Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X
Volume 3, Issue 7, 2025

INFERSITY V1: A RETRIEVAL-AUGMENTED GENERATION (RAG) BASED CHATBOT FOR INTELLIGENT ACCESS TO UNIVERSITY RESOURCES

Muhammad Talha Jahangir[*1], Arslan Hussain[2], Muhammad Humza Khan[3], Maaz Khalil[4], Muhammad Faizan Elahi Nonari[5]

[*1,2,4,5] *Department of Computer Science, MNS-University of Engineering and Technology, Multan, Pakistan.*
[3] *Department of Electrical Engineering, MNS-University of Engineering and Technology, Multan, Pakistan.*

[*1]mtalhajahangir@mnsuet.edu

DOI: https://doi.org/10.5281/zenodo.16400496

Keywords
Chatbot, RAG, University Helpdesk, Gemini, LLM, Web Scraping

Article History
Received on 24 April 2025
Accepted on 09 July 2025
Published on 24 July 2025

Corresponding Author:
Muhammad Talha Jahangir

Abstract

University websites and documents like UG Rules and Prospectus can be hard to navigate, especially for students searching for specific info like admission or scholarships. Student services are not on 24-hour operation, which may slow down the process of information access and frustrate students. The proposed research paper develops the Infersity v1 chatbot that would improve the user experience regarding academic material requests. It's based on a large language model (LLM) working with Retrieval Augmented Generation (RAG). This chatbot's goal is to solve a common issue that students encounter, the inability to get accurate and prompt responses from university administration about academic departments, programs, campus amenities, and student services. To provide precise and contextually aware responses, the suggested method integrates a domain-specific knowledge base sourced from university data into a RAG pipeline using Gemini 2.0-flash. The Retrieval Augmented Generation Assessment Score (RAGAS) was used in a case study conducted at Muhammad Nawaz Sharif University of Engineering and Technology in Multan (the First Public Sector Engineering University of South Punjab) to evaluate the system's performance. With a high RAGAS score of 0.96 for questions about prospectuses, 0.95 for questions about undergraduate regulations, and 0.95 for questions about websites, the chatbot showed excellent performance. These findings show that it is a dependable and easy to use academic assistant that has the potential to be widely used in higher education institutions.

1. INTRODUCTION

Universities know the importance of official websites for promoting campus resources and information about the university, so they spend a lot of time and money-making improvements. By providing details on admissions, academic departments, student services, and campus services, these websites act as the university's online public face. [1]. Although there is continuous improvement in the design and user interface (UI) of various websites, there are still a lot of institutional websites that cannot handle the rising demand in real-time and user-centric interaction. This is especially true when users are required to find certain information in a quick and efficient way. Lack of communication between the administrative departments of any organization and its stakeholders and in particular, the students and possible applicants is a common issue to the higher educational institutions. A confusing site is hard and tedious to operate through when seeking important details such as academic schedules, admission regulations, tuition fees and campus facilities. During peak times, such as exam schedules or admission seasons, administrative offices are typically slowed down with redundant requests, and this causes slowness in response, and unsatisfied users. Lack of readily available and instant support may lead to confusion and isolation of the university environment among guests and new students.

Conversational chatbots driven by artificial intelligence are a reasonable alternative to these problems, as they enhance user experience and accelerate knowledge distribution. Due to the presence of automated assistance, 24/7 service provision capability, and individual user experiences, chatbots have been especially effective in a range of sectors, such as healthcare [2], [3], cybersecurity [4], [5], retail [6], and hospitality [7]. Chatbots will allow the reducing of the amount of work done by administrative employees within the framework of a higher education institution by answering routine questions and ensuring stable and fast access to institutional information. This feature enhances the responsiveness and accessibility of university communication channels while also increasing operational efficiency.

The meteoric rise [8] in interest in using chatbots by industries at present is attributed to the overwhelming success of ChatGPT. In fact, the global chatbot market size was valued at 5.39 billion dollars in 2023 which is expected to reach 42.83 billion dollars by 2033, according to a market research report published [9] by Spherical Insights & Consulting.

This study presents Infersity v1, an enhanced and domain-specific chatbot system that will operate in an educational setting. This system combines a Large Language Model (LLM) and a Retrieval-Augmented Generation (RAG) structure to provide a high level of precision and contextualization of responses that can suit the specific needs of students and faculty and administrative personnel. In contrast to the traditional chatbots which work on pre-written, hard-coded scripts, Infersity v1 operates on dynamically extracted information based on a curated, structured knowledge base that is based on the specific institution academic and administrative data.

The chatbot uses generative capabilities in combination with retrieval-based grounding to make sure that its outputs are contextually appropriate and reliable in terms of facts presented. This design enables Infersity v1 to be flexible to different query types (curriculum, academic schedules and institutional policies) providing current, consistent answers as the underlying information changes over time. The system improves the user experience by reducing the hallucinations that are typical of generative models.

In this way, Infersity v1 can not only enhance access to information in learning organizations but also show how the latest AI methods can be applied in a specific real-life scenario when accuracy, relevance, and usability are the key issues. As shown in figure 1, without RAG, ChatGPT cannot access up-to-date or specific information, so it fails to answer questions like Mr. Talha Jahangir's research interests. With RAG, the system retrieves relevant documents, adds context to the prompt, and generates accurate answers, stating Mr. Jahangir’s department and listing his research areas, such as AI, Machine Learning, and Generative AI.

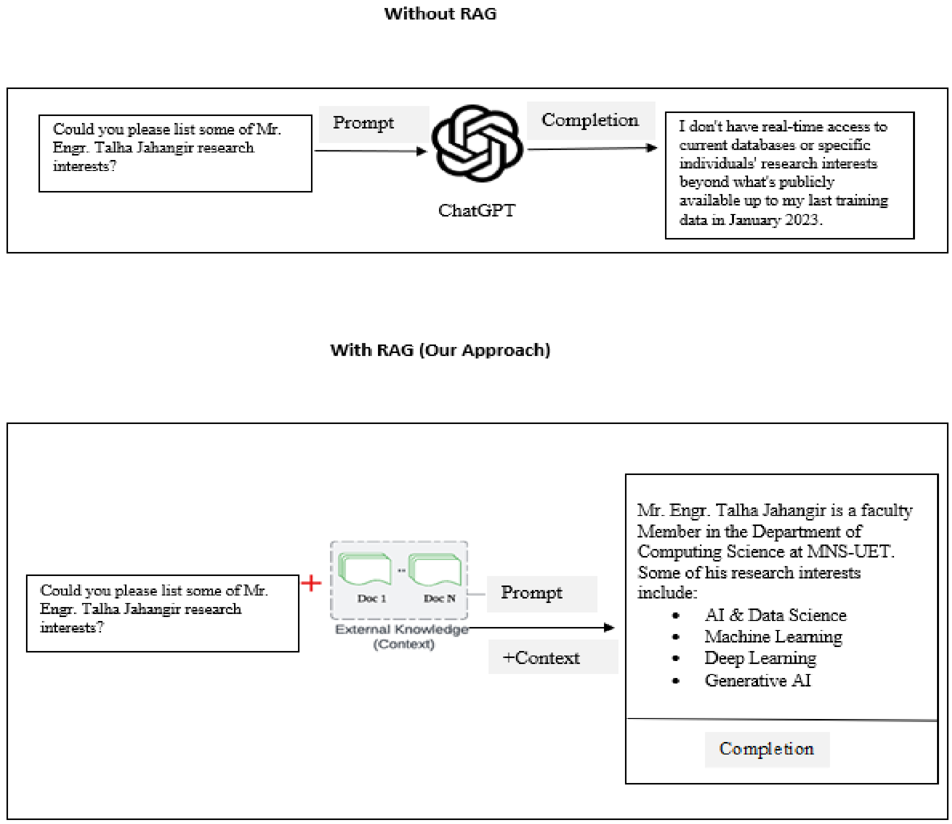

Fig. 1. Comparative example of response generation using a traditional approach versus a Retrieval-Augmented Generation (RAG) approach.

2. Contributions

- To design and create an intelligent RAG chatbot based system that can help to handle a large variety of queries related to the university by using automated processing of institutional documents like university prospectus, undergraduate rules and the contents of the official websites.
- Automation of information search on both static sources such as the prospectus and UG rules and dynamic source via the site scraping so that the chatbot will make correct and current responses in real-time without a human being managing the process.

Demonstration of the practical effectiveness of the chatbot with the help of the evaluation based on the RAGAS framework that proved the high performance of the chosen chatbot, the factual accuracy of its answers, and the quality of responses based on user-relevant queries of various types.

The rest of this article is divided into nine connected sections, In Section 3, We Compare and analyze Related Works. In Section 4, we explain the dataset used. In section 5, we explain the preprocessing done on our dataset. In Section 6, we discuss the methodology and techniques used. In Section 7 we explain the evaluation metric used and the formula each metric contains. In section 8, we provide a detailed analysis of our experiment and result. In section 9, we discussed the work done. Finally, in Section 10, we conclude our paper. In section 11 are the references used in paper.

3. Related Work

Recent research on educational chatbots encompasses various domains, including their application fields, objectives, learning experiences, design methodologies, technological frameworks, evaluation strategies, and additional experiments. Studies have identified the use of educational AI-chatbots across areas such as health advocacy, language learning, and self-advocacy. These systems may be either flow-based or AI-driven, performing functions like answering frequently asked questions (FAQs), administering quizzes, recommending activities, and disseminating info about events according to Islam et al [10].

Educational chatbots have been shown to enhance students' studying experiences by increasing motivation, sustaining engagement, and providing immediate online support according to Wollny et al. [11]. Furthermore, they add to making education more accessible and readily available according to Okonkwo et al. [12]. The design elements of chatbots, particularly their roles and appearances, significantly influence their effectiveness as pedagogical tools as shown by Martha et al. [13]. Winkler et al. [14] said that chatbots can incorporate speech recognition features and are created using a variety of techniques, such as AI-based and flow-based approaches. Dialogflow and ChatFuel are two popular technologies used in chatbot implementation; each affects the chatbot's functionality and overall quality, so careful selection is required during the design and development stages. According to Pérez et al. [15] AI-based chatbots use machine learning (ML) and natural language processing (NLP) to provide more flexible and adaptive dialogues, whereas flow-based chatbots, like those constructed with Dialogflow, provide structured, script-driven interactions.

To determine the efficency of educational chatbots, researchers use different types of evaluation: survey, controlled experiments, and longitudinal studies. The main criteria of such assessments include such main aspects as the usability, motivation of users, and system acceptance. To illustrate, Hobert et al. 16] and Hwang et al. [17] have performed experimental studies in which they considered the impact that chatbot use has on academic performance. These types of experiments enable the identification of how chatbots can affect the performance and interest of students regarding learning. In the meantime, surveys are typical to collect feedback about the users both instructors and students, providing information on the perception of various stakeholders regarding the utility and convenience of chatbot systems.

The other important direction of research is the study of the interaction styles, especially the comparison of user-driven and chatbot-driven dialogue [14][10]. The interactions that occur between the user and the system are more personalized and user led as opposed to the interactions that occur between the user and the chatbot which tend to be more structured and automatic. Finding a balance between these two modes may result in more intuitive, natural and more efficient interactions, and better overall user experience.

To maximize the educational chatbots potential, one has to also discuss several contextual and technical issues. These are the problems of personalization, flexibility to suit various learning settings, compatibility with the existing systems, and the provision of significant activities. These aspects are essential to address making chatbot use in education as effective and scalable as possible and eventually lead to better learning outcomes and better support of both students and educators. Also, the ethical aspect is vital, particularly in the case of user data security and compliance with learning standards. According to Ungless et al. [18] to be able to address these problems more successfully, new development methodologies are necessary. Nevertheless, there are also persistent challenges when it comes to the complexity of programming and the need to make chatbots relevant in the context of ongoing changes in education. [19], Adamopoulou et al. [20] said that the robustness and flexibility of educational chatbots should be enhanced by employing the technology breakthroughs.

More possibilities and challenges are available with more developed language models. As an example, Wen et al. [21] analyzed how models, such as BERT, identify links between expressions and queries, which inform the refinement of responses relevance and interaction patterns. This research study is important to the extent that it emphasizes the importance of using context-aware language models to make complex and meaningful responses of chatbots. Kuhail et al. [22] address the ever-present gap between user intention and the chatbot-generated answers, especially in complex academic environments where key issues are context awareness and knowledge in specific domains. Such papers also highlight the significance of applying high-level models which can provide accurate and contextual answers to complicated questions.

There is much potential in chatbots improving the user experience and learning results when incorporated into university systems and metaverse spaces. As an example, Xie et al. [23] demonstrate the way chatbots embedded into the university-based systems based on the metaverse can instantly and personally help in such operations as answering the most frequently asked questions and course navigation. These technologies enhance academic support through simplifying the process of information search and reducing administrative loads with the help of NLP and machine learning.

The article by Chandra et al. [24] tells us about the development of a question-answering chatbot to support the Indonesian university admissions process where the sequence-to-sequence learning can be applied in more specific institutional settings. This application demonstrates that chatbots may be tailored to perform specific administrative activities. Oliveira et al. [25] also introduces a dynamic chatbot that helps students to communicate in such fields as event information, academic support, and admissions. This system gives priority to user feedback to enhance accuracy, dependability and data security. The fact that it is updated regularly ensures its relevance.

Moreover, the article of Martinez-Araneda et al. [26] titled "Designing a chatbot to support problem-solving in a programming course" was published in the INTED2024 Proceedings. TutorBot+, a chatbot-based educational platform based on ChatGPT and other large language models (LLMs) is introduced in IATED, 2024, pp. 966975. A quasi-experimental analysis demonstrates that TutorBot+ has a positive effect on student computational reasoning skills, and this example shows the transformative capacity of AI-based edtech in challenging subject areas, such as computer science

Table 1. Comparative analysis of all the related work to our research.

Ref. No.	Year	Corpus	Methodology	Results
[24]	2019	The corpus consists of conversation data from Telkom University admission, including 2,506 training data and 397 testing data.	sequence-to-sequence model with an attention mechanism, trained on WhatsApp conversation data	The model achieved a BLEU score of 44.68, indicating improved response accuracy by attention mechanisms and reversed sentence input
[14]	2023	Course content from an online university course in pedagogy	Developed a chatbot with scaffolding-based dialogue strategies integrated into Moodle using Dialogflow and Google Cloud tools	Students using Sara showed improved engagement and better learning outcomes compared to control groups
[10]	2024	4,000+ queries from UNNES helpdesk (2021–2023).	Hybrid model using TF-IDF for FAQs and Llama RAG for complex queries	TF-IDF handled 78% of queries; RAG handled 22%, improving response time and user satisfaction by 3.5.
	2025	Undergraduate Rule, Prospectus, and University Website (Beautifulsoup4)	Document and Web Extraction, Multi-Source Data Integration, Vector Store, Intelligent Chunk Retrieval, Generative AI Answering, RAGAS Evaluation Framework	Faithfulness, Context Precision, Context Recall, Semantic Similarity and Answer Correctness at RAGAS of 1.00. There was also high Answer Relevancy with 0.86. It takes an average of 1.7 seconds to respond, with the first token having a latency time of 0.37 seconds.

4. Corpus

The dataset, which was used in the development of our Retrieval-Augmented Generation (RAG) chatbot, consists of the Prospectus, the Undergraduate Rules and Regulations, and text obtained from the official website of the MNS University of Engineering and Technology (MNS-UET), Multan. These sources combined represent a comprehensive and up-to-date way the university addresses the academic policies, administrative processes, and institutional information within the university and are useful in enabling the chatbot to give accurate and relevant responses.

MNS University of Engineering and Technology prospectus is an all-inclusive evaluation of the programs provided and the management structure of the university. The prospectus is opened with the remarks made by the Chancellor and the Vice Chancellor that focuses on the vision of the university in terms of academic excellence, innovation, and commitment to national development. The history of Multan is presented with the emphasis on educational and cultural importance of the city. The prospectus continues to entail a detailed history of MNS-UET, the purpose and goals statements as well as the strategic objectives of the organization. The academic section elaborates many undergraduate and graduate programs in business, science, engineering, and technology. The prospectus contains information on program size, eligibility criteria, curriculum patterns, as well as semester-by-semester course content and credit hours in each department. Specific attention is given to such flagship departments as Chemical Engineering and Civil Engineering Technology, with their lab description, faculty, and research purpose. This includes administrative information such seat allocation, merit list routine and the documentation that needs to be done to gain admission. Consequently, current as well as future students can use the document as the primary source of information.

The second document that complements the prospectus is UG Rules and Regulations which explains the administrative and academic policies that govern the student life in MNS-UET. It also sets out policies concerning admissions, exams, degree requirements, and disciplinary actions along with defining institutional roles such as the faculty, the deans and the department heads. Credit hours, contact hours and grading schemes are well explained in one of the major areas of focus which is the semester system. The document also classifies the subjects in three categories which include lecture based, project based and practical based. Grading policy addresses the course repetition, requirements of an academic probation, methods of CGPA calculation, and distribution of grades and grade points. It also details the process of withdrawal, incomplete grades, expulsion and re-admission in the university along with the re-admission procedure. Other areas include merit positions, degree honors, and eligibility to medals.

The official site of MNS-UET is a trustworthy and official source of extra information and an active and often updated platform which provides institutional information. It contains up to date, important details concerning academic calendars, departmental organization, faculty directories, admissions process, event announcements, and policy announcements. Since this web-based content is made part of the dataset, the chatbot can keep abreast of the changes happening in the institution to offer precise, timely, and contextual appropriate responses

5. Preprocessing

In the case of the preprocessing step of our RAG chatbot development, we introduced a multi-level pipeline to convert the unstructured academic data into a structured, query-ready dataset. This started by extraction of text in image and text-based PDF files. The university prospectus contained scanned images that were largely composed of visual data that had to be converted to machine readable text using Optical Character Recognition (OCR). We used advanced OCR software to minimize the problem of semantic distortion and keep the document structure, such as headings, paragraphs, and tables.

In the case of PDFs with text, such as the UG rules and regulations of the university, we resorted to PyMuPDF and PDFPlumber. These tools allowed to parse tables and other structured material correctly and maintain the logical document layout. The raw text obtained was then normalized by using formatting standardization, elimination of artifacts, and whitespace correction in order to ensure that the data was the same across all the documents.

Following the first cleaning stage, a text was divided according to a special text splitting strategy which aims at maintaining contextual coherence. This dynamic chunking scheme guaranteed that the resulting segments had enough semantic information content in them and hence could be efficiently retrieved in downstream tasks. This approach was directly designed to adhere to the token budget of large language models as it maximizes context-keeping and performance.

Concurrently, web data was handled with the library BeautifulSoup4. The script interpreted HTML data and retrieved useful textual contents using special tags like `` `<p>` ``, `` `<h1>` ``, `` `<h2>` ``, `` `<li>` ``, and `` `<span>` ``. These tags were selected since they are the ones that usually capture meaningful data in web documents. Once these elements were isolated, the command of `get_text (strip=True) was used to get rid of any HTML tags left behind and any unnecessary whitespace, leaving clean and proper-structured text.

Preprocessing pipelines played an important role in converting text data (document or Web-based) to a structured and semantically significant format and embedded it in an index. This is a critical step to the success of any Retrieval-Augmented Generation (RAG) system, which is in the quality of the responses directly dependent on the quality and structure of the input data that it is based on.Raw text data is usually full of noise, irrelevant data, and inconsistent presentation and structures that would complicate retrieval, as well as generation tasks. To counter this, the pipeline has done a few cleaning and structuring processes, including the removal of redundant content, the division of content into content-relevant sections, and the maintenance of metadata, including source kind, page count, and topic titles. Such enhancements made the data optimized to semantic embedding, and that it can be retrieved efficiently based on the user query.

With Retrieval-Augmented Generation (RAG) systems, in which a language model produces answers because of retrieved context, text structure and preprocessing greatly improve retrieval precision and provide more contextually based answers. The given preprocessing step is the core of the overall performance of the system, since it has a direct impact on the relevance and reliability of the answers created. With the input data clean, coherent, and semantically meaningful, the system will be more able to give accurate, context-sensitive answers to a broad variety of user questions, thereby becoming more effective and trustworthy in domain-specific real-world tasks.

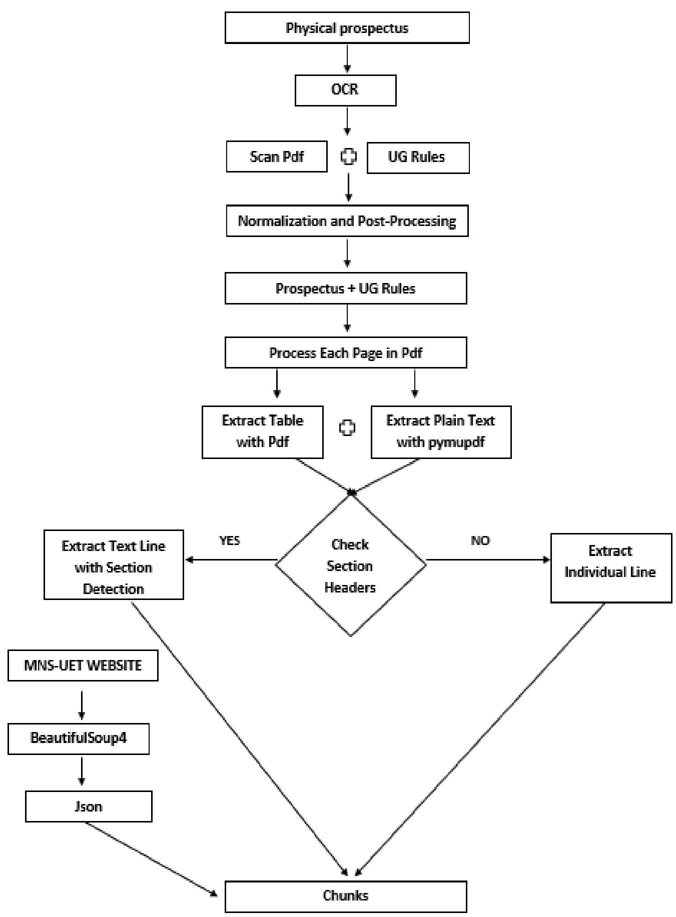

Fig. 2. Preprocessing pipeline demonstration

6. Proposed Methodology:

The proposed system utilizes a powerful Retrieval-Augmented Generation (RAG) framework in order to enable smart and accurate question answering from document datasets, specifically PDF documents. The approach also provides that user requests are responded by using information found in the source documents by integrating large language model (LLM) generation with strong semantic search.

The first step in the procedure is preprocessing, where documents are in PDF format, that have been uploaded onto the system, are parsed to break them into contextually sensible sections. This segmentation retains important metadata, such as the name of the source file, the page number, paragraph index and a unique chunk identifier. These metadata elements are also necessary when tracing the origin of the information that has been fetched and making it easier to create effective references in the end results.

Each text chunk is then converted to high-dimensional vector representation using the BAAI/bge-small-en-v1.5 model of Sentence Transformers library. We chose this model since it finds a good trade-off between semantic richness and computational effectiveness that qualifies it to be used in real-time applications. The user interface shows the progress of the operation in a dynamic progress bar and embedding is performed in batches to preserve memory and to improve responsiveness. To support fast, scalable approximate nearest neighbor (ANN) searches on inner product similarity, the created vectors will be normalized to L2 and indexed in an FAISS IndexFlatIP index. The system semantic retrieval engine is based on this index.

In parallel, web data is collected using the requests and BeautifulSoup4 libraries. Starting from a base URL, the system crawls the website within its domain, discovering and visiting internal links up to a defined maximum number of pages. For each page, meaningful content is extracted from specific tags like <p>, <h1>, <h2>, and others. The get_text (strip=True) method ensures that text is cleaned of HTML tags and whitespace. The crawler enforces a delay between requests to prevent overloading servers.

When a user submits a question, the query is encoded into a vector using the same embedding model. The top k most semantically similar document chunks are then found using the FAISS index and this query vector. The system carries out a context expansion step to improve the returned results' informativeness. Large amounts of pertinent content can be sent to the model in a single prompt by utilizing Gemini 2.0 Flash's capacity to handle up to 1 million tokens. Following their retrieval, the fragments are sorted according to their original locations, grouped by document and page number, and put together into a structured context block. This block is added to a thoughtfully constructed prompt template that helps Gemini 2.0 Flash produce clear, concise, and well-organized answers. In addition, the prompt specifies that you must include a "Sources" section with the names and page numbers of the original documents to synthesize complex information, such as specifics about people, roles, or sections.

Upon the generation of an answer, the system then marks and aligns any sources cited to the matching chunks in the original paper. This consistency will make every segment of the response traceable, and this will provide enhanced visibility and credibility to the response. The sources are then displayed on an intuitive and easy to use Streamlit interface. The interface consists of the tabbed document views that enable the user to browse among various files that have been uploaded, highlighted references that demonstrate exactly where the material that has been being referred to is located and expandible previews that also enable users to check the original context without clicking into different panes. Moreover, it is possible to tune chat and retrieval parameters (chunk size, top-k results, or LLM temperature settings), finally having more control and customization directly in the interface.

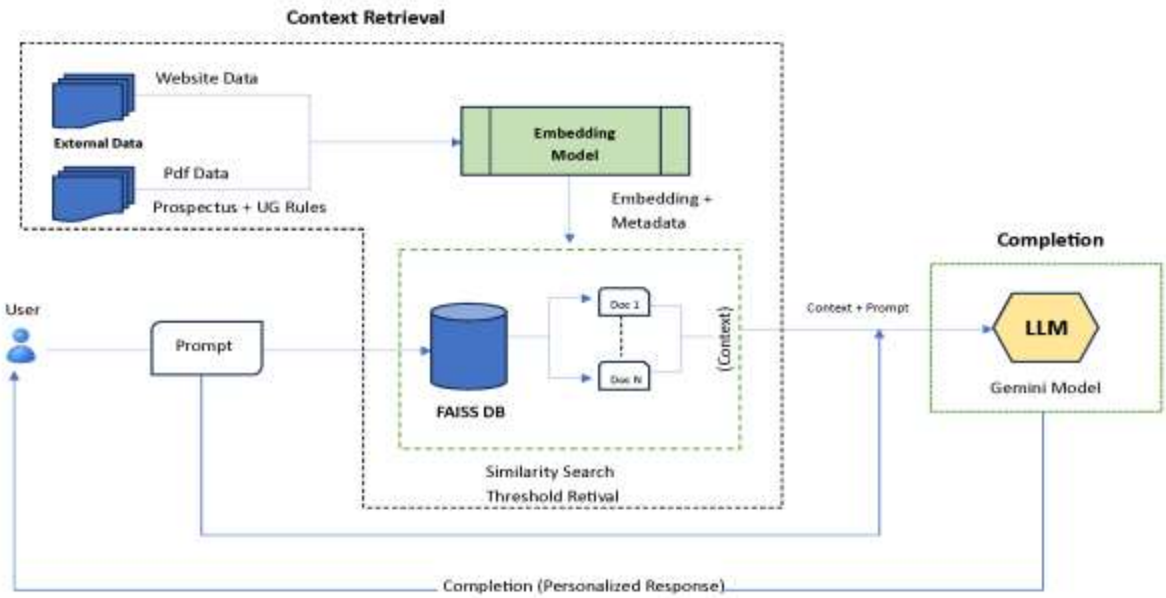

Fig. 3. Retrieval Augmented Generation System Demonstration.

Prompt: what documents do i need for admission?

