iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.78524

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Artificial Intelligence Chatbot for College Using Retrieval Augmented Generation

Aditya Sadafale[1], Adesh Gawande[2], Sumit Muddgal[3], Sheikh Abdul Kadir Sheik Salim[4], Darshan Parbhat[5], Gitesh Hikar[6], Sunpreet Kaur Nanda[7]

[1, 2, 3, 4, 5, 6]*Students,* [7] *Assistant Professor, Electronics and Communication Engineering Department, P.R. Pote Patil College of Engineering and Management, Amravati, India*

I. INTRODUCTION

The increasing volume of digital information available on institutional websites has created challenges for users attempting to retrieve specific details efficiently. College websites typically contain information about departments, faculty members, academic programs, admission procedures, and administrative resources. However, this information is often distributed across multiple web pages and documents, making it difficult for users to locate relevant content quickly.

Traditional website search mechanisms rely primarily on keyword matching. These systems fail to capture the semantic meaning of user queries and often produce incomplete or irrelevant results. With the advancement of Natural Language Processing (NLP) and Large Language Models (LLMs), conversational interfaces have emerged as a promising solution for improving information accessibility.

This paper proposes an AI chatbot capable of understanding natural language queries and retrieving relevant institutional information using semantic similarity techniques. The system uses a Retrieval- Augmented Generation (RAG) framework that combines vector-based semantic retrieval with language model–based response generation. The approach ensures that responses are grounded in actual institutional data rather than relying solely on generative model knowledge.

II. LITERATURE REVIEW

Several studies have explored the development of AI-based chatbots for information retrieval and conversational assistance. Early chatbot systems relied on rule-based mechanisms and predefined responses, limiting their ability to handle complex queries. With the development of machine learning techniques, chatbots began to incorporate NLP models for improved language understanding.

Recent advancements in transformer-based architectures have significantly enhanced conversational AI systems. Models such as BERT, GPT, and other transformer-based networks have enabled semantic understanding and contextual response generation. However, purely generative models often suffer from hallucination, where the model produces inaccurate or fabricated information.

Retrieval-Augmented Generation (RAG) has emerged as a hybrid approach that combines information retrieval with language generation. In RAG systems, relevant documents are first retrieved from a knowledge base and then used as context for response generation. This approach improves response accuracy by grounding the generated output in real data.

Previous work has demonstrated the effectiveness of semantic embeddings and vector similarity techniques in building efficient information retrieval systems. Embedding models convert textual data into numerical vectors that capture semantic relationships between words and sentences. These vectors can then be compared using similarity metrics such as cosine similarity to identify relevant information.

III. METHODOLOGY

The proposed system follows a modular architecture consisting of data collection, text processing, semantic embedding, retrieval, and response generation components. Initially, institutional data is collected using automated web scraping techniques. The scraper extracts textual information from the college website, including departmental descriptions, faculty details, and other academic content. The extracted data is stored in a structured format for further processing. The collected textual data is then processed using a chunking mechanism that divides large documents into smaller segments. This segmentation improves the efficiency and accuracy of the retrieval system by allowing more granular representation of information. Each text segment is converted into a numerical vector representation using a transformer-based embedding model. The

