

A Retrieval-Augmented Multilingual Chatbot for Intelligent College Information Services

**Arun Babu P.¹, Pendyala Bhavya Sree², Lingala Anusha³,
Shaik Bepari Suhana Afreen⁴, Mirza Riyaz Baig⁵**

¹Assistant Professor, ^{2,3,4,5}Student, Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning), Rajeev Gandhi Memorial College of Engineering and Technology, Nandyal, India.

E-mail: ¹arunbabu1208@gmail.com, ²bhavyasree1412@gmail.com, ³lingalaanusha598@gmail.com,

⁴Shaikbeparisuhanaafreen@gmail.com, ⁵mriyazbaig765@gmail.com

Abstract

The study proposes an artificial intelligence-based multilingual university chatbot by utilizing the Retrieval-Augmented Generation (RAG) approach for intelligent institutional information retrieval. The developed chatbot leverages Sentence Transformer embeddings, FAISS-enabled semantic search, and the Llama-3 large language model for generating context-specific answers that depend on the institutional knowledge base. To ensure high availability, the chatbot offers support for multiple languages by enabling language translation and voice communication based on the Whisper speech-to-text model. Other functionalities of the chatbot include user authentication, image and document search, chat history management, and campus navigation. The proposed system was developed for Rajeev Gandhi Memorial College of Engineering and Technology (RGM CET), and the experiment was carried out with representative queries from the institution. The experimental results obtained were 92% response accuracy, 90% retrieval relevance, 1.8 s response time on average, and 91% user satisfaction, thus proving the efficiency of the suggested architecture in providing reliable and scalable academic information services.

Keywords: Chatbot, Retrieval-Augmented Generation, Large Language Models, Multilingual NLP, Educational Information Systems.

1. Introduction

The advancement of AI and NLP has contributed much to the development of conversational systems by providing a means for intelligent interaction between humans and computers through natural language. Chatbots have been widely used in various fields such as healthcare, customer support, finance, and education for providing automated responses without involving humans. Chatbots have also been used in the educational field for providing academic and administrative information.

Institutions have wide range of information in relation to admissions, courses offered, faculty members, infrastructure of university, exams conducted, and other services for students. Even though all the aforementioned information is readily available on the official website of the universities, it is also true that navigating this vast amount of information poses difficulties to people due to its ever-changing nature.

Recent developments in the field of LLMs have greatly enhanced the effectiveness of conversational AI through the generation of contextually appropriate responses. Yet, independent LLMs are likely to generate responses that are either inaccurate or out of date for specific domain queries since these depend mainly on pre-trained knowledge. The combination of retrieval and generation of language is achieved by RAG where the responses generated are based on the documents retrieved from a knowledge source other than LLM.

This paper proposes an AI-based multilingual chatbot for Rajeev Gandhi Memorial College of Engineering and Technology (RGM CET). The proposed system uses Sentence Transformer embeddings, semantic search with the help of FAISS, and the Llama-3 language model in a RAG model for providing precise information about the institution. In addition, the proposed chatbot offers multilingual communication, voice-based interaction using the Whisper model, authentication, retrieval of images and documents, and navigation through the campus. The proposed system architecture can prove to be an efficient and scalable solution for information access in educational institutions.

2. Related Works

Conversational AI has evolved in recent years to move away from rule-based techniques to more advanced techniques based on retrieval-augmented language models that can generate context-aware and factual responses. Lewis et al. [1] proposed the concept of Retrieval Augmented Generation (RAG), an approach where dense document retrieval is performed followed by language modeling. With the incorporation of external documents in the generation process, RAG improves the factual accuracy and decreases the probability of hallucinations.

Semantic retrieval is an essential part of a RAG system. The researchers from Facebook proposed the Facebook AI Similarity Search (FAISS) [2], an efficient library that can be used for indexing and searching dense vector representations. The FAISS allows for fast retrieval of the nearest neighbors in a high-dimensional space and is commonly used for the purpose of semantic search and conversational AI systems. In addition to the above-mentioned paper, the researchers from Google proposed BERT [9], while Devlin et al. proposed Sentence-BERT (SBERT) [3].

Knowledge-enriched conversational models have also helped in making chatbots more reliable. The work done by Ghazvininejad et al. [4] proved that the use of external knowledge in conversation generation results in greater relevancy of the generated responses than in case of using purely generative conversation models. Moreover, Brown et al. [8] presented an example of amazing language understanding ability of transformers.

Previous research on chatbots in educational contexts has established the utility of such tools in the context of education. The work by Winkler and Söllner [5] pointed to the efficacy of chatbots for providing administrative and student assistance, whereas the work by Lalwani et al. [10] showcased the development of chatbots using NLP. However, the latter was done using traditional NLP only.