Completion: To complete your admission at MNS-UET, you need to pay the university dues and submit these documents:

- Paid Original Bank Challan (keep photocopies).
- Original Domicile Certificate.
- Original applicable certificates and degrees (Matriculation/"O"-Level, F.Sc./"A"-Level, B.Sc., Diploma of Associate Engineering (DAE), B.Tech (Pass) or any equivalent qualifications).
- Six sets of photocopies of the above-mentioned documents.
- Two attested copies of CNIC/ "B" Form.
- Six recent passport-size photographs.
- Bio-data card Form-IX (duly completed).
- Medical Certificate Form (Dully signed and stamped by a PMDC registered Medical Practitioner).
- Duly attested Current Income certificate of the parent/ guardian.
- Undertaking (Sample Form —XI) on a Rs. 100/-judicial paper (duly completed).

Fig. 4. A graphical depiction of the response to the query related to prospectus.

Prompt: I want to change discipline after admission?

Completion: Yes, a student can opt for a change in discipline after the first year of study at MNS-UET.

A student, after the first year of study at the University, may opt for a change in discipline. The minimum admission merit of the discipline, being opted for, must be equal to or lower than the merit of the opting student at the time of his admission. The student opting to change his discipline after the first year must apply to the Vice-Chancellor through his Chairperson and the Dean. Acceptance shall depend on the availability of seats in the opted discipline. On acceptance of his request, the student shall start afresh with credit being given for any University core subjects studied in the first year in the original discipline. He will be issued a new registration number, and his maximum permissible duration count will start afresh. Students cannot opt for change in discipline after their second year at the University.

Fig. 5. A graphical depiction of the response to the query related to UG Rules.

Prompt: What are the jobs available?

Completion: Available jobs and related notices can be found on the "Jobs" page of the MNS UET website.

Based on the information provided, here are the types of jobs and notices available:

- Notices regarding Central Testing Services (CTS) answer keys.
- Important notices about interviews.
- Lists of successful candidates in written skill tests.
- Information about contacting the Recruitment Complaint Redressal Cell (RCRC) for complaints.
- Notices regarding the recruitment test paper distribution for teaching & non-teaching staff.
- Information about typing/skill tests for positions such as Assistant, Computer Operator, Store Keeper, and Junior Clerk.
- Roll number slips for written MCQs / typing skill tests.

Additionally, there is a research-related job opportunity for MS-enrolled students in Chemical Engineering Program:

- Two research assistants are required for the HEC-funded Project "Efficient electrodes for vanadium redox flow battery to tackle the intermittent nature of power from solar energy"

Fig. 6. A graphical depiction of the response to the query related to Website content of MNS UET.

7. Evaluation metric

The evaluation metric used for this research are: [27]

7.1. Context Precision

Context Precision measures how well the relevant ground-truth information is ranked within the retrieved context. Ideally, all relevant chunks should appear at the top of the ranking. The metric ranges from 0 to 1, with 1 indicating perfect ranking of relevant content.

$$\text{context precision@k} = \frac{\sum \text{precision@k}}{\text{total number of relevant items in the top k results}}$$

Where k is the number of chunks in context

7.2. Context Recall

Context Recall evaluates how much of the ground-truth information (as annotated in the answer) is successfully captured in the retrieved context. It is calculated by comparing the retrieved context with the ground truth, with scores ranging from 0 to 1—higher values reflecting better alignment and coverage.

To evaluate context recall based on the ground truth answer, each sentence is examined to see if it can be linked to the retrieved context. Ideally, every sentence in the ground truth answer should be supported by information found in the retrieved context.

$$\text{context recall} = \frac{|\text{GT sentences that can be attributed to context}|}{|\text{Number of sentences in GT}|}$$

7.3. Faithfulness

This measure gauges the factual correctness of the produced answer by determining the ratio of the correct statements that are justified by the context offered. It is calculated by the division of the number of correct answers by the total number of statements in the answers using the question, the retrieved contexts, and the generated response to be factually accurate.

$$\text{Fauthfullness} = \frac{|\text{Number of claims in the generated answer that can be infe}}{|\text{Total numbers of claims in the generated a}}$$

8. Results

8.1. Experiment Setup

The experimental environment was set up with common software tools and consumer-level hardware to facilitate easier development and testing. The hardware of the system was an Intel Core i3 11th Generation processor, an 8 GB RAM, and a 512 GB solid-state drive (SSD).

As far as software that was used, Docker and Visual Studio Code (VS Code) were utilized. Docker was used to generate containerized settings that ensured isolation, consistency, and ease of deployment during the various stages of the experiment. This enabled one to package runtime components, libraries, and dependencies into reproducible and mobile containers. This approach decreased configuration conflicts and increased the dependability of the experimental results.

The visual studio code, an integrated development environment (IDE), has been very helpful in project management, debugging, and programming. It was more efficient to develop, as it could be extended and easily integrated with Docker and version control systems such as Git.

8.2. Result

The current performance of Infersity v1 chatbot demonstrates that it is a reliable, intelligent and efficient academic information retrieval system as it has achieved excellent results in the initial rounds of evaluation. The system is designed on a workflow basis under Retrieval-Augmented Generation (RAG) framework and employs
generative and semantic search features to provide highly accurate, context-sensible and user relevant answers to a wide variety of institutional requests. It gives direct access to crucial information that is retrieved of complex documents and active web-based sources, which is its main purpose of serving academic stakeholders, namely, faculty, administrative personnel, and students.

The 6 main performance measures to determine the effectiveness of RAG systems are Faithfulness, Context Precision, Context Recall, Semantic Similarity, Answer Correctness and Answer Relevancy. RAGAS (Retrieval-Augmented Generation Assessment Suite) was employed in order to evaluate the chatbot. In combination, the metrics evaluate the chatbot in terms of its capacity to find contextually relevant information, generate factually and semantically accurate answers, and respond according to the user intent.