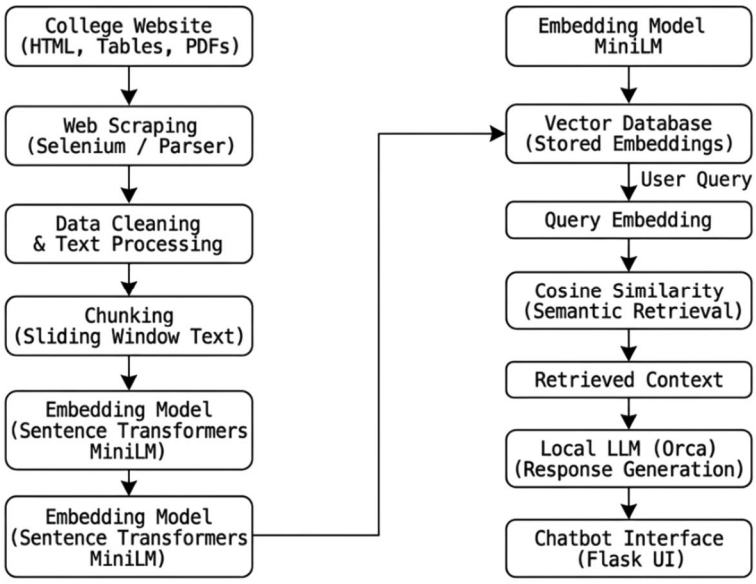


Figure 1 : Methodology of the AI Chatbot for College Information System

embedding model maps semantically similar sentences into nearby points in a high-dimensional vector space. When a user submits a query, the system generates an embedding vector for the query and compares it with stored document embeddings using cosine similarity. The similarity scores determine which text segments are most relevant to the query. The top-ranked segments are retrieved and passed to a local Large Language Model. The model receives both the retrieved context and the user query as input and generates a structured response. This combination of retrieval and generation forms the basis of the RAG architecture used in the system.

IV. SYSTEM IMPLEMENTATION

The chatbot system is implemented using Python and several open-source libraries. Web scraping is performed using Selenium to extract relevant information from dynamic web pages. The scraped content is stored in JSON format and processed using a text segmentation module. Semantic embeddings are generated using a Sentence Transformer model, which converts each text chunk into a high-dimensional vector representation. These embeddings form the system's knowledge base.

The retrieval module performs similarity comparisons between the query embedding and stored embeddings using cosine similarity. The most relevant text segments are selected based on their similarity scores. The generation module integrates a local Large Language Model through the GPT4All framework. The model processes the retrieved context and user query to generate natural language responses. The chatbot interface is implemented using a Flask-based backend and a simple web user interface. Users can interact with the system by entering queries, and the chatbot responds with relevant information retrieved from the institutional knowledge base.

V. RESULT AND EVALUATION

The developed system was evaluated by testing multiple user queries related to departmental information, faculty details, and academic programs. The semantic retrieval approach demonstrated improved accuracy compared to traditional keyword-based search methods.

For example, when users submitted queries containing synonyms or natural language phrasing, the system was able to identify relevant

information even when exact keywords were not present in the source documents. The integration of semantic embeddings allowed the chatbot to understand the intent behind queries and retrieve appropriate content.

The Retrieval-Augmented Generation framework ensured that generated responses were grounded in retrieved institutional data. This approach reduced the likelihood of hallucinated responses and improved the reliability of the chatbot.

