International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 06 | June - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

RAG based Chatbot using LLMs

Ananya G
Department of ISE
R V College of Engineering
Bengaluru, India

Dr. Vanishree K
Department of ISE
R V College of Engineering
Bengaluru, India

***Abstract*—Historically, Artificial Intelligence (AI) was used to understand and recommend information. Now, Generative AI can also help us create new content. Generative AI builds on existing technologies, like Large Language Models (LLMs) which are trained on large amounts of text and learn to predict the next word in a sentence. Generative AI can not only create new text, but also images, videos, or audio. This project focuses on the implementation of a chatbot based the concepts of Generative AI and Large Language Models which can answer any query regarding the content provided in the PDFs. The primary technologies utilized include Python libraries like LangChain, PyTorch for model training, and Hugging Face's Transformers library for accessing pre-trained models like Llama2, GPT-3.5 (Generative Pre-trained Transformer) architectures. The responses are generated using the Retrieval Augmented Generation (RAG) approach. The project aims to develop a chatbot which can generate the sensible responses from the data in the form of PDF files. The project demonstrates the capabilities and applications of advanced Natural Language Processing (NLP) techniques in creating conversational agents that can be deployed across various platforms in the corporation, to enhance user interaction and support automated tasks.**

***Index Terms*—Generative AI, Artificial Intelligence, Natural Language Processing, Large Language Model, Llama2, Transformers, Document Loaders, Retrieval Augmented Generation, Vector Database, Langchain, Chainlit**

I. Introduction

In today's digital age, the demand for intelligent conversational agents, known as chatbots, has surged dramatically. These chatbots, powered by cutting-edge technologies such as Large Language Models (LLMs) and advanced Natural Language Processing (NLP) techniques, have revolutionized how businesses and organizations interact with their customers and users. In line with this technology, the project aims to develop a sophisticated chatbot utilizing LLMs and related technologies, specifically trained on a set of emails. Leveraging the Retrieval-Augmented Generation (RAG) approach within the Python programming language, the chatbot will be capable of understanding user queries, retrieving relevant information from a corpus of email data, and generating contextually appropriate responses. The utilization of LLMs, such as Llama2, Llama3, Mistral, GPT (Generative Pretrained Transformer), combined with the RAG architecture, offers unparalleled capabilities in natural language understanding and generation. By training the chatbot on a specific set of emails, it is ensured that the chatbot is tailored to the domain-specific needs and queries encountered in real-world email communications. This approach enables the chatbot to provide accurate and relevant responses to user inquiries, thereby enhancing user experience and streamlining communication processes.

By harnessing the power of LLMs, the project aims to create a chatbot that can understand natural language queries, generate contextually relevant responses, and provide valuable assistance to users within the company. The project idea is proposed in the desire to leverage cutting-edge AI advancements to enhance user interactions and streamline communication processes. Understanding LLMs and NLP is essential for developing advanced AI systems, chatbots, language models, and applications that require robust natural language understanding and generation capabilities. These technologies are revolutionizing how computers interact with and process human language, enabling a wide range of innovative applications across industries, which opens a wide range of learning opportunities.

II. Literature Review

[1] The review suggested that chatbots can be used everywhere because of its accuracy, lack of dependability on human resources and 24x7 accessibility. In recent years, advancements in technologies such as Artificial Intelligence (AI), Big Data, and Internet of Things (IoT) have revolutionized various industries. Among these innovations, Chatbots, or conversational AIs, have emerged as a significant application. Chatbots, powered by AI and Natural Language Processing (NLP), simulate human conversation, offering automation and efficiency across diverse domains like education, healthcare, and business. Through a review of existing literature, this study explores the types, advantages, and disadvantages of chatbots, highlighting their versatility, accuracy, and ability to operate continuously without reliance on human resources.

[2] The paper presents a college inquiry chatbot as a solution to challenges in locating specific information, especially for non-affiliated visitors in the college website. While GUI and web-based interfaces are mainstream, alternative interfaces occasionally emerge to address specific needs. Powered by AI and NLP algorithms, the developed chatbot intelligently handle queries related to various college activities, including examination cell, admission, academics, attendance, placement, and more.