To ensure sound testing, we developed an evaluation protocol consisting of 150 questions distributed equally in three major categories, namely, the university prospectus, undergraduate (UG) rules, and the official university site. In order to portray a realistic distribution of information requests that could be made by staff and students, 50 queries were hand-selected in each domain. Some of the topics covered were admission requirements, degree program structures, academic policies, semester regulations, fee details and web-based announcements or services. This domain-specific testing allowed the chatbot to go through a series of use cases with varying degrees of contextual and structural sophistication.

The outcomes of the assessment were very good. The chatbot scored an ideal score of 1.00 in five of the six RAGAS metrics: Faithfulness, Context Precision, Context Recall, Semantic Similarity and Answer Correctness. Such perfect scores mean that the chatbot is always able to recall the appropriate contextual units, base its responses on such units, maintain the intended meaning of the source text, and produce answers that can be shown to be correct. Specifically, the Context Precision and Recall scores of 1.00 indicate the capability of the system to not only pick but also use the most relevant parts of source documents, be they PDF-based or HTML-based. In a parallel fashion, Faithfulness 1.00 guarantees that the chatbot will not hallucinate or make up content, all the sentences it produces will be traceable to the source, which is proven to be correct.

The only score a bit lower was Answer Relevancy with 0.81. Although this score is still high, it indicates that sometimes generated responses may be improved to be more consistent with more complex user intent. One of the possible improvements that would be possible to make in future versions would be implementation of more sophisticated reranking methods, improvement in intent classification, or feedback loop to the user.

The chatbot achieved consistent results, by domain. It scored 0.95, 0.96, and 0.95 on questions touching on prospectuses, undergraduate regulations, and websites, respectively. These results indicate that the chatbot can operate both with static and dynamic types of content and prove that it is domain adaptable. It proves helpful as a scalable digital assistant in academic setting through its ability to search and respond with information that is institutionally correct, either through academic calendars, policy documents or real-time web-based information.

Infersity v1 chatbot is also outstanding concerning performance efficiency besides accuracy. It is powered by the Gemini 2.0 Flash model, which is a language model with the latest technology that is optimized in terms of performance and quality. This gives the system the capability of making responses in low latency thus making it ideal in real time interactions. The average response time is approximately 1.7 seconds with the first-token latency of 0.37 seconds. Such reaction times ensure smoothness and engaging user experience which is the key to maintain attention and assurance in interactive systems. Moreover, the chatbot is able to address even complex or extended

queries with impressive speed since it responds at the throughput of approximately 236 tokens per second.
The combination of Gemini 2.0 Flash and optimized vector retrieval and document indexing methods also makes it possible to retrieve the context quickly when dealing with larger corpora of institutional documents. In order to ensure semantic similarity between document segments and user queries, we organized the vector space based on normalized indexing by FAISS and advanced embedding models (including BGE-small).
This architecture enables future growth to other departments, campuses or education institutions because it supports scalable and modular deployment.
In summary, it can be said that the application of RAG-based AI systems in higher education has taken a new step with the introduction of the Infersity v1 chatbot. It is a disruptive instrument of online academic assistance due to its low latency level, versatility, and high factual accuracy. It has an architectural design that allows further enhancements and changes, and the results of the evaluation have shown that it is ready to be deployed into the real world. Systems like Infersity provide a convincing illustration of how generative AI can be both practical and reliable in academic settings, as institutions increasingly rely on AI-driven solutions to assist students and expedite information access.

Table 2. Overview of Results

Category	Retrieval		Generation		RAGAS Score	Evaluation	
	Precision	Recall	Faithfulness	Relevancy	Mean	Answer Similarity	Answer Corecteness
UG rules	1.00	1.00	1.00	0.81	0.95	1.00	1.00
Prospectus	1.00	1.00	1.00	0.86	0.96	1.00	1.00
Website	1.00	1.00	1.00	0.81	0.95	1.00	1.00

9. Discussion

Our RAG chatbot first used a locally hosted language model that generated embeddings and made deductions using Ollama. However, we encountered limitations in performance responses were significantly delayed, and the embedding quality was insufficient for accurate retrieval. Consequently, we transitioned to using Google Gemini, specifically the Gemini 2.0-Flash model, which offered faster response times and more reliable semantic embeddings. The chatbot pipeline extracts text and tables from PDF documents using PyMuPDF and pdfplumber and also scrapes structured content from university websites via BeautifulSoup4. Extracted content is chunked and embedded using the BGE-small model, then indexed using FAISS with L2 normalization. For retrieval, the most relevant chunks are selected and contextually expanded.
These are then passed to Gemini 2.0-Flash for answer generation. This change substantially improved the chatbot's answer relevancy and response latency, making it more suitable for real-time academic query resolution.

Table 3. Relative Comparison of our paper.

Year	Name	Corpus	Methodology	Results
2024	From Questions to Insightful Answers: Building an Chatbot For university Resources	Web Scraping from Official University web Site: Use 42 Departments Data	Web Extraction, Vector Store, Intelligent Chunk Retrieval, Generative AI Answering, RAGAS Evaluation Framework	Achieving RAGAS scores 0.96 Answer Correctness 0.89 Answer Similarity 0.84 Precision 0.98 Recall 0.99 Faithfulness 0.99 Relevancy 0.99

2025	Infersity v1: A Retrieval-Augmented Generation (RAG) Based Chatbot for intelligent Access to University Resources	Undergraduate Rule, Prospectus, and University Website (Beautifulsoup4)	Document and Web Extraction, Multi-Source Data Integration, Vector Store, Intelligent Chunk Retrieval, Generative AI Answering, RAGAS Evaluation Framework	Achieving RAGAS scores of 1.00 in Faithfulness 1.00, Context Precision 1.00, Context Recall 1.00, Answer Similarity 1.00, and Answer Correctness 1,00. It also showed strong Answer Relevancy with a score of 0.86. On average, it responds in about 1.7 seconds, with latency for the first token as low as 0.37 seconds.

