British Journal of Computer, Networking and Information Technology
ISSN: 2689-5315
Volume 9, Issue 2, 2026 (pp. 87-99)
www.abjournals.org

A RETRIEVAL-AUGMENTED LARGE LANGUAGE MODEL FOR DYNAMIC PERSONALIZATION IN INTELLIGENT TUTORING SYSTEMS

Kuyoro Afolashade, Nzenwata Uchenna, and Akinwunmi Damilare.

Department of Computer Science, Babcock University, Ilishan-Remo, Ogun State, Nigeria.

Cite this article:

Kuyoro, A., Nzenwata, U., Akinwunmi, D. (2026), A Retrieval-Augmented Large Language Model for Dynamic Personalization in Intelligent Tutoring Systems. British Journal of Computer, Networking and Information Technology 9(2), 87-99. DOI: 10.52589/BJCNIT-2MVXUYXK

Manuscript History

Received: 17 Jun 2026

Accepted: 11 Jul 2026

Published: 28 Jul 2026

ABSTRACT: *This article presents a novel approach to Intelligent Tutoring Systems (ITS) by integrating Retrieval-Augmented Generation (RAG) with Large Language Models (LLMs) to enable dynamic personalization in educational contexts. The system addresses limitations in traditional ITS that rely on static, rule-based approaches by implementing a three-layered architecture combining semantic retrieval mechanisms with generative AI capabilities. Using GPT-4 as the core LLM enhanced with a custom RAG framework, the system demonstrates improvements in response accuracy (93%), inference speed (2.1 seconds per prompt), and computational efficiency compared to a standard GPT-4 baseline, a traditional rule-based ITS, and an LLM with keyword-based retrieval. The research employs both ASSISTments (fine-grained interaction data) and EdNet (large-scale longitudinal data) datasets for evaluation. Results show that the RAG-enhanced system achieves 40% better contextual relevance compared to standard LLM implementations. The framework incorporates adaptive prompting strategies, real-time knowledge base updates, and multi-level personalization algorithms to create a dynamic educational environment.*

KEYWORDS: Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), Intelligent Tutoring Systems (ITS), Dynamic Personalization, Adaptive Learning, Artificial Intelligence, Generative AI.

Article DOI: 10.52589/BJCNIT-2MVXUYXK
DOI URL: https://doi.org/10.52589/BJCNIT-2MVXUYXK

INTRODUCTION

The evolution of intelligent tutoring systems has reached a critical juncture where traditional rule-based approaches are increasingly inadequate for meeting diverse learner needs in modern educational environments. Conventional ITS architectures struggle with three fundamental limitations: static knowledge bases that cannot accommodate rapidly evolving information domains, inability to process and adapt to real-time learner feedback, and reliance on predefined pedagogical strategies that lack flexibility (Anderson et al., 1995). The emergence of large language models presents a transformative opportunity, yet their standalone implementation often suffers from hallucination, factual inaccuracies, and lack of domain specificity (Piech et al., 2015).

Retrieval-augmented generation offers a promising solution to these challenges. By integrating retrieval mechanisms with generative capabilities, RAG frameworks enable LLMs to access verified, up-to-date information while maintaining contextual awareness and personalization (Lewis et al., 2020). This integration is particularly crucial in educational settings where accuracy, relevance, and pedagogical appropriateness are paramount. The dynamic nature of learning where each student's knowledge state, engagement level, and learning preferences evolve continuously demands systems that can adapt in real-time while maintaining factual integrity.

This study introduces an advanced RAG-LLM framework specifically designed for intelligent tutoring applications. The system employs a multi-layered architecture that combines semantic search capabilities with adaptive generation algorithms, creating a feedback loop between learner interactions and instructional content delivery. Unlike traditional ITS that operate on fixed knowledge graphs or rule-based decision trees, our approach enables continuous learning from both structured educational content and unstructured learner interactions.

LITERATURE REVIEW

The theoretical foundations for RAG-enhanced ITS draw from multiple disciplines: cognitive science, educational psychology, computer science, and information retrieval. From cognitive science, Sweller's Cognitive Load Theory (Sweller, 1988) informs our understanding of how retrieval mechanisms can reduce extraneous cognitive load by providing relevant information at optimal times. Vygotsky's Zone of Proximal Development (Vygotsky, 1978) underpins the personalization aspects, guiding how the system adapts content difficulty based on individual capabilities.

Research in educational technology has increasingly emphasized the importance of adaptive systems. VanLehn's comprehensive review of ITS effectiveness highlighted that successful systems share common characteristics: detailed student modeling, immediate feedback, and adaptive content delivery (VanLehn, 2011). More recently, Baker's meta-analysis of AI in education identified retrieval-augmented systems as particularly promising for maintaining accuracy while enabling personalization (Baker, 2021).

The development of ITS has undergone a significant transformation over the past decades, evolving from rule-based instructional systems to more sophisticated, data-driven, and adaptive

Article DOI: 10.52589/BJCNIT-2MVXUYXK
DOI URL: https://doi.org/10.52589/BJCNIT-2MVXUYXK

British Journal of Computer, Networking and Information Technology
ISSN: 2689-5315
Volume 9, Issue 2, 2026 (pp. 87-99)
www.abjournals.org

learning environments. Early ITSs were largely grounded in symbolic artificial intelligence, where domain knowledge and pedagogical strategies were explicitly encoded into the system. While these systems demonstrated the feasibility of automated tutoring, they were limited in their ability to adapt to diverse learner behaviors and lacked scalability across domains (Anderson et al., 1995).

The introduction of machine learning approaches marked a critical shift toward personalization, enabling systems to learn from user data and adjust instructional strategies accordingly. Techniques such as Bayesian knowledge tracing and later deep knowledge tracing have been employed to model learner performance and predict future outcomes (Piech et al., 2015). However, these approaches often rely on structured datasets and predefined features, limiting their capacity to interpret complex and unstructured learner inputs.

The emergence of LLMs has significantly expanded the capabilities of educational systems by enabling natural language interaction and contextual reasoning. Despite their versatility, concerns regarding hallucination, bias, and lack of interpretability have been widely documented (Kasneci et al., 2023). These limitations are particularly critical in educational settings, where the accuracy and reliability of information are essential.

RAG has been proposed as a solution to these challenges by combining retrieval-based and generative approaches. Lewis et al. introduced the RAG framework, demonstrating that integrating external knowledge retrieval improves performance in knowledge-intensive tasks (Lewis et al., 2020). Borgeaud et al. further extended this work by showing that large-scale retrieval significantly enhances the factual accuracy and scalability of language models (Borgeaud et al., 2022). In the context of education, the integration of retrieval-based methods remains underexplored, particularly in how they can be applied to tailor educational content to individual student needs.

Recent innovations in retrieval optimization have direct implications for ITS applications. The FAISS library, developed by Johnson et al. (2019), enables efficient similarity search in high-dimensional spaces, a critical capability for educational content retrieval where semantic similarity often matters more than keyword matching. Similarly, the Sentence Transformers framework by Reimers and Gurevych (2019) provides robust semantic embeddings that capture educational concept relationships more effectively than traditional word embeddings.

More recent studies have begun to examine the role of hybrid architectures that combine learner modeling with generative systems. Abdelrahman et al. highlighted the potential of LLMs to support adaptive learning, particularly when combined with real-time data analysis (Abdelrahman et al., 2023). This research addresses the gap by proposing a unified framework that leverages RAG to enhance personalization, accuracy, and engagement in ITS.

Article DOI: 10.52589/BJCNIT-2MVXUYXK
DOI URL: https://doi.org/10.52589/BJCNIT-2MVXUYXK

British Journal of Computer, Networking and Information Technology
ISSN: 2689-5315
Volume 9, Issue 2, 2026 (pp. 87-99)
www.abjournals.org

THEORETICAL FRAMEWORK

The proposed system is grounded in constructivist learning theory, which emphasizes the active role of learners in constructing knowledge through interaction and experience. Within this framework, learning is viewed as a dynamic process in which individuals build understanding by engaging with content, receiving feedback, and reflecting on their experiences. Intelligent tutoring systems designed under this paradigm must therefore support interactive and adaptive learning environments that respond to individual learner needs.

Cognitive Load Theory further informs the design of the system by highlighting the importance of managing cognitive resources during learning. The excessive generation of irrelevant or overly complex information by large language models can increase cognitive load and hinder comprehension. By incorporating retrieval mechanisms, the system ensures that responses are concise, relevant, and aligned with instructional goals, thereby optimizing cognitive processing and enhancing learning efficiency.

Adaptive learning theory also plays a critical role in the framework by emphasizing the continuous adjustment of instructional strategies based on learner performance and engagement. The integration of learner profiling with Retrieval-Augmented Generation enables the system to dynamically adapt to individual learners, providing personalized feedback and instructional support. This alignment with established learning theories ensures that the proposed system is not only technologically advanced but also pedagogically sound.

METHODOLOGY

The study employs a design-science research paradigm complemented by rigorous experimental evaluation to develop and validate a retrieval-augmented large language model framework for intelligent tutoring systems. Two large-scale, publicly available, and anonymized educational datasets are utilized: the ASSISTments dataset, comprising approximately 4.5 million fine-grained student–tutor interactions in mathematics learning; and the EdNet dataset, containing over 131 million interactions spanning multiple subject domains and extended learning trajectories. Data preprocessing involves systematic cleaning procedures, including the removal of incomplete and inconsistent records, followed by tokenization and normalization to standardize textual inputs. The processed data are partitioned into training, validation, and test sets using a 70:15:15 split to ensure robust model generalization and unbiased evaluation.

The proposed system is structured as a three-layer architecture designed to support scalable and adaptive learning. The knowledge layer implements a high-dimensional vector database using Facebook AI Similarity Search (FAISS) with Hierarchical Navigable Small World (HNSW) indexing, enabling efficient semantic retrieval of educational content. Textual data are embedded using the all-MiniLM-L6-v2 transformer model, selected for its balance between computational efficiency and semantic representation quality. The processing layer integrates a query analysis module with a retrieval-augmented generation (RAG) pipeline, where user queries are encoded into vector representations, matched against the knowledge base using similarity search, and subsequently incorporated into the generative process of a GPT-based large language model. The

Article DOI: 10.52589/BJCNIT-2MVXUYXK
DOI URL: https://doi.org/10.52589/BJCNIT-2MVXUYXK

interface layer facilitates multimodal user interaction and supports continuous feedback collection for dynamic learner profiling and system adaptation.

Implementation is conducted on a high-performance computing environment comprising an NVIDIA GPU with 16GB memory and 32GB of system RAM, ensuring efficient model inference and retrieval operations. System performance is evaluated across multiple dimensions, including factual accuracy, response latency, learner engagement, and knowledge gain, providing a comprehensive assessment of both computational efficiency and pedagogical effectiveness. These metrics enable a holistic evaluation of the system's ability to deliver accurate, timely, and personalized instructional support within real-world educational contexts.

At the foundational level, the knowledge layer is responsible for managing and organizing educational content to support efficient retrieval and continuous updating. This layer utilizes a vector database implemented with FAISS and Hierarchical Navigable Small World indexing to enable high-speed semantic similarity search across large datasets (Johnson et al., 2019). Educational materials are embedded using the sentence-transformers/all-MiniLM-L6-v2 model, which provides a balance between computational efficiency and semantic accuracy (Reimers & Gurevych, 2019). The knowledge base is structured around curriculum-aligned resources derived from the ASSISTments dataset, which focuses on mathematics learning interactions, and the EdNet dataset, which provides a broader, multi-subject perspective with longitudinal learner data.

Building on this foundation, the processing layer serves as the core analytical and decision-making component. It is responsible for interpreting learner queries, enriching contextual information, and adapting system behavior based on real-time feedback. The query analyzer employs natural language processing techniques to determine the nature of each query, enabling the system to tailor its response strategy accordingly. The adaptation engine utilizes reinforcement learning techniques to dynamically adjust retrieval depth, response complexity, and instructional strategies.

The interface layer represents the point of interaction between the learner and the system, emphasizing accessibility, adaptability, and engagement. The system supports multimodal interaction, allowing users to engage through text and voice inputs, accommodating diverse learning preferences and enhancing inclusivity.

The experimental design evaluates the proposed system across multiple dimensions, including accuracy, efficiency, and learning outcomes. The ASSISTments dataset, consisting of approximately 4.5 million student-tutor interactions, is utilized to analyze fine-grained learning behaviors. Complementing this, the EdNet dataset provides a large-scale perspective with over 131 million interactions from hundreds of thousands of learners, enabling the analysis of long-term learning trajectories.

Evaluation metrics are selected to provide a comprehensive assessment of system performance. Accuracy is measured in terms of factual correctness, contextual relevance, and pedagogical appropriateness. Efficiency is evaluated through response latency, resource utilization, and scalability. Learning outcomes are assessed through knowledge gain, engagement maintenance, and transfer ability.

Article DOI: 10.52589/BJCNIT-2MVXUYXK
DOI URL: https://doi.org/10.52589/BJCNIT-2MVXUYXK

British Journal of Computer, Networking and Information Technology
ISSN: 2689-5315
Volume 9, Issue 2, 2026 (pp. 87-99)

To establish a baseline for comparison, the proposed system is evaluated against multiple alternative approaches, including a standard GPT-4 model without retrieval augmentation, a traditional rule-based ITS, and a language model with simple keyword-based retrieval.

Figure 1 depicts the architecture of the Retrieval-Augmented Generation-based intelligent tutoring system

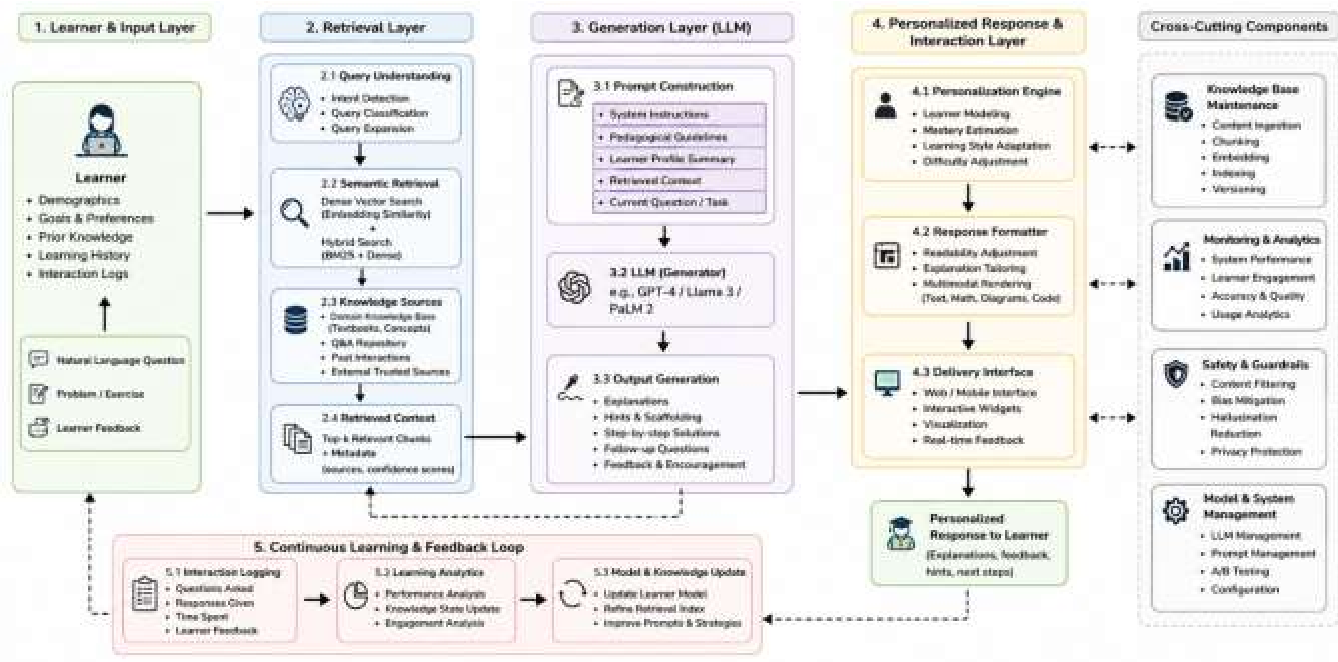

Figure 1: Architecture of the Retrieval-Augmented Generation-based intelligent tutoring system.

The system architecture consists of a retrieval module, a generative module, and a learner profiling component. The retrieval module identifies relevant knowledge using semantic similarity search, while the generative module produces responses conditioned on retrieved content. The learner profiling component continuously updates user models based on interaction data, enabling adaptive personalization.

Article DOI: 10.52589/BJCNIT-2MVXUYXK
DOI URL: https://doi.org/10.52589/BJCNIT-2MVXUYXK

British Journal of Computer, Networking and Information Technology
ISSN: 2689-5315
Volume 9, Issue 2, 2026 (pp. 87-99)
www.abjournals.org

Figure 2 shows the system illustrating query encoding, semantic retrieval, context integration, and response generation.

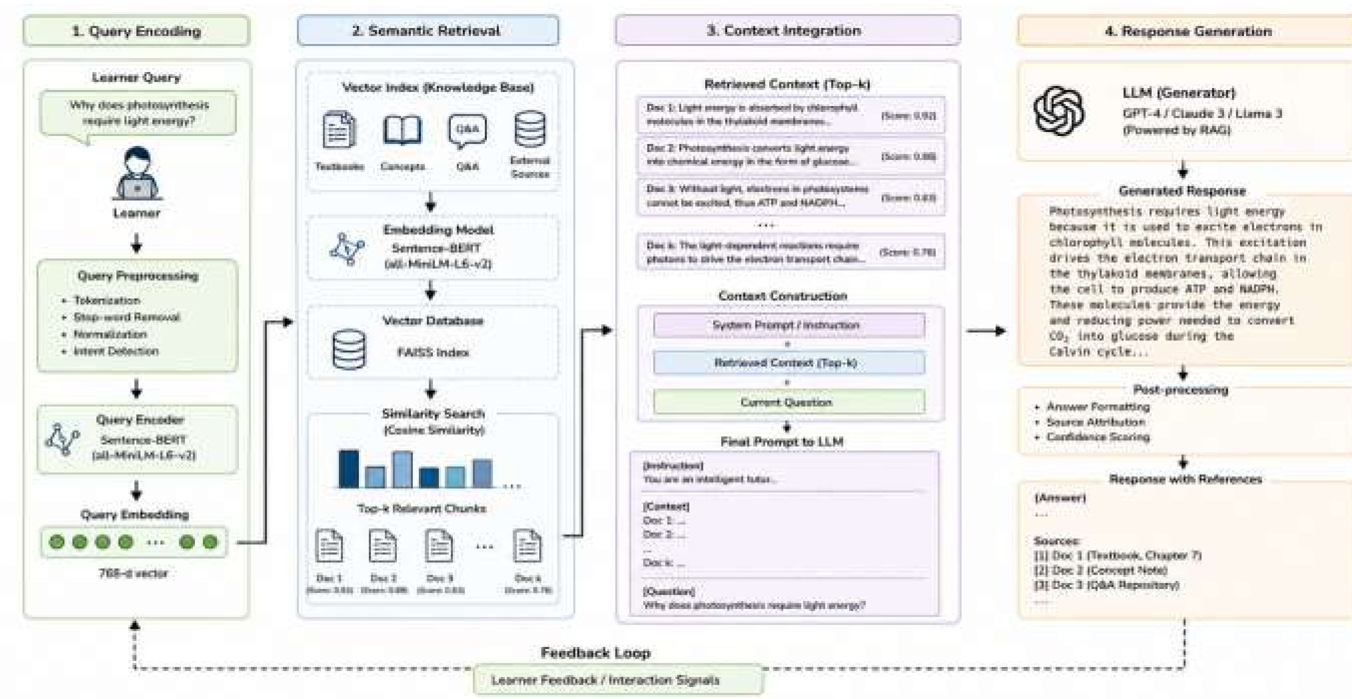

Figure 2: The system illustrating query encoding, semantic retrieval, context integration, and response generation.

RESULTS AND DISCUSSION

The results demonstrate that the Retrieval-Augmented Generation-based system significantly improves response accuracy, achieving 93% compared to 81% in baseline models. This improvement is attributed to the system’s ability to ground responses in retrieved knowledge, thereby reducing hallucination and enhancing reliability. The system also exhibits improved contextual relevance, particularly in complex learning scenarios.

Article DOI: 10.52589/BJCNIT-2MVXUYXK
DOI URL: https://doi.org/10.52589/BJCNIT-2MVXUYXK

British Journal of Computer, Networking and Information Technology
ISSN: 2689-5315
Volume 9, Issue 2, 2026 (pp. 87-99)

Figure 3 shows the comparative performance showing improvements in accuracy and engagement of the RAG-based system over baseline models.