[3] The paper talks about the challenges posed by the pandemic, accessing health-care services has become increasingly difficult. To address this issue, a chatbot application

leveraging Natural Language Processing (NLP) and machine learning concepts is proposed. This chatbot system, developed using supervised machine learning, aims to provide disease diagnosis and treatment recommendations with detailed descriptions of various illnesses before consulting with a doctor. The application features a GUI-based text assistant for user-friendly interaction, allowing users to input symptoms and risk factors for their condition. The chatbot then offers personalized suggestions, including analgesics and advice on when to seek physical medical attention.

[4] This paper introduces a Retrieval Augmented Generation (RAG) approach for constructing a chatbot that addresses user queries using Frequently Asked Questions (FAQ) data. Leveraging Large Language Models (LLMs), particularly a paid ChatGPT model, the system utilizes contextual question answering capabilities acquired through training. The paper outlines the training of an in-house retrieval embedding model using infoNCE loss, showcasing its superior performance over a general-purpose public embedding model in terms of retrieval accuracy and Out-of-Domain (OOD) query detection. Furthermore, the paper explores the optimization of LLM token usage and associated costs using Reinforcement Learning (RL), proposing a policy-based model external to the RAG pipeline. This model interacts with the pipeline through policy actions, updating policies to optimize costs.

[5] This study addresses the challenge of integrating Large Language Models (LLMs) into corporate environments where internal data utilization is limited. It proposes a method for implementing generative AI services using LLMs within the LangChain framework. The study explores various strategies to leverage LLMs, focusing on fine-tuning and direct use of document information. It details information storage and retrieval methods, employing the Retrieval Augmented Generation (RAG) model for context recommendation and Question-Answering (QA) systems. By enhancing understanding of generative AI technology, the study enables active utilization of LLMs in corporate service implementation, offering valuable insights for practical applications.

III. METHODOLOGY

The methodology of the project involves preprocessing pdf data, segmenting text, and generating embeddings for semantic understanding. Leveraging Retrieval Augmented Generation (RAG), Retrievers bridge generative models and external knowledge sources. the users interact with the LLM using a web application which integrates with the database and the generative model.

A. Preprocessing of PDF Data

The data is provided to the model in the form of PDF files. PDF documents contain text data that needs to be extracted for analysis. Text extraction techniques are used to convert the textual content of PDFs into a format that can be processed by the chatbot. Preprocessing steps may be applied to the extracted text to clean and standardize it. This can involve removing irrelevant content, such as headers, footers, and non-text elements, as well as handling special characters and formatting issues.

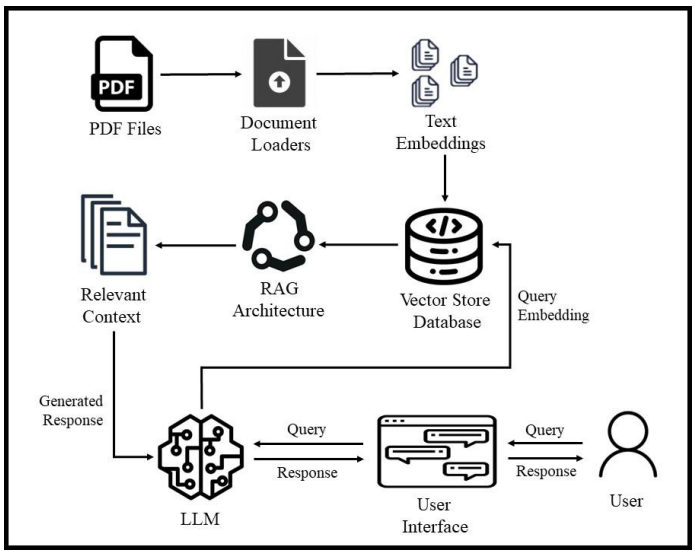

Fig. 1. Methodology for the proposed system

B. Text Segmentation and Embedding Generation

The extracted text from the text files undergoes segmentation into smaller units, enhancing the efficiency of subsequent processing and analysis. This segmentation divides the text into manageable chunks, enabling the system to focus on specific aspects of the information. Following segmentation, the system generates embeddings for the segmented text. Embeddings are numerical representations of text that capture the semantic meaning of the information. By encoding the underlying context and relationships within the text, embeddings enable the system to interpret and understand the content more effectively. This transformation of textual data into numerical vectors facilitates various downstream tasks, such as semantic search and contextually enriched content generation.