With the advent of multilingual technologies, the possibilities of using chatbots have been enhanced even more. Zhang and Zong [6] examined the use of neural machine translation for multilingual communication, while Radford et al. [7] presented the Whisper model as a robust multilingual speech recognizer. Using such technologies, Medina et al. [11] came up with EduX-RAG, which is a framework for multilingual educational chatbot that uses

multilingual retrieval with RAG. Similarly, Yafie et al. [12] have also built an academic chatbot using RAG for campus information.

Despite these advancements, limited research integrates semantic retrieval, multilingual interaction, speech recognition, secure authentication, and campus navigation within a unified educational chatbot framework. The proposed work addresses this gap by combining these technologies into a comprehensive AI-driven multilingual college chatbot for accurate and context-aware institutional information retrieval.

3. Proposed Work

3.1 System Architecture

The proposed chatbot is an AI-enabled multilingual chatbot for colleges, which will offer contextual, relevant, and up-to-date information related to the institution to its students, parents, faculties, and visitors of the campus. The bot utilizes a framework called RAG (Retrieval-Augmented Generation), where the process of semantic document retrieval using a language model and LLM (Large Language Model) helps in producing relevant responses based on institutional knowledge. In contrast to only considering the pre-trained knowledge of the language model, the process starts by retrieving the most relevant data from a knowledge base.

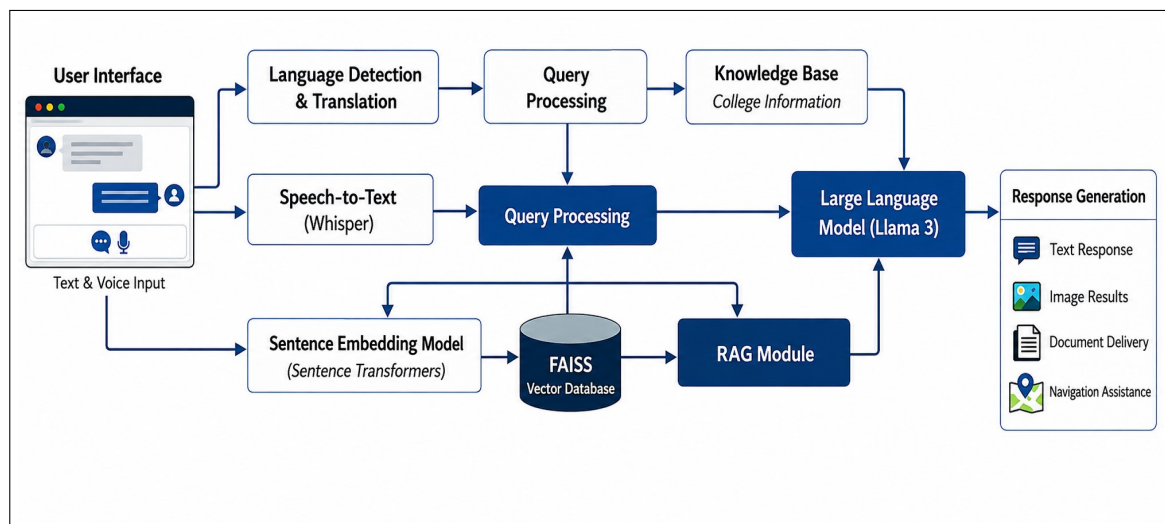


Figure 1. Architecture of the proposed AI-driven multilingual college chatbot

Figure 1 shows that the architecture comprises a number of connected modules such as user interface, language detection and translation module, speech-to-text module, query

processor, sentence embedding model, FAISS vector database, RAG model, knowledge repository of the institution, language model based on Llama, and response generator module. The modular structure facilitates the processing of the text and voice queries in different languages independently while maintaining scalability and maintainability of the entire system. Additionally, the architecture incorporates other services like image search, document delivery, and campus navigation.

3.2 User Interaction and Functional Design

The functional behavior of the chatbot is shown in the use case diagram presented in Figure 2 below. Two main actors are supported by the system; they are users and administrators. Users are allowed to register themselves, log in, ask questions using text or voice, see responses of the chatbot, get multilingual support, search for institution documents and pictures, navigate on campus and have access to past chat sessions. Administrator is involved in management of institution data and updating the knowledge base.

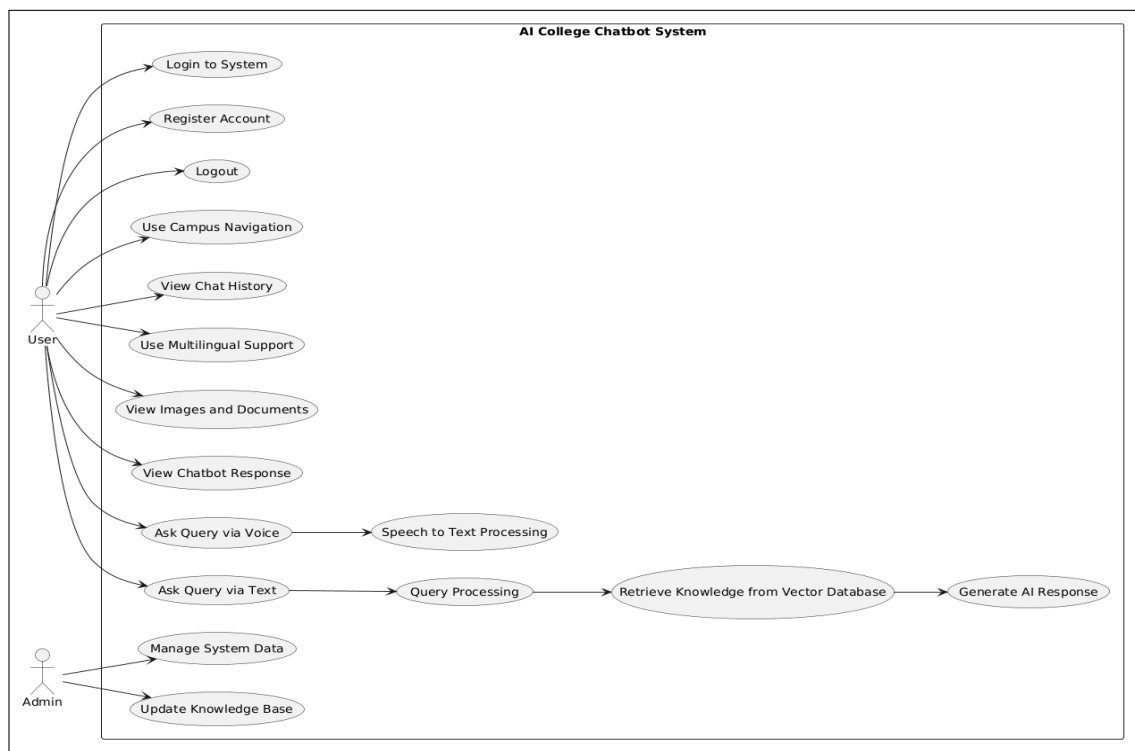


Figure 2. Use case diagram of user and administrator interactions with the proposed system

The segregation of the user and administrative roles ensures enhanced system security without the loss of up-to-date institutional knowledge. It is the task of the user authentication mechanism to guard individualized services like the chat history, while administrative tasks

are reserved for authorized individuals who will manage the knowledge database. The use case model therefore captures both operational functionality and system management within a unified framework.

3.3 Query Processing and Embedding Generation

When a user queries a question Q , the system initially does pre-processing tasks that include tokenization, normalization, and language detection. In the case of voice queries, the Whisper model for speech recognition transcribes the voice query to text. Let the processed query be represented as

$$Q = \{w_1, w_2, w_3 \dots w_n\} \quad (1)$$

where w_i denotes the individual tokens in the query.

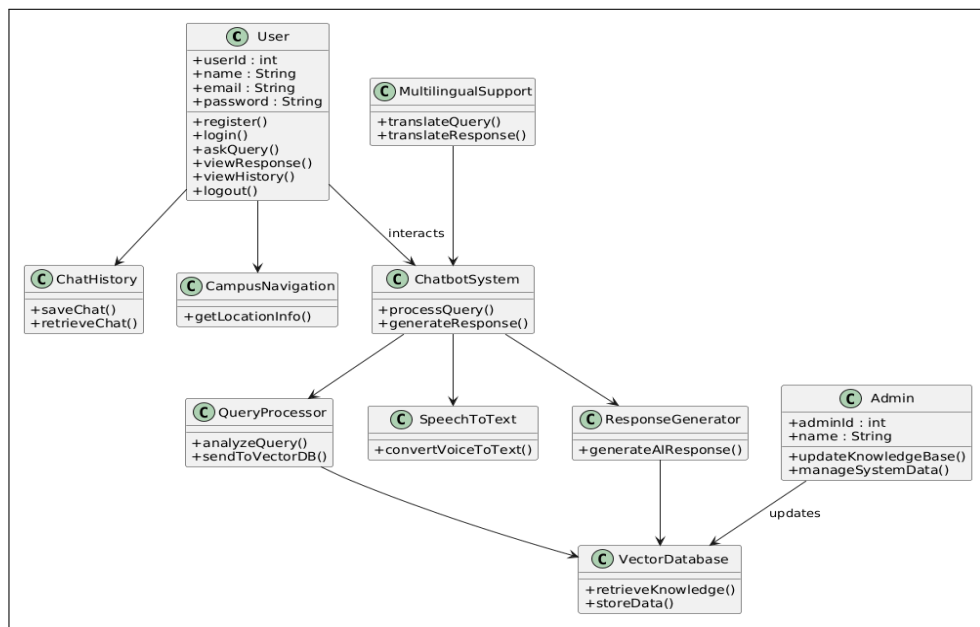


Figure 3. Class diagram - Relationships among the chatbot system components

The processed query is transformed into a dense semantic embedding using a Sentence Transformer model,

