Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.2, May : 2024

# MULTI DOCUMENT CHAT-BOT

**Ms. Deepali Dev**, Assistant Professor, Department of CSE-AIML, ABES Engineering College
**Kartik Saini**, **Kishan Jayan, Adarsh Verma,** Student, Department of CSE-AIML, ABES Engineering College

**Abstract**
This research introduces a project, "Multi-Document Chat-bot," leveraging Generative AI models, particularly Large Language Models. The system enables users, including teachers, students, doctors, lawyers, scientists, and corporate professionals, to upload multiple documents and extract insights or generate document summaries through natural language queries. Users will not need to spend too much time manually reading the whole document. Furthermore, this project involves the use of Langchain which is an open source framework designed to simplify the development of applications utilizing large language models and OpenAI's gpt-3.5 LLM API.
Keywords: Langchain, LLMs, GPT, Embeddings, Vector Database

## 1. Introduction

A large Language Model or LLM is an AI model that is trained on a large amount of textual data and is used in generative AI applications where users can input a query in natural language and receive output from LLM in a natural language. LLMs can be used in any type of natural language tasks like sentiment analysis, Chatbots, and question-answering systems. To simplify the development of LLM applications Langchain is introduced which is an open source framework in the Generative AI field. Langchain allows developers to integrate external components with different Large Language Models and private data that eases the development of complex applications. Private data can include documents, reports, and databases affiliated with a specific enterprise or organization. Our project "Multi-Document Chat bot" emphasizes upon easy information retrieval from public or private data. The data can be in the form of PDF, DOCX, txt files, and images. This project is based on the Python framework Langchain which offers functionality for integrating vector databases, LLMs, embeddings, conversation memory, document loaders, text splitting, prompt templates, and chaining.

## 2. Literature

Oguzhan [1] discussed the core features of Langchain and its architecture which enabled developers to swiftly create generative AI applications. The applications Langchain shows its potential in innovation regarding sophisticated LLM technologies. Eduardo [2] showcases the application of ChatGPT and LangChain in developing Natural Language Interfaces for Databases (NLIDBs). It explores two alternatives: one involving SQL query generation from natural language questions and the other involving keyword extraction for Keyword Search (KwS) tools. Through experiments, it was found that ChatGPT-KwS-Prompt, which incorporates prompts with contextual information and examples to aid in keyword extraction, emerged as the most effective NLIDB solution. This demonstrates the potential of integrating ChatGPT and LangChain to enhance NLIDB performance. Shubham [3] discussed the development of an AI system tailored for Indian Legal Question Answering (AILQA) within the criminal domain. Various combinations of embedding and QA models were explored, focusing on assessing their efficacy in addressing Indian legal queries. The study utilized the OpenAI GPT model, specifically the Davinci variant, and integrated prompts to enable the AILQA system to comprehend natural language queries and provide accurate responses. Through empirical analysis and expert evaluations, the research uncovered valuable insights into the challenges and potential solutions for creating effective AILQA systems. Results indicated that the architectures utilizing GPT-3-based models with appropriate prompts consistently exhibited strong contextual understanding, surpassing even human lawyer answers in terms of accuracy, relevance, and reliability. [4] This project presents an innovative approach to PDF chatbots, utilizing advanced technologies to

enhance response accuracy, fluency, and relevance while streamlining development and deployment processes. These enhancements have the potential to increase the adoption and utilization of chatbots across various applications. Future efforts may involve training the chatbot on a larger dataset of PDF files and evaluating its performance across a greater variety of documents. The MIT-GPT [5] project introduces an innovative college query assistant that leverages the OpenAI API and LangChain to deliver precise responses to inquiries about higher education. Engineered to swiftly and accurately address user queries while safeguarding user data privacy, the system's architecture is underpinned by a deep learning model trained on a substantial dataset encompassing college-related information. It draws upon both publicly available data from sources like government databases and college websites, as well as proprietary data curated by the development team. For decades, chatbots have been an integral part of global technological evolution, finding applications across various fields. This paper [6] classifies chatbot services into customer-centric and public administration-oriented categories, examining their potential for enhancement and implementation. Customer-centric chatbots serve as personal assistants for tasks such as food ordering, appointment scheduling, and flight booking, seamlessly integrating into daily life and enhancing convenience. Public administration-oriented chatbots aid citizens in resolving queries and accessing government services without requiring the direct involvement of administrative staff.

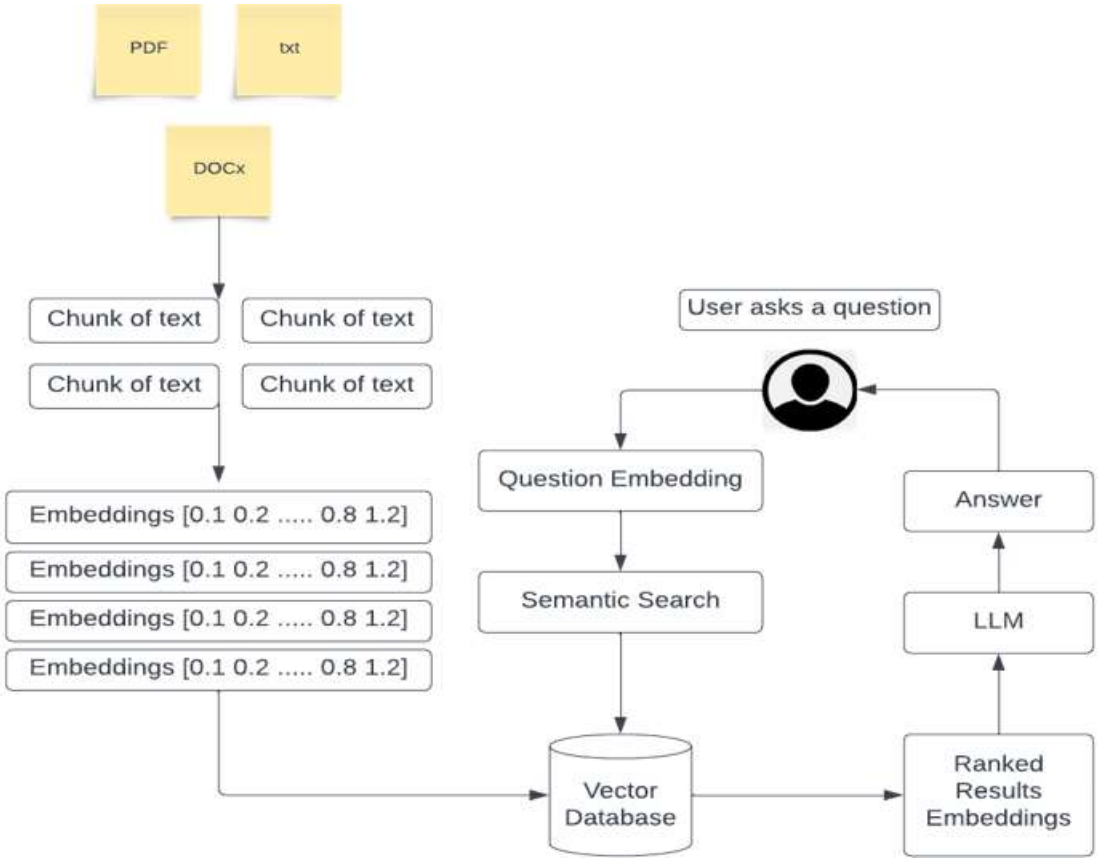
## 3. Methodology

The essential components of our project are embeddings, vector databases, LLM, Streamlit, and Langchain. When a user uploads documents, the text from those documents is split into chunks. The smaller the size of the chunks better the response of the chatbot. Those chunks are further converted into embeddings which are generally numerical representations of textual data from the documents. Those embeddings are then stored in the vector database using the indexing method. Similar embeddings will be indexed nearer to each other in the data for faster information retrieval.

When a user asks a question in natural language, the question is also converted into embeddings. Then vector similarity search is initiated to get a group of the most relevant answers or embeddings. Similarity search refers to searching according to the context, meaning, or intent of the user's query and not just comparing keywords only. Cosine similarity is used to compare the relevancy of the query embedding and vector DB embeddings. If the dot product of the query embedding and vector DB embeddings is closer to 1 more relevant will be the answer. The answer embedding is passed to the LLM and the LLM generates the final response in natural language to the user.

To enhance user experience a Python-based framework Streamlit is used which eases the development of Python-based GUI applications. Conversation memory functionality is also implemented in this chatbot so that the bot can answer questions from previous chats too. Langchain is one of the most essential components of this project which helped us to integrate other external components.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.2, May : 2024




## 4. Result and Discussion

We have successfully implemented a project which is a document-based chatbot. We have tested many types of LLMs, embeddings, and vector databases to select the best out of them. Results showed that using OpenAI's embeddings, OpenAI's gpt-3.5, and Chroma vector database gives more stability to our project. The response of the Chatbot mainly depends on factors like the types of LLM, embeddings, vector database used, and also on how the texts were split into chunks. We also tested our project with open-source LLMs namely Meta's LLaMA2 and Mistral and observed that in some cases LLaMA2 model was giving incomplete responses but it had greater accuracy than the Mistral model. Overall OpenAI's gpt-3.5 turbo model performed better than LLaMA2 and Mistral model.

| MODEL | CORRECT RESPONSES | INCORRECT RESPONSES | ACCURACY |
|---|---|---|---|
| GPT-3.5 TURBO | 13 | 4 | 76.47 % |
| LLAMA2 7B | 12 | 5 | 70.59 % |
| MISTRAL 7B | 9 | 8 | 52.94 % |

Table 1. Accuracy of LLMs of models

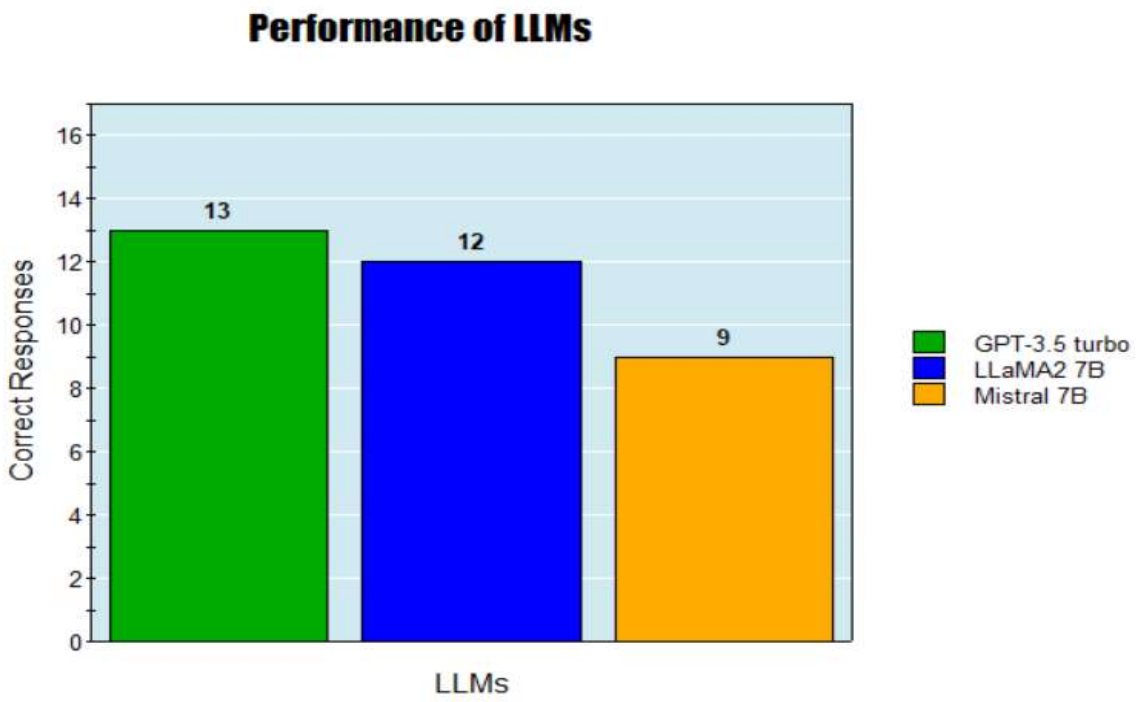



Fig.1 Performance of LLMs

## 5. Conclusion

In this paper, we introduced a project "Multi-Document Chatbot" a document assistant that can assist its users in gaining important insights from their documents, generating summaries from a variety of articles and books. This eliminates the need to train an LLM from scratch on private documents or databases. We also discussed about core components and technologies needed to make this project and also the methodology and architecture of this project.

For future scope, we will be implementing this project by using more robust technologies. We will be adopting Next Js to convert this project into a SaaS product, testing more LLMs, embeddings, and vector databases for more accurate responses.

References

[1] G. G. W. V. M. L. Y. I. R. G. L. L. M. C. Eduardo Nascimento, "A Family of Natural Language Interfaces for," in *Instituto Tecgraf and Departamento de Informática*, PUC-Rio, Rio de Janeiro, 2023.

[2] N. K. S. V. S. P. Krishna Kumar Nirala, "A survey on providing customer and public administration," *MACHINE VISION THEORY AND APPLICATIONS FOR CYBER* , vol. 81, no. 16, p. 32, 2022.

[3] S. K. M. A. K. M. N. S. A. B. Shubham Kumar Nigam, "Comparative Analysis of Artificial Intelligence," in *arXiv:2309.14735v1 [cs.CL]* , Kanpur, Uttar Pradesh, 2023.

[4] T. S. G. A. R. T. Arjun Pesaru, "AI ASSISTANT FOR DOCUMENT MANAGEMENT USING LANG," *International Research Journal of Modernization in Engineering Technology and Science* , vol. 05, no. 06, p. 4, 2023.

[5] T. C. A. Oguzhan Topsakal, "Creating Large Language Model Applications Utilizing LangChain: A," in *5th International Conference on Applied* , Konya, Turkey, 2023.

[6] E. B. H. O. C. K. C. E. C. J. O. J. J. A. Mark Edward M. Gonzales, "From Unstructured to Structured: Transforming Chatbot," in *arXiv:2305.04258v1 [cs.DB]*, Manila, Philippines, 2023.

[7] V. H. A. S. R. B. A. D. R. P. Vrishani Shah, "Using GPT-3 to Create General Purpose," *International Journal of Scientific Research & Engineering Trends*, vol. 9, no. 4, p. 4, 2023.