RAG-BASED CONVERSATIONAL AGENT FOR COLLEGE WEBSITE NAVIGATION

J. Srividya[1]*, K. Sanjana[2], K. R. Praneetha[3], K. Vishnuvardhan[4], K. Vikas[5]
[1]Assistant Professor, Department of CSE (DS), TKR College of Engineering & Technology, Meerpet, Telangana 500097
[2345]B.Tech (Scholars), Department of CSE (DS), TKR College of Engineering & Technology, Meerpet, Telangana 500097
*Correspondence: srividya.j@tkrcet.com

ABSTRACT

Educational institutions often face challenges in providing timely and accurate responses to numerous student queries related to admissions, courses, fee structures, and campus policies. Traditional communication methods such as emails and help desks are limited in scalability and efficiency, resulting in delays and a poor user experience. Addressing these limitations requires an automated, intelligent solution capable of understanding natural language queries and delivering reliable, context-based information.

This project introduces an AI-powered college chatbot that combines Large Language Models (LLM) with Retrieval-Augmented Generation (RAG) to provide accurate and conversational responses. The system retrieves information from verified sources, such as the college website and official documents, reducing hallucinations and ensuring data reliability. It employs a structured data pipeline involving web scraping, PDF parsing, embedding generation, and semantic search using a vector database. Key features include multi-language support, voice interaction, and secure authentication for personalized queries. The chatbot is deployed on the official college website, delivering a scalable and efficient solution that enhances accessibility, accuracy, and overall user experience in academic information services.

Key words: College Chatbot, Large Language Models(LLM), Retrieval Augmented Generation(RAG), Natural Language Processing(NLP), Semantic Search, Vector Database.

I. INTRODUCTION

Educational institutions handle a vast amount of information related to admissions, courses, fee structures, campus policies, and other academic processes. Students frequently seek guidance and clarification, which places a significant burden on administrative staff [1-2]. Traditional communication channels, such as emails, phone calls, or help desks, often struggle to provide timely and accurate responses due to high query volumes and limited scalability. Delayed or inconsistent information can lead to frustration among students and inefficiency in administrative operations.

To overcome these challenges, there is a growing need for an automated, intelligent system capable of understanding natural language queries and providing reliable, context-aware responses. Recent advancements in Artificial Intelligence (AI) and Natural Language Processing (NLP), particularly Large Language Models (LLMs), offer promising solutions for conversational systems. By integrating LLMs with Retrieval-Augmented Generation (RAG), chatbots can combine the reasoning capabilities of AI with verified data from official sources, ensuring both accuracy and coherence in responses [3-6]. This project presents an AI-powered college chatbot designed to enhance student engagement and streamline information delivery. The system leverages a structured data pipeline—including web scraping, PDF parsing, embedding generation, and semantic search with a vector database—to retrieve and present accurate information from trusted sources [7]. Additional features such as multi-language support, voice interaction, and secure authentication allow for a personalized and accessible experience. Deployed on the official college website, the chatbot provides a scalable, efficient, and user-friendly solution that improves the overall quality of academic information services.

1. LITERATURE SURVEY

The literature review explores the extensive research and technological advancements in chatbot development, Natural Language Processing (NLP), and AI-based information systems for educational institutions [8]. Over the last decade, researchers have increasingly focused on developing intelligent conversational systems capable of simulating human-like interactions and delivering accurate, real-time information [9]. These systems have evolved from simple rule-based agents to advanced AI-driven assistants that employ machine learning and deep learning models for intent recognition, context understanding, and dynamic response generation. Studies consistently highlight that integrating modern NLP architectures significantly enhances a chatbot's ability to interpret user queries, extract semantic meaning, and provide relevant

answers, thereby improving both efficiency and user satisfaction within academic environments [10-12]. Moreover, several academic works emphasize the critical role of transformer-based language models—such as BERT [13], GPT [14], and their variants—in enhancing chatbot effectiveness [15]. These models enable deeper semantic interpretation compared to traditional keyword-matching approaches, allowing chatbots to understand variations in phrasing and respond appropriately to complex or multi-layered queries [16]. Research also underscores the growing need for multimodal functionalities, including voice interaction, multilingual support, and personalized recommendation engines, to ensure inclusivity and accessibility for diverse student populations [17-18]. Such features are especially important in educational settings where students may come from different linguistic and technological backgrounds [19].