$$E_q = f(Q) \quad (2)$$

where $f(\cdot)$ represents the embedding function and E_q denotes the semantic embedding of the user query. The vector form represents the context of the query in question and makes semantic similarity matching possible between the query and institutional documents, rather than the traditional keyword matching method. The embedded query will be used as the input

of the semantic retrieval engine. Figure 3 depicts the interaction between the three components.

3.4 Knowledge Base Representation and Semantic Retrieval

Institutional information is collected from official college resources and partitioned into semantically coherent text chunks. Each document chunk D_i is encoded using the same embedding model employed for query representation,

$$E_i = f(D_i) \quad (3)$$

where E_i denotes the embedding corresponding to document chunk D_i . All document embeddings are indexed within the FAISS vector database to support efficient nearest-neighbor retrieval.

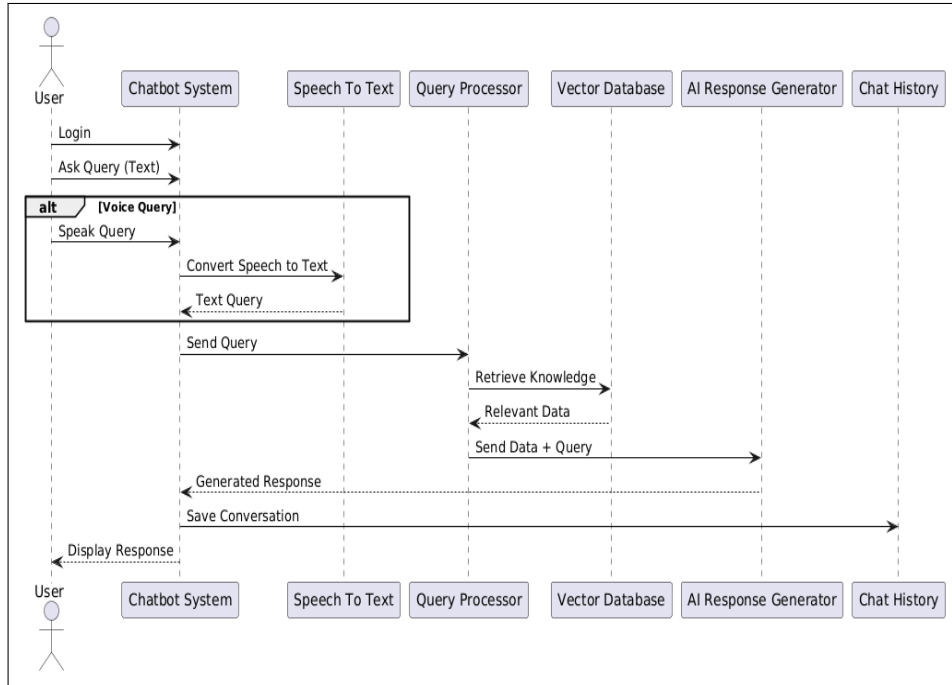


Figure 4. Sequence diagram of end-to-end query processing workflow

The similarity between the query embedding and each document embedding is computed using cosine similarity,

$$Sim(E_q, E_i) = \frac{E_q \cdot E_i}{\|E_q\| \|E_i\|} \quad (4)$$

The system retrieves the $top - k$ document embeddings with the highest similarity scores,

$$R = TopK(Sim(E_q, E_i)) \quad (5)$$

where R represents the retrieved document set. These retrieved knowledge chunks provide contextual evidence for response generation within the RAG framework. The overall retrieval workflow is illustrated in Figure 4.

3.5 Retrieval-Augmented Response Generation

The retrieved document set is represented as

$$R = \{D_1, D_2, D_3 \dots D_k\} \quad (6)$$

where k denotes the number of retrieved knowledge chunks. The augmented prompt comprising the user query and retrieved documents is supplied to the Llama-based language model to generate the final response,

$$A = LLM(Q, R) \quad (7)$$

where A represents the generated answer. Since the language model receives institution-specific contextual information, the generated responses are more accurate, reliable, and less prone to hallucination than those produced by standalone language models.