C. Creation of Vector Store Databases

Vector Store Databases serve as foundational components of the chatbot system, enabling efficient storage and retrieval of textual embeddings. These databases store the numerical representations of text, known as embeddings, which encapsulate the semantic meaning of the information.The embeddings stored in Vector Store Databases enable the Retriever-Augmented Generation (RAG) systems to retrieve and integrate relevant information into the generated outputs. RAG systems leverage the complementary strengths of retrieval-based and generation-based approaches to produce more contextually accurate and informative responses compared to traditional generation models.

D. Retrievers in RAG Framework

In the RAG framework, Retrievers serve as essential components that bridge the gap between the generative model and external knowledge sources. Their role is pivotal in enriching

the content generation process by facilitating access to relevant information from external sources. Retrievers accomplish this by employing various techniques such as semantic search and information retrieval to identify and retrieve pertinent information based on user queries. By accessing external knowledge sources, Retrievers enhance the comprehensiveness and accuracy of the generated responses.

E. User Interaction with Large Language Model (LLM)

User Interaction with the Large Language Model (LLM) is facilitated through a dedicated web-based interface tailored for seamless communication. Upon receiving user queries, the LLM undertakes comprehension, transforming them into query embeddings that encapsulate the semantic essence of the inquiries. Leveraging these embeddings, the system conducts semantic searches to retrieve pertinent context, subsequently crafting responses that adeptly address the users' queries. Through this iterative process, the LLM ensures effective and contextually relevant interactions, enhancing user satisfaction and system usability

IV. Implementation and Results

Python's versatility, combined with its robust community support and cross-platform compatibility, has made itself widely utilized in training Large Language Models (LLMs). Python 3.x (Python 3.8 or higher) is used for development in this project.

Deep Learning Libraries like PyTorch, LangChain as the primary deep learning framework for model development and training. LangChain is a deep learning framework primarily focused on natural language processing (NLP) tasks. It provides a set of tools and utilities specifically tailored for NLP applications, including text preprocessing, tokenization, sequence modeling, and language generation. LangChain aims to simplify the development and deployment of NLP models by offering high-level abstractions and pre-built components for common NLP tasks.

Transformers are the architectural backbone that powers LLMs, enabling them to process and understand text at scale. Transformers Library like Hugging Face Transformers library, open-source library developed by Hugging Face, a company specializing in natural language processing (NLP) technologies, which provides easy-to-use interfaces for working with transformer-based models, including both pre-trained models and tools for fine-tuning them on custom datasets. The library supports a wide range of transformer architectures, including BERT, GPT, RoBERTa, T5, and more.

Chainlit is the open-source Python libraries that allows to create web applications for machine learning and data science projects with minimal effort. It's designed to make it easy for developers to build interactive web apps without requiring expertise in web development.

Utilizing the mentioned technologies, the chatbot has been developed which takes the PDF as input and answers any queries asked by the user. The following mentions the features of the developed web application:

- The web application enables users to upload any PDF file they wish to query. The PDF data undergoes parsing to extract the relevant content. This involves removing unnecessary elements such as headers, footers, and any other extraneous details.
- The pre-processed text is segmented into smaller units or chunks to facilitate efficient processing and analysis. This segmentation helps in managing large volumes of text data. Embeddings, which are vector representations of the text, are generated using libraries like sentence-transformers. These embeddings encode the semantic meaning of the text, making it suitable for retrieval and generation tasks.
- The generated embeddings are stored in vector store databases like FAISS. These databases serve as repositories for the embeddings, allowing quick and efficient retrieval based on semantic similarity.
- The embeddings in the vector store enable the Retriever-Augmented Generation (RAG) system to retrieve relevant information, enhancing the contextuality and accuracy of the chatbot's responses. When a user query is received, it is converted into query embeddings, which are used to perform a semantic search in the vector store to retrieve relevant context.
- The project utilizes pre-trained LLM, Llama2-7B model. The model is obtained from the Hugging Face Transformers library, which provides tools for fine-tuning and deployment.
- The retrieved context, along with the user query, is fed into the LLM to generate coherent and contextually relevant responses. The system uses RAG to integrate external knowledge sources seamlessly.
- A user-friendly web-based interface is developed using framework, Chainlit. This interface allows users to interact with the chatbot in real-time.

