AI-Powered Finance Chatbot

Introduction

The client works in the finance industry and we developed an AI-powered chatbot for their website. The goal was to streamline the process of accessing and delivering the exact information based on queries by the end users. By harnessing the power of OpenAI technology, our chatbot drastically reduced the time that was getting consumed earlier for this. To achieve this, we used a Large Language Model (LLM) based AI chatbot. We used OpenAI APIs to seamlessly communicate with the finance data that was provided to us by the client in the form of pdf files. The process involves feeding these files into the AI model, by chunking the texts and then converting it into vector form. OpenAI Embedding is used to feed the information with OpenAI APIs. The method has enabled us to significantly reduce users' search time by an impressive 90%, sparing them the ordeal of having to manually navigate through numerous website pages to arrive at the information they need.

We trained the chatbot to analyze user queries and extract relevant information from the pdf files provided. The process ensures that the responses offered are those that are extracted solely from the information available in the pdfs only; with nothing coming from the internet. Through seamless integration via API and robust data storage capabilities, our chatbot not only enhances user experience but also provides valuable insights that can be harnessed for analytics downstream.

Client details:

Name: Confidential | **Industry:** Financial Services | **Location:** USA

Technologies:

Python, FastAPI, OpenAI, LangChain, FAISS vector DB, Chroma DB, PyPDF2, python-dotenv, Linux, PostgreSQL, OpenAI Embeddings

AI-Powered Finance Chatbot

Project Description:

Application Overview:

We have used OpenAI and LangChain to manage chat sessions. A text splitter is used to handle large amount of texts that is available in multiple pdf files. The backend is addressed using Python (3.10) with the new Lightweight API framework **FastAPI.** All the REST APIs are written in FastAPI and raw queries have been used for Database connectivity. Backed by OpenAI APIs, the chatbot responds in very less time to users' queries.

Variables impacting the performance of the application:

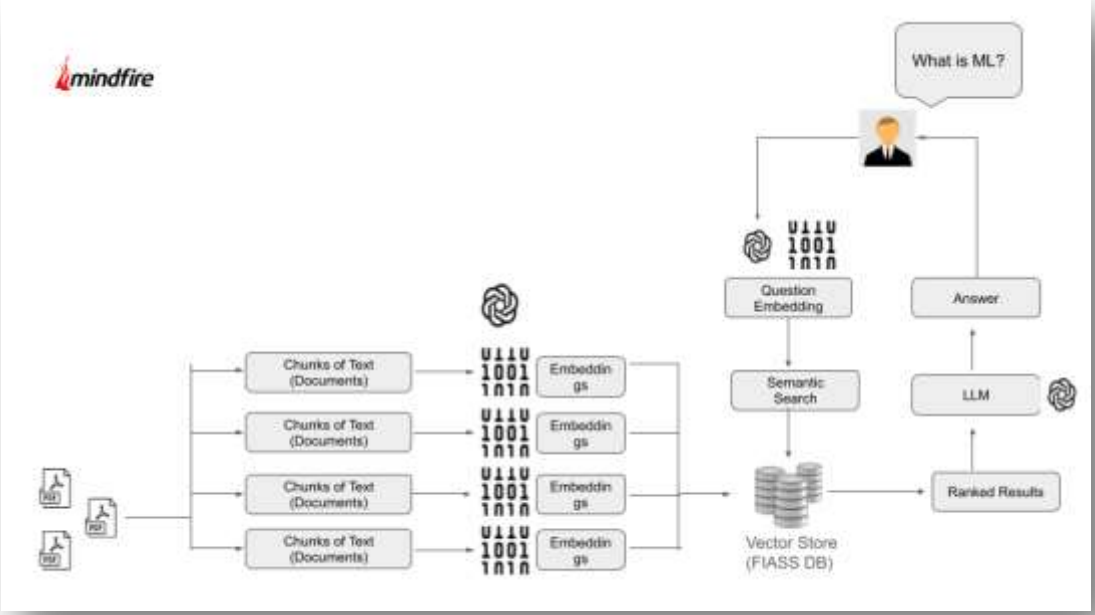
- GPU type and number of Cuda cores
- GPU Memory
- Size of CPU
- System Memory size
- Network Bandwidth
- Batch Size of questions
- Parallel Queries

Key Challenges in Optimization:

- Converting N number pdf documents and their multiple pages by extracting the text from the files.
- Using a text splitter to convert the large extracted texts into smaller chunks with overlap.
- Converting the chunks in the vector database with the help of OpenAI Embeddings.
- Handling the memory of the chat history to escape from max length errors.
- Optimizing applications to have less memory footprint.

AI-Powered Finance Chatbot

Architecture Diagram



Screenshots: