International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 04 | Apr 2026 www.irjet.net p-ISSN: 2395-0072

A Survey of Retrieval-Augmented Generation Techniques for Intelligent Educational Assistants

Lakshya Singh[1], Er. Alok Singh Jadaun[2], Prof. Brajesh Kumar Singh[3]

[1]*Department of Computer Science and Engineering*
Raja Balwant Singh Engineering Technical Campus
Bichpuri, Agra, Uttar Pradesh, India
[2] *Department of Computer Science and Engineering*
Raja Balwant Singh Engineering Technical Campus
Bichpuri, Agra, Uttar Pradesh, India
[3] *Department of Computer Science and Engineering*
Raja Balwant Singh Engineering Technical Campus
Bichpuri, Agra, Uttar Pradesh, India

Abstract - *Retrieval-Augmented Generation (RAG) has emerged as a transformative approach for improving the factual reliability and contextual grounding of large language models, particularly in knowledge-intensive domains such as education. Standalone language model-based educational assistants frequently suffer from hallucinated responses, insufficient curriculum alignment, and limited explainability, which collectively impede their deployment in formal learning environments. This paper presents a systematic review of recent advances in Retrieval-Augmented Generation techniques, with a focused lens on intelligent educational assistants. Core RAG architectures—including naive, advanced, and modular designs—are examined alongside their educational applications spanning tutoring, examination support, and academic assistance. A structured comparative analysis of representative studies highlights critical design trade-offs in retrieval strategies, language model selection, and evaluation methodologies. The paper further delineates key open challenges related to evaluation standardization, knowledge base maintenance, scalability, pedagogical quality, and responsible deployment. Future research directions are outlined to advance the development of reliable, scalable, and learner-centric RAG-based educational systems.*

Key Words: **Retrieval-Augmented Generation, Large Language Models, Intelligent Educational Assistants, Artificial Intelligence in Education, Knowledge Grounding, Educational Chatbots, Information Retrieval**

1. INTRODUCTION

The rapid advancement of Large Language Models (LLMs) has fundamentally reshaped natural language processing, enabling sophisticated capabilities in question answering, text summarization, and conversational interaction [2]. Within education, LLM-powered assistants have attracted growing interest for their potential to deliver personalized tutoring, instant feedback, and scalable academic support. Despite these capabilities, standalone LLMs exhibit critical limitations including hallucinated responses, reliance on outdated training data, and absence of domain grounding—all of which pose acute risks in high-stakes educational settings.

Factual accuracy and curriculum alignment are nonnegotiable requirements for educational systems. Unlike general-purpose chatbots, educational assistants must deliver precise, explainable, and context-aware answers grounded in verified instructional materials such as textbooks and lecture notes. Since conventional LLMs generate responses from probabilistic patterns alone—without runtime access to external knowledge—they frequently produce confident yet incorrect answers [7], [18].

To overcome these limitations, Retrieval-Augmented Generation(RAG) has emerged as a principled framework that couples the generative fluency of LLMs with dynamic retrieval from external knowledge sources. By conditioning response generation on retrieved documents—such as course slides, research articles, and curated curricula—RAG-based systems achieve substantially improved factual accuracy, transparency, and domain relevance [2], [17]. This paper provides a comprehensive review of RAG techniques with a specific focus on intelligent educational assistants, examining core architectures, retrieval strategies, educational applications, comparative design choices, open challenges, and future directions.

2. BACKGROUND AND MOTIVATION

Educational question-answering systems have evolved through three broad phases: rule-based retrieval, statistical machine learning, and deep transformer-based models. Early assistants relied on keyword matching and static knowledge bases, offering high precision within constrained domains but lacking conversational flexibility and adaptability [4], [7].

The arrival of large-scale transformer models introduced substantially richer interaction patterns, yet their direct deployment in education introduced new problems. LLMs generate responses from learned statistical associations without verifying claims against authoritative sources, producing hallucinated or partially correct answers that can mislead learners and undermine trust [5], [13]. Furthermore, domain-specific curricula, learning objectives, and instructional sequencing constraints are not inherently encoded in general-purpose language models.