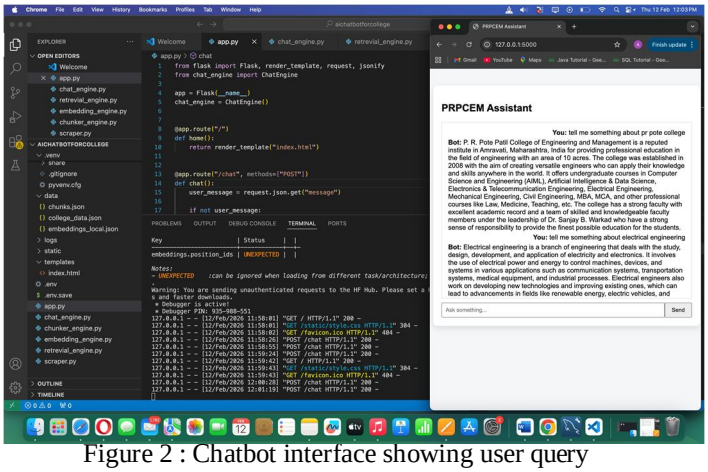


Figure 2 : Chatbot interface showing user query

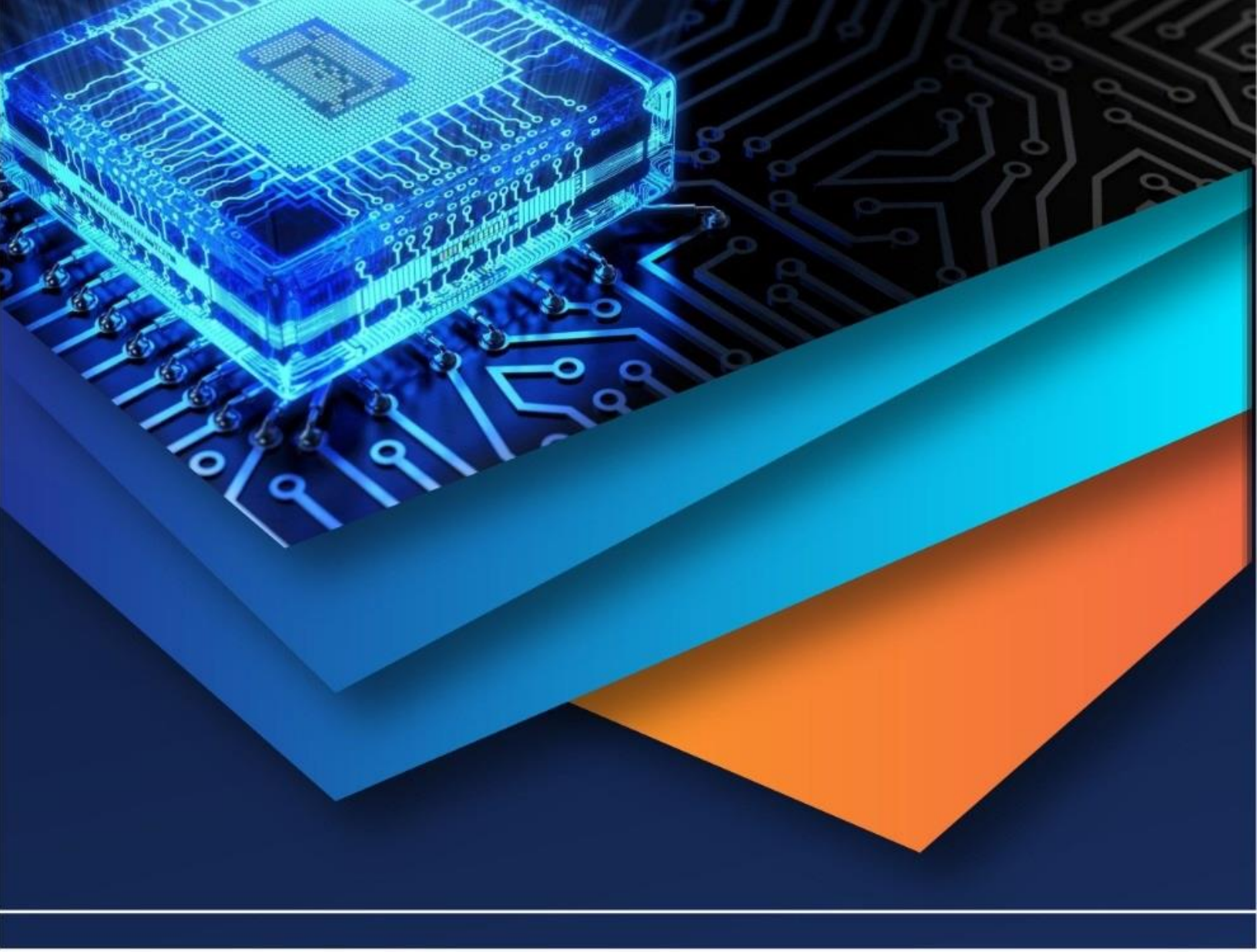
VI. CONCLUSION

This paper presented the design and implementation of an AI chatbot for retrieving institutional information using a Retrieval-Augmented Generation architecture. The system integrates web scraping, semantic embeddings, cosine similarity retrieval, and a local Large Language Model to deliver context-aware responses to user queries. The results demonstrate that combining semantic retrieval with generative models provides a more effective approach to institutional information access than traditional keyword-based search systems. The modular architecture of the system allows for future scalability and integration with advanced vector databases and cloud-based language models.

Future work may include implementing scalable vector indexing methods such as FAISS, integrating cloud-based LLM APIs for improved response quality, and enabling automated data ingestion pipelines to keep the knowledge base updated.

REFERENCES

[1] P. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” Advances in Neural Information Processing Systems (NeurIPS), 2020.
[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Proceedings of NAACL-HLT, 2019.
[3] A. Vaswani et al., “Attention Is All You Need,” Advances in Neural Information Processing Systems, 2017.
[4] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT Networks,” Proceedings of EMNLP, 2019.
[5] T. Brown et al., “Language Models are Few-Shot Learners,” Advances in Neural Information Processing Systems, 2020.
[6] Meta AI, “LLaMA: Open and Efficient Foundation Language Models,” 2023.
[7] HuggingFace, “Sentence Transformers: Multilingual Sentence Embeddings,” https:// www.sbert.net
[8] GPT4All Documentation, “Running Local Large Language Models,” https://docs.gpt4all.io



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY