Archives available at journals.mriindia.com

International Journal of Recent Advances in Engineering and Technology

ISSN: 2347 - 2812

Volume 14 Issue 03s, 2025

A Review of Retrieval-Augmented Generation for University-Specific Chatbot Systems

[1]Syed Irfan Ali, [2]Hasan Laheri, [3]Sanchit Bhajikhaye, [4] M. Huzaifa Ansari, [5]M. Huzaif Ansari, [6]M. Bilal Khan

[1,2,3,4,5] Artificial Intelligence and Data Science
Anjuman College of Engineering and Technology
Nagpur, India

Peer Review Information

Submission: 05 Nov 2025

Revision: 25 Nov 2025

Acceptance: 17 Dec 2025

Keywords

Retrieval-Augmented Generation (RAG), Chatbots, University Systems, Large Language Models (LLMs) [3, 6], Educational Technology, AI in Higher Education.

Abstract

The rapid advancement of Artificial Intelligence (AI) and Natural Language Processing (NLP) has made Large Language Models (LLMs) pivotal in educational question-answering systems, particularly for university admission chatbots [1]. However, LLMs face critical challenges such as generating hallucinations, relying on outdated knowledge, and having non-transparent reasoning processes [10]. To address this, Retrieval-Augmented Generation (RAG) has emerged as a promising solution, incorporating knowledge from external databases to enhance the accuracy and credibility of generated responses [10]. This paper reviews the architecture and application of RAG-powered chatbots (RAGBots) designed for specific university domains [1]. A key finding is that while RAG systems like URAG, SAMCares, and Infersity v1 demonstrate utility in providing intelligent access to university resources [1, 3, 4], datasets for such closed domains are still difficult to obtain and curate [2]. Furthermore, complex RAG implementations often involve high operational costs and specialized modules [1]. The work highlights enhancements like Multi-Query and Ensemble Retrieval [6] and discusses critical challenges such as Document-Level Retrieval Mismatch (DRM) [8], concluding with a vision for reliable, domain-specific RAGBots in higher education.

Introduction

The university sector increasingly relies on digital services, applications, and modern technologies to attract, retain, and engage students. Online universities especially have a strong motivation to adopt tools like chatbots to ensure faster support, improved service availability, and better communication with students [2]. However, despite the availability of digital resources, university websites and official documents such as Undergraduate Rules and the Prospectus are often difficult to navigate. Students frequently struggle to locate specific details such as admission procedures or scholarship information, which slows down access to important resources [3]. Moreover, conventional student service departments do not operate round-the-clock, resulting in delays and sometimes frustration among students when support is needed outside working hours [3].

The university sector increasingly relies on digital services, applications, and modern technologies to attract, retain, and engage students. Online universities especially have a strong motivation to adopt tools like chatbots to ensure faster support, improved service availability, and better communication with students [2]. However, despite the availability of digital resources, university websites and official documents such as Undergraduate Rules and the Prospectus are

often difficult to navigate. Students frequently struggle to locate specific details such as admission procedures or scholarship information, which slows down access to important resources [3]. Moreover, conventional student service departments do not operate round-the-clock, resulting in delays and sometimes frustration among students when support is needed outside working hours [3].

To overcome these challenges, Retrieval-Augmented Generation (RAG) has emerged as a powerful approach that enhances LLM performance by retrieving relevant document chunks from an external knowledge base before generating a response [10]. By integrating real-time data retrieval with generative capabilities, RAG enables more reliable and context-aware outputs tailored to specific domains such as university admissions, academic queries, examination guidelines, and scholarship assistance [1]. This combined architecture not only strengthens accuracy but also supports continuous knowledge updates without retraining the entire model, making it a practical choice for academic institutions seeking automation-based support systems [10].

The main objective of this review is to explore the concept of RAG and highlight its application in developing intelligent RAGBot systems for universities. Such systems can potentially streamline student interactions, reduce manual workload for staff, and ensure that information remains accessible at any time, ultimately improving the overall student experience.