3.6 Multilingual and Voice Processing

In order to make the application accessible to users having different linguistic backgrounds, the suggested chatbot adopts a multilingual query processing technique and speech interface facility. In the beginning, the language of the user query is detected by the language detector. If the language being used is not English, then the query is translated into English.

Let Q_L denote the query in the user's native language Q_E represent its English translation. The translation process is expressed as:

$$Q_E = T(Q_L) \quad (8)$$

where $T(\cdot)$ denotes the multilingual translation function. All inputs that arrive in different languages are converted to a standard language, and thus the semantic embedding model and the vector retrieval engine work the same way without considering the user's preferred language. The output produced by the language model will then be converted back to the user's preferred language before delivering it via the chatbot interface.

Besides being able to interact via text, the chatbot also has the ability to process voice

queries due to the incorporation of the Whisper speech recognition model. As soon as the user makes the query in voice form, the voice signal is converted into text and then processed. The speech-to-text conversion is represented as

$$T = \text{Whisper}(S(t)) \quad (9)$$

where $S(t)$ denotes the input speech signal and T represents the transcribed text. This generated text is then pre-processed, represented using the Sentence Transformer model, and passed to the semantic retrieval component. With multilingual translation alongside speech recognition capability, the suggested chatbot presents a user-friendly interface that can help users from multiple languages interact with the system, without making any changes to the process of retrieval and response generation.

3.7 Campus Navigation and Backend deployment

In addition to Q&A, the proposed system utilizes an intelligent campus navigation feature that helps students, parents, and visitors to find the location of any particular place within the campus. The intelligent navigation system uses the Haversine formula to calculate the shortest distance between the user's present location and the selected destination. The distance is calculated as

$$d = 2r \sin^{-1} \left(\sqrt{\sin^2 \left(\frac{\phi_2 - \phi_1}{2} \right) + \cos(\phi_1) \cos(\phi_2) \sin^2 \left(\frac{\lambda_2 - \lambda_1}{2} \right)} \right) \quad (10)$$

where d represents the distance between two locations, r is the Earth's radius, ϕ_1 and ϕ_2 denote the latitudes, and λ_1 and λ_2 represent the longitudes of the source and destination, respectively. The computed distance enables the chatbot to provide location-aware navigation assistance and improve the overall user experience for campus visitors.

The back-end architecture of the suggested solution is built upon Flask, which facilitates coordination in the interaction between the user interface, query execution engine, semantic search engine, and the Large Language Model. The database is built on top of MongoDB, which stores user credentials, authentication information, chat history, and organization-related data.