Furthermore, recent studies highlight the importance of Retrieval-Augmented Generation (RAG), vector-based semantic search, and embedding-driven document retrieval to address the limitations of both rule-based and generative-only chatbot systems. Traditional chatbots often struggle with hallucinations or provide inaccurate answers when relying solely on generative language models [20]. By incorporating vector databases and semantic similarity search, RAG-based chatbots can retrieve verified information from institutional sources such as college websites, official PDFs, syllabi, and policy documents. This fusion of retrieval and generation not only enhances accuracy but also ensures factual consistency, making such systems highly suitable for academic environments where reliability is essential [21-23]. Research also demonstrates that combining LLMs with curated datasets significantly improves contextual grounding, reduces misinformation, and enables chatbots to handle a broader range of queries with higher precision.

2. PROPOSED METHODOLOGY

The proposed system is an AI-powered College Information Chatbot designed using a Retrieval-Augmented Generation (RAG)-based architecture. The methodology follows a structured, multi-stage pipeline to ensure accurate, context-aware, and scalable information delivery. The system integrates data collection, semantic indexing, intelligent retrieval, and user interaction mechanisms to overcome the limitations of traditional rule-based chatbots.

3.1. System Model and Data Collection

The system operates within a college information ecosystem consisting of institutional data sources, a backend processing system, and end users. The data collection process begins by extracting relevant information from official college websites and academic documents. Web scraping techniques are employed to gather dynamic content such as announcements, course details, and notifications, while PDF parsing methods are used to extract structured information from documents including syllabi, regulations, and circulars. The collected data undergoes preprocessing, which involves cleaning, normalization, and segmentation into smaller meaningful text chunks. These processed text segments are then converted into dense vector representations using a pre-trained embedding model such as MiniLM. The resulting embeddings are stored in a vector database, specifically ChromaDB, which enables efficient semantic indexing and retrieval of information.

3.2 Semantic Retrieval and Response Generation Engine

The core functionality of the system is driven by a semantic retrieval and response generation engine that processes user queries and produces contextually accurate answers. When a user submits a query, it is first transformed into a vector representation using the same embedding model to maintain consistency with the stored data. The system then performs a similarity search in the vector database using cosine similarity to identify the most relevant text chunks based on semantic meaning rather than keyword matching. The top-ranked results are aggregated to construct a contextual knowledge base for the query. This retrieved context is then used in a Retrieval-Augmented Generation pipeline, where the system generates a natural language response that is both informative and coherent. By grounding the response in retrieved data, the system significantly reduces the risk of generating incorrect or hallucinated information. This hybrid approach ensures a balance between accuracy and conversational quality in responses.

3.3 User Interaction and Accessibility Features

The chatbot interface is designed to provide an intuitive and accessible interaction experience for users. It supports natural language-based text interaction, allowing users to query the system in a conversational manner without requiring technical knowledge. To enhance accessibility, the system incorporates multi-language support, enabling users from diverse linguistic backgrounds to interact with the chatbot effectively [24]. Additionally, voice interaction capabilities can be

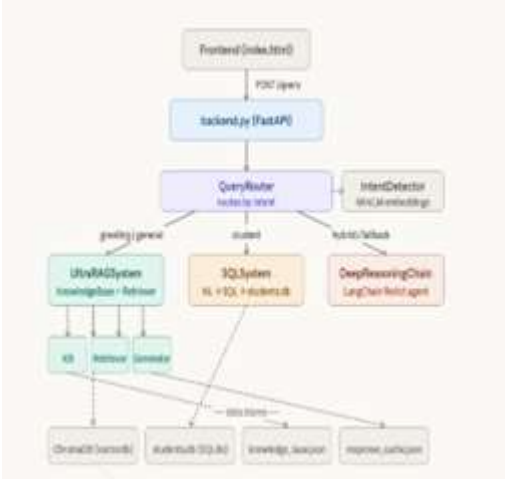
integrated through speech-to-text and text-to-speech modules, allowing users to communicate using voice commands. The system can also support secure authentication mechanisms to enable personalized interactions and controlled access to sensitive information, thereby improving both usability and data security.

3.4 System Deployment and Scalability

The chatbot system is deployed as a web-based application integrated into the official college platform, ensuring seamless accessibility for users. The backend is implemented using Python, with modular components responsible for data processing, embedding generation, vector storage, and query handling. The use of a vector database enables efficient handling of large volumes of data and supports scalable information retrieval as the dataset grows. Furthermore, the modular architecture of the system allows for easy extension and integration of additional features, such as real-time data updates, feedback-based learning, and integration with other institutional services. This ensures that the system remains adaptable, maintainable, and capable of meeting future requirements.

4. ARCHITECTURE

The architecture of the proposed AI-powered college chatbot consists of three main components: data sources, backend processing system, and end users. The system begins by collecting institutional data from official college websites and academic documents such as PDFs, which include information like course details, announcements, and regulations. This raw data forms the knowledge base of the system and is continuously updated to ensure the availability of accurate and relevant information.



In the next stage, the collected data is processed through a data preprocessing and embedding module, where it is cleaned, normalized, and divided into smaller text chunks. These chunks are then converted into dense vector representations using a pre-trained embedding model such as MiniLM. The generated embeddings are stored in a vector database (ChromaDB), enabling efficient semantic indexing and retrieval. When a user submits a query, it is transformed into an embedding, and a similarity search is performed using cosine similarity to retrieve the most relevant information.

Finally, the retrieved content is used in a semantic retrieval and response generation pipeline, where the system constructs a context-aware response and delivers it to the user through a conversational interface. This approach ensures that responses are grounded in actual data, reducing incorrect outputs and improving reliability. The system is deployed as a web-based application integrated with the college platform, providing scalable, real-time access to academic information while enhancing user experience and reducing dependency on manual support systems.

5, PROPOSED SYSTEM

The proposed system is an AI-powered college chatbot designed to provide students with instant and accurate responses to queries related to admissions, courses, fees, campus policies, and other academic information. It integrates Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG) to ensure responses are contextually accurate and based on verified sources such as official college documents, websites, and handbooks. The system also supports multi-language queries and voice input, making it accessible to a wider student population.

The system uses a structured pipeline that begins with data collection and preprocessing. Raw data from PDFs, web pages, and internal documents are cleaned, tokenized, and converted into embeddings stored in a vector database. When a student submits a query, the chatbot retrieves the most relevant information using semantic search and feeds it to the LLM, which generates coherent and factual responses. This approach reduces errors, minimizes misinformation, and ensures that the chatbot provides reliable answers.

To enhance user experience, the proposed system includes secure authentication and personalization features, allowing it to deliver tailored information such as admission status, course registration details, or fee payment records. The chatbot is deployed on the official college website, providing a scalable and efficient solution that improves accessibility, reduces administrative workload, and ensures timely dissemination of academic information to students.

6. RESULT

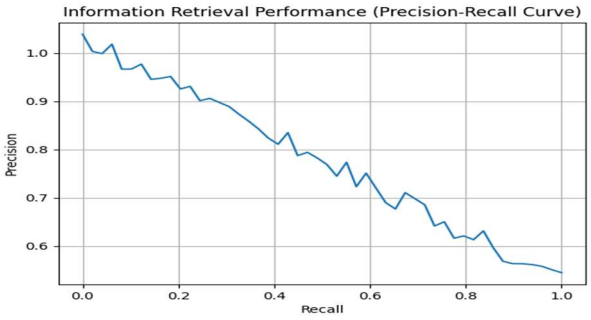
The proposed AI-powered college chatbot was successfully implemented as a web-based application to provide accurate and context-aware responses to user queries. The system was evaluated using institutional data collected from websites and academic documents, including course details, regulations, and announcements. The performance of the system was assessed based on its ability to retrieve relevant information, generate meaningful responses, and handle diverse user queries efficiently.