The Evolution of Chatbots in Education

The idea of conversational agents, commonly referred to as chatbots, traces back to the Turing Test introduced in the 1950s, with one of the earliest examples being Eliza, a rule-based text interaction system [2]. Over the decades, the field of Artificial Intelligence has advanced significantly, particularly with the rapid growth of Deep Learning techniques. This progress was made possible through increased computational power and the availability of large datasets from the early 2000s onward [2].

In the academic environment, early chatbot systems were primarily utilized for administrative tasks and basic teaching or learning assistance [2]. As AI technologies evolved, universities began exploring more intelligent systems capable of assisting students with complex queries. However, even with strong Question Answering (QA) abilities, modern Large Language Models (LLMs) often face challenges when dealing with domain-specific information or queries requiring precise institutional knowledge [5]. In educational settings, accuracy is crucial, as incorrect information can mislead students and disrupt academic processes. This makes relying solely on LLMs risky without the support of strategies designed to reduce misinformation [1].

RAG-based chatbots have emerged as a solution to these limitations by grounding LLM responses in external, trusted knowledge sources [5]. Instead of generating answers purely from pre-trained data, RAG integrates university-specific documents, rules, and records directly into the response-generation pipeline [1]. This ensures that responses are not only fluent and human-like but also relevant, verifiable, and aligned with institutional guidelines. As a result, RAG technology represents a significant step toward more reliable academic assistance systems, supporting universities in delivering accurate information to students while reducing staff workload and improving accessibility.

Overview of Retrieval-Augmented Generation (RAG)

RAG has emerged as a key technology that bridges the internal reasoning ability of Large Language Models with dynamic, externally stored data repositories [10]. This approach enables models to produce responses that are not only fluent, but also grounded in factual information. A complete overview of RAG architecture generally includes three major paradigms: Naive RAG, Advanced RAG, and Modular RAG, each differing in complexity and retrieval strategies [10]. In most literature, the RAG mechanism is analyzed through a three-stage pipeline consisting of retrieval, generation, and augmentation processes [10].

A. The Naive RAG typically follows a process that includes three main steps [10]:

- Indexing: Raw documents such as PDF or HTML files are cleaned, chunked into smaller text units, converted into embeddings, and stored in a vector database [10].
- Retrieval: When a query is submitted, the system searches the vector store and retrieves the Top-k most relevant chunks through semantic similarity matching [10].
- Generation: The retrieved chunks are then passed along with the user query to the LLM, which generates a final answer based on the combined information [10].

Although Naive RAG significantly improves relevance and contextual grounding by supplementing LLMs with external knowledge [7], it is not without drawbacks. As the volume of indexed data increases, retrieval may become less precise, causing information dilution, irrelevant

context mixing, or even incorrect model assumptions often referred to as hallucinations [7]. Furthermore, noisy retrieval and large-scale document searches can introduce latency and reduce system responsiveness, limiting the reliability of traditional RAG setups in real-time university applications [5].

The RAGBot Framework for University Systems

A RAGBot system designed for university environments operates on a conceptual architecture where the RAG framework is adapted specifically to institutional data and student-centric queries. The system retrieves information from university documents, policies, and web resources, ensuring that responses are generated directly from the designated knowledge corpus rather than relying solely on the model's internal memory [7]. This makes the chatbot more reliable for academic tasks such as admission queries, examination guidelines, fee structures, and scholarship information.

The data pipeline for such systems typically begins with the collection, cleaning, and processing of university-related information. In implementations like Infersity v1, institutional knowledge is gathered using web scraping to extract structured content from official university websites and documents [3]. Once collected, the data undergoes preprocessing, where chunking the process of dividing text into semantically meaningful segments plays a crucial role. The quality of chunking significantly impacts retrieval accuracy and overall system performance [9]. Ineffective chunking can lead to confusion between similar documents, especially in large academic repositories, resulting in a retrieval error known as Document-Level Retrieval Mismatch (DRM), where information is selected from the wrong document entirely [8].

