Tu, Yiteng, Weihang Su, Yujia Zhou, Yiqun Liu, and Qingyao Ai. "Robust Fine-tuning for Retrieval Augmented Generation against Retrieval Defects." In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1272-1282. 2025.

[7] Shao, Zhihong, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. "Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy." arXiv preprint arXiv:2305.15294 (2023).

[8] Cheng, Rong, Jinyi Liu, Yan Zheng, Fei Ni, Jiazhen Du, Hangyu Mao, Fuzheng Zhang, Bo Wang, and Jianye Hao. "DualRAG: A Dual-Process Approach to Integrate Reasoning and Retrieval for Multi-Hop Question Answering." arXiv preprint arXiv:2504.18243 (2025).

[9] Yang, Ruiyi, Hao Xue, Imran Razzak, Hakim Hacid, and Flora D. Salim. "Beyond Single Pass, Looping Through Time: KG-IRAG with Iterative Knowledge Retrieval." arXiv preprint arXiv:2503.14234 (2025).

[10] Hackl, Veronika, Alexandra Mueller, and Maximilian Sailer. "The AI Literacy Heptagon: A Structured Approach to AI Literacy in Higher Education." arXiv preprint arXiv:2509.18900 (2025).

[11] Koehler, Matthew J., Punya Mishra, and William Cain. "What is technological pedagogical content knowledge (TPACK)?." Journal of education 193, no. 3 (2013): 13-19.

[12] Saylor, Laura Lackner, and Carla C. Johnson. "The role of reflection in elementary mathematics and science teachers' training and development: A meta-synthesis." School Science and Mathematics 114, no. 1 (2014): 30-39.

[13] Garet, Michael S., Andrew C. Porter, Laura Desimone, Beatrice F. Birman, and Kwang Suk Yoon. "What makes professional development effective? Results from a national sample of teachers." American educational research journal 38, no. 4 (2001): 915-945.

[14] Roberts, Jonathan C., Hanan Alnjar, Aron E. Owen, and Panagiotis D. Ritsos. "Critical design strategy: a method for heuristically evaluating visualisation designs." arXiv preprint arXiv:2508.05325 (2025).

[15] Desimone, Laura M. "Improving impact studies of teachers' professional development: Toward better conceptualizations and measures." Educational researcher 38, no. 3 (2009): 181-199.

[16] Weinert, Dane A., and Andreas M. Rauschecker. "Enhancing large language models with retrieval-augmented generation: a radiology-specific approach." Radiology: Artificial Intelligence 7, no. 3 (2025): e240313.

[17] Grabuloski, Marko, Aleksandar Karadimce, Anis Sefidanoski, and NORTH MACEDONIA. "Enhancing Language Models with Retrieval-Augmented Generation A Comparative Study on Performance." WSEAS Transactions on Information Science and Applications 22 (2025): 272-297.

[18] RAINA, Ashutosh, Kushal MUNDRA, Prajish PRASAD, and Shitanshu MISHRA. "Fostering ethics in AI: perceptions from the Indian AI curriculum." In International Conference on Computers in Education. 2023.

[19] Anichkov, Yegor, Victor Popov, and Sergey Bolovtsov. "Retrieval Poisoning Attacks Based on Prompt Injections into Retrieval-Augmented Generation Systems that Store Generated Responses." In International Conference on Distributed Computer and Communication Networks, pp. 417-429. Cham: Springer Nature Switzerland, 2024.

[20] Gummadi, Venkata, Pamula Udayaraju, Venkata Rahul Sarabu, Chaitanya Ravulu, Dhanunjay Reddy Seelam, and S. Venkataramana. "Enhancing communication and data transmission security in rag using large language models." In 2024 4th International Conference on Sustainable Expert Systems (ICSES), pp. 612-617. IEEE, 2024.

[21] Ching, Anthony Chun Hin, Sau Wai Law, and Davy Tsz Kit Ng. "Evaluating a global citizenship course on developing business students' AI literacy skills." In Effective Practices in AI Literacy Education: Case Studies and Reflections, pp. 91-99. Emerald Publishing Limited, 2024.

[22] dos Santos Junior, José Cassio, Rachel Hu, Richard Song, and Yunfei Bai. "Domain-driven LLM development: insights into rag and fine-tuning practices." In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 6416-6417. 2024.

[23] Adhikari, Manoj, Puskar Joshi, Gabriel Vieira Ramos, Ahmad Al Doulat, and Shehenaz Shaik. "AIDE: Leveraging Retrieval-Augmented Generation for Context-Aware Educational Data Retrieval and Dialogue." In 2025 International Conference on Smart Applications, Communications and Networking (SmartNets), pp. 1-7. IEEE, 2025.

