LangchainIQ: Intelligent Content and Query Processing

Chinmay Pichad
Department of Computer Engineering,
Sardar Patel Institute of Technology,
Mumbai, India

Roshan Sawant
Department of Computer Engineering,
Sardar Patel Institute of Technology,
Mumbai, India

Ganesh Supe
Department of Computer Engineering,
Sardar Patel Institute of Technology,
Mumbai, India

ABSTRACT
LangchainIQ is an educational platform that uses Artificial Intelligence (AI) and Natural Language Processing (NLP) technologies to uplift the learning experience. This platform is designed to enhance their content processing and query processing on a wide range of input formats with additional assessment capabilities by QnA generation with use of input text. LangchainIQ's AI-powered chatbot provides a wide array of content formats, including PDFs, Excel sheets, and YouTube videos. It breaks the given input data, then converts it into chuck and stores it in embedded form, ultimately increasing security. With use of LLM models the content processing power is enhanced.

One of the groundbreaking features of LangchainIQ is its proficiency in creating knowledge bases from PDF files. This knowledge base facilitates efficient content retrieval and processing, enabling learners to quickly access and understand the information they need. Additionally, for CSV files, LangchainIQ processes queries by creating dataframes, making it a versatile tool for handling different data formats.

General Terms
Artificial Intelligence (AI), Natural Language Processing (NLP), Langchain, OpenAI, Streamlit

Keywords
Langchain, OpenAI, Vector Store, Faiss, Embedding, Dataframe, AI, Transcripts, LLM

1. INTRODUCTION
In the education field, the quest for efficient knowledge assessment and retrieval has been greatest challenge for educators and students alike. Both stakeholders encounter a multifaceted dilemma: the scarcity of high-quality questions for evaluating comprehension, coupled with the arduous task of pinpointing specific answers or essential concepts buried within voluminous textbooks. Additionally, the modern educational landscape presents new hurdles, such as navigating complex tabular data and unearthing relevant information from extensive lecture videos. The need for a comprehensive, innovative solution is evident, one that transcends the conventional constraints of educational materials and empowers both educators and students in their quest for effective learning and knowledge evaluation.

To confront this ubiquitous challenge head-on, our research project introduces a novel application designed to revolutionize the way knowledge is assessed and retrieved in educational settings. This application seeks to address these pressing issues by enabling the retrieval of specific answers from PDF documents, generating multiple-choice questions from textual content, conducting in-depth analysis of Excel files, and extracting key concepts from lengthy video lectures. By amalgamating cutting-edge technology with the ever-evolving educational landscape, this application serves as a beacon of hope for educators and students, aiming to create a comprehensive solution for their diverse needs and challenges.

The implications of our research project are profound. Educators will benefit from a ready repository of high-quality questions that can be utilized for assessments, fostering an environment that promotes critical thinking and knowledge retention. Furthermore, the application's ability to extract specific answers from PDF documents empowers educators and students with a precise, time-efficient, and context-aware tool, streamlining the process of information retrieval.

For students, the application's capacity to generate multiple-choice questions from textual material not only enhances comprehension but also offers a valuable self-assessment resource. Likewise, the application's proficiency in analyzing Excel files and extracting key concepts from video lectures provides a more streamlined approach to studying and research, saving precious time and energy.

In essence, our research project endeavors to bridge the existing gap between knowledge assessment and retrieval in the education sector. By creating an innovative application, we aspire to redefine the educational landscape, providing a comprehensive solution to the challenges faced by both educators and students. Through this research, we aim to empower the educational community with a powerful tool that enhances learning, encourages critical thinking, and simplifies the process of knowledge evaluation.

2. RELATED WORK
Chatbots in Education System Vijaya Lakshmi, Y* and Ishfaq majid** (2022) provided an introductory exploration of chatbot applications in educational institutions. Despite its valuable insights, it falls short in offering a specific methodology for implementation. However, the paper offers a foundational overview of how chatbots can be integrated into the educational domain.

AI Assistant for document management using Langchain and Pincone (2023) provides us the system methodology and system diagram. Thai paper explain how components works to perform specific task and get the result.

A Survey of Large Language Models" (2022) delves into the inner workings of LLMs and their embedding processes, shedding light on the technology's core operations. Nevertheless, the paper lacks a direct comparison of different LLM models, making it challenging to assess their relative strengths and weaknesses.

An Effective Query System Using LLMs and LangChain" (2022) does not explicitly specify its disadvantages but

suggests future work related to processing CSV data for analysis. This hints at the potential for improving its data processing capabilities, making it a point of interest for further research.

An efficient integration and indexing method based on feature patterns and semantic analysis for big data" (2022) is dedicated to the methodology of indexing and similarity checks in the context of big data, making it a valuable contribution to data management practices.

"The Agents of AI: Data Analysis with LLMs and LangChain Agents" (2022) presents a system designed for CSV data analysis, demonstrating the role of LLMs and LangChain in data analysis applications.

RETA-LLM: A Retrieval-Augmented Large Language Model Toolkit" (2023) does not specify its disadvantages but offers suggestions for effectively utilizing LLM models, making it an interesting prospect for researchers seeking guidance on LLM implementation.

Faiss: Efficient Similarity Search and Clustering of Dense Vectors" (2021) primarily serves as a data store, contributing to the efficiency of similarity search and clustering of dense vectors. It does not explicitly outline its advantages or disadvantages.

3. Implementation Methodology

The implementation methodology consists of following aspects

3.1 PDF Answering

The PDF content and query processing system takes a PDF document as input and breaking it into text chunks, each of approximately 2000 characters with a 100-character overlap factor. These chunks are processed and then transformed into numerical vectors using a language model - Generative Pre-trained Transformer (GPT). The resulting embeddings are stored in a directory. These embeddings are store in pkl(pickle) format. The pickle module is used arrange Python objects periodical and save them to a file.

To efficiently extract data an indexing tool like Faiss is used in our system. It facilitates the storage and rapid retrieval of embedded vectors, Langchain framework is used for whole system building which enhances the search process by providing a semantic understanding of the content.The Figure 1shows to take the input of PDF, CSV file and break it into chuck and embed it and processed with LLM model

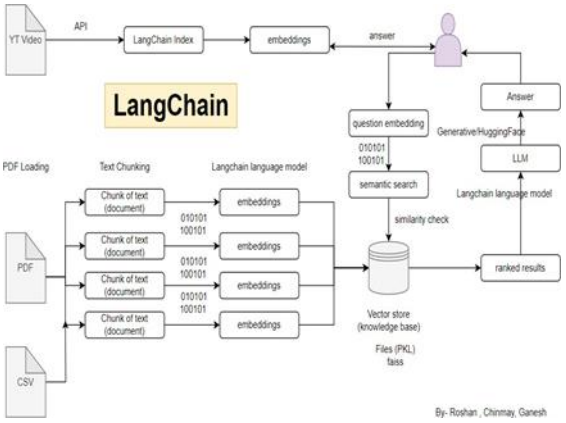

Fig. 1: Takes input of PDF, CSV file and break it into chuck and embed it and processed with LLM model

When a user submits a query, it is embedded using the same language model used for the PDF content. Faiss is employed for similarity-based search, quickly identifying chunks that match the query. The Langchain, in conjunction with a language model (LLM), further refines the search by comprehending the query's context and intent. The final output is generated by presenting the relevant text chunks based on their relevance and context, offering users an efficient means to access information within the PDF document. This integrated approach optimizes content retrieval, ensuring accurate and context-aware results.

Facebook AI Similarity Search (FAISS) - It is a library used for quickly searching relevant data in a given document. It provides semantic search for searching through documents.

