iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: III Month of publication: March 2025

DOI: https://doi.org/10.22214/ijraset.2025.67435

Open-Source RAG Chatbot

Mr. Palanivel Ramaswamy[1], Abhishek Verma[2], Aditya Kumar[3], Vinay Kumar[4]
[1] Assistant Professor, Dept. of Artificial Intelligence and Data Science, Nitte Meenakshi Institute of Technology, Bengaluru, India
[2, 3, 4]Dept. of Artificial Intelligence and Data Science, Nitte Meenakshi Institute of Technology, Bengaluru, India

Abstract: This project presents an open-source, document-based chatbot that utilizes retrieval-augmented generation (RAG) techniques to enhance user interactions. Built on the FLAN-T5 language model and leveraging the LangChain framework, the chatbot is capable of generating contextually relevant responses based on the content of uploaded documents. By combining information retrieval with text generation, the system improves its ability to handle complex queries, offering accurate and efficient conversational responses. The framework is designed to support various document formats and can be customized for specific domains or applications. This makes the chatbot highly adaptable for use cases such as customer support, education, and knowledge management, where document-based interactions are essential. The project aims to provide an efficient, scalable, and flexible solution for integrating AI-driven conversational agents with document retrieval capabilities.
Keywords: Retrieval-Augmented Generation (RAG), FLAN-T5, Natural Language Processing (NLP), LangChain Framework, FAISS, Embeddings, Streamlit, Conversational AI, MLOps.

I. INTRODUCTION

The rapid evolution of artificial intelligence (AI) has paved the way for more advanced and efficient chatbot systems, capable of understanding and responding to user queries in a highly contextual manner. One such advancement is the integration of Retrieval-Augmented Generation (RAG), which combines the strengths of information retrieval and natural language generation to enhance conversational agents. This technology allows chatbots to generate contextually relevant and accurate responses by retrieving and analyzing information from large sets of documents or databases.

In particular, the FLAN-T5 language model, developed by Meta, offers powerful capabilities in text generation, enabling chatbots to engage in dynamic and meaningful conversations. By leveraging frameworks like LangChain and tools such as FAISS for efficient document retrieval, RAG-based systems can process unstructured data more effectively, making them highly suitable for applications in domains like customer support, education, and knowledge management [1].

This project aims to develop an Open-Source RAG Chatbot that utilizes the FLAN-T5 model and LangChain framework to provide a scalable and flexible solution for document-driven interactions. The system supports multiple document formats (e.g., PDF), enabling users to upload files and receive context-aware responses based on the content. The integration of embeddings, vector stores, and conversational memory ensures that the chatbot not only retrieves relevant data but also maintains an ongoing, coherent interaction with users.

II. LITERATURE REVIEW

The literature review highlights the evolution and application of key tools and technologies in the development of Retrieval-Augmented Generation (RAG)-based chatbots. Transformer-based language models like FLAN-T5 and have emerged as foundational technologies for natural language understanding and text generation, enabling intelligent, context-aware, and dynamic interactions. The integration of RAG methodologies, supported by frameworks such as LangChain and vector retrieval tools like FAISS, has significantly enhanced the accuracy, efficiency, and scalability of conversational systems.

[2] Research studies demonstrate the effectiveness of embedding techniques and document loaders in managing diverse file formats, while platforms like Streamlit simplify user interface development for seamless interaction. These advancements are critical for addressing challenges in unstructured data retrieval and domain-specific chatbot deployment. The studies also highlight the growing focus on ethical AI practices, empathy in responses, and scalability through MLOps. Future research is oriented toward improving response quality, exploring multimodal data processing, and deploying robust cross-platform solutions, ensuring these systems remain versatile and impactful across industries.

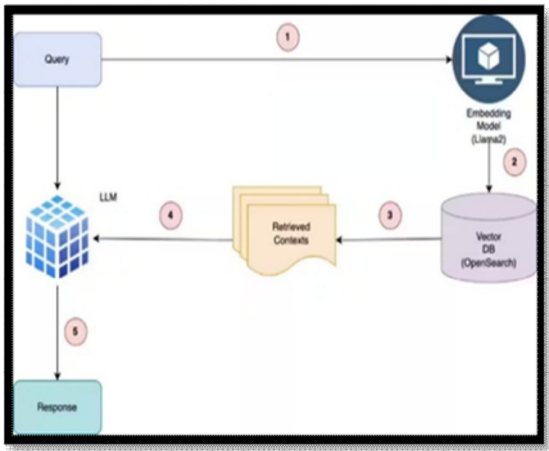
The development of Retrieval-Augmented Generation (RAG) systems for document-driven chatbots has been extensively studied in recent years. Key advancements in AI-driven retrieval systems and language models have laid the groundwork for building robust and scalable conversational agents.