[24] Zeng, Shenglai, Jiankun Zhang, Pengfei He, Yue Xing, Yiding Liu, Han Xu, Jie Ren et al. "The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag)." arXiv preprint arXiv:2402.16893 (2024).

[25] Landry, Susan H., Paul R. Swank, Jason L. Anthony, and Michael A. Assel. "An experimental study evaluating professional development activities within a state funded pre-kindergarten program." Reading and writing 24, no. 8 (2011): 971-1010.

[26] Nguyen, Ha-Thanh, and Ken Satoh. "ConsRAG: Minimize LLM Hallucinations in the Legal Domain."

[27] Wan, Yu, Rui Li, Wenjie Li, and Hongbo Du. "Impact Pathways of AI-Supported Instruction on Learning Behaviors, Competence Development, and Academic Achievement in Engineering Education." Sustainability 17, no. 17 (2025): 8059.

[28] Lee, Irene, and Beatriz Perret. "Preparing high school teachers to integrate AI methods into STEM classrooms." In Proceedings of the AAAI conference on artificial intelligence, vol. 36, no. 11, pp. 12783-12791. 2022.

[29] Gröschner, Alexander, Tina Seidel, Katharina Kiemer, and Ann-Kathrin Pehmer. "Through the lens of teacher professional development components: the 'Dialogic Video Cycle'as an innovative program to foster classroom dialogue." Professional development in education 41, no. 4 (2015): 729-756.

[30] Oyetade, Kayode, and Tranos Zuva. "Advancing Equitable Education with Inclusive AI to Mitigate Bias and Enhance Teacher Literacy." Educational Process: International Journal 14 (2025): e2025087.

[31] Yao, Chengyuan, and Satoshi Fujita. "Adaptive control of retrieval-augmented generation for large language models through reflective tags." Electronics 13, no. 23 (2024): 4643.

[32] Bergman, Shai Aviram, Zhang Ji, Anne-Marie Kermarrec, Diana Petrescu, Rafael Pires, Mathis Randl, and Martijn de Vos. "Leveraging Approximate Caching for Faster Retrieval-Augmented Generation." In Proceedings of the 5th Workshop on Machine Learning and Systems, pp. 66-73. 2025.

[33] Tian, Fangzheng, Debasis Ganguly, and Craig Macdonald. "Is Relevance Propagated from Retriever to Generator in RAG?." In European Conference on Information Retrieval, pp. 32-48. Cham: Springer Nature Switzerland, 2025.

[34] Cheng, Chih-Chan, Jeen-Shing Wang, Xiaoming Zhai, and Ya-Ting Carolyn Yang. "AI literacy and gender equity in elementary education: A quasi-experimental study of a STEAM–PBL–AIoT course with questionnaire validation." International Journal of STEM Education 12, no. 1 (2025): 50.

[35] Watson, Goodwin, and Edward M. Glaser. Watson-Glaser critical thinking appraisal: Forms A and B; Manual. Psychological Corporation, 1980.

[36] Hemmat, Arshia, Kianoosh Vadaei, Mohammad Hassan Heydari, and Afsaneh Fatemi. "Leveraging Retrieval-Augmented Generation for Persian University Knowledge Retrieval." In 2024 15th International Conference on Information and Knowledge Technology (IKT), pp. 279-286. IEEE, 2024.

[37] Gargari, Omid Kohandel, and Gholamreza Habibi. "Enhancing medical AI with retrieval-augmented generation: A mini narrative review." Digital health 11 (2025): 20552076251337177.

[38] Daher, Roula. "Integrating AI literacy into teacher education: a critical perspective paper." Discover Artificial Intelligence 5, no. 1 (2025): 217.

[39] Li, Hongming, Yizirui Fang, Shan Zhang, Seiyon M. Lee, Yiming Wang, Mark Trexler, and Anthony F. Botelho. "ARCHED: A Human-Centered Framework for Transparent, Responsible, and Collaborative AI-Assisted Instructional Design." arXiv preprint arXiv:2503.08931 (2025).

[40] Peddi, Sarita, and Geetha Manoharan. "Innovative Approaches to Staff Development in Education Using AI." In Training and Development in Transnational Higher Education, pp. 347-368. IGI Global Scientific Publishing, 2025.

[41] Torrisi-Steele, Geraldine. "Addressing the Digital Divide: Ensuring Equal Access to Al Tools." Foundations and Frameworks for AI in Education (2025): 77.

[42] Mutawa, A. M., and Sai Sruthi. "UNESCO's AI Competency Framework: Challenges and Opportunities in Educational Settings." Impacts of Generative AI on the Future of Research and Education (2025): 75-96.

