

# Rag based model using AI foundry open AI models for embedding and Conversation:

## 1. Introduction

### Retrieval-Augmented Generation (RAG) Overview:

Retrieval-Augmented Generation (RAG) is an advanced natural language processing (NLP) technique that combines the power of information retrieval with language generation models. Rather than relying solely on a pre-trained language model's internal knowledge, RAG enhances responses by retrieving relevant external information (e.g., from a document database or knowledge base) and incorporating it into the generation process.

In a typical RAG pipeline:

1. **Retrieval:** A user's query is embedded and used to search a vector database for the most semantically relevant documents or passages.
2. **Augmentation:** Retrieved content is added as context.
3. **Generation:** A language model (e.g., GPT-4o) uses the context and query to generate a more accurate, grounded response.

This architecture improves accuracy, ensures better factual grounding, and enables dynamic question answering based on up-to-date or custom datasets.

### Use Case: Q&A from Uploaded PDF Documents

This project implements a RAG-based solution focused on extracting knowledge from PDF documents uploaded by users. The goal is to enable users to:

- Upload one or more PDF documents.
- Ask natural language questions about the document content.
- Receive relevant, accurate answers based on the PDF's text.

By converting document text into embeddings, storing them in a FAISS vector index, and using Azure's powerful AI models (including text embedding and GPT-4o chat), the system provides a scalable and intelligent solution for enterprise document search and understanding.

## 2. Objectives

This project aims to build an intelligent document question-answering system using a Retrieval-Augmented Generation (RAG) architecture. The key objectives are as follows:

## Enable Semantic Search from PDF Documents

- Extract and process text from uploaded PDF files.
- Split text into semantically meaningful chunks to preserve context.
- Generate embeddings to represent the meaning of each chunk for efficient similarity search.

## Use Azure AI for Embedding Generation

- Utilize Azure AI Foundry's text embedding models to transform text chunks into high-dimensional vector representations.
- Leverage the scalability and performance of Azure AI for secure and reliable embedding computation.

## Leverage FAISS for Fast Vector Retrieval

- Store text embeddings in a **FAISS (Facebook AI Similarity Search)** index.
- Use FAISS for efficient approximate nearest neighbor (ANN) search to retrieve the most relevant document chunks based on the user's query.



## Use GPT-4o for Answering User Queries

- Employ **Azure OpenAI GPT-4o** to process user questions.
- Provide GPT-4o with retrieved document chunks as context to generate accurate and grounded answers.
- Ensure high-quality, context-aware natural language responses.

# 3. Architecture Overview

This section outlines the architecture of the RAG-based document question-answering system built using Azure AI services, FAISS, and Python. The solution is modular, scalable, and designed for performance and flexibility.



## High-Level Components

### 1. User Interface

- Allows users to upload PDF documents and input natural language questions.
- Can be implemented using a web app (e.g., Streamlit, React, Flask frontend).

### 2. Document Ingestion Pipeline

- Receives and stores uploaded PDFs.
- Extracts raw text using PDF parsers such as **PyMuPDF**.
- Splits the text into semantically meaningful chunks.

### 3. Embedding Generation

- Each text chunk is sent to **Azure AI Foundry's Text Embedding Model**.
- Embeddings are returned as numerical vectors representing the meaning of the chunk.

### 4. Vector Storage & Search

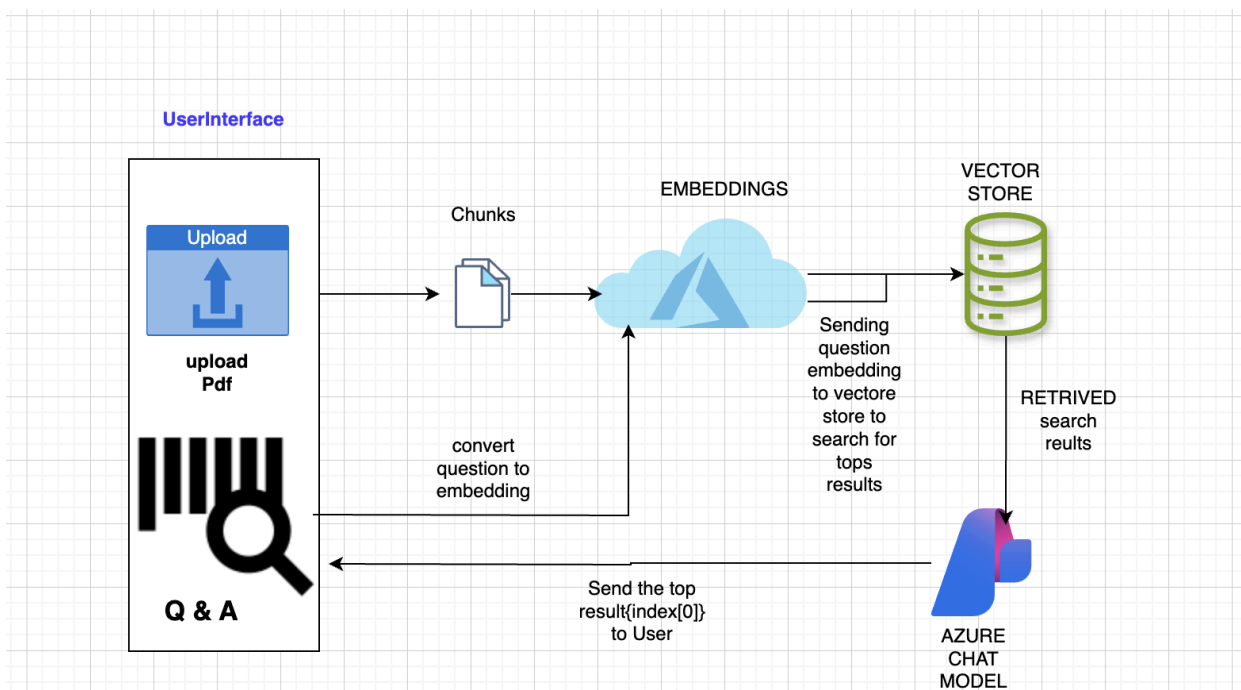
- All chunk embeddings are stored in a **FAISS index**.
- On receiving a user query, the system:
  - Embeds the query using the same Azure model.
  - Searches FAISS for top-N most similar text chunks.