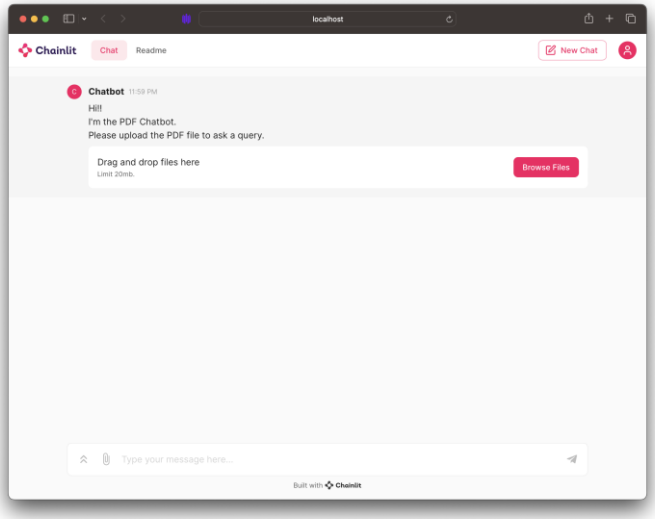

Fig. 2. User interface for the Chatbot

The implementation of the RAG framework significantly improved the chatbot's ability to provide accurate and contextually relevant responses. This approach used Reinforcement Learning to minimize the number of LLM tokens required, reducing the overall computational cost. The web-based interface provided a seamless and interactive user experience. Users could query the chatbot and receive prompt responses, enhancing their overall interaction with the system.

Figure 2 Shows the chatbot interface using which the users can interact with the LLM. The chatbot responds to the query using the data provided in the PDF files.

V. Conclusion

The chatbot is designed to engage in natural language conversations, providing intelligent responses to the queries related to uploaded PDFs. The chatbot is expected to answer the queries based on the the PDF data. The responses are generated using the Retrieval Augmented Generation (RAG) approach.

In conclusion, the implementation of the chatbot using LLMs and the RAG framework demonstrated the potential of advanced NLP techniques in creating efficient and effective conversational agents. The project achieved significant improvements in response accuracy and efficiency by employing the RAG framework, which integrated external knowledge sources to enrich the chatbot's contextual understanding. The use of a policy-based model for optimizing LLM token usage demonstrated substantial cost savings while maintaining high response quality. The results of this project highlight the effectiveness of combining LLMs with retrieval mechanisms to create sophisticated conversational agents capable of handling complex queries. The chatbot not only automated routine query responses but also provided a scalable solution for future expansion and enhancement. The implementation sets a foundation for future research and development in the field of AI-driven conversational systems, paving the way for more sophisticated and efficient automated support solutions.

References

[1] S. Meshram, N. Naik, M. VR, T. More and S. Kharche, "Conversational AI: Chatbots," 2021 International Conference on Intelligent Technologies (CONIT), Hubli, India, 2021, pp. 1-6, doi: 10.1109/CONIT51480.2021.9498508.

[2] Lalwani, Tarun and Bhalotia, Shashank and Pal, Ashish and Rathod, Vasundhara and Bisen, Shreya, Implementation of a Chatbot System using AI and NLP (May 31, 2018). International Journal of Innovative Research in Computer Science & Technology (IJIRCST) Volume-6, Issue-3, May-2018.

[3] Bal, Sauvik & Jash, Kiran & Mandal, Lopa. (2024). An Implementation of Machine Learning-Based Healthcare Chabot for Disease Prediction (MIBOT). 10.1007/978-981-99-6866-4-32.

[4] Kulkarni, Mandar, Praveen Tangarajan, Kyung Kim, and Anusua Trivedi. "Reinforcement learning for optimizing rag for domain chatbots.".arXiv preprint arXiv:2401.06800.(2024).