The literature underscores the importance of RAG systems in bridging the gap between retrieval and generative capabilities of AI. Innovations in transformer models, embeddings, and frameworks like LangChain have made it feasible to build advanced, domain-specific chatbots. These studies form the foundation for designing scalable and user-friendly RAG chatbots.

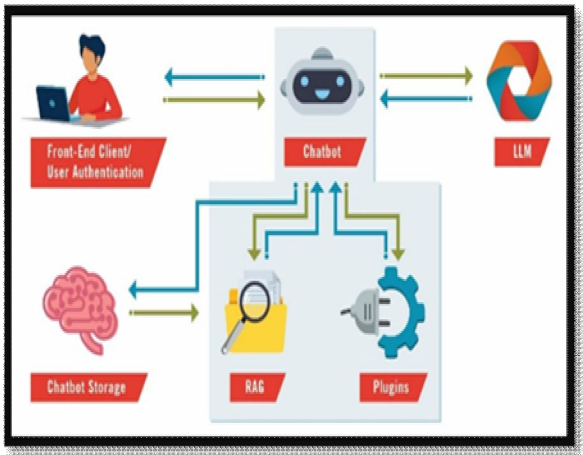
III. METHODOLOGY

The methodology outlines the structured approach to developing the Open-Source RAG Chatbot, focusing on the integration of document retrieval and language generation.

1) *Problem Definition:* Address the need for a chatbot capable of understanding and interacting with unstructured document-based data.
2) *System Design:* Incorporate a modular framework that includes a conversational LLM (FLAN-T5) and a retrieval mechanism (FAISS). Ensure scalability and adaptability to various domains, emphasizing a user-friendly interface.
3) *Document Processing:* Enable users to upload multi-format documents (PDF).Utilize loaders and parsers to extract text content for further processing.
4) *Text Splitting and Embedding:* Split documents into manageable chunks using text splitters (e.g., CharacterTextSplitter).Generate embeddings using HuggingFace sentence transformers to encode the text into vector representations.
5) *Vector Store Creation:* Store embeddings in FAISS to enable efficient similarity searches for user queries.
6) *Conversational Chain Setup:* Configure the conversational retrieval chain using LangChain. Integrate memory management for maintaining conversation history.
7) *Interaction Flow:* Develop an interactive interface with Streamlit for document uploads and chat functionalities [3,4].



Process for generating responses to queries using an LLM and a vector database.



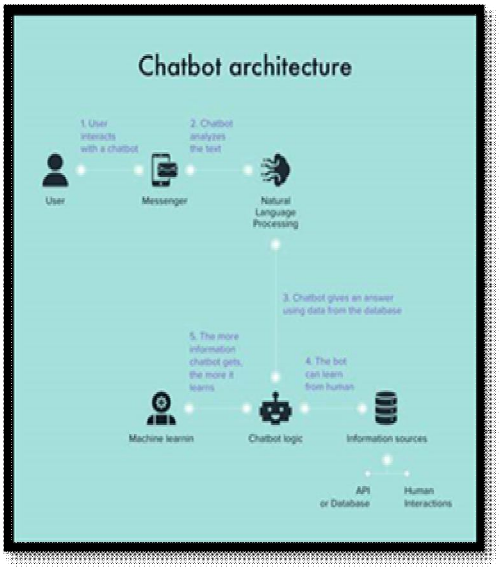
Architecture of a chatbot system, highlighting the key components and their interactions.

IV. IMPLEMENTATION

The implementation phase involves the technical execution of the designed methodology:

1) *Environment Setup:* Install and configure essential libraries, including LangChain, HuggingFace Transformers, FAISS, and Streamlit. Use Python for backend logic and model integration.
2) *Document Upload and Preprocessing:* Develop an upload feature in Streamlit to handle multi-format documents. Parse and clean the text content, preparing it for text splitting and embedding.
3) *Embedding and Vector Store Creation:* Use HuggingFace sentence-transformers to generate embeddings from document chunks. Store embeddings in FAISS for fast and accurate retrieval.
4) *LLM Integration and Conversational Chain:* Load the FLAN-T5 model using LangChain and configure it for conversational AI. Link the retrieval chain to the LLM, ensuring it uses the vector store for context-based responses.
5) *User Interface:* Build a user-friendly Streamlit interface for interaction. Display conversational history, response outputs, and document processing progress.
6) *Testing and Optimization:* Evaluate performance on various queries and optimize response time and accuracy. Fine-tune system parameters (e.g., token limits, chunk size, overlap) to balance performance and resource usage [5].

This phased methodology and structured implementation ensure the chatbot is functional, scalable, and capable of delivering meaningful, context-aware interactions with users.


Chatbot architecture
1. User interacts with a chatbot
2. Chatbot analyzes the text
User
Messenger
Natural Language Processing
3. Chatbot gives an answer using data from the database
5. The more information chatbot gets, the more it learns
4. The bot can learn from human
Machine learnin
Chatbot logic
Information sources
API or Database
Human Interactions

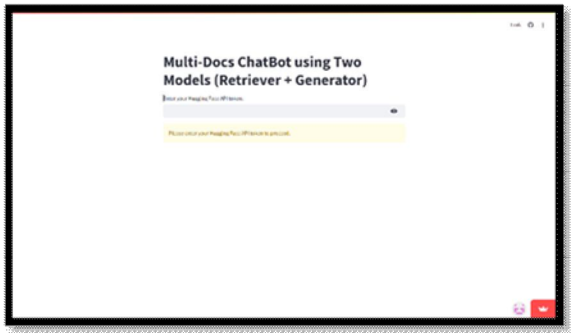
Architecture of a chatbot, breaking down its components and how they interact to generate responses.

V. RESULTS

The Open-Source RAG Chatbot achieved significant performance in accurately retrieving relevant information from uploaded documents and generating context-aware responses. During testing, the chatbot successfully processed various document formats such as PDF, demonstrating robust compatibility and versatility. Users benefited from a seamless interface built using Streamlit, which simplified document uploads and facilitated interactive conversations. The system maintained low response times, ensuring real-time query handling suitable for diverse applications [6,7].

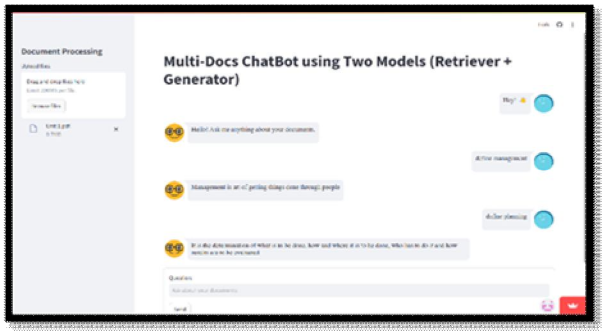
A key strength of the chatbot is its integration of FLAN-T5 with LangChain, leveraging retrieval-augmented generation (RAG) to enhance accuracy and minimize hallucinations. This design effectively bridges the gap between static document storage and dynamic conversational AI. The modular framework also supports scalability, making it adaptable for specific use cases, including customer support, education, and corporate training. However, challenges remain, such as occasional parsing errors due to poor document quality and the need for more domain-specific fine-tuning to improve contextual understanding.

[8] Looking ahead, enhancements such as advanced memory features, cross-platform deployment, and improved mobile compatibility are anticipated. Security improvements for handling sensitive data will further increase the chatbot’s applicability in critical fields like healthcare and legal services. These developments will solidify the chatbot's role as a reliable and scalable solution for document-driven conversational tasks.



Multi-Docs ChatBot using Two Models (Retriever + Generator)

Initial interface of the Streamlit application.



Document Processing

Multi-Docs ChatBot using Two Models (Retriever + Generator)

Hello! Ask me anything about your documents.

Management is art of getting things done through people

Updated interface provides a more interactive and user-friendly experience for interacting with the multi-Docs ChatBot.

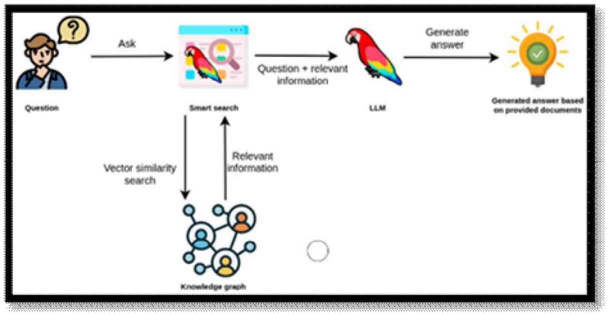
VI. APPLICATIONS

1) *Customer Support:* Efficiently answer user queries by leveraging company policy documents, user manuals, and knowledge bases.
2) *Education:* Serve as an academic assistant, providing answers to questions derived from textbooks, research papers, and notes.
3) *Corporate Training:* Facilitate onboarding and training by delivering context-aware responses based on corporate policies and guidelines.
4) *Healthcare:* Assist in medical documentation and answer queries based on healthcare protocols and research articles.
5) *Legal Consultation:* Provide information extracted from legal documents, case studies, and policy files, supporting legal professionals.
6) *Knowledge Management:* Streamline document-based query handling in organizations, improving internal knowledge accessibility.
7) *Research Assistance:* Aid researchers by extracting and summarizing relevant information from large datasets and reports.
8) These applications highlight the broad scope of RAG chatbots in transforming information-driven industries by enhancing accessibility and efficiency [9,10].