### 5. Answer Generation

- The query and retrieved chunks are packaged as context input for **Azure OpenAI's GPT-4o model**.
- GPT-4o generates a human-like, context-aware response based on the retrieved content.

### 6. Response Delivery

- The generated answer is returned to the user via the UI.



## 4. Implementation

### 1.Connect the models in Azure

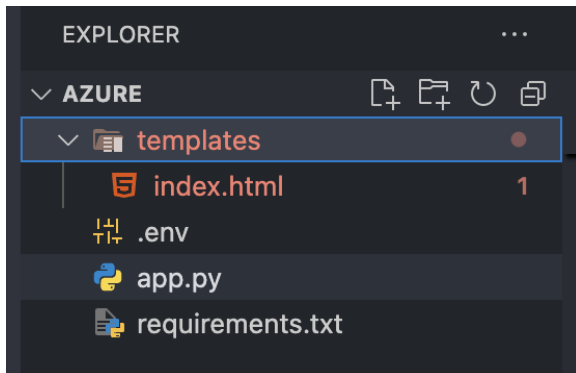
Deploy the required models to the AI foundry:

Login to azure console.In the search bar search for azure ai services then,Go to Azure AI Foundry in azure open AI.

1. Select **Deployments** on the left menu.

2. Select **Deploy model > Deploy base model**.
3. Select **text-embedding-3-large** from the dropdown list and confirm the selection.
4. Accept the defaults.
5. Select **Deploy**.
6. Repeat the previous steps for **gpt-4o**. Which will be used in later steps.

2. Structure of the Code: This is how the code has been organized in the vs code. I.e UI for now will be implemented just using Index.html and the backend is [app.py](#) which is python based.



### 3. Requirement.txt

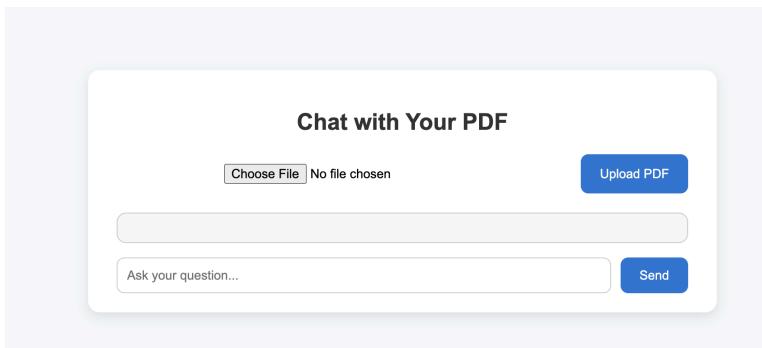
```
requirements.txt
1 flask
2 python-dotenv
3 openai>=1.0.0
4 faiss-cpu
5 PyMuPDF
```

4. .env file : we get this information from the deployment models in Ai foundry

```
AZURE_OPENAI_API_KEY=9a0vt06X9SAJKRZbTHQTFVc8K02omlKMVmtVuIp0L
AZURE_OPENAI_ENDPOINT=https://triosoftllc.openai.azure.com/
AZURE_EMBEDDING_DEPLOYMENT=text-embedding-3-large
AZURE_CHAT_DEPLOYMENT=gpt-4o
```