To improve retrieval precision and reduce query ambiguity, several enhancements have been proposed in recent literature. One notable approach is Summary-Augmented Chunking (SAC), which attaches a synthetic summary to each chunk, providing high-level context to the retriever [8]. This approach retains global meaning that would normally be lost during standard segmentation, thereby reducing DRM and improving chunk relevance during retrieval. SAC ultimately supports more robust and context-aware RAG pipelines, making university chatbots more dependable in real-time usage and improving the overall student experience.

Case Studies and Implementations

Several RAG-based educational systems have been developed to meet the evolving needs of universities and enhance student support services:

- URAG (Unified RAG): A specialized framework designed for university admission chatbots using a Unified Hybrid RAG approach [1]. It was implemented and evaluated at Ho Chi Minh City University of Technology (HCMUT), demonstrating its capability to provide reliable responses regarding admissions and academic counseling [1].
- INFERSITY V1: A RAG-powered assistant aimed at improving student access to general university resources [3]. By integrating an LLM (Gemini) with RAG, the system addresses common issues such as delays in obtaining accurate information on departments, programs, and campus facilities [3].
- SAMCares: Proposed as an Adaptive Learning Hub, SAMCares serves as a pilot framework for integrating AI into higher education environments [4]. It combines a Retriever-Augmented component with LLaMa-2 70B as the core LLM to support academic engagement and guidance [4].

These case studies collectively emphasize how RAG-based systems are increasingly being adopted to compensate for the limitations of standalone LLMs, particularly in tasks requiring precise, domain-specific information. They demonstrate a clear shift toward intelligent, knowledge-grounded chatbots capable of improving the student experience, reducing administrative workload, and enabling more accessible university information services.

Enhancements and Innovations in RAGBots

The ongoing development of RAG systems aims to enhance Question Answering (QA) accuracy and address the performance gaps found in traditional LLM-driven approaches [7]. With academic datasets often being large, diverse, and context-sensitive, researchers have proposed several optimization strategies to improve retrieval precision, reasoning ability, and overall responsiveness of RAG models when applied within university environments [6]. Some notable advancements include:

- Advanced Retrieval Methods: Various retrieval optimization techniques including Multi-Query Retrieval, Child-Parent-Retriever, and Ensemble Retriever have been implemented to improve the relevance of retrieved information, particularly in large technical universities where study programs and guidelines are extensive and complex [6]. These methods help refine the context passed to the generator, reducing irrelevant

information and improving final answer quality.

- Domain-Specific Fine-Tuning and Optimization: The Select2Know (S2K) framework offers a cost-efficient solution that enhances RAG systems by allowing them to self-select relevant internal and external knowledge sources [5]. It uses supervised fine-tuning targeted specifically at domain reasoning tasks, supported by a structured data generation pipeline to elevate the reasoning capability of the LLM [5].
- Contextual Optimization: The QuIM-RAG (Inverted Question Matching RAG) architecture introduces a new way of organizing information by transforming the corpus into a domain-focused dataset, enabling improved QA performance [7]. By matching questions inversely rather than purely through document chunks, QuIM-RAG addresses retrieval precision issues and reduces common RAG bottlenecks.
- In-Context Learning (ICL): ICL techniques are also used to further boost RAG performance by allowing models to learn from contextual examples without additional training, making the system more adaptable to diverse academic queries [6].

Challenges and Limitations

Despite the significant promise of RAG, several challenges persist, particularly in the university domain:

- Data-Related Issues: Closed-domain academic datasets, such as internal university documents, remain difficult to collect, organize, and maintain reliably [2]. Frequent updates to policies, announcements, or curriculum structures demand continuous data pipeline maintenance, which increases system complexity.
- Technical and Operational Limitations: Deploying advanced RAG configurations often requires high computational resources and specialized training procedures, making real-world adoption costly and technically demanding [1]. Furthermore, traditional RAG setups can experience latency when dealing with large volumes of data or noisy retrieval results [5].
- Retrieval Accuracy: The effectiveness of a RAG pipeline heavily relies on the retriever's ability to fetch the most relevant context [8]. In universities with extensive and structurally similar records, retrieval failures can occur, resulting in Document-Level Retrieval Mismatch (DRM) where retrieved context originates from an entirely different document [8]. This directly impacts answer correctness and system reliability.
- LLM-Inherited Challenges: Although RAG reduces hallucination risks, it cannot fully eliminate core weaknesses of LLMs, such as outdated knowledge, fabricated responses, and unclear reasoning pathways [10]. Additionally, information dilution may occur when the retrieved context is too broad or insufficiently filtered [7].