RAG directly addresses these limitations by enabling dynamic, inference-time retrieval from curated knowledge bases, thereby grounding generation in verifiable and curriculum aligned content. This combination preserves the fluency of LLMs while adding factual accountability, making RAG a foundational technology for trustworthy educational AI [2], [17], [21].

3. RETRIEVAL-AUGMENTED GENERATION (RAG)

Retrieval-Augmented Generation is a hybrid framework that enhances factual accuracy by integrating external knowledge retrieval into the language model generation pipeline. Rather than depending solely on parametric knowledge acquired during pre-training, RAG systems retrieve relevant documents from an external knowledge base at inference time and condition the generator on this retrieved context [2], [18].

A canonical RAG architecture comprises two core components. The retriever identifies and fetches the most semantically relevant passages from indexed sources—such as lecture notes, textbooks, or course materials—using dense vector embeddings stored in a vector database. Upon receiving a user query, the system encodes it into an embedding, performs similarity search, and retrieves the top-k relevant text chunks. The generator, typically a transformer-based LLM, then produces a response conditioned on both the original query and the retrieved context, yielding outputs that are linguistically coherent and grounded in traceable sources [1], [3].

RAG approaches span three levels of architectural sophistication. Naive RAG follows a straightforward pipeline: retrieve documents and append them directly to the prompt. Advanced RAG incorporates query rewriting, re-ranking, and multi-hop retrieval to improve contextual precision. Modular RAG extends this further by enabling flexible composition of specialized retrievers, memory modules, re-rankers, and generators, allowing the system to be customized for specific tasks or domains [8], [18]. Fig. 1 illustrates the general RAG workflow.

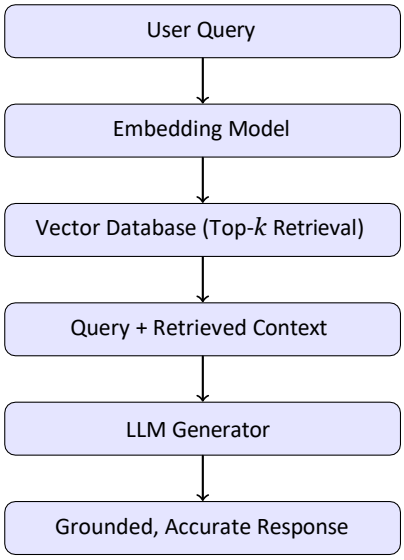

Fig -1: General workflow of a RAG-based intelligent educational assistant.

4. RAG IN EDUCATIONAL ASSISTANTS

The integration of RAG into intelligent educational assistants has gained considerable momentum due to its capacity to fulfill the accuracy and grounding requirements of learning environments. By leveraging curated academic resources— lecture slides, textbooks, research articles, and course-specific content—RAG-powered systems deliver context-aware and curriculum-aligned responses across diverse instructional scenarios including personalized tutoring, concept clarification, doubt resolution, and automated academic support [12], [17], [21].

Recent empirical work demonstrates RAG effectiveness across multiple educational domains. In intelligent tutoring systems, RAG enables step-by-step explanations by retrieving relevant instructional content prior to response generation [16], [19]. For examination preparation and formative assessment, RAG-based models generate practice questions and constructive feedback directly grounded in course material [9], [11]. Virtual teaching assistants in higher education employ RAG to handle student queries spanning lectures, assignments, and administrative information [13], [22].

A particularly notable advantage of RAG is that it enables compact, locally deployed language models to achieve performance levels comparable to large proprietary systems— substantially reducing computational cost while preserving data privacy [1], [11]. This scalability advantage makes RAG especially attractive for resource-constrained educational institutions. Nevertheless, effectiveness depends critically on knowledge base quality, coverage, and organization; poorly curated content can degrade retrieval performance and generation quality [10], [14].

From a student experience perspective, RAG-based systems offer a measurable improvement in transparency and trust. Because responses can be traced back to specific retrieved documents, learners can verify the source of information— a feature that distinguishes RAG from opaque standalone LLMs and aligns with academic integrity expectations in formal education. Studies such as Salminen et al. [23] and Alsafari et al. [21] highlight that students demonstrate higher satisfaction and trust when AI responses are linked to verifiable instructional materials, suggesting that source attribution is not merely a technical feature but a pedagogically meaningful design choice.

5. LITERATURE SEARCH METHODOLOGY

To ensure systematic coverage of the relevant literature, this review followed a structured search methodology consistent with established systematic review practices. Academic databases including Google Scholar, IEEE Xplore, and Scopus were queried using combinations of the following keywords: "Retrieval-Augmented Generation," "RAG," "Large Language Models in Education," "Intelligent Educational Assistants," "Educational Chatbots," "LLM tutoring systems," and "knowledge-grounded dialogue systems."

The search was bounded to publications from 2022 to 2025 to capture recent developments in this rapidly evolving field. An initial corpus of approximately 80 papers was identified. After removing duplicates and applying inclusion criteria— namely, papers explicitly addressing RAG architectures or LLM-based educational applications with empirical or architectural contributions—25 papers were retained for detailed analysis, of which 23 are directly cited in this review. Papers were excluded if they addressed LLMs without educational context, lacked sufficient architectural or evaluative detail, or appeared in non-peer-reviewed venues without substantial technical contribution. This process is consistent with established systematic literature review practices [4].

The retained studies span a range of educational domains including computer science education, statistics, healthcare, physics, and general university support. The distribution of publication venues includes ACM conferences (SIGCSE, IUI, CHI, L@S), Elsevier journals (Computers & Education, Procedia Computer Science), and Springer journals (Advanced Intelligent Systems, Information), ensuring a diverse and credible evidence base for the comparative analysis presented in this review.

6. COMPARATIVE ANALYSIS OF EXISTING APPROACHES

Existing RAG-based educational systems reveal significant variation across retrieval strategy, language model selection, domain focus, and evaluation methodology. Early systems paired simple dense retrieval with large proprietary models, while recent work increasingly prioritizes modularity, cost efficiency, and domain adaptability [2], [18]. Table I provides a structured comparison of representative studies across these dimensions.

Retrieval strategies range from traditional dense vector similarity search to more sophisticated hybrid and graph-based retrieval. With respect to language model configurations, a clear trend emerges toward locally deployed smaller models, motivated by privacy constraints and cost reduction [1], [11]. Evaluation

TABLE I

Comparative Analysis of RAG-Based Educational Assistants				
Study	**Domain**	**Retrieval Strategy**	**LLM Configuration**	**Evaluation Method**
Gao et al. [18]	General NLP	Dense Vector Retrieval	GPT (Proprietary)	ROUGE, Human Evaluation
Li et al. [17]	Education Apps	Hybrid Retrieval	Multiple LLMs	Accuracy, Relevance
Mishra et al. [1]	University Systems	Dense + Local Retrieval	Local LLM (Ollama)	Accuracy, Latency
Yigci et al. [13]	Higher Education	Dense Retrieval	GPT-4	User Study
Yu et al. [11]	CS Education	Dense Retrieval	Small LM (Local)	Student Feedback
Nemeth et al. [22]	Statistics	RAG Pipeline	GPT-based TA	Expert Assessment
Khan et al. [12]	University VA	Dense Retrieval	LLM + RAG Framework	Response Quality
Wang et al. [6]	Healthcare Edu.	RAG-Enhanced Pipeline	GPT-4	Clinical Accuracy
Tyndall et al. [9]	Exam Support	RAG + AI Agents	Multiple LLMs	Exam Performance
Cao [19]	CS1 Tutoring	Retrieval + ITS	LLM Tutor	Student Outcomes

methodologies remain inconsistent, spanning automated NLP metrics (BLEU, ROUGE, cosine similarity), human expert judgment, and user studies—reflecting a lack of standardized pedagogical benchmarks [10].