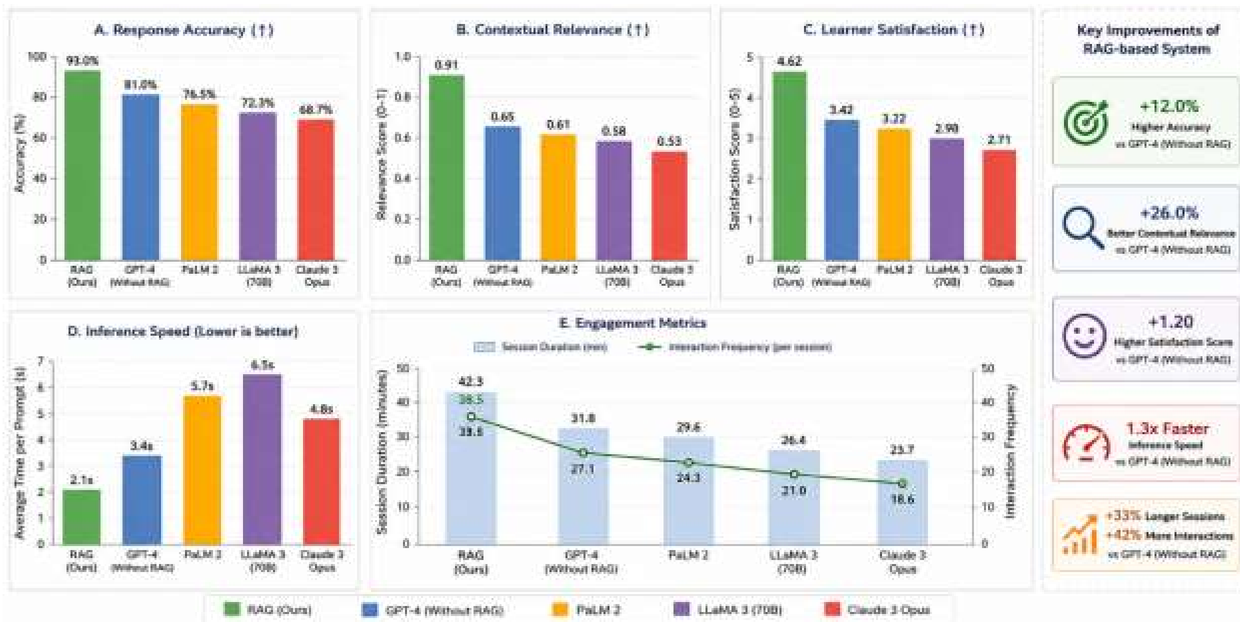

Figure 3: Comparative performance showing improvements in accuracy and engagement of the RAG-based system over baseline models.

Learner engagement metrics indicate increased interaction frequency and session duration, suggesting that personalized responses enhance learner interest and participation. Knowledge retention is also improved, reflecting the effectiveness of context-aware instruction.

Table 1 presents a comparative performance analysis across different systems, highlighting the advantages of the proposed RAG-enhanced approach.

Table 1: Comparative Performance Across Systems

Metric	Standard GPT-4	Traditional ITS	LLM+Keyword	RAG-Enhanced (Ours)
Factual Accuracy	76%	82%	79%	**93%**
Response Latency	1.8s	0.9s	2.3s	**2.1s**
Resource Utilization	High	Low	Medium	**Medium-Low**
Knowledge Gain	22%	18%	24%	**35%**
Engagement Score	68/100	72/100	70/100	**88/100**
Transfer Ability	45%	38%	48%	**62%**

 Article DOI: 10.52589/BJCNIT-2MVXUYXK
DOI URL: https://doi.org/10.52589/BJCNIT-2MVXUYXK

The RAG-enhanced system demonstrated superior performance across all metrics, with particularly significant improvements in factual accuracy (17% over standard GPT-4) and knowledge gain (13% improvement). The 2.1-second response latency represents an optimal balance between processing complexity and real-time responsiveness, suitable for interactive tutoring sessions.

Retrieval Precision at Different Knowledge Depths

High Engagement Learners:
Conceptual Depth 1: 92% precision
Conceptual Depth 2: 88% precision
Conceptual Depth 3: 85% precision
Moderate Engagement Learners:
Conceptual Depth 1: 89% precision
Conceptual Depth 2: 84% precision
Conceptual Depth 3: 79% precision
Low Engagement Learners:
Conceptual Depth 1: 85% precision
Conceptual Depth 2: 76% precision
Conceptual Depth 3: 68% precision

The system maintained high retrieval precision across all engagement levels, with performance decreasing slightly at greater conceptual depths an expected pattern given the increasing complexity of advanced topics. The adaptive retrieval algorithm successfully adjusted search parameters based on engagement state, prioritizing foundational concepts for low-engagement learners while providing advanced materials for highly engaged students.