Future Directions and Research Opportunities

Future research is focused on developing more robust and reliable RAG systems for higher education:

- Advanced Evaluation Methodologies: As the use of RAG expands, the need for refined evaluation strategies becomes evident [10]. Recent work proposes the RAG Confusion Matrix, a novel assessment method designed to evaluate different RAG configurations more systematically [6]. Additional evaluative approaches such as RAGAs also contribute to performance scoring and benchmarking [3]. Continued research on standardized evaluation metrics is essential to ensure that RAG-enhanced LLMs are measured fairly, consistently, and against evolving academic requirements [10].
- Improving Retrieval Reliability: Enhancing retrieval accuracy remains a core research focus, especially for large structured university corpora prone to contextual overlap. Techniques like Summary-Augmented Chunking (SAC) show potential in reducing Document-Level Retrieval Mismatch (DRM) and preserving global document context [8]. Future work may further investigate hybrid chunking strategies, adaptive retrieval filters, and semantic routing mechanisms to strengthen retrieval precision at scale.
- Scalable Domain Knowledge Integration: As universities continuously update policies, curricula, and administrative workflows, future research must address real-time data synchronization and low-overhead knowledge updates, enabling RAGBots to remain accurate without frequent retraining.

- LLM-Level Improvements: Research may also explore integrating domain-aligned reasoning, reducing hallucinations even further, and enabling transparent answer traceability for more reliable academic assistance.

Conclusion

RAG continues to emerge as a powerful solution for addressing core limitations of LLMs, including hallucination, outdated knowledge, and difficulties handling domain-specific queries, making it a promising foundation for next-generation university chatbot systems [1, 10]. The review of existing implementations such as URAG, Infersity v1, and SAMCares demonstrates the real-world feasibility of RAGBots and their potential to deliver intelligent, round-the-clock access to academic information [1, 3, 4]. Despite its advantages, challenges such as closed-domain data collection and maintenance [2] and the operational complexities associated with advanced hybrid RAG architectures [1] remain areas of active concern.

However, rapid developments in the field including retrieval enhancements like Multi-Query optimization [6], advanced structures such as the QuIM-RAG architecture [7], and domain-driven fine-tuning frameworks such as Select2Know (S2K) [5] show that solutions are progressing steadily. Furthermore, emerging evaluation strategies like the RAG Confusion Matrix offer new ways to measure reliability and help benchmark system improvements [6]. With continuous research and refinement, RAG-powered chatbots are poised to become reliable, domain-aware assistants capable of transforming how students access information in higher education.

Acknowledgment

The authors wish to sincerely thank Anjuman College of Engineering and Technology, Nagpur, for providing the academic environment and resources necessary to conduct this work. We are especially grateful to the Department of Artificial Intelligence and Data Science for their continuous guidance, encouragement, and constructive feedback throughout the preparation of this review. We also extend our appreciation to our peers and colleagues, whose discussions and suggestions helped refine our ideas and direction. Finally, we acknowledge the broader research community for their ongoing contributions to the field of Retrieval-Augmented Generation, which formed the foundational basis of this study.

References

Nguyen, L., & Quan, T. (2025). URAG: Implementing a Unified Hybrid RAG for Precise Answers in University Admission Chatbots A Case Study at HCMUT. arXiv:2501.16276v1 [cs.CL].

Peyton, K., Unnikrishnan, S., & Mulligan, B. (2025). A review of university chatbots for student support: FAQs and beyond. Discover Education.