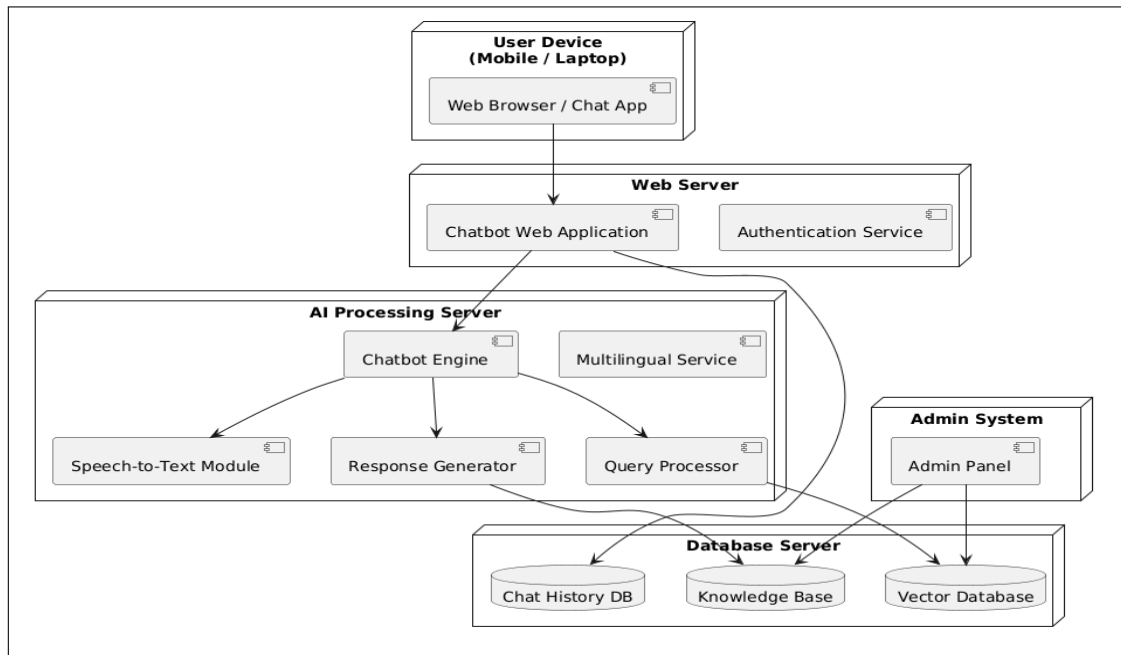


Figure 5. Deployment architecture of the proposed chatbot system

As shown in Figure 5, the deployment architecture is designed using a layered architecture including the client layer, the web server, AI processing server, and the database server. The AI processing server will include the chatbot engine, multilingual service, query processor, response generator, and speech-to-text processing module while the database server will consist of the institutional knowledge database, FAISS vector database, and the chat history database.

4. Results and Discussion

The AI-based multilingual chatbot designed for the college has been tested with the help of a prototype version created for Rajeev Gandhi Memorial College of Engineering and Technology (RGM CET). The test was conducted with an objective to establish the effectiveness of the Retrieval-Augmented Generation (RAG) model in retrieving institution-specific data and formulating relevant responses. Relevant queries have been tested for admissions, departments, faculty, facilities, placements, and administrative services.

4.1 Experimental Validation

The implementation of the chatbot was done based on the architecture discussed in Section 3, where institutional documents were first encoded as semantic embeddings based on the Sentence Transformer model and were stored in a vector database FAISS. In the course

of evaluation, every query was encoded as an embedding vector, and the nearest matching documents from the indexed knowledge base were obtained using the cosine similarity search. The returned document fragments served as context to generate grounded answers using the Llama-3 language model. Thus, the evaluation proved that retrieval provided relevant information from the institution, which reduced the hallucinations of the language model.

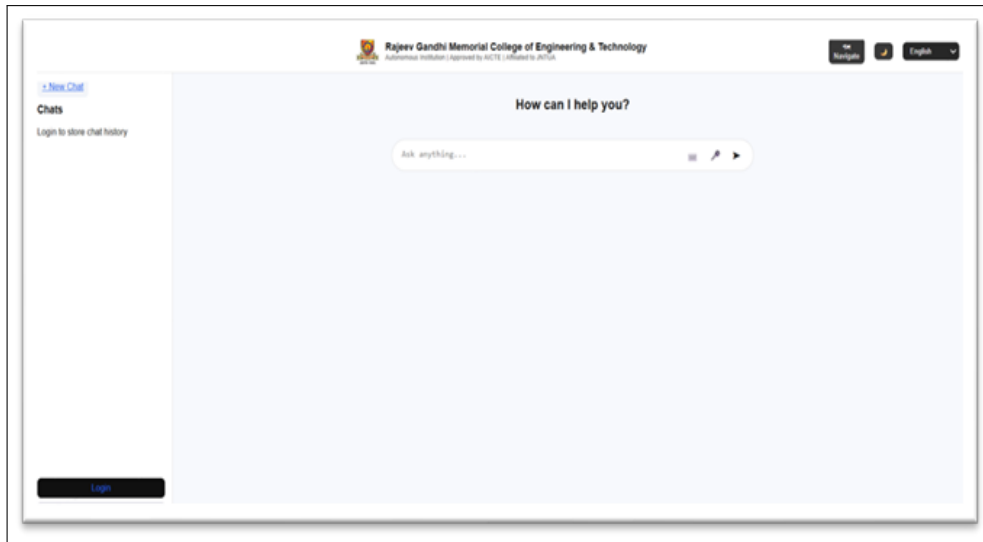


Figure 6. User interface of the proposed multilingual college chatbot

The graphical user interface illustrated in Figure 6 shows the conversation interface that was used to submit text and speech queries. This interface makes the interaction process easier by abstracting away the complexities of the retrieval pipeline.