VII. CONCLUSION

The Open-Source RAG Chatbot represents a significant advancement in integrating document retrieval and conversational AI. By leveraging FLAN-T5, LangChain, and retrieval-augmented generation (RAG) techniques, the chatbot bridges the gap between static information storage and dynamic user interaction. It provides accurate, context-aware responses by processing documents in various formats, making it a versatile tool for domains such as customer support, education, and corporate training.

The modular and scalable design of the chatbot ensures adaptability to diverse use cases, while its open-source nature fosters collaborative development and customization. Despite its current strengths, including high accuracy and seamless user interaction, the system presents opportunities for improvement, such as enhanced domain-specific fine-tuning, long-term memory integration, and cross-platform compatibility [11]. In conclusion, this project highlights the potential of RAG-based conversational systems in transforming how organizations and individuals' access and interact with document-based information. With continued advancements and adoption, such chatbots can play a pivotal role in streamlining processes, improving productivity, and democratizing AI-driven solutions [12].



Process for generating answers to user questions using a knowledge graph and an LLM.

REFERENCES

[1] Harshit Kumar Chaubey, G. Tripathi, R. Ranjan, and S. k Gopalaiyengar, “Comparative Analysis of RAG, Fine-Tuning, and Prompt Engineering in Chatbot Development,” pp. 169–172, Aug. 2024, doi: https://doi.org/10.1109/icftss61109.2024.10691338.

[2] Harcharan Singh Kabbay, “Streamlining AI Application: MLOps Best Practices and Platform Automation Illustrated through an Advanced RAG based Chatbot,” pp. 1304–1313, Jul. 2024, doi: https://doi.org/10.1109/icscss60660.2024.10625230.

[3] Ananya G, “RAG based Chatbot using LLMs,” INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT, vol. 08, no. 06, pp. 1–5, Jun. 2024, doi: https://doi.org/10.55041/ijsrem35600.

[4] S. Vakayil, D. Sujitha Juliet, Anitha. J, and Sunil Vakayil, “RAG-Based LLM Chatbot Using Llama-2,” Apr. 2024, doi: https://doi.org/10.1109/icdcs59278.2024.10561020.

[5] Rohan Paul Richard, Ebenezer Veemaraj, Juanith Mathew Thomas, J. Mathew, C. Stephen, and Richie Suresh Koshy, “A Client-Server Based Educational Chatbot for Academic Institutions,” pp. 1–5, Jun. 2024, doi: https://doi.org/10.1109/conit61985.2024.10627567.

[6] B. Saha and U. Saha, "Enhancing International Graduate Student Experience through AI-Driven Support Systems: A LLM and RAG-Based Approach," 2024 International Conference on Data Science and Its Applications (ICoDSA), Kuta, Bali, Indonesia, 2024, pp. 300-304, doi: 10.1109/ICoDSA62899.2024.10651944.

[7] D. Saraswat, “AI - Driven Pedagogies: Enhancing Student Engagement and Learning Outcomes in Higher Education,” International Journal of Science and Research (IJSR), vol. 13, no. 11, pp. 1152–1154, Nov. 2024, doi: https://doi.org/10.21275/sr241119221041.

[8] Mutiara Auliya Khadija, A. Aziz, and Wahyu Nurharjadmo, “Automating Information Retrieval from Faculty Guidelines: Designing a PDF-Driven Chatbot powered by OpenAI ChatGPT,” Oct. 2023, doi: https://doi.org/10.1109/ic3ina60834.2023.10285808.

[9] V. Bhat, S. D. Cheerla, J. R. Mathew, N. Pathak, G. Liu and J. Gao, "Retrieval Augmented Generation (RAG) Based Restaurant Chatbot with AI Testability," 2024 IEEE 10th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService), Shanghai, China, 2024, pp. 1-10, doi: 10.1109/BigDataService62917.2024.00008.Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[10] G. Gamage et al., "Multi-Agent RAG Chatbot Architecture for Decision Support in Net-Zero Emission Energy Systems," 2024 IEEE International Conference on Industrial Technology (ICIT), Bristol, United Kingdom, 2024, pp. 1-6, doi: 10.1109/ICIT58233.2024.10540920.

[11] A. T. Neumann, Y. Yin, S. Sowe, S. Decker and M. Jarke, "An LLM-Driven Chatbot in Higher Education for Databases and Information Systems," in IEEE Transactions on Education, doi: 10.1109/TE.2024.3467912.

[12] N. S. Amarnath and R. Nagarajan, "An Intelligent Retrieval Augmented Generation Chatbot for Contextually-Aware Conversations to Guide High School Students," 2024 4th International Conference on Sustainable Expert Systems (ICSES), Kaski, Nepal, 2024, pp. 1393-1398, doi: 10.1109/ICSES63445.2024.10762977.