Brilliance
Research of Artificial Intelligence

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.6719

Development of an Academic Services Chatbot Based on Retrieval-Augmented Generation (RAG)

Mohammad Labib Husain[1], Yudi Wibisono[2*], Ani Anisyah[3]
[1,2,3]Universitas Pendidikan Indonesia, Indonesia
[1]labibhusain14@upi.edu, [2] yudi@upi.edu, [3]anianisyah@upi.edu

open access

***Corresponding Author**

Article History:
Submitted: 31-07-2025
Accepted: 09-08-2025
Published: 12-08-2025

Keywords:
Academic Services; AI assistant; Chatbot; Large Language Model; Retrieval-Augmented Generation; RAGAS;

ABSTRACT

Higher education institutions struggle to provide accurate and accessible academic information. Traditional chatbots are often limited in capability, while standard Large Language Models (LLMs) pose a significant risk of factual "hallucinations," rendering them unsuitable for official university use where trustworthiness is paramount. This study aims to increase the accessibility and effectiveness of academic services by developing a trustworthy chatbot. The primary objective is to implement the Retrieval-Augmented Generation framework to create a reliable AI assistant that is factually grounded in a verified, domain-specific knowledge base. A knowledge base was constructed from official FPMIPA UPI documents and structured using hierarchical chunking. The system employs a multi-stage RAG pipeline featuring query contextualization and reranking before generation with Gemini 2.5 Pro. Performance was evaluated using metrics from the RAGAS framework on a 100-question dataset categorized into factual, reasoning, and out-of-context queries. The evaluation revealed strong performance on factual queries, achieving a Faithfulness score of 0.9100. A significant performance decrease was observed for reasoning tasks, with Context Recall dropping to 0.5926. Crucially, the system successfully handled 81.25% of out-of-context questions by correctly refusing to answer, thereby effectively preventing hallucination. The RAG framework provides a viable and empirically-validated blueprint for creating a trustworthy conversational AI for higher education. The model proves to be an effective tool for factual information delivery and has strong potential to modernize how student support and academic services are delivered.

INTRODUCTION

Within the contemporary higher education sector, the imperative for digital transformation has become increasingly evident, driven by relentless technological advancements and their profound societal influence (Singun, 2025). Established communication methods like telephone calls and emails frequently prove ineffective, often resulting in unsatisfactory responses from both administrative staff and faculty. Therefore, deploying AI-powered chatbot platforms emerges as a promising strategy for significantly enhancing student support services (Mahlatse, 2024).

Chatbots enable students to easily access important information such as admission procedures, scholarships, and tuition fees, without needing to visit information counters or wait for staff availability (Okonkwo, 2021; Rahim et al., 2022). Over the past few years, leading colleges and universities worldwide have adopted AI-based chatbots on their websites to serve as 24/7 virtual guides, offering a cost-effective alternative to hiring multiple staff members (Rahim et al., 2022). These tools not only improve service delivery but also help reduce the repetitive workload on administrative staff, allowing them to focus on more complex tasks.

Initially, non-machine learning approaches such as rule-based systems were the first attempts to build chatbots due to their simplicity. These systems rely on pattern-matching rules to generate responses but often produce inaccurate answers due to their limited contextual understanding (Singh & Namin, 2025). The inherent limitations of these early systems created a clear need for more advanced conversational models, driving the shift towards the technologies discussed next.

The answer to this need emerged with the recent advancement of Large Language Models (LLMs). These are natural language processing (NLP) models trained on massive amounts of text to develop a broad general knowledge applicable to various fields (Brown et al., 2020), enabling the creation of far more intelligent and capable conversational agents than their rule-based predecessors. However, deploying these powerful models directly for institutional purposes presents critical challenges. The most significant is the risk of "hallucination," where the model generates plausible but incorrect or fabricated information (Ji et al., 2023). Furthermore, their nature as general-purpose models means they lack knowledge of specific, private, or real-time institutional data, making them unreliable for providing information that must be precise and trustworthy, such as an official academic policy or a specific curriculum deadline.