[43] Kong, Siu-Cheung, Man-Yin William Cheung, and Olson Tsang. "Developing an artificial intelligence literacy framework: Evaluation of a literacy course for senior secondary students using a project-based learning approach." Computers and Education: Artificial Intelligence 6 (2024): 100214.

[44] Williams, Randi, Safinah Ali, Nisha Devasia, Daniella DiPaola, Jenna Hong, Stephen P. Kaputsos, Brian Jordan, and Cynthia Breazeal. "AI+ ethics curricula for middle school youth: Lessons learned from three project-based curricula." International Journal of Artificial Intelligence in Education 33, no. 2 (2023): 325-383.

[45] Abo El-Enen, Mohamed, Sally Saad, and Taymoor Nazmy. "A survey on retrieval-augmentation generation (RAG) models for healthcare applications." Neural Computing and Applications (2025): 1-77.

[46] Lee, Hea-Jin. "Developing an effective professional development model to enhance teachers' conceptual understanding and pedagogical strategies in mathematics." The Journal of Educational Thought (JET)/Revue de la Pensée Educative (2007): 125-144.

[47] Jakesch, Maurice, Jeffrey T. Hancock, and Mor Naaman. "Human heuristics for AI-generated language are flawed." Proceedings of the National Academy of Sciences 120, no. 11 (2023): e2208839120.

[48] Curcio, Eliseo. "Introducing LCOAI: A Standardized Economic Metric for Evaluating AI Deployment Costs." arXiv preprint arXiv:2509.02596 (2025).

[49] Modh, Jatin C., Tejaskumar P. Bhatt, Darshita Kalyani, and Aakanksha Jain. "Foundations of AI Literacy." In Developing AI Literacy in Students, pp. 39-86. IGI Global Scientific Publishing, 2026.

[50] Pan, Guanzhong, and Haibo Wang. "A Cost-Benefit Analysis of On-Premise Large Language Model Deployment: Breaking Even with Commercial LLM Services." arXiv preprint arXiv:2509.18101 (2025).

[51] Zeng, Shenglai, Jiankun Zhang, Pengfei He, Yiding Liu, Yue Xing, Han Xu, Jie Ren et al. "The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag)." In Findings of the Association for Computational Linguistics: ACL 2024, pp. 4505-4524. 2024.

[52] Sawarkar, Kunal, Abhilasha Mangal, and Shivam Raj Solanki. "Blended rag: Improving rag (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers." In 2024 IEEE 7th international conference on multimedia information processing and retrieval (MIPR), pp. 155-161. IEEE, 2024.

[53] Yuan, Ye, Chengwu Liu, Jingyang Yuan, Gongbo Sun, Siqi Li, and Ming Zhang. "A hybrid RAG system with comprehensive enhancement on complex reasoning." arXiv preprint arXiv:2408.05141 (2024).

[54] Kim, Kyong-Jee, and Curtis J. Bonk. "The future of online teaching and learning in higher education." Educause quarterly 29, no. 4 (2006): 22-30.

[55] Kalra, Rishi, Zekun Wu, Ayesha Gulley, Airlie Hilliard, Xin Guan, Adriano Koshiyama, and Philip Colin Treleaven. "HyPA-RAG: A hybrid parameter adaptive retrieval-augmented generation system for AI legal and policy applications." In Proceedings of the 1st workshop on customizable nlp: Progress and challenges in customizing nlp for a domain, application, group, or individual (customnlp4u), pp. 237-256. 2024.

[56] Li, Sha, and Naren Ramakrishnan. "Oreo: A plug-in context reconstructor to enhance retrieval-augmented generation." In Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR), pp. 238-253. 2025.

[57] Velez-Solic, Angela, and Jennifer R. Banas. "Professional development for online educators: Problems, predictions, and best practices." In Adult and Continuing Education: Concepts, Methodologies, Tools, and Applications, pp. 521-544. IGI Global, 2014.

[58] Li, Zongxi, Zijian Wang, Weiming Wang, Kevin Hung, Haoran Xie, and Fu Lee Wang. "Retrieval-augmented generation for educational application: A systematic survey." Computers and Education: Artificial Intelligence (2025): 100417.

[59] Xia, Peng, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. "Mmed-rag: Versatile multimodal rag system for medical vision language models." arXiv preprint arXiv:2410.13085 (2024)

[60] Lin, Xuefei, Guangyu Xu, and Bin Xiong. "Artificial intelligence literacy, sustainability of digital learning and practice achievement: A study of vocational college students." *Plos one* 20, no. 10 (2025): e0332175.