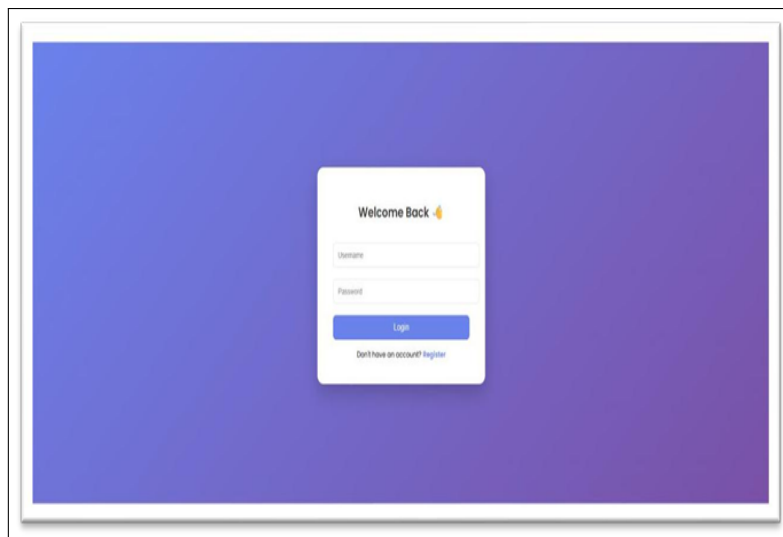
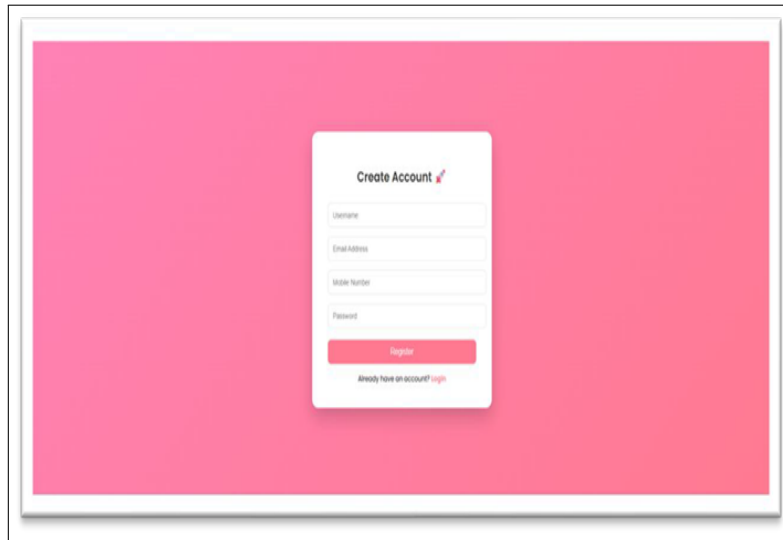


Figure 7. Secure user login interface for chatbot authentication

The authentication process was tested using the secured login interface as shown in

Figure 7, to ensure that personalized applications such as chat history and user profile could be accessed by only authorized users.



The image shows a 'Create Account' form with a pink background. The form is white and contains the following fields: Username, Email Address, Mobile Number, and Password. Below the Password field is a red 'Register' button. At the bottom of the form, there is a link that says 'Already have an account? login'.

Figure 8. User registration interface

Similarly, the registration module illustrated in Figure 8 managed to create user accounts and store encrypted credentials in MongoDB, proving the robustness of the authentication module.

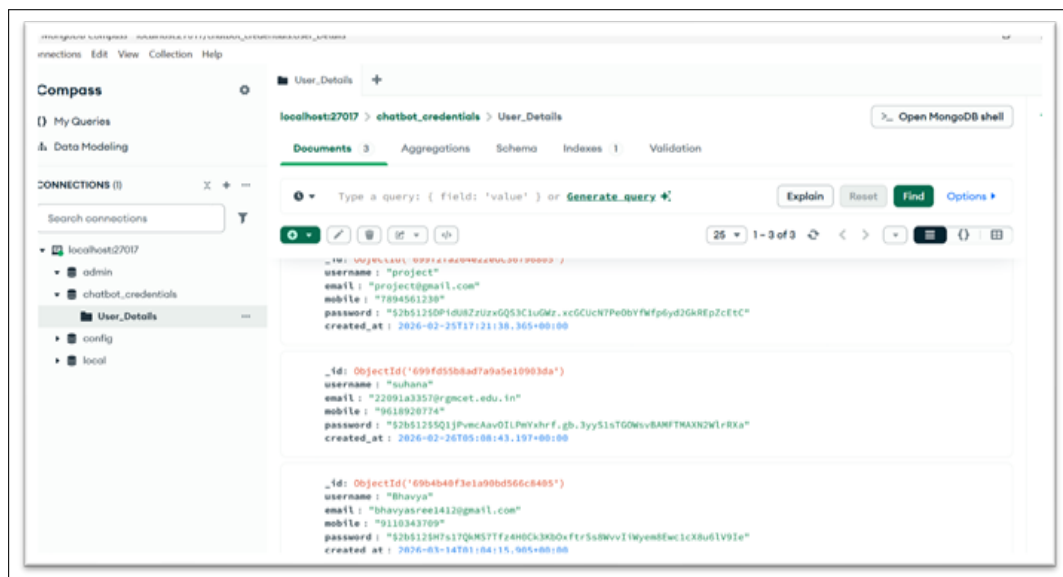


Figure 9. MongoDB database storing user authentication information

The successful storage of user information is verified in Figure 9, which presents the MongoDB database containing registered user records

4.2 Retrieval Performance and Response Analysis

To evaluate the effectiveness of the RAG model, the relevance of the retrieved knowledge and correctness of the generated response to institution-specific queries were considered. In this case, semantic embedding of every query was generated using the Sentence Transformer model and compared with document embeddings using FAISS. Document chunks that were retrieved had relevance to the information sought despite being dissimilar in terms of wording between the institution documents and user query.