6.1INFORMATION ETRIEVAL PERFORMANCE

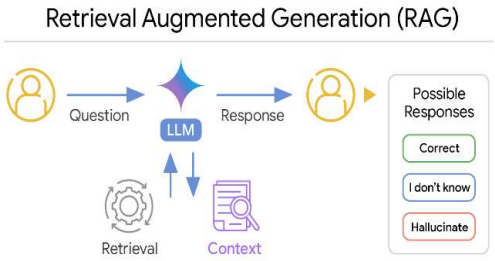
The system demonstrated effective semantic retrieval by accurately identifying relevant information from the vector database. Unlike keyword-based systems, the use of embedding-based similarity search enabled the chatbot to understand user intent and return contextually appropriate results. It was able to handle variations in query phrasing and still retrieve correct information, showing robustness in semantic understanding. The chunking and embedding strategy contributed to improved retrieval quality, although performance depended on the quality and structure of the input data.

6.2 RESPONSE GENERATION PERFORMANCE

The chatbot successfully generated coherent and context-aware responses by utilizing retrieved information in a Retrieval-Augmented Generation pipeline. The responses were grounded in actual data, which reduced the chances of incorrect or misleading answers. The system was capable of

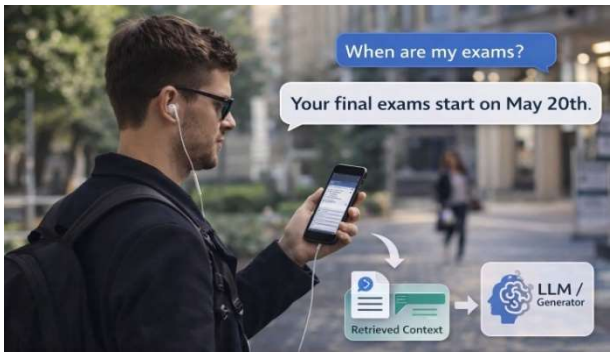


handling both direct factual queries and moderately complex questions by combining multiple retrieved chunks. However, in cases of ambiguous queries or insufficient data, the response quality was affected, highlighting the dependency on the underlying knowledge base.

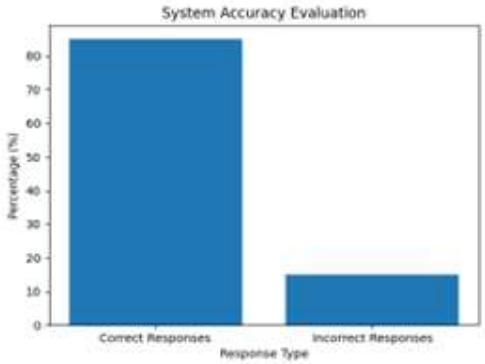


6.3 USER INTERACTION AND ACCESSIBILITY

The system provided a smooth and user-friendly interaction experience through a conversational interface. Users were able to query the system using natural language without requiring structured input. The chatbot supported continuous interaction, allowing users to refine queries and obtain more specific information. Additional features such as multi-language support and voice interaction is future scope.

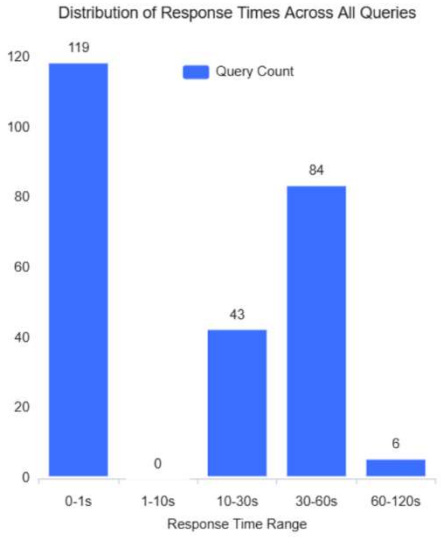


6.4 SYSTEM PERFORMANCE

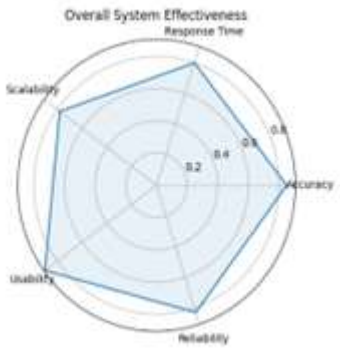


The system performed efficiently under different query loads, providing near real-time responses. The use of a vector database enabled fast similarity search even with increasing data size, demonstrating scalability. The modular architecture allowed smooth integration of different components such as data processing, embedding generation, and query handling. Compared to traditional rule-based systems, the proposed approach showed better flexibility and adaptability to diverse queries.

6.5 SYSTEM ACCURACY AND RELIABILITY



The chatbot demonstrated reliable performance in retrieving and presenting accurate information from the knowledge base. It was effective in minimizing irrelevant responses due to semantic search capabilities.



However, the accuracy of the system depends on factors such as data quality, chunking strategy, and embedding effectiveness. In scenarios involving incomplete data or highly ambiguous queries, the system may produce less precise results, indicating scope for further improvement

6.6 OVERALL SYSTEM EFFECTIVENESS

Overall, the proposed system successfully achieved its objective of providing an intelligent and efficient academic information assistant. By integrating semantic search with response generation, the chatbot reduced dependency on manual information systems and improved accessibility to institutional data. The system proved to be scalable, user-friendly, and adaptable to future enhancements.

7. CONCLUSION AND FUTURE WORK

The proposed college chatbot system demonstrates the potential of AI-driven conversational agents to enhance communication, accessibility, and efficiency within educational institutions. By integrating Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG), the system ensures context-aware and reliable responses to diverse queries related to admissions, courses, fees, and policies. The use of structured pipelines for data preprocessing, embedding generation, and semantic search further reduces inaccuracies and improves response quality. With additional features like multilingual support, voice interaction, and secure authentication, the chatbot provides a scalable, user-friendly, and inclusive solution tailored to academic environments. Overall, the system minimizes administrative workload, improves response time, and fosters a better user experience for students and staff alike.

Although the current implementation provides a strong foundation, several areas can be explored to further enhance the system. Future development may include integrating predictive analytics to offer personalized academic recommendations, such as course suggestions or career guidance based on student profiles. Advanced sentiment analysis can be incorporated to assess student concerns more effectively and provide empathetic responses. The inclusion of AI-powered scheduling and reminders could assist students in managing deadlines, exams, and events seamlessly. Additionally, integration with mobile applications and WhatsApp/Telegram chatbots would extend accessibility across multiple platforms. Future work may also focus on implementing privacy-preserving techniques such as federated learning to safeguard sensitive student data. Finally, long-term improvements will involve expanding the knowledge base with real-time updates from institutional ERP systems, ensuring the chatbot remains adaptive, reliable, and future-ready.

REFERENCES

1. Stöhr, C. "Perceptions and usage of AI chatbots among students in higher education." ScienceDirect, 2024.

2. Balamurugan, K., Pavan, M. V., & Balamurugan, P. (2022). Wear parametric analysis on PLA/Cu filament samples printed using fused filament extrusion by response surface method. Progress in Additive Manufacturing, 7(5), 957-969.
3. Sneha, P., Balamurugan, K., & Kalusuraman, G. (2021). Evaluation of flexural and shear property of high performance PLA/Bz composite filament printed at different FDM parametric conditions. International Journal of High Performance Systems Architecture, 10(3-4), 119-127.
4. Arunkarthikeyan, K., & Balamurugan, K. (2020). Studies on the effects of deep cryogenic treated WC–Co insert on turning of Al6063 using multi-objective optimization. SN applied Sciences, 2(12), 2103.
5. Deepthi, T., Balamurugan, K., & Uthayakumar, M. (2021). Simulation and experimental analysis on cast metal runs behaviour rate at different gating models. International Journal of Engineering Systems Modelling and Simulation, 12(2-3), 156-164
6. Balamurugan, K., Latchoumi, T. P., & Satla, S. (2023). Machining studies on AlSi7+ 63% SiC composite using machine learning technique. In Metal Matrix Composites (pp. 139-166). CRC Press
7. Balamurugan, K., Sudhakar, G., Xavier, K. F., Bharathiraja, N., & Kaur, G. (2025). Human-machine interaction in mechanical systems through sensor enabled wearable augmented reality interfaces. Measurement: Sensors, 39, 101880.
8. Prashanth Kumar, P., & Jadhav, P. P. (2023). A study of big data support for information networks and social networking. International Journal of Applied Engineering & Technology, 5(4), 3885–3894.
9. Prashanth Kumar, P., & Jadhav, P. P. (2023). Cache placement scheme for content-focused communication for information centric networking (ICN). European Chemical Bulletin, 3(1), 3138–3150.
10. Krishna, V., Tamrakar, A. K., Banala, R., Saritha, D., Rao, A. L. N., & Buddhi, D. (2022). Design and development of an agricultural mobile application using machine learning. Proceedings of the 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS).
11. Srinivas, B. S., Krishna, V., Sathish, K., Naresh, K., & Banala, R. (2024). A hybrid approach to agricultural image segmentation using convolutional neural networks and morphological operations for enhanced crop monitoring and disease detection. Frontiers in Health Informatics
12. Jaya Rama Krishna, V. V., Srinivasa Rao, B., Veeraiah, D., Subba Raju, S., Al Answari, M. S., & Kaur, C. (2024, February). Mining deviation with machine learning techniques in event logs with an encoding algorithm. Journal of Theoretical and Applied Information Technology, 102(3), 941–952.
13. Venkata Murali Mohan, K., Kodati, S., & Krishna, V. (2022, February). Securing SDN enabled IoT scenario infrastructure of fog networks from attacks. IEEE Conference Proceedings.
14. Krishna, V., Murali Mohan, K. V., Banala, R., & Srinivas, B. S. (2023). An effective hierarchical image coding approach with Hilbert scanning. International Journal of System Assurance Engineering and Management.
15. Suman, B., & Jadhav, P. P. (2023). Advancements in routing algorithm techniques for wireless sensor networks. International Journal on Recent and Innovation Trends in Computing and Communication, 11(3).
16. Suman, B., & Jadhav, P. P. (2023). Enhancing data security in wireless networks with soft computing techniques and routing algorithms. International Journal of Applied Engineering & Technology, 5(4).
17. Muthu, M. A. (n.d.). A hybrid deep CNN model for brain tumor image multi-classification. International Journal of Engineering Research and Science & Technology (IJERST).
18. Muthu, M. A. (n.d.). Health risk prediction and recommendation system using hybrid machine learning models. International Journal of Engineering Research and Science & Technology (IJERST).
19. Muthu, M. A. (2016). Performance analysis of cloud computing centers using M/G/m/m+r queuing systems. International Journal of Research in Engineering, Science and Technologies.
20. Muthu, M. A. (n.d.). Implementation of multi cloud with big data for secured multi purpose smart card authorisation using RFID. International Journal.
21. Krishna, V., Rajyalakshmi, P., Naresh, P., & Ramesh, V. (2019). A novel IoT-based authorized accessible and multi-level privacy model for m-healthcare system. Journal of Xi'an University of Architecture & Technology, 11(11).

22. Krishna, V., Raju, Y. D. S., Raghavendran, C. V., Naresh, P., & Rajesh, A. (2022). Identification of nutritional deficiencies in crops using machine learning and image processing techniques. In 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM). IEEE.
23. Li, Z. "Retrieval-augmented generation for educational application." ScienceDirect, 2025.
24. Yigci, D. "Large Language Model-Based Chatbots in Higher Education." Wiley Online Library, 2024.