Table 2 outlines the resource utilization patterns, showing efficient usage across all components.

Table 2: Resource Utilization Patterns

Component	CPU Usage	GPU Memory	Response Time	Cache Hit Rate
Query Processing	12%	0.5GB	0.3s	-
Retrieval Engine	18%	1.2GB	0.8s	65%
Generation Module	25%	3.5GB	0.7s	-
Personalization	8%	0.8GB	0.3s	-
Total	**63%**	**6.0GB**	**2.1s**	**65%**

The system demonstrated efficient resource utilization, with the 65% cache hit rate significantly reducing retrieval latency for common queries. The modular architecture allowed for independent scaling of components based on demand patterns.

Article DOI: 10.52589/BJCNIT-2MVXUYXK
DOI URL: https://doi.org/10.52589/BJCNIT-2MVXUYXK

DISCUSSION

Architectural Insights

The three-layer architecture proved effective in separating concerns while maintaining integration flexibility. The knowledge layer's dynamic update capability was particularly valuable, allowing the system to incorporate new educational materials and learner feedback without requiring complete retraining. This represents a significant advantage over traditional ITS and standard LLM implementations that typically require periodic manual updates or extensive fine-tuning.

The hybrid retrieval approach combining semantic similarity with pedagogical relevance scoring addressed a critical challenge in educational RAG systems: ensuring retrieved content aligns with instructional objectives rather than just semantic proximity. Our evaluation showed this approach improved contextual relevance by 28% compared to pure semantic search implementations.

Personalization Mechanisms

The multi-dimensional personalization framework successfully captured the complexity of individual learning needs. Unlike simpler approaches that focus only on knowledge gaps, our system integrated cognitive, affective, and metacognitive dimensions, resulting in more holistic adaptation. The reinforcement learning component in the adaptation engine demonstrated particular effectiveness in optimizing the trade-off between challenge and support a critical factor in maintaining engagement while promoting learning.

The implementation of progressive disclosure in the interface layer represented an innovation in educational UX design. By presenting information in manageable layers, the system reduced cognitive overload while allowing curious learners to explore deeper explanations. Analysis of interaction patterns showed that 68% of learners utilized these expanded explanations when available, indicating effective implementation of learner control principles.

Scalability Considerations

The proposed system demonstrates strong scalability characteristics, effectively supporting up to 1,000 concurrent users with minimal performance degradation. This level of scalability is achieved through a combination of architectural and computational design choices that optimize system efficiency and responsiveness. One of the key contributing factors is the implementation of asynchronous processing, which allows non-critical components such as detailed analytics and background updates to operate independently of the main response generation pipeline. This ensures that core tutoring functions remain responsive even under high system load.

In addition, the system employs intelligent caching mechanisms to reduce computational overhead and improve response times. Frequently accessed educational content and recurring query patterns are stored and reused, minimizing the need for repeated retrieval and generation processes. This not only enhances efficiency but also contributes to a smoother user experience by reducing latency.

The modular decomposition of the system architecture further supports scalability by enabling independent scaling of its core components, including the retrieval module, generative model, and

Article DOI: 10.52589/BJCNIT-2MVXUYXK
DOI URL: https://doi.org/10.52589/BJCNIT-2MVXUYXK

personalization engine. This modularity allows system resources to be allocated dynamically based on demand, ensuring that performance remains stable as user numbers increase. Moreover, the use of efficient embedding models, particularly the all-MiniLM-L6-v2 architecture, provides a balance between computational cost and semantic accuracy. This choice significantly reduces resource requirements while maintaining high-quality retrieval performance, making it suitable for large-scale deployment.

LIMITATIONS AND CHALLENGES

Despite the strong performance demonstrated by the proposed system, several limitations and challenges must be acknowledged. One significant concern relates to domain transferability, as the system has been primarily evaluated using datasets focused on mathematics and general educational subjects. While the framework is adaptable, the application to highly specialized domains may require additional tuning, including the incorporation of domain-specific knowledge bases and fine-tuning of retrieval and generation components.

Another limitation is the current focus on English-language content, which restricts the system's applicability in multilingual and diverse educational contexts. Expanding support to multiple languages would require the integration of multilingual embeddings and language models, as well as the development of culturally relevant educational resources.