[5] C. Jeong, "Generative AI service implementation using LLM application architecture: based on RAG model and LangChain framework," Journal of Intelligence and Information Systems, vol. 29, no. 4, pp. 129–164, Dec. 2023.

[6] Jeong, Cheonsu. (2023). A Study on the Implementation of Generative AI Services Using an Enterprise Data-Based LLM Application Architecture. Advances in Artificial Intelligence and Machine Learning. 3. 1588-1618. 10.54364/AAIML.2023.1191.

[7] Afzal, Anum & Kowsik, Alexander & Fani, Rajna & Matthes, Florian. (2024). Towards Optimizing and Evaluating a Retrieval Augmented QA Chatbot using LLMs with Human-in-the-Loop.

[8] Bacciu, A.; Cocunasu, F.; Siciliano, F.; Silvestri, F.; Tonellotto, N.; and Trappolini, G. 2023. RRAML: Reinforced Retrieval Augmented Machine Learning.

[9] Chen, Jiawei & Lin, Hongyu & Han, Xianpei & Sun, Le. (2024). Benchmarking Large Language Models in Retrieval-Augmented Generation. Proceedings of the AAAI Conference on Artificial Intelligence. 38. 17754-17762. 10.1609/aaai.v38i16.29728.

[10] Li, Xianzhi & Chan, Samuel & Zhu, Xiaodan & Pei, Yulong & Ma, Zhiqiang & Liu, Xiaomo & Shah, Sameena. (2023). Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks. 408-422. 10.18653/v1/2023.emnlp-industry.39.

[11] Zhihan Lv, Generative artificial intelligence in the metaverse era, Cognitive Robotics, Volume 3, 2023, Pages 208-217, ISSN 2667-2413, https://doi.org/10.1016/j.cogr.2023.06.001. (https://www.sciencedirect.com/science/article/pii/S2667241323000198)

[12] Zant, Tijn & Kouw, Matthijs & Schomaker, Lambert. (2012). Generative Artificial Intelligence. 10.1007/978-3-642-31674-6-8.

[13] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[14] Rackauckas, Zackary. "RAG-Fusion: a New Take on Retrieval-Augmented Generation." arXiv preprint arXiv:2402.03367 (2024).

[15] Khan, Salman, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. "Transformers in vision: A survey." ACM computing surveys (CSUR) 54, no. 10s (2022): 1-41.

[16] Goldman, Sharon M. "Transformers." Journal of Consumer Marketing 27, no. 5 (2010): 469-473.

[17] Alan, Ahmet Yusuf, Enis Karaarslan, and Omer Aydin. "A RAG-based Question Answering System Proposal for Understanding Islam: MufassirQAS LLM." arXiv preprint arXiv:2401.15378 (2024).

[18] Feuerriegel, Stefan & Hartmann, Jochen & Janiesch, Christian & Zschech, Patrick. (2023). Generative AI.

[19] Quidwai, Mujahid Ali, and Alessandro Lagana. "A RAG Chatbot for Precision Medicine of Multiple Myeloma." medRxiv (2024): 2024-03.

[20] Bras¸oveanu, Adrian MP, and Ra˘zvan Andonie. "Visualizing transformers for nlp: a brief survey." In 2020 24th International Conference Information Visualisation (IV), pp. 270-279. IEEE, 2020.

[21] Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac et al. "Huggingface's transformers: State-of-the-art natural language processing." arXiv preprint arXiv:1910.03771 (2019).

[22] Fill, Hans-Georg & Fettke, Peter & Ko¨pke, Julius. (2023). Conceptual Modeling and Large Language Models: Impressions From First Experiments With ChatGPT. Enterprise Modelling and Information Systems Architectures. 18. 1-15. 10.18417/emisa.18.3.

[23] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J. and Wang, H., 2023. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997.

[24] Li, J., Yuan, Y. and Zhang, Z., 2024. Enhancing LLM Factual Accuracy with RAG to Counter Hallucinations: A Case Study on Domain-Specific Queries in Private Knowledge-Bases. arXiv preprint arXiv:2403.10446.