Embedding Process begins with creating a directory named "embeddings" if it doesn't already exist, serving as the storage location for the embeddings vectors.

Then system takes a file and its original filename as input and stores the document embeddings. It first writes the file to a temporary location and then determines its file extension. Depending on the file type (CSV, PDF, or TXT), it loads the data using the appropriate loader (CSVLoader, PyPDFLoader, or TextLoader) and splits the text into chunks. It then uses the OpenAIEmbeddings from OpenAI api to obtain embeddings and stores them in FAISS.

Then retrieval of document embeddings is performed with Langchain . Finally, it loads the vectors from the pickle file and returns them.

3.2 CSV query processing

Processing of CSV Content and query processing is preformed similarly as PDF Processing where content is embedded ans stored. Additional step in CSV Processing is creation of Dataframe which is done with use of Langchain and OpenAI api. When user input the query with use of GPT model query is rephrase and embedded. After that with use of LLM keyword are picked and dataframe is created by using OpenAI api. After that input content is processed with respect to dataframe.

3.3 MCQ generation and answering

This module interacts with the OpenAI API to generate multiple-choice questions based on user-provided text or a YouTube video transcript. The system allows the user to input text or a YouTube video URL, then generates questions related to the content using OpenAI's text-davinci-003 engine.

The system sends requests to the OpenAI API to obtain questions related to the provided content.

The application maintains a session state to store user inputs, question data, and options state. It includes functions to handle question rendering, checkbox interactions, and updating the user's score based on their answers. Additionally, there are functions to handle YouTube-specific features, such as retrieving video transcripts and highlighting relevant segments.

The system is build using YouTube Transcripts API to fetch video transcripts, and it includes functionality to play relevant video segments when a user answers a question related to a YouTube video.

3.4 Video Query Answering

The development of a Video Query Answering Model represents a significant advancement in natural language processing and interactive content retrieval.
Input of Video input is taken, Youtube video can be taken as input. Transcripts from this Video is extracted by using Youtube API and then extracted text is taken and processed similarly as PDF.

4. RESULTS AND DISCUSSIONS

After implementing the above proposed models, following results have been found which are shown in Fig.2 to Fig.5.

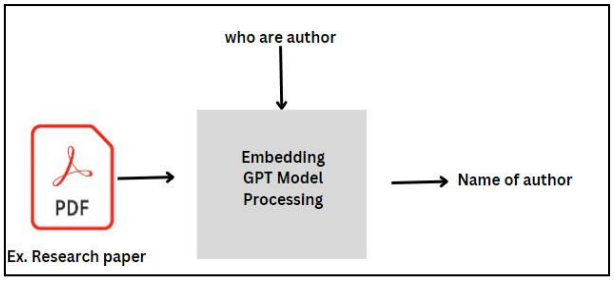

Fig. 2: Input of PDF document is a taken and query processed with GPT model and output is given

Fig. 3: Example were input of research paper is provided with query as who have written this paper. Output of names of writer is provided

Fig. 4: Uploading input file by drag and drop option or by browsing from media

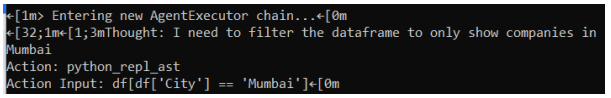

Fig. 5: Creation of dataframe of given input query for input of csv file

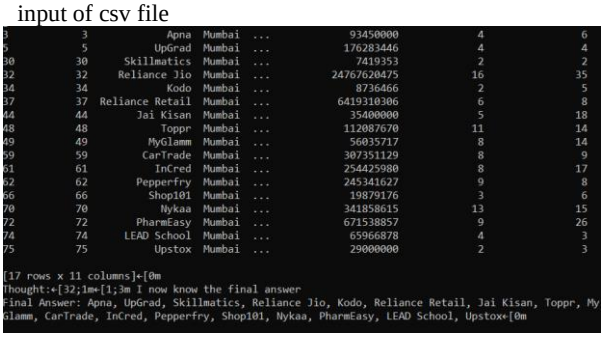

Fig. 6: Scanning through csv file and getting output based on dataframe which is build with langchain and openAI api

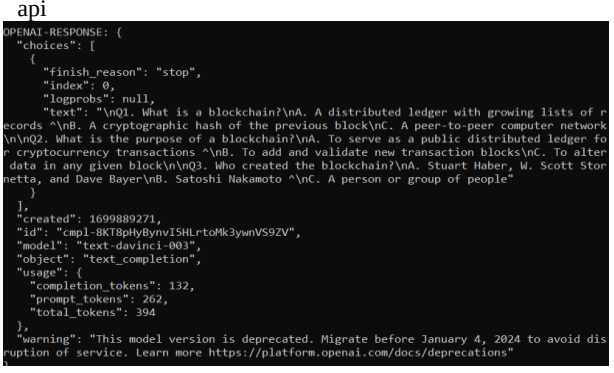

Fig. 7: Scanning text data for mcq generation, terminal output of mcq generated and model used(text-davinci-003) provided by openAI

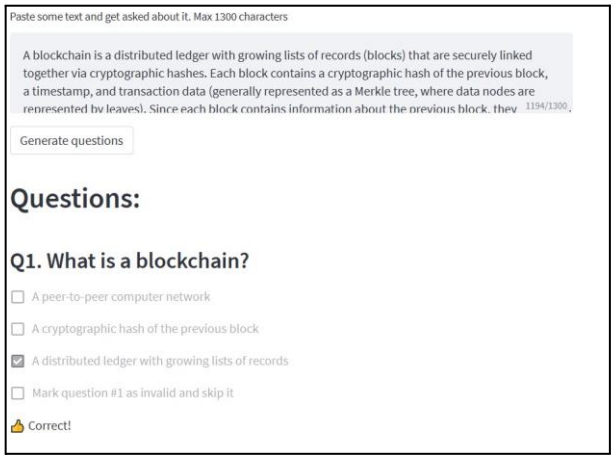

Fig. 8: UI output of mcq generated

Fig. 9: Example were input of youtube video is provided with query as. Output of given query is procure

5. ACKNOWLEDGMENTS

The authors would like to express their sincere appreciation to the LangChain team for their invaluable support and contribution to the project. We extend our gratitude to OpenAi for providing access to their GPT models, which formed the cornerstone of out AI chatbot, CSV processing model, AI video answering system and MCQ-type Question and answer generator.

Additionally, we acknowledge the assistance and feedback received from our colleagues and peers, which significantly enhanced the quality of our work. This research was made possible in part by the support of Prof. Kiran Gawande and Prof. Reeta Koshy, for which we are thankful

We also want to express our gratitude to the reviewers for their insightful comments and suggestions, which have greatly improved the final version of this paper.

6. REFERENCES

[1] Braun, S., & Tsay, J. (2022). A chatbot for PDFs: Using LangChain and Pinecone to build a conversational AI assistant for document management. arXiv preprint arXiv:2201.08244.

[2] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[3] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 9-13.

[4] Wolf, T., Debut, L., Sanh, V., Chaurasia, R., Devlin, J., & Ruder, S. (2020). Huggingface transformers: State-of-the-art natural language processing. arXiv preprint arXiv:2005.14165.

[5] Kumar, A., & Raschka, S. (2021). Pinecone: A simple and efficient framework for large language model inference. arXiv preprint arXiv:2103.10811.

[6] Adith Sreeram A S, Pappuri Jithendra Sai: "An Effective Query System Using LLMs and LangChain", IJERT, olume 12, Issue 06 (June 2023).

[7] NR Tejaswini, Vidya S, Dr. T Vijaya Kumar : "Langchain-Powered virtual assistant for PDF Communication", IJERT, Issue 2023