Technical Documentation

Overview

The Generative AI application is designed to assist users in synthesizing and extracting information from large and complex documents using an interactive Q&A system. This application significantly reduces the time required to analyze extensive documents by offering concise and relevant responses to user queries. Unlike publicly available applications like ChatGPT, this AI platform designed specifically for an organization that allows users to process confidential documents in a controlled environment. Its target audience includes professionals such as researchers, policy makers, underwriters, and lawyers, who regularly handle substantial amounts of documentation.

Purpose

The primary purpose of the AI application is to:

- **Streamline document analysis:** It allows users to ask specific questions about the content and receive accurate, synthesized responses without reading the entire document.
- **Document summarization:** It provides users with the gist of the document, helping to pinpoint essential information.
- **Information extraction:** Users can extract and summarize information from complex tables and sections within documents.

Target Audience

- **Researchers:** Simplifies literature review by summarizing large amounts of academic papers.
- **Policy Makers:** Assists in interpreting policy documents to retrieve crucial insights.
- **Underwriters:** Extracts relevant clauses from insurance policies for easier decision-making.
- **Lawyers:** Analyzes legal documents, summarizing and answering specific legal queries.

Unique Features

- **Q&A-based interaction:** Users can engage with the AI model through questions, making the process interactive and user-centric.
- **Document Parsing:** Supports PDF files, allowing users to upload documents directly and start querying right away.
- **Table Extraction:** The AI can extract and summarize data from tables within documents, providing a comprehensive analysis of complex data formats.
- **Data Privacy and Security**: It ensures that sensitive documents are processed within a secure and controlled environment, making it particularly useful for professionals dealing with confidential information. The feature of uploading the document for Q&A is stored within the cloud environment of the organization. Organizations can safely upload sensitive materials without the risk of exposing them to a publicly

available platform, ensuring compliance with internal and external privacy regulations.

Use Cases

- **Insurance Policies & Legal Documents:** Provides quick answers regarding the contents of insurance policies or legal contracts, highlighting critical clauses and sections.
- **Investment Analysis:** Assists in the analysis of financial documents like cash flow statements, summarizing key financial metrics.
- **Academic Research:** Summarizes and synthesizes research papers, assisting in literature reviews and academic writing.
- **Personal Finance Data:** Analyzes personal finance records and documents, helping users manage and reflect on their financial health.


Technical Architecture

AI Model

- **Large Language Model:** GPT-4o mini (OpenAI's large language model).
- **Embedding Model:** OpenAI's text embedding 3 small (for document parsing and context extraction).

System Architecture

- **Frontend:** The application is built using **Streamlit**, providing a user-friendly web interface for document uploads and interaction.
- **Backend:**
  - **Azure Blob Storage:** Stores the documents uploaded by users in the cloud, ensuring scalability and security.
  - **LlamaParse:** Responsible for parsing the PDF files and generating the initial text data for processing.
  - **Llama Index:** Handles the indexing, transformation, and storage of the parsed document content.
  - **OpenAI APIs:** Powers the Q&A functionalities using OpenAI's models for language understanding and response generation.
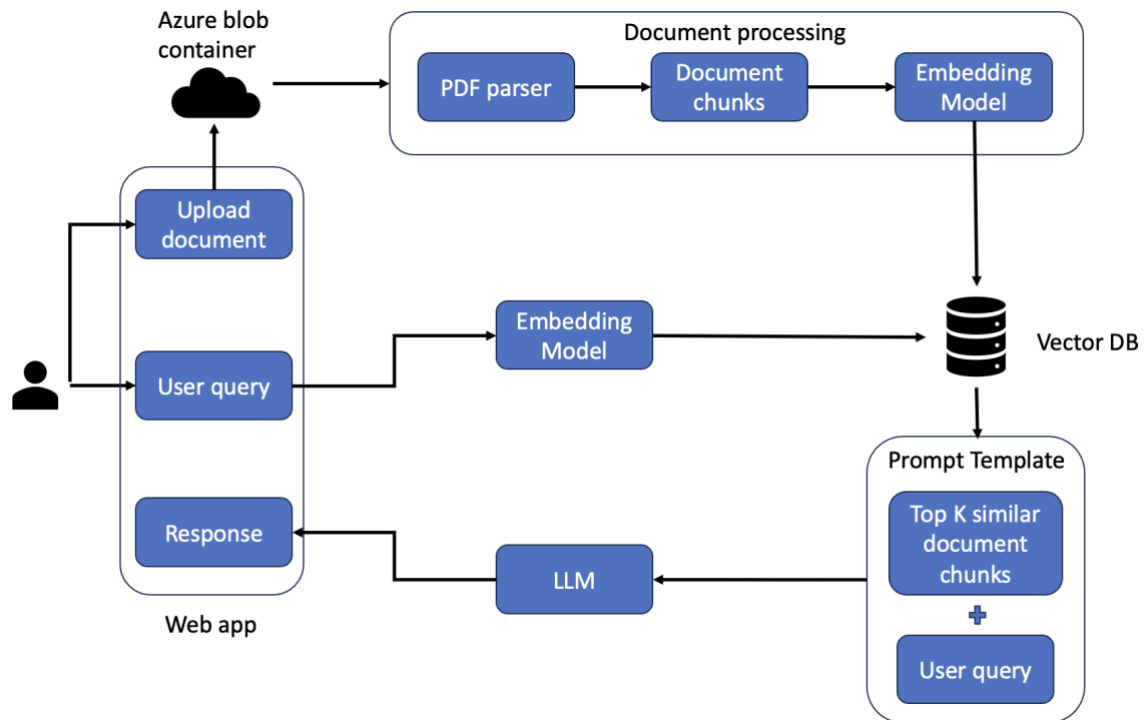
API Endpoints

The application integrates with several key API endpoints:

- **OpenAI**: Provides access to the GPT-4o model for generating responses and the text embedding model for parsing documents.
- **Llama Cloud**: Manages the parsing of uploaded PDF documents.
- **Azure Blob Storage**: Serves as the storage solution for all user-uploaded files.

Dependencies

- **Streamlit:** Used for building the interactive web application interface.
- **Azure Blob Storage:** Manages cloud storage for uploaded documents.
- **Llama Cloud:** Responsible for document parsing services.
- **Llama Index:** Manages indexing and storing parsed documents.
- **OpenAI API**: Provides the core functionality for both the embedding model and the Q&A agent.

Conclusion

This Generative AI application offers an innovative solution for professionals working with extensive documents. By leveraging powerful AI models for Q&A, document parsing, and information extraction, it enhances efficiency, reduces time spent on manual document review, and provides targeted insights tailored to the user's needs.