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.6719

This study identifies a clear research gap: the need for a chatbot architecture that combines the advanced conversational abilities of LLMs with the factual accuracy required for a specialized academic domain. The Retrieval-Augmented Generation (RAG) model offers a state-of-the-art solution to this problem (Lewis et al., 2020). This approach enhances an LLM by first retrieving relevant information from a verified, domain-specific knowledge base and then using that context to generate a factually grounded response, effectively mitigating the risk of hallucination (Gao et al., 2023). Therefore, this research focuses on implementing this framework to develop a chatbot for academic administration services, which is expected to increase accessibility and effectiveness in obtaining academic information. This study applies the RAG architecture within the specific context of the Faculty of Mathematics and Natural Sciences Education (FPMIPA) at Universitas Pendidikan Indonesia (UPI). The primary contribution of this work is the creation of a robust and reliable chatbot model that provides a practical framework for other educational institutions to leverage advanced AI for effective service automation.

LITERATURE REVIEW

Large Language Models (LLMs)

The advent of Large Language Models (LLMs) represents a paradigm shift in natural language processing. Built upon the Transformer architecture, which leverages self-attention mechanisms to process language with unprecedented contextual awareness (Vaswani et al., 2017), LLMs have demonstrated remarkable capabilities in generating fluent, coherent, and human-like text across a wide range of tasks (Zhao et al., 2023). Furthermore, their integration into everyday applications has been made possible by the easy and affordable access to LLMs—through web APIs, sandbox environments, and open-source toolkits. This accessibility has empowered a new generation of developers and researchers to experiment with and implement modern LLM-based AI at a scale and speed previously unseen (Erickson, 2025).

However, despite their power and widespread adoption, LLMs suffer from critical limitations when deployed in domain-specific, knowledge-intensive applications. The most significant of these is "hallucination," the tendency to generate factually incorrect or nonsensical information with high confidence (Ji et al., 2023). This, combined with their inherent knowledge cutoff (being unable to access information beyond their last training date), makes them unreliable for tasks requiring factual accuracy and real-time data (Dhingra et al., 2022), such as providing academic administrative information.

Retrieval-Augmented Generation (RAG)

To address the limitations of LLMs, the Retrieval-Augmented Generation (RAG) framework was introduced (Lewis et al., 2020). According to a recent survey, the workflow of RAG is generally understood as a three-stage process. The first stage is retrieval, where relevant information is sourced from external databases based on the given input. Following this, the second stage is fusion, where the retrieved information is integrated with the original input or the model's intermediate states. Finally, the third stage is generation, where a generator model synthesizes a final answer based on the combined input and the corresponding retrieved data (Wu et al., 2024).

Advancements in RAG Architectures

The field has rapidly evolved beyond the classical RAG framework, with a primary focus on enhancing the quality of the retrieved context through two main strategies: pre-retrieval and post-retrieval. In the pre-retrieval stage, techniques such as indexing optimization (e.g., finer-grained segmentation or adding metadata) and query transformation (like rewriting or expanding the user's question) are applied to improve the quality of the search signal. Subsequently, in the post-retrieval stage, the retrieved information is further processed before being passed to the LLM, for instance, by reranking the documents to highlight the most relevant information or by compressing the context to remove noise and irrelevant content (Gao et al., 2022).

RAGAS: A Framework for Evaluating RAG Systems

Evaluating the performance of a RAG system is a complex task, as traditional metrics often fail to capture its multi-dimensional nature. To address this, this study utilizes metrics inspired by frameworks like RAGAS (Retrieval-Aided Generation Assessment), which offers a comprehensive suite for evaluating RAG pipelines (Es et al., 2023). The evaluation focused on two key aspects: the quality of the final generated answer and the effectiveness of the retrieval component.

For the quality of the generated answer, several metrics were employed. Faithfulness was used to measure the factual consistency of the response against the retrieved context, ensuring the model does not invent information. Answer Relevancy assessed how pertinent the generated answer was to the original query. To compare the response against a benchmark, Answer Semantic Similarity was calculated; this metric evaluates the semantic resemblance between the generated answer and a ground truth answer using a bi-encoder model. Furthermore, Factual Correctness was used to evaluate the factual accuracy of the response compared to the ground truth reference. This is achieved by using an LLM to first break down the response and reference into individual claims and then using natural language inference to determine the factual overlap between them.

Crucially, the effectiveness of the retrieval component itself was also evaluated. This was measured using Context Recall, which ensures that all the necessary information from the knowledge base required to answer the query has been successfully retrieved. Together, this suite of metrics provides a robust methodology for the multi-faceted evaluation of the RAG system developed in this research, covering both the accuracy of the retrieval process and the quality of the final generated response.

METHOD

This section details the systematic methodology for the development and evaluation of the RAG-based academic chatbot. The process is segmented into several key stages: data collection and preprocessing, knowledge base structuring and indexing, the retrieval and generation pipeline, conversational memory implementation, evaluation, and user interface development. The overall methodology of this research, as depicted in the system architecture in Figure 1, is divided into six main stages, starting from data collection and preprocessing to the final user interface development. Each stage will be explained in detail in the following subsections.

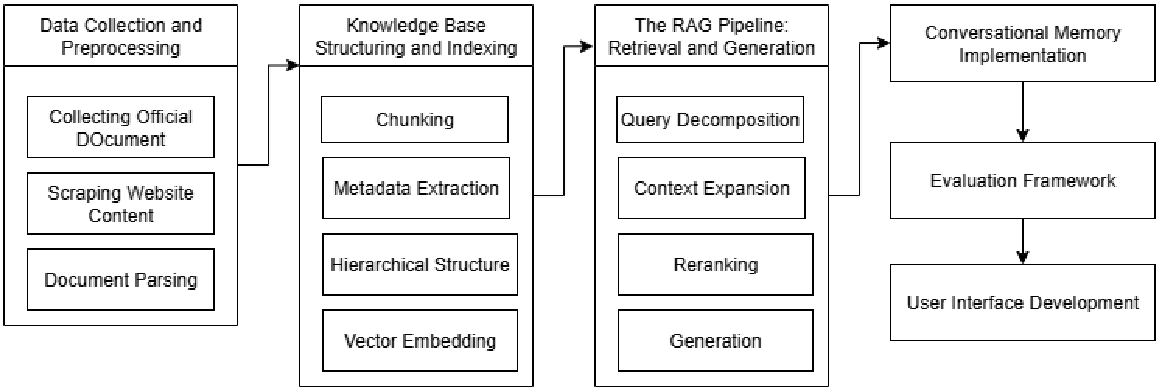

Data Collection and Preprocessing
Collecting Official DOcument
Scraping Website Content
Document Parsing
Knowledge Base Structuring and Indexing
Chunking
Metadata Extraction
Hierarchical Structure
Vector Embedding
The RAG Pipeline: Retrieval and Generation
Query Decomposition
Context Expansion
Reranking
Generation
Conversational Memory Implementation
Evaluation Framework
User Interface Development

Fig. 1 The Overall Research Methodology Workflow

Data Collection and Preprocessing

The foundation of the system is a comprehensive knowledge base, constructed by gathering data from multiple official sources. This process involved scraping content from the primary university website, the academic and student affairs (akmawa) portal of FPMIPA, and collecting official documents. A sophisticated preprocessing pipeline was then employed to extract and structure this raw data. The docling library was utilized to parse content from both websites and various document formats. To handle non-textual information, the system integrated the Qwen2-VL-2B-Instruct vision-language model, which generates relevant textual descriptions from images. Furthermore, tables within documents were automatically identified based on their visual structure, and the information within their cells was extracted and linearized into a textual format, ensuring no critical data was lost.

Knowledge Base Structuring and Indexing

Following preprocessing, the cleaned text was segmented using docling's HybridChunker. This approach combines semantic structure analysis with a token size limit to produce coherent and optimally-sized chunks. Key parameters were set, including a maximum token limit and the use of the meta-llama/Llama-3.1-8B tokenizer for accurate token counting. To enrich the data for advanced retrieval, metadata was extracted for each chunk, including a chunk_id, filename, and page_numbers.

To address the issue of context loss in isolated chunks, a hierarchical node structure was implemented. This strategy establishes parent-child relationships, where a parent_node contains a list of its child_ids and each child_node has a parent_id field. This allows the system to retrieve a specific chunk while also accessing its broader parent context. Sequential relationships were also maintained via previous_id and next_id fields. The complete set of processed nodes, rich with text, metadata, and hierarchical information, was then prepared for indexing. For the embedding stage, Google's gemini-embedding-001 model was selected due to its strong performance in retrieval tasks and ease of implementation via API. Finally, the indexing process involved ingesting all nodes into a Qdrant vector store, where each chunk's vector embedding and its corresponding metadata payload were stored as a single object.

E-ISSN : 2807-9035
Volume 5, Number 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.6719

The RAG Pipeline: Retrieval and Generation

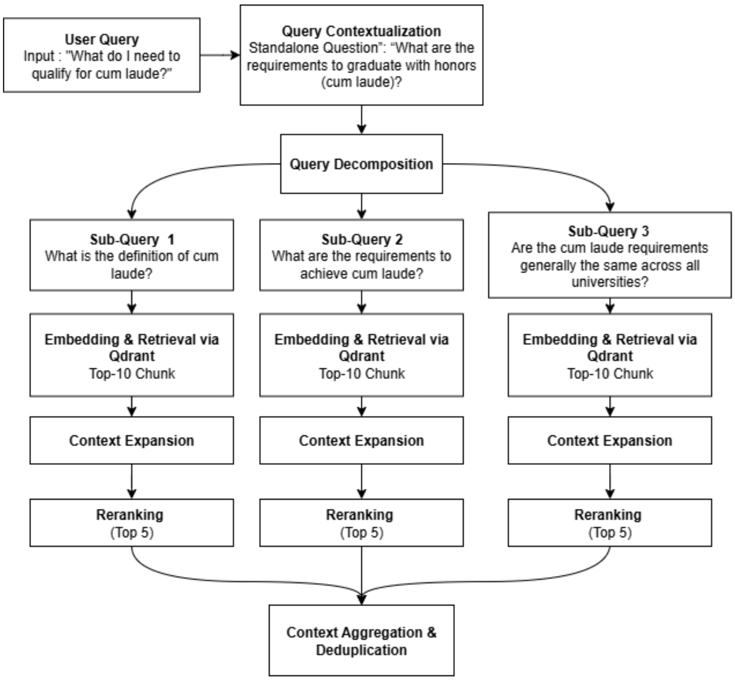

User Query
Input : "What do I need to qualify for cum laude?"
Query Contextualization
Standalone Question": "What are the requirements to graduate with honors (cum laude)?
Query Decomposition
Sub-Query 1
What is the definition of cum laude?
Sub-Query 2
What are the requirements to achieve cum laude?
Sub-Query 3
Are the cum laude requirements generally the same across all universities?
Embedding & Retrieval via Qdrant
Top-10 Chunk
Embedding & Retrieval via Qdrant
Top-10 Chunk
Embedding & Retrieval via Qdrant
Top-10 Chunk
Context Expansion
Context Expansion
Context Expansion
Reranking
(Top 5)
Reranking
(Top 5)
Reranking
(Top 5)
Context Aggregation & Deduplication

Fig. 2 Retrieval Process

The information retrieval process, as illustrated in Figure 2, was designed as a multi-stage pipeline to ensure optimal results. Leveraging the conversational memory, the process begins with Query Contextualization. This initial step rephrases the user's latest input, which might be an ambiguous follow-up question, into a self-contained, standalone question before any further processing. Following this, the contextualized query undergoes Query Decomposition, where the now-complete question is broken down by an LLM into several simpler, focused sub-questions. The system then queries the Qdrant vector store to fetch the top 15 relevant chunks for these sub-questions. This is followed by a Context Expansion phase. Using the pre-established hierarchical metadata (parent_id, child_ids, next_id), the system retrieves adjacent and parent nodes to enrich the context. It also identifies the titles of the retrieved chunks and scans the entire docstore to pull in all other chunks sharing the same title, creating a comprehensive set of candidate information.

To refine this large set of candidates, a reranking stage is employed using the FlashrankRerank method, powered by the ms-marco-MiniLM-L-12-v2 model. This reranker re-evaluates and re-sorts the candidates, selecting only the top 5 chunks with the highest relevance scores. This curated context is then passed to the generation stage. For text generation, the Gemini 2.5 Pro model was utilized, selected for its advanced reasoning and language synthesis capabilities. The interaction is managed through a structured PromptTemplate, which combines the final retrieved context with the user's original query. The entire pipeline, from retriever to LLM, is unified into a RetrievalQA chain using the LangChain framework, which automates the process and returns a final answer along with its source documents.

Conversational Memory Implementation

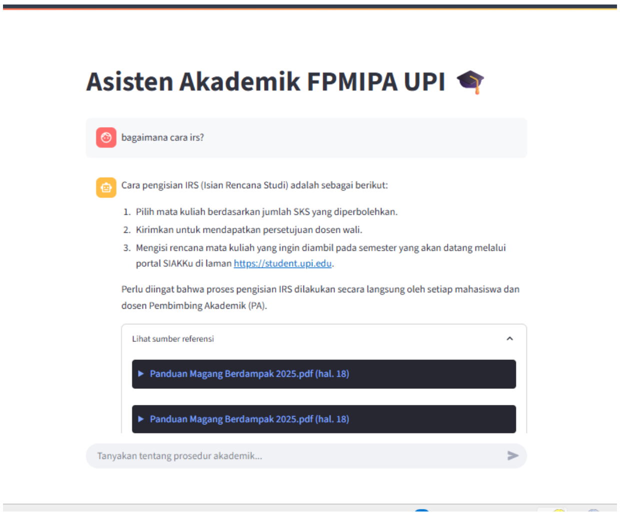
To overcome the stateless nature of a basic RAG system, a conversational memory mechanism was implemented to handle follow-up questions naturally. The system was designed to combine the retrieved document context, the user's current question, and the entire conversation history into the final prompt. This allows the LLM to understand the ongoing conversational flow and provide contextually appropriate answers. This process is automated using LangChain's RunnableWithMessageHistory wrapper, which stores the turn-by-turn history for each user session in an in-memory store, enabling continuous and coherent dialogue.

Evaluation Framework

The chatbot's performance was assessed using a custom-generated evaluation dataset of 100 unique questions, designed to test various system capabilities. To ensure a comprehensive evaluation, this dataset was strategically divided into three distinct categories. The first and largest category consisted of 60 factual questions, where the answers could be directly found within the knowledge base. The second category included 20 reasoning questions, which were more complex queries requiring the system to retrieve, combine, and reason over multiple pieces of information to form a comparison or conclusion. Finally, a set of 20 out-of-context questions was included, for which the answers were intentionally absent from the knowledge base. This last category was crucial for testing the system's ability to handle ignorance gracefully and to avoid hallucination when faced with unknown topics. Each of the 100 questions was processed by the chatbot, and the resulting answers and retrieved contexts were recorded for both quantitative analysis using standard RAG metrics and in-depth qualitative analysis of specific case studies.

User Interface Development

A user-friendly interface for the chatbot was developed using Streamlit. This framework was chosen primarily for its rapid development capabilities and seamless integration with the existing Python and LangChain scripts, eliminating the need for separate backend and frontend development. The final user interface design is illustrated in Figure 3.



Asisten Akademik FPMIPA UPI 🎓

bagaimana cara irs?

Cara pengisian IRS (Isian Rencana Studi) adalah sebagai berikut:

1. Pilih mata kuliah berdasarkan jumlah SKS yang diperbolehkan.
2. Kirimkan untuk mendapatkan persetujuan dosen wali.
3. Mengisi rencana mata kuliah yang ingin diambil pada semester yang akan datang melalui portal SIAKKu di laman https://student.upi.edu.

Perlu diingat bahwa proses pengisian IRS dilakukan secara langsung oleh setiap mahasiswa dan dosen Pembimbing Akademik (PA).

Lihat sumber referensi
▸ Panduan Magang Berdampak 2025.pdf (hal. 18)
▸ Panduan Magang Berdampak 2025.pdf (hal. 18)
Tanyakan tentang prosedur akademik...

Fig. 3 Interface Chatbot

RESULT

This section presents the empirical findings from the evaluation of the RAG chatbot. The system's performance was assessed based on a 100-question test dataset, which was strategically divided into three categories: factual, reasoning, and out-of-context questions. The results for each category are presented below, combining quantitative metrics with illustrative qualitative case studies.

Quantitative Performance Summary

The overall quantitative performance of the system across the three distinct question categories is summarized in Table 1. This comparative view highlights the chatbot's varying effectiveness depending on the complexity and nature of the user's query.

Table 1. Performance Metrics by Question Category

Relationships	Factual Question	Reasoning Question	Out-of-Context Questions
Faithfulness	0.9100	0.9048	N/A
Answer Relevancy	0.6862	0.4182	N/A
Context Recall	0.8519	0.5926	N/A
Semantic Similarity	0.7682	0.7373	0.7572
Factual Correctness (Precision)	N/A	N/A	0.8125
Factual Correctness (Recall)	N/A	N/A	0.6415
Factual Correctness (F1)	N/A	N/A	0.6875

N/A (Not Applicable) indicates that the metric was not tested for a specific category, as it was not relevant to that category's primary evaluation goal. For out-of-context questions, for instance, the goal is to correctly refuse to answer, not to measure the faithfulness or relevancy of an answer.

Performance on Factual Questions

This category, comprising 60 questions, tested the system's core ability to retrieve and present information that is directly stated in the knowledge base. The system demonstrated strong performance in this area, achieving a Faithfulness score of 0.9100 and a high Context Recall of 0.8519. These scores indicate that for straightforward queries, the chatbot is highly reliable in finding the correct source material and presenting the information accurately without

E-ISSN : 2807-9035
Volume 5, **Number** 2, November 2025
https://doi.org/10.47709/brilliance.v5i2.6719

fabrication. While the Answer Relevancy score was lower at 0.6862, qualitative analysis suggests this was often due to answers being overly comprehensive rather than incorrect. A successful case study below illustrates this typical performance.

Table 2. Example of a Factual Question

User Question	Generated Response	Reference
How much is the administrative fee for applying for academic leave if I still have an active study period?	The administrative fee for applying for academic leave while you still have an active study period is Rp250,000.00.	Students who apply for leave during the current semester must pay an administrative fee of Rp250,000.00.

This is a classic example of a successful factual retrieval in Table 2. The chatbot correctly identified the key entities ("academic leave," "administrative fee") and extracted the precise numerical value from the source document, providing a direct and accurate answer.

Performance on Reasoning Questions

This category consisted of 20 more complex questions designed to test the system's ability to synthesize information from multiple sources. The results revealed a significant challenge in this area. While Faithfulness remained high at 0.9048, indicating the system did not fabricate information, Context Recall dropped sharply to 0.5926. This was mirrored by a low Answer Relevancy score of 0.4182. This pattern suggests that while the chatbot is honest with the information it finds, it struggles to retrieve all the necessary, disparate pieces of context required to form a complete and relevant answer for complex reasoning tasks.

Table 3. Example of a Reasoning Question

User Question	Generated Response	Reference
For an undergraduate student in the Japanese Language Education Program, what are the language requirements for the thesis abstract and the language proficiency certificate for graduation?	Based on the available information, here is the answer to your question as an undergraduate student in the Japanese Language Education Study Program: 1. Thesis Abstract Language: Your thesis abstract must be written in three languages:Indonesian, English, Japanese (in accordance with the requirements for foreign language study programs). Language Proficiency Certificate: As a graduation requirement, you must possess a Japanese language proficiency certificate that demonstrates both active and passive language skills equivalent to level N3.	The thesis abstract must be written in both Indonesian and English. However, since you are from a foreign language study program, you must also include an abstract in the language relevant to your program, which is Japanese. As a graduation requirement, you must possess both active and passive Japanese language skills equivalent to level N3 (JLPT).

Table 3 shows that although the average score in this category is lower, this example illustrates the system’s capability for complex reasoning. It successfully retrieves two distinct rules from the knowledge base (one regarding abstracts and another concerning proficiency certificates) and synthesizes them into a coherent, well-structured, and accurate response.

Performance on Out-of-Context Questions

This final category of 20 questions tested the system's ability to handle queries for which no information exists in the knowledge base. This is a crucial test of the system's safety and its ability to avoid hallucination. The primary metric here was Factual Correctness (Precision), which measured the system's success rate in correctly stating it did not know the answer. The system achieved a score of 0.8125, meaning it correctly identified and refused to answer unknown questions in 81.25% of cases. In these instances, it provided a predefined, helpful response.

Table 4. Example of an Out-of-Context Questions

User Question	Generated Response	Reference
On which floor is the cafeteria in the FPMIPA building?	Sorry, the information regarding that matter was not found in the documents I have.	Information about the location of the cafeteria in the FPMIPA building was not found in the provided documents.

As shown in Table 4, the chatbot correctly cross-referenced the query with its knowledge base, determined that no information regarding the cafeteria's location was available, and responded with a polite and honest refusal. This behavior exemplifies the ideal response to an out-of-context question, as it avoids hallucination and helps maintain user trust. Table 4 highlights this as a strong example of factual robustness in the absence of relevant context.

DISCUSSION

This section provides an in-depth analysis and interpretation of the evaluation findings. The performance of the RAG chatbot across the different question categories is contextualized within existing literature, followed by an acknowledgment of the study's limitations and a discussion of its broader implications and directions for future research.

Interpretation of Key Findings

The evaluation results reveal a nuanced performance profile for the developed chatbot. The system's high scores on factual questions (Faithfulness of 0.9100, Context Recall of 0.8519) confirm that the core RAG architecture is robust and highly effective for its primary task: retrieving and accurately presenting directly stated information. This demonstrates the system's fundamental reliability for a large portion of typical student queries.

The most significant insight comes from the performance on reasoning questions. The sharp decline in Context Recall (0.5926) and Answer Relevancy (0.4182), despite Faithfulness remaining high (0.9048), highlights a critical challenge. This pattern suggests that while the system is faithful to the context it manages to find, its retrieval mechanism struggles to locate and integrate all the necessary, disparate pieces of information required for complex synthesis. The low answer relevancy is a direct symptom of this incomplete context retrieval, not a failure of the generation model itself.

Furthermore, the performance on out-of-context questions is a strong indicator of the system's safety and reliability. With a Factual Correctness (Precision) of 0.8125, the chatbot successfully identified and refused to answer unknown queries in the vast majority of cases. This ability to "gracefully fail" is crucial for building user trust and preventing the spread of misinformation, a key advantage over ungrounded LLMs.

Comparison with Existing Literature

These findings both align with and contribute to the broader body of research on conversational AI. The system's high performance on factual tasks confirms the effectiveness of the RAG architecture (Lewis et al., 2020) as a practical method to mitigate the hallucination risks detailed in surveys like Ji et al. (2023). The challenges observed in the reasoning tasks provide empirical validation for claims made in recent surveys (e.g., Wu et al., 2024), which identify multi-hop retrieval and complex reasoning as key open challenges and active areas of research in the RAG field. Our results quantify this challenge within a specific academic domain. Finally, the 81.25% success rate in handling out-of-context queries demonstrates a significant improvement in reliability compared to the unpredictability of early chatbot systems (Singh & Namin, 2025).

Limitations of the Study

It is important to acknowledge several limitations of this research. First, the chatbot's knowledge is strictly dependent on the provided documents. Its performance, particularly on reasoning tasks, is constrained by how information is structured and distributed within the source materials. Any errors or ambiguities in the original documents will be propagated by the system. Second, the evaluation was conducted on a generated dataset. While structured and diverse, it may not fully capture the nuance and unpredictability of real-world student queries. Third, the system's performance is tied to the specific models chosen (e.g., Gemini 2.5 Pro, the reranker model). Using different models would likely yield different results. Lastly, the study's scope is limited to a single faculty (FPMIPA UPI), and its applicability to other domains would require further testing.