Several additional observations emerge from the comparative analysis. First, domain-specific systems consistently outperform general-purpose configurations when the knowledge base is well-curated and aligned with instructional objectives—underscoring the importance of knowledge engineering in educational RAG design. Second, systems that incorporate re-ranking mechanisms report notably improved retrieval precision compared to naive top-k approaches, suggesting that post-retrieval filtering is a worthwhile investment even at added computational cost [8], [18]. Third, evaluation
gaps are particularly pronounced in studies targeting K-12 education, where learner age, cognitive development stage, and curriculum constraints introduce evaluation dimensions absent from higher education settings [4], [14].

7. RESEARCH GAPS AND CHALLENGES

Despite notable progress, several critical gaps constrain the maturity of RAG-based educational systems. First, there is a striking absence of standardized evaluation frameworks tailored to educational contexts. Most studies employ generic NLP metrics that fail to capture pedagogical effectiveness, learning gains, or long-term engagement [10]. Without unified benchmarks, cross-study comparison remains unreliable. Metrics such as BLEU and ROUGE measure surface-level text similarity but cannot assess whether a response genuinely supports conceptual understanding, motivates the learner, or aligns with the instructional goal of a given lesson.

Second, knowledge base construction and maintenance present persistent challenges. Static, manually curated repositories do not scale across evolving curricula, diverse subjects, or multiple institutions [12], [22]. Inconsistent content structuring, missing pedagogical metadata, and inadequate alignment with learning objectives further degrade retrieval quality. Automated

knowledge base construction from raw course materials—including slides, transcripts, and problem sets—remains an open research problem requiring advances in both document parsing and semantic indexing.

Third, scalability remains an open concern. Retrieval latency over large corpora can disrupt real-time classroom interactions [9], while context window limitations restrict how much retrieved content can be effectively utilized. Balancing retrieval depth with response speed is particularly challenging for institutions with limited computational infrastructure [1], [11]. On-device or edge deployment of RAG systems, which would benefit privacy-sensitive educational environments, remains technically demanding and largely unexplored.

Fourth, pedagogical and ethical dimensions remain underexplored. Many systems optimize for factual accuracy while neglecting instructional quality—specifically, clarity of explanation, adaptability to learner proficiency, and support for critical thinking [14], [15]. Issues of bias in training data, overreliance on automated feedback, and transparency in retrieval decisions further underscore the need for human-in-the-loop RAG designs [5], [14]. The risk of learners uncritically accepting AI-generated content without developing independent reasoning skills is a particularly pressing concern that current RAG architectures do not adequately address.

8. FUTURE RESEARCH DIRECTIONS

Future research should advance RAG along three converging fronts. Technically, hybrid dense-sparse retrieval, graph-based knowledge representations [8], and multi-hop retrieval hold promise for improving contextual precision and handling complex educational queries. Graph-based retrieval, in particular, is well-suited for educational knowledge domains where concepts are inherently interconnected—for example, understanding calculus presupposes knowledge of algebra and limits, relationships that a graph structure can explicitly encode and exploit during retrieval. Domain-adapted embedding models trained on educational corpora can better align retrieval with curricular content [17].

Pedagogically, future RAG systems must embed learner centered design by adapting retrieved content and explanation style to individual learner profiles, prior knowledge levels, and learning objectives [15], [16]. Frameworks that incorporate scaffolding, formative feedback, and adaptive difficulty have strong potential to enhance learning outcomes [19], [20]. Human-in-the-loop mechanisms—allowing educators to guide and validate system outputs—can further balance automation with instructional accountability [21], [23]. The integration of learner modeling techniques, which dynamically track student knowledge state and misconceptions, into RAG retrieval pipelines represents a particularly promising direction for personalized educational AI.