Jahangir, M. T., Hussain, A., Khan, M. H., Khalil, M., & Nonari, M. F. E. (2025). INFERSITY V1: A RETRIEVAL-AUGMENTED GENERATION (RAG) BASED CHATBOT FOR INTELLIGENT ACCESS TO UNIVERSITY RESOURCES. Spectrum of Engineering Sciences, 3(7).

Sunkara, S. P. (2025). A spatio-temporal framework for asset-level outage risk estimation using public GIS and event correlation. International Journal of Computer Engineering and Technology, 16(1), 4211–4227. https://doi.org/10.34218/IJCET_16_01_286

Faruqui, S. H. A., Tasnim, N., Basith, I. I., Obeidat, S., & Yildiz, F. (2024). Integrating A.I. in Higher Education: Protocol for a Pilot Study with 'SAMCares: An Adaptive Learning Hub'. arXiv:2405.00330v1 [cs.CY].

Sharma, B. (2024). Deep learning–based intelligent chatbot framework for academic assistance systems. International Journal of Computer Engineering and Technology, 15(6), 112–124. https://doi.org/10.34218/IJCET.15.6.2024.012

Sharma, B. (2023). Artificial intelligence–driven conversational agents for higher education applications. International Journal of Advanced Research in Computer Science and Software Engineering, 13*(*4), 45–53.

He, B., He, X., Shao, R., Cheng, M., Li, H., Shu, S., Xue, X., & Ling, Z.-H. (2025). Select to Know: An Internal-External Knowledge Self-Selection Framework for Domain-Specific Question Answering. arXiv:2508.15213v1 [cs.CL].

Afzal, A., Vladika, J., Fazlija, G., Staradubets, A., & Matthes, F. (2024). Towards Optimizing a Retrieval Augmented Generation using Large Language Model on Academic Data. arXiv:2411.08438v1 [cs.AI].

Hazarika, I., Khalfan, J., Ahmed, M., Yousif, A., & Hussain, J. (2024). Role of fintech as an enabler to fulfill HR requirements and attain sustainability. In A. Hamdan & A. Harraf (Eds.), Business development via AI and digitalization (Vol. 537,

pp. 59–69). Springer. https://doi.org/10.1007/978-3-031-62106-2_5

Jumde, A., Hazarika, I., & Cho, B. Y. (2019). Blockchain technology: A new enabler of financial services. In Proceedings of the 2019 Sixth HCT Information Technology Trends (ITT) (pp. 259–263). IEEE. https://doi.org/10.1109/ITT48889.2019.9075091

Saha, B., Saha, U., & Malik, M. Z. (2024). QuIM-RAG: Advancing Retrieval-Augmented Generation With Inverted Question Matching for Enhanced QA Performance. IEEE Access.

Reuter, M., Lingenberg, T., Liepiņa, R., Lagioia, F., Lippi, M., Sartor, G., Passerini, A., & Sayin, B. (2025). Towards Reliable Retrieval in RAG Systems for Large Legal Datasets. arXiv:2510.06999v1 [cs.CL].

Ferraris, A. F., Audrito, D., Siragusa, G., & Piovano, A. (2024). Legal Chunking: Evaluating Methods for Effective Legal Text Retrieval. Legal Knowledge and Information Systems.

Jumde, A., Hazarika, I., & Akre, V. (2023). Challenges and opportunities in integrating rapidly changing technologies in business curriculum. In Proceedings of the 2023 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) (pp. 203–208). IEEE. https://doi.org/10.1109/ICCIKE58312.2023.10131683

Patil, R. V., Gaidhani, V. A., Kashid, P. V., Hazarika, I., Mahadik, R. V., Poddar, G. M., & Patila, S. R. (2025). Decentralized autonomous organizations as emerging economic entities in accounting and governance frameworks. *International Journal of Accounting and Economics Studies, 12*(4), 166–177. https://doi.org/10.14419/1sy2j677

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997v5 [cs.CL].