Document Based Question Answering System Using GPT 3.5 and FAISS

Jeet Muley
IT Final Year Student
Dr Babasaheb Ambedkar Technological University
Lonere, India.
20330331246019@dbatu.ac.in

Prasanna Meshram
IT Final Year Student
Dr Babasaheb Ambedkar Technological University
Lonere, India.
20330331246030@dbatu.ac.in

Pranita Jadhav
Associate Professor
Dr Babasaheb Ambedkar Technological University
Lonere, India.
ppjadhav@dbatu.ac.in

Sapana Barphe
Associate Professor
Dr Babasaheb Ambedkar Technological University
Lonere, India.
ssbarphe@dbatu.ac.in

Ektaa Meshram
Associate Professor
Dr Babasaheb Ambedkar Technological University
Lonere, India.
ekta.meshram13@dbatu.ac.in

***Abstract*— Document based question answering is task of getting queries from the user and based on those queries extracting relevant information from the collection of documents and processing the information to provide perfect and concise answers to the users. This paper proposes a document-based question answering system built using open-source python libraries: LangChain, FAISS and Streamlit and GPT-3.5 LLM developed by OpenAI. Langchain is framework to develop applications based on LLMs. Facebook AI Similarity Search (FAISS) is a library for rapid searching and clustering of dense vectors. System uses LangChain to enable GPT-3.5 to interact with document-store created using FAISS. LangChain's components, such as document loaders and retrieval strategies, are utilized to manage the document store and streamline the process of fetching relevant information in response to user questions. Preliminary results indicate an enhancement in the precision and speed of answering complex queries over existing system.**

Keywords*— *NLP, ChatGPT, Vector Similarity, LangChain, FAISS, DBQA, LLM, AI

I. Introduction

In today's digital world, there is a huge amount of data available due to the widespread use of the internet and digitization[1]. With so much information out there, we need tools that can summarize and extract the important details from documents. The increasing size of these documents makes it necessary to develop automated question answering systems that can handle large texts efficiently. Recent advancements in language models have opened up new possibilities for creating such systems.

However, current text based question answering systems still have some issues[2]. Firstly, they often struggle to understand common sense and reason about things the way humans do. Secondly, dealing with large volumes of data can be slow and inefficient, making it difficult to provide quick and relevant answers. Thirdly, these systems sometimes fail to grasp the context of the query or the information in the documents, leading to incomplete or incorrect responses

The proposed system aims to tackle these problems by using GPT-3.5, a powerful language model, and FAISS, a tool for efficiently searching and retrieving data. The goals of this system are:
To use GPT-3.5 to better understand context and reason like humans do. As a language model trained on vast amounts of data, GPT-3.5 has shown promise in comprehending and generating human-like responses, which could help address the issues of common-sense reasoning and contextual understanding.

To use FAISS to store and quickly retrieve large documents. FAISS is great at finding similar data quickly, which means the system can handle massive amounts of documents without slowing down or compromising on performance.

By combining GPT-3.5's ability to understand context and reason, with FAISS's efficient data storage and retrieval, this proposed system aims to provide accurate and relevant answers to queries, overcoming the limitations of current question answering systems.

II. Literature Review

1) J. Zhang "Application Research of Similarity Algorithm in the Design of English Intelligent Question Answering System"

The paper discusses the importance of sentence similarity algorithms in question answering systems and proposes a method for calculating sentence similarity based on

WordNet, a semantic dictionary [3]. It introduces the use of WordNet to calculate the semantic information of sentences, employs a longest word matching method to find similar sentences. While the paper focuses on a WordNet-based approach, the core ideas of leveraging semantic information and efficient similarity calculation methods are relevant to our research on question answering system using GPT-3.5 and FAISS, which employ more advanced techniques for semantic understanding

2) S. Acharya, K. Sornalakshmi, B. Paul and A. Singh ” Question Answering System using NLP and BERT”

The paper proposes a question answering system that creates a customized dataset from user inputs using Named Entity Recognition to extract relevant information. It then employs BERT and Closed Domain Question Answering techniques to provide answers from this dataset [4]. While sharing the goal of question answering, our research explores more advanced methods, leveraging the powerful GPT-3.5 language model and FAISS similarity search library for efficient document-based question answering. The paper's approach to dataset creation is insightful, but our techniques potentially offer improved performance and capabilities.

3) Tan, Yiming Min, Dehai Li, Yu Li, Wenbo Hu, Nan Chen, Yongrui Qi, Guilin “Evaluation of ChatGPT as a Question Answering System for Answering Complex Questions”

The research paper evaluates the effectiveness of ChatGPT, a large language model, as a Question Answering (QA) system for complex queries [5]. The authors suggest that integrating ChatGPT into existing database QA (DBQA) systems could enhance their performance. They highlight ChatGPT's ability to understand and generate human-like responses, making it a valuable tool for accurate and efficient information retrieval in complex question answering scenarios. The insights from this research guided us in effectively combining the strengths of GPT-3.5 and FAISS for efficient semantic search and retrieval of relevant information from document databases.

4) Oguzhan Topsaka and Tahir Cetin Akinci ”Creating Large Language Model Applications Utilizing LangChain: A Primer on Developing LLM Apps Fast”

The article provides a comprehensive guide on leveraging LangChain to rapidly develop applications powered by large language models (LLMs) like ChatGPT[6]. It outlines the process of integrating LLMs into a database question-answering (DBQA) system, showcasing how LangChain can facilitate seamless interaction between the LLM and databases. The authors demonstrate how LangChain’s modular design and pre-built components can expedite the creation of sophisticated DBQA systems. They provide practical insights into handling the complexities of LLM integration, such as managing context and ensuring accurate responses. Overall, the article served as a valuable resource for us looking to harness the capabilities of LLMs like ChatGPT for efficient and effective DBQA system development.

5) Jegou, Herv´ e , Douze, Matthijs , Johnson, Jeff´ , Hosseini, Lucas , Deng, Chengqi “FAISS: Similarity search and clustering of dense vectors library”

The article underscores the advantages of employing the FAISS library[7]. It emphasizes FAISS's efficiency in searching and clustering large sets of high-dimensional vector data, a crucial requirement for database applications dealing with such data. The library's ability to handle vectors that exceed RAM capacity and its support for various similarity metrics contribute to its versatility in similarity search tasks. This research article provided insight into the potential of FAISS as document store and for similarity search.

III. Proposed System

The core of our proposed document-based question answering system is the GPT-3.5 large language model (LLM). We evaluated multiple LLMs and selected GPT-3.5 due to its strong performance on a variety of natural language tasks and its publicly available nature as a freeware model. Since one of the goals of this research is to showcase a system that can be compiled free of cost using open-source software, we utilized only freeware components.

For document storage and similarity search to retrieve relevant documents for a given query, we employed the FAISS library. FAISS provides efficient similarity search capabilities which allow identifying the most relevant documents from a corpus to serve as context for answering a query.

To integrate the document retrieval from FAISS with the GPT-3.5 LLM, we used LangChain as a framework to orchestrate the components. LangChain obtains the relevant context documents from FAISS for a given query, and then performs prompt engineering to construct an effective prompt consisting of the query and context documents. This prompt is passed to the GPT-3.5 LLM to generate a final answer.

The Overall workflow is as follows:

- Input: The system accepts documents of various types consisting following formats: PDF, TXT, JSON, Web Pages,
- Text Splitter: This is preprocessing step. LangChain’s text splitter module is used to divide

original text into smaller chunks to generate embeddings in later step.

- Embedding Generator: Utilizing text-embeddings-small embedder this component generates embeddings from the text chunks. Embeddings are vector representations that capture the semantic meaning of the text.
- FAISS: Short for Facebook AI Similarity Search, FAISS is used to store and search through these embeddings in a Vector DB. It enables quick retrieval of relevant documents based on similarity.
- Query Processing: Queries are processed in two ways: standalone and within the context of retrieved relevant documents.
- QnA Model: OpnAI's GPT-3.5 it can answer any complex question with help of context generated by FAISS.

This architecture aims to streamline the process of finding accurate answers to user queries by combining advanced NLP models with efficient similarity search algorithm.

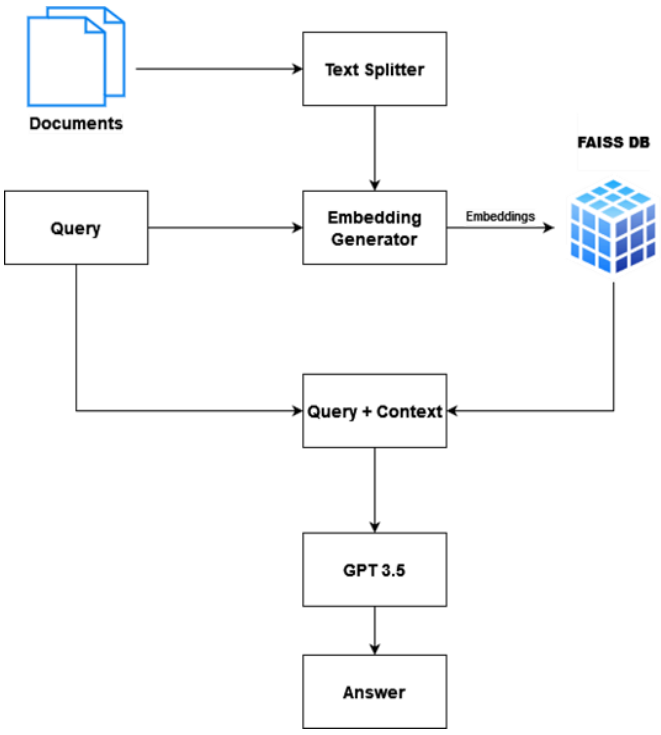

Fig. 1 System Architecture

IV. RESULTS

The SQuAD 2.0 dataset is a popular benchmark dataset for evaluating question-answering systems. It consists of over 100,000 question-answer pairs based on Wikipedia articles, with the answers being spans of text from the articles it includes questions that are unanswerable given the provided context, making the task more challenging and realistic[8]. The dataset is widely used in the natural language processing community to train and evaluate models for extractive question answering, where the goal is to identify the relevant span of text that answers a given question based on the provided context.

The proposed system was evaluated on the testing set of the SQuAD 2.0 dataset. To test the complete system, the contexts of questions from the JSON file of the dataset were copied to a text file which then used as document. The system was then tested on 5,000 questions from the dataset in a zero-shot manner without any fine-tuning.

BERT (Bidirectional Encoder Representations from Transformers) was selected as the model for comparison due to its widespread adoption and state-of-the-art performance in various natural language processing tasks, including question answering. BERT has been extensively fine-tuned and adapted for question answering tasks, providing a strong baseline for evaluating the performance of the proposed system

The performance of the proposed system was compared with a similar system using a BERT model fine-tuned on the SQuAD 2.0 dataset. The results are as follows:

TABLE 1

Model	Precision	Recall	F1-score
BERT (Fine-tuned) + FAISS	25.12%	37.32%	30.03%
GPT-3.5 + FAISS (Proposed)	49.33%	75.73%	59.74%

The results demonstrate that the proposed system using GPT-3.5 and FAISS offers significant advantages over the BERT model fine-tuned on the SQuAD 2.0 dataset. Specifically, the proposed system achieved higher precision, recall, and F1-score values, indicating improved performance on answering questions based on a document.

To evaluate the performance of our document question answering system beyond just metrics like those used for SQuAD 2.0, we conducted human evaluation. This allowed us to assess the quality of the full generated answers, which can span multiple sentences or paragraphs, on key criteria like clarity, correctness, and fluency.

The evaluators were asked to rate the system's answer on a scale of 1-5 for the following criteria:

1. Clarity: How clear and understandable is the answer?
2. Correctness: How factually accurate is the answer based on the source text?
3. Fluency: How fluent and natural is the language used in the answer?

We randomly selected 10 questions based on different Wikipedia articles covering a range of topics. These questions were presented to human evaluators through a Google Form. For each question, the evaluators were shown the question, the top result generated by our system, and the relevant source text from Wikipedia.

A total of 25 volunteer evaluators completed the ratings across all 10 questions. The average ratings were:

TABLE 2

Criterion	**Average Rating in Percentage**
Clarity	84.%
Correctness	82%
Fluency	88%

These results indicate that overall, the answers generated by our system were judged to be clear, correct based on the source material, and fluent. However, there was some variance across the individual questions.

V. Conclusion

We have developed a document-based question answering system that leverages the power of FAISS and large language models (LLMs) such as GPT 3.5. Our results show that our system can achieve high accuracy and relevance for document-based question answering, as well as provide more context-specific, up-to-date, and adaptable answers than fine-tuned models. Our system also demonstrates the potential of LangChain as a framework and FAISS for similarity search for building advanced use cases around LLMs, such as chatbots, summarization, and more. We believe that our system can be a useful tool for information retrieval and knowledge extraction from unstructured data sources.

Acknowledgment

Our thanks to project guide Prof. Pranita Jadhav and Prof. Sapana Barphe for their continued support for this paper. We are thankful to the IT Lab, Department of Information Technology, Dr. Babasaheb Ambedkar Technological University, Lonere, for providing facility for carrying out the research.

References

[1] H. A. Pandya and B. S. Bhatt, “Question Answering Survey: Directions, Challenges, Datasets, Evaluation Matrices.” arXiv, Dec. 07, 2021. doi: 10.48550/arXiv.2112.03572.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] B. Ojokoh and E. Adebisi, “A Review of Question Answering Systems,” *Journal of Web Engineering*, vol. 17, pp. 717–758, Jan. 2019, doi: 10.13052/jwe1540-9589.1785..

[3] J. Zhang, “Application Research of Similarity Algorithm in the Design of English Intelligent Question Answering System,” in *2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC)*, Dec. 2022, pp. 1–4. doi: 10.1109/ICMNWC56175.2022.10031708.

[4] S. Acharya, K. Sornalakshmi, B. Paul, and A. Singh, “Question Answering System using NLP and BERT,” in 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), Oct. 2022, pp. 925–929. doi: 10.1109/ICOSEC54921.2022.9952050.

[5] “Evaluation of ChatGPT as a Question Answering System for Answering Complex Questions,” ar5iv. Accessed: Mar. 10, 2024. [Online]. Available: https://ar5iv.labs.arxiv.org/html/2303.07992

[6] O. Topsakal and T. C. Akinci, “Creating Large Language Model Applications Utilizing LangChain: A Primer on Developing LLM Apps Fast,” *International Conference on Applied Engineering*