Evaluatively, the field urgently needs standardized benchmarks that incorporate pedagogical indicators, learner performance outcomes, and longitudinal user studies, moving beyond conventional NLP metrics [4], [10]. Explainability features that link retrieved sources to generated responses will be essential for building trust with both learners and educators [5]. Responsible deployment principles—including bias mitigation, data governance, and ethical oversight—must be integrated into system design from the outset [5], [14]. Collaborative efforts between AI researchers, educational technologists, curriculum designers, and ethicists will be necessary to translate technical RAG advances into pedagogically sound, equitable, and sustainable educational tools.

9. CONCLUSION

This paper presented a systematic review of Retrieval-Augmented Generation techniques for intelligent educational assistants, synthesizing evidence from 23 recent studies spanning the period 2022 to 2025. The review demonstrated that RAG effectively addresses core limitations of standalone LLMs—particularly hallucination, domain misalignment, and lack of source transparency—making it a foundational paradigm for trustworthy educational AI. Core architectures (naive, advanced, and modular RAG), educational applications across tutoring, examination support, and virtual assistance, and comparative design trade-offs were

examined alongside a structured literature search methodology.

Open challenges in evaluation standardization, knowledge base management, scalability, and pedagogical quality were identified, with corresponding future research directions outlined spanning technical, pedagogical, and evaluative dimensions. A limitation of this review is its focus on English language publications from 2022 to 2025; multilingual and earlier foundational work may offer additional insights not captured here. Furthermore, the rapidly evolving nature of both LLM capabilities and RAG architectures means that findings may require periodic revision as the field advances.

Overall, RAG represents a promising and rapidly maturing technology for building reliable, learner-centric educational assistants. Its ability to ground responses in verifiable sources, adapt to domain-specific knowledge bases, and enable cost effective local deployment positions it as a particularly viable solution for diverse educational institutions worldwide— provided that technical advances are complemented by pedagogically informed design, rigorous evaluation, and responsible deployment practices.

REFERENCES

[1] A. Mishra and N. Brahmanapally, "A Comparative Performance Analysis of Locally Deployed Large Language Models Through a Retrieval-Augmented Generation Educational Assistant Application for Textual Data Extraction," AI, vol. 6, no. 6, p. 119, Jun. 2025, doi: 10.3390/ai6060119.

[2] W. Fan et al., "A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models," in Proc. 30th ACM SIGKDD, Barcelona, Spain: ACM, Aug. 2024, pp. 6491–6501, doi: 10.1145/3637528.3671470.

[3] H. Li, Y. Su, D. Cai, Y. Wang, and L. Liu, "A Survey on Retrieval-Augmented Text Generation," arXiv preprint arXiv:2202.01110, 2022.

[4] V. Liu, E. Latif, and X. Zhai, "Advancing Education through Tutoring Systems: A Systematic Literature Review," arXiv:2503.09748, Mar. 2025, doi: 10.48550/arXiv.2503.09748.

[5] D.-M. Cordova-Esparza, "AI-Powered Educational Agents: Opportunities, Innovations, and Ethical Challenges," Information, vol. 16, no. 6, p. 469, May 2025, doi: 10.3390/info16060469.

[6] D. Wang et al., "Enhancement of the Performance of Large Language Models in Diabetes Education through Retrieval-Augmented Generation: Comparative Study," J Med Internet Res, vol. 26, p. e58041, Nov. 2024, doi: 10.2196/58041.

[7] B. D. Nye, D. Mee, and M. G. Core, "Generative Large Language Models for Dialog-Based Tutoring: An Early Consideration of Opportunities and Concerns," in Proc. AIED Workshop on Empowering Education with LLMs, CEUR-WS, vol. 3487, pp. 78–88, 2023.

[8] B. Peng et al., "Graph Retrieval-Augmented Generation: A Survey," ACM Trans. Inf. Syst., vol. 44, no. 2, pp. 1–52, Feb. 2026, doi: 10.1145/3777378.

[9] E. Tyndall, C. Gayheart, A. Some, J. Genz, T. Wagner, and B. Langhals, "Impact of retrieval augmented generation and large language model complexity on undergraduate exams created and taken by AI agents," Data & Policy, vol. 7, p. e57, 2025, doi: 10.1017/dap.2025.10024.