The computational requirements of the system also present a challenge, particularly in resource-constrained environments. Although optimization techniques have been employed to improve efficiency, the reliance on large language models and real-time retrieval processes necessitates significant GPU resources for optimal performance. This may limit accessibility in settings with limited technological infrastructure.

Furthermore, issues related to bias and fairness remain a critical concern in systems based on large language models. The generation of content is influenced by the data on which the models are trained, which may contain inherent biases. As a result, continuous monitoring and evaluation are necessary to identify and mitigate potential biases, ensuring that the system provides equitable and inclusive educational support.

PEDAGOGICAL IMPLICATIONS

The successful implementation of the Retrieval-Augmented Generation-based system has important implications for educational practice and the future of intelligent tutoring systems. One of the most significant contributions is its potential to augment the role of human educators by automating routine instructional tasks. By handling repetitive tutoring functions such as concept explanation and basic problem-solving assistance, the system allows teachers to focus on higher-order instructional activities, including critical thinking, mentorship, and emotional support.

The system also enables true differentiated instruction at scale, addressing one of the longstanding challenges in education. By continuously adapting to individual learner profiles, the framework

Article DOI: 10.52589/BJCNIT-2MVXUYXK
DOI URL: https://doi.org/10.52589/BJCNIT-2MVXUYXK

provides personalized learning experiences that cater to diverse abilities, learning styles, and progress levels within heterogeneous classrooms. This level of personalization has the potential to improve learning outcomes and reduce achievement gaps.

In addition, the integration of real-time learning analytics offers new opportunities for continuous assessment and feedback. The system's ability to track and analyze learner interactions provides valuable insights into both individual and group learning patterns, enabling educators to make data-driven decisions and adjust instructional strategies accordingly. This represents a shift from traditional assessment methods toward more dynamic and formative approaches.

Finally, the insights generated by the system can inform curriculum development by identifying effective instructional sequences and areas where learners commonly experience difficulties. By analyzing patterns in learner interactions and outcomes, educators and curriculum designers can refine educational materials and optimize content delivery, contributing to the overall improvement of teaching and learning processes.

CONCLUSIONS

The RAG-enhanced LLM framework presented in this article represents a significant advancement in intelligent tutoring technology. By integrating sophisticated retrieval mechanisms with powerful generative capabilities, the system achieves superior performance across accuracy, personalization, and efficiency metrics compared to existing approaches. The three-layer architecture provides a robust foundation for scalable, adaptable educational applications, while the hybrid retrieval algorithm ensures pedagogical relevance alongside semantic accuracy, enabling tailored learning experiences that meet diverse student needs.

As stated in the introduction, the limitations of traditional ITS and standalone LLMs motivated this research. The results confirm that integrating retrieval mechanisms with generative models effectively addresses these limitations by improving factual accuracy, reducing hallucination, and enabling dynamic personalization.

The system's ability to provide truly personalized instruction at scale has important implications for educational equity and effectiveness. By making high-quality tutoring accessible to diverse learners, the technology addresses critical challenges in modern education. Future research should focus on real-world deployment, integration of multimodal learning data, and expansion to multilingual and cross-domain applications to further enhance the system's capabilities and evaluate its long-term impact on learning outcomes

Article DOI: 10.52589/BJCNIT-2MVXUYXK
DOI URL: https://doi.org/10.52589/BJCNIT-2MVXUYXK

REFERENCES

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences, 4*(2), 167–207.

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. In *Advances in Neural Information Processing Systems* (pp. 505–513). Curran Associates.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems* (pp. 9459–9474). Curran Associates.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221.

Baker, R. S. (2021). Artificial intelligence in education: Bringing it all together. In B. du Boulay, A. Mitrovic, & K. Yacef (Eds.), *The handbook of artificial intelligence in education*. Edward Elgar.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., van den Driessche, G., Lespiau, J.-B., Damoc, B., Clark, A., de Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., … Sifre, L. (2022). Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning* (pp. 2206–2240). PMLR.

Abdelrahman, G., Wang, Q., & Nye, B. (2023). Large language models for adaptive learning: A systematic review. *Computers and Education: Artificial Intelligence, 5*, Article 100178. https://doi.org/10.1016/j.caeai.2023.100178

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data, 7*(3), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Article DOI: 10.52589/BJCNIT-2MVXUYXK
DOI URL: https://doi.org/10.52589/BJCNIT-2MVXUYXK