10. Conclusion

The development and deployment of the Infersity v1 chatbot using Retrieval-Augmented Generation (RAG) and Gemini 2.0-flash represents a significant advancement in enhancing student access to university resources. By leveraging a domain-specific knowledge base and delivering accurate, context-aware responses, the system addresses a critical need for timely and relevant information in academic settings. The high RAGAS score of 0.95 obtained during the case study at Muhammad Nawaz Sharif University of Engineering and Technology demonstrates the chatbot's reliability, effectiveness, and user-friendliness. This research highlights the potential of RAG-based AI assistants to transform university helpdesks and improve administrative efficiency, paving the way for

REFERENCES

Meyer, K. A., & Jones, S. (2011). Information found and not found: what university websites tell students. Online journal of distance learning administration, 14(3), 1-10.

Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., & Lau, A. Y. (2018). Conversational agents in healthcare: a systematic review. Journal of the American Medical Informatics Association, 25(9), 1248-1258.

Neupane, S., Mitra, S., Mittal, S., Gaur, M., Golilarz, N. A., Rahimi, S., & Amirlatifi, A. (2025). Medinsight: A multi-source context augmentation framework for generating patient-centric medical responses using large language models. ACM Transactions on Computing for Healthcare, 6(2), 1-19.

Franco, M. F., Rodrigues, B., Scheid, E. J., Jacobs, A., Killer, C., Granville, L. Z., & Stiller, B. (2020). SecBot: A business-driven conversational agent for cybersecurity planning and management. 2020 16th international conference on network and service management (CNSM),

Mitra, S., Neupane, S., Chakraborty, T., Mittal, S., Piplai, A., Gaur, M., & Rahimi, S. (2024) Localintel: Generating organizational threat intelligence from global and local cyber knowledge. arXiv preprint arXiv:2401.10036.

Chung, M., Ko, E., Joung, H., & Kim, S. J. (2020). Chatbot e-service and customer satisfaction regarding luxury brands. Journal of business research, 117, 587-595.

Leung, X. Y., & Wen, H. (2020). Chatbot usage in restaurant takeout orders: A comparison study of three ordering methods. Journal of Hospitality and Tourism Management, 45, 377-386.

Tinn, R., Cheng, H., Gu, Y., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2023). Fine-tuning large neural language models for biomedical natural language processing. Patterns, 4(4).

Zhou, T., Liu, Y., & Kumar, M. (2024). AI-powered chatbots for improving interactive user experience: state-of-the-art.

Islam, M., Warsito, B., & Nurhayati, O. (2024). AI-driven chatbot implementation for enhancing customer service in higher education: A case study from Universitas Negeri Semarang. Journal of Theoretical and Applied Information Technology, 102(14), 5690-5701.

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are we there yet?-a systematic literature review on chatbots in education. Frontiers in artificial intelligence, 4, 654924.

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

Martha, A. S. D., & Santoso, H. B. (2019). The design and impact of the pedagogical agent: A systematic literature review. Journal of educators Online, 16(1), n1.

Winkler, R., Hobert, S., Salovaara, A., Söllner, M., & Leimeister, J. M. (2020). Sara, the lecturer: Improving learning in online education with a scaffolding-based conversational agent.

Pérez, J. Q., Daradoumis, T., & Puig, J. M. M. (2020). Rediscovering the use of chatbots in education: A systematic literature review. Computer Applications in Engineering Education, 28(6), 1549-1565.

Hobert, S., & Meyer von Wolff, R. (2019). Say hello to your new automated tutor–a structured literature review on pedagogical conversational agents.

Hwang, G.-J., & Chang, C.-Y. (2023). A review of opportunities and challenges of chatbots in education. Interactive Learning Environments, 31(7), 4099-4112.

Ungless, E. L., Vitsakis, N., Talat, Z., Garforth, J., Ross, B., Onken, A., Kasirzadeh, A., & Birch, A. (2024). Ethics Whitepaper: Whitepaper on Ethical Research into Large Language Models. arXiv preprint arXiv:2410.19812.

Adamopoulou, E., & Moussiades, L. (2020a). Chatbots: History, technology, and applications. Machine Learning with applications, 2, 100006.

Adamopoulou, E., & Moussiades, L. (2020b). An overview of chatbot technology. Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part II 16,

Wen, X., Chen, J., Chen, Y., & Sheng, S. (2021). Chinese Sentence Semantic Matching with Multi-Granularity Based on Siamese Neural Network. SEKE,

Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. Education and Information Technologies, 28(1), 973-1018.

Xie, Q., Lu, W., Zhang, Q., Zhang, L., Zhu, T., & Wang, J. (2023). Chatbot integration for metaverse-a university platform prototype. 2023 ieee international conference on omni-layer intelligent systems (coins),

Chandra, Y. W., & Suyanto, S. (2019). Indonesian chatbot of university admission using a question answering system based on sequence-to-sequence model. Procedia Computer Science, 157, 367-374.

Oliveira, P. F., & Matos, P. (2023). Introducing a chatbot to the web portal of a higher education institution to enhance student interaction. Engineering Proceedings, 56(1), 128.

Martinez-Araneda, C., Gutiérrez, M., Maldonado, D., Gómez, P., Segura, A., & Vidal-Castro, C. (2024). Designing a Chatbot to support problem-solving in a programming course. INTED2024 Proceedings.

Evaluation of RAG (Retrieval-Augmented Generation) using RAGAs - On Human annotated and Synthetic Test Data Generation: Complete guide from theory to implementation. Available: https://medium.com/@rupeshyadav153/evaluation-of-retrieval-augmented-generation-rag-using-ragas-on-human-annotated-and-synthetic-09c4b825c298broader adoption in higher education institutions.