Figure 10. Sample institutional response generated by the chatbot

The retrieved contextual data was used along with the original user query before the data was passed to the Llama-3 language model. The responses generated from this process were consistent with the retrieved institutional document and were able to answer user queries related to admissions, courses, faculty members, and facilities at the institution. Figure 10 shows an example of a chatbot response generated through the proposed retrieval pipeline framework.

4.3 Performance Evaluation

The performance metrics illustrated in Table 1 reveal the efficiency of the RAG framework in providing reliable institutional knowledge using semantic retrieval and context-based response generation.

Table 1. Performance evaluation of the proposed chatbot system

Metrics	Results
Accuracy	92%
Response Time	1.8s
Retrieval Relevance	90%
User Satisfaction	91%

As seen from the performance results, the combination of Sentence Transformer embeddings with the FAISS vector search is successful in finding semantically relevant knowledge from the repository, which helps in the generation of answers by the language model that stay in line with the context obtained during the retrieval stage. Thus, the performance proves that the grounding of the response generation stage is successfully achieved through the retrieval stage.

5. Conclusion

This research study has introduced an AI-powered multilingual college chatbot with a Retrieval-Augmented Generation architecture for intelligent institutional knowledge retrieval. Using the combination of Sentence Transformer embeddings, FAISS vector search, and Llama-3 language model, the developed system is able to provide context-aware answers based on institutional knowledge verification. The addition of multilingual translation, Whisper speech recognition, secure user authentication, image/document search, and campus navigation makes the system more accessible for users. Results of experimental evaluation prove that the proposed approach is capable of retrieving relevant knowledge and providing reliable answers with interactive performance. Further research will focus on the implementation of knowledge base synchronization, personalization, mobile application support, and multilingual solutions.

References

- [1] Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." Advances In Neural Information Processing Systems 2020, vol. 33, 9459-9474.

- [2] Jégou, Hervé, Matthijs Douze, Jeff Johnson, Lucas Hosseini, and Chengqi Deng. "FAISS: Similarity Search and Clustering of Dense Vectors Library." *Astrophysics Source Code Library* 2022: ascl-2210.
- [3] Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence Embeddings Using Siamese Bert-Networks." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th international Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, 3982-3992.
- [4] Ghazvininejad, Marjan, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. "A Knowledge-Grounded Neural Conversation Model." In *Proceedings of the AAAI conference on artificial intelligence* 2018, vol. 32, no. 1.
- [5] Winkler, Rainer, and Matthias Söllner. "Unleashing the Potential of Chatbots in Education: A State-of-the-Art Analysis." In *Academy of Management Proceedings* 2018, Briarcliff Manor, NY 10510: Academy of Management, vol. 2018, no. 1, 15903.
- [6] Zhang, Jiajun, and Chengqing Zong. "Neural Machine Translation: Challenges, Progress and Future." *Science China Technological Sciences* 2020, vol. 63, no. 10, 2028-2050.
- [7] Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. "Robust Speech Recognition via Large-Scale Weak Supervision." In *International Conference on Machine Learning*, PMLR 2023, 28492-28518.
- [8] Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. "Language models are Few-Shot Learners." *Advances in Neural Information Processing Systems* 2020, vol. 33, 1877-1901.
- [9] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 2019, vol. 1, 4171-4186.
- [10] Lalwani, Tarun, Shashank Bhalotia, Ashish Pal, Vasundhara Rathod, and Shreya Bisen. "Implementation of a Chatbot System using AI and NLP." *International Journal of Innovative Research in Computer Science & Technology (IJIRCST)* 2018, vol. 6, no. 3, 26-30.

- [11] Medina, Johanne, Soon Gyo-Jung, Joni Salminen, Kholoud Khalil Aldous, and Bernard J. Jansen. "EduX-RAG: Retrieval Augmented Generation Framework for Cross-Lingual Educational Chatbots." 3rd International Conference on Foundation and Large Language Models (FLLM), IEEE 2025, 465-473.
- [12] Yafie, Haddad Alwi, Achmad Udin Zailani, Widang Muttaqin, Muhammad Sheva Atallah Daffansyah, and Muhammad Rafi Ramzi. "Academic Chatbot for Campus Information Services Using Retrieval-Augmented Generation." Sinkron: jurnal dan penelitian teknik informatika 2026, vol. 10, no. 3, 1330-1339.