[10] Z. F. Han et al., "Improving Assessment of Tutoring Practices using Retrieval-Augmented Generation," in Proc. AAAI Workshop on AI for Education, 2024.

[11] Z. Yu, S. Liu, P. Denny, A. Bergen, and M. Liut, "Integrating Small Language Models with Retrieval-Augmented Generation in Computing Education: Key Takeaways, Setup, and Practical Insights," in Proc. 56th ACM SIGCSE, Pittsburgh, PA: ACM, Feb. 2025, pp. 1302–1308, doi: 10.1145/3641554.3701844.

[12] U. H. Khan, M. H. Khan, and R. Ali, "Large Language Model based Educational Virtual Assistant using RAG Framework," Procedia Computer Science, vol. 252, pp. 905–911, 2025, doi: 10.1016/j.procs.2025.01.051.

[13] D. Yigci, M. Eryilmaz, A. K. Yetisen, S. Tasoglu, and A. Ozcan, "Large Language Model-Based Chatbots in Higher Education," Advanced Intelligent Systems, vol. 7, no. 3, p. 2400429, Mar. 2025, doi: 10.1002/aisy.202400429.

[14] D. Hooshyar et al., "Problems With Large Language Models for Learner Modelling: Why LLMs Alone Fall Short for Responsible Tutoring in K-12 Education," arXiv:2512.23036, Dec. 2025, doi: 10.48550/arXiv.2512.23036.

[15] Y. Zhang, E. L. Ouh, A. Ho, S. L. Lo, K. W. Tan, and F. Lin, "PromptTutor: Effects of an LLM-Based Chatbot on Learning Outcomes and Motivation in Flipped Classrooms," in Proc. 30th ACM ITiCSE, Nijmegen, Netherlands: ACM, Jun. 2025, pp. 445–451, doi: 10.1145/3724363.3729095.

[16] C. Xia, "Research on the Application of Large Language Models in Intelligent Tutoring System," in Proc. ICIAAI 2025, vol. 122, Dordrecht: Atlantis Press, 2025, pp. 993–1003, doi: 10.2991/978-94-6463-823-3_97.

[17] Z. Li, Z. Wang, W. Wang, K. Hung, H. Xie, and F. L. Wang, "Retrieval-augmented generation for educational application: A systematic survey," Computers and Education: Artificial Intelligence, vol. 8, p. 100417, Jun. 2025, doi: 10.1016/j.caeai.2025.100417.

[18] Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv preprint arXiv:2312.10997, 2023.

[19] C. Cao, "Scaffolding CS1 Courses with a Large Language Model-Powered Intelligent Tutoring System," in Proc. 28th ACM IUI, Sydney, Australia: ACM, Mar. 2023, pp. 229–232, doi: 10.1145/3581754.3584111.

[20] A. Lieb and T. Goel, "Student Interaction with NewtBot: An LLM-as-tutor Chatbot for Secondary Physics Education," in Ext. Abstracts CHI 2024, Honolulu, HI: ACM, May 2024, pp. 1–8, doi: 10.1145/3613905.3647957.

[21] B. Alsafari, E. Atwell, A. Walker, and M. Callaghan, "Towards effective teaching assistants: From intent-based chatbots to LLM-powered teaching assistants," Natural Language Processing Journal, vol. 8, p. 100101, Sep. 2024, doi: 10.1016/j.nlp.2024.100101.

[22] R. Németh, A. Tátrai, M. Szabó, and Á. Tamási, "Using a RAG-enhanced large language model in a virtual teaching assistant role: Experiences from a pilot project in statistics education," Hungarian Statistical Review, vol. 7, no. 2, pp. 3–27, 2024, doi: 10.35618/hsr2024.02.en003.

[23] J. Salminen et al., "Using Cipherbot: An Exploratory Analysis of Student Interaction with an LLM-Based Educational Chatbot," in Proc. ACM L@S 2024, Atlanta, GA: ACM, Jul. 2024, pp. 279–283, doi: 10.1145/3657604.3664690.