Implications of the Study

Despite its limitations, this study has significant practical and academic implications. Practically, it provides a robust, end-to-end blueprint for developing a reliable AI academic assistant, offering a tangible solution for universities seeking to automate information services and improve student support. Academically, the clear identification and quantification of reasoning as a key performance bottleneck for RAG systems in a real-world domain contributes valuable empirical data to the field, highlighting specific areas where innovation is most needed.

Future Work

Several clear avenues for future work emerge from these findings. The most pressing need is to address the identified weakness in complex reasoning. Future research should therefore focus on improving performance on these tasks, which could involve exploring more advanced query decomposition strategies or implementing multi-step retrieval agents capable of iterative searching. Another direction is to enhance the system's safety guardrails, aiming to improve upon the 81.25% success rate for out-of-context detection, perhaps by training a dedicated classifier for this purpose. Finally, the most critical next step is to move beyond a static dataset and deploy the chatbot in a live environment to conduct a comprehensive user study, which would allow for an evaluation of its real-world performance, usability, and impact on student satisfaction.

CONCLUSION

This research successfully addressed the objective of enhancing the accessibility and effectiveness of academic information through the development and rigorous evaluation of a sophisticated Retrieval-Augmented Generation (RAG) chatbot. The empirical evaluation confirmed the system's high reliability for its core tasks, particularly in answering factual queries with a faithfulness score of 0.9100 and in correctly handling out-of-context questions in the majority of cases. The primary contribution of this work is an end-to-end, empirically-validated framework that provides a practical blueprint for building trustworthy, domain-specific AI assistants in a higher education setting, effectively mitigating the known hallucination risks of standard Large Language Models. While challenges in complex reasoning persist as a clear area for future work, this study affirms that a well-designed RAG system is a powerful and viable tool for modernizing student support services within academic institutions.

REFERENCES

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

Dhingra, B., Cole, J. R., Eisenschlos, J. M., Gillick, D., Eisenstein, J., & Cohen, W. W. (2022). Time-aware language models as temporal knowledge bases. Transactions of the Association for Computational Linguistics, 10, 257-273.

Erickson, J. S., Santos, H., Pinheiro, V., Mccusker, J. P., & Mcguinness, D. L. (2025). LLM experimentation through knowledge graphs: towards improved management, repeatability, and verification. Journal of Web Semantics, 85, 100853.

Es, S., de Jong, J., Al-Itejawi, T., & de Rijke, M. (2023). RAGAS: automated evaluation of retrieval augmented generation. arXiv preprint arXiv:2309.15217.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Sun, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems, 33 (pp. 9459-9474).

Mahlatse, S. H. E. K. G. O. L. A. (2024). Utilising Artificial Intelligence Chatbots for Student Support at Comprehensive Open Distance E-learning Higher Learning Institutions in the Fifth Industrial Revolution (Doctoral dissertation, University of South Africa).

Okonkwo, C. W., & Ade-Ibijola, A. (2021). Chatbots applications in education: A systematic review. Computers and Education: Artificial Intelligence, 2, 100033.

Rahim, N. I. M., Iahad, N. A., Yusof, A. F., & Al-Sharafi, M. A. (2022). AI-based chatbots adoption model for higher-education institutions: A hybrid PLS-SEM-neural network modelling approach. Sustainability, 14(19), 12726.

Singh, S. U., & Namin, A. S. (2025). A survey on chatbots and large language models: testing and evaluation techniques. Natural Language Processing Journal, 9, 100128.

Singun, A. J. (2025). Unveiling the barriers to digital transformation in higher education institutions: a systematic literature review. Discover Education, 4(1), 37.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, 30.

Wu, S., Xiong, Y., Cui, Y., Wu, H., Chen, C., Yuan, Y., ... & Xue, C. J. (2024). Retrieval-augmented generation for natural language processing: A survey. arXiv preprint arXiv:2407.13193.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.