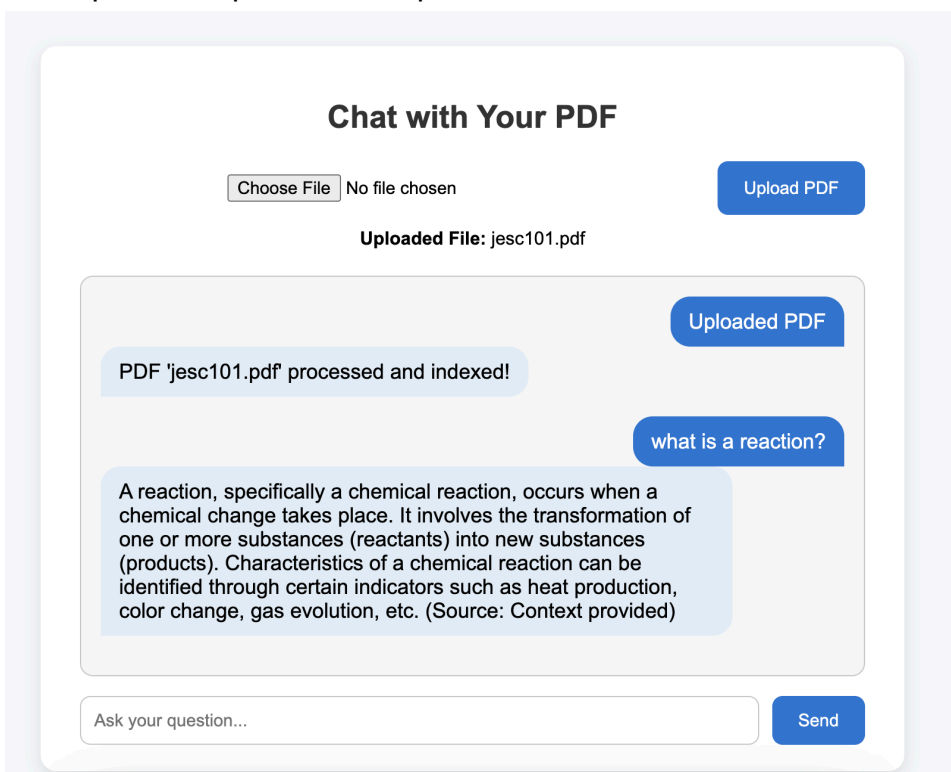
5. Code can be viewed in the code section in this git repo

## 5. Outputs:




The screenshot shows a web interface titled "Chat with Your PDF". At the top, there is a "Choose File" button and the text "No file chosen". To the right is a blue "Upload PDF" button. Below these is a text input field. At the bottom, there is a text input field with the placeholder "Ask your question..." and a blue "Send" button.

Now upload the pdf and ask questions



The screenshot shows the same "Chat with Your PDF" interface, but now with a file uploaded. The "Choose File" button is still present, but the text next to it says "No file chosen". Below this, it says "Uploaded File: jesc101.pdf". To the right of this text is a blue "Upload PDF" button. Below the file name, there is a blue button labeled "Uploaded PDF". In the chat area, there is a light blue bubble that says "PDF 'jesc101.pdf' processed and indexed!". Below this, there is a blue bubble that says "what is a reaction?". Below the question bubble, there is a light blue bubble containing the answer: "A reaction, specifically a chemical reaction, occurs when a chemical change takes place. It involves the transformation of one or more substances (reactants) into new substances (products). Characteristics of a chemical reaction can be identified through certain indicators such as heat production, color change, gas evolution, etc. (Source: Context provided)". At the bottom, there is a text input field with the placeholder "Ask your question..." and a blue "Send" button.

## 6. Code Output:

 Top 3 Retrieved Chunks: These are the top **k=3** document chunks retrieved from the vector store (FAISS) based on similarity to the user's question.

- The app uses **Azure OpenAI embeddings** to convert text into vectors.
- Each chunk is a segment of the PDF split using a fixed **chunk\_size** (e.g., 500 characters).
- FAISS performs a **similarity search** (L2 distance) to return the top 3 most relevant chunks.

Chunk 1:

-----

always be balanced.

n

In a combination reaction two or more substances combine to form a new single substance.

n

Decomposition reactions are opposite to combination reactions. In a decomposition reaction, a single substance decomposes to give two or more substances.

n

Reactions in which heat is given out along with the products are called exothermic reactions.

n

Reactions in which energy is absorbed are known as endothermic reactions.

n

When an element displaces another element from its compound

-----

Chunk 2:

-----

our previous classes. Whenever a chemical change occurs, we can say that a chemical reaction has taken place.

You may perhaps be wondering as to what is actually meant by a chemical reaction. How do we come to know that a chemical reaction has taken place? Let us perform some activities to find the answer to these questions.

Figure 1.1

Burning of a magnesium ribbon in air and collection of magnesium oxide in a watch-glass

Activity 1.1

Activity 1.1

Activity 1.1

## Activity 1.1

## Activity 1.1

### CAUTION:

-----

### Chunk 3:

-----

say that when two or more substances  
(elements or compounds) combine to form a single product, the reactions  
are called combination reactions.

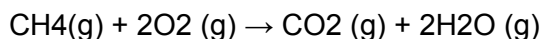
In Activity 1.4, we also observed that a large amount of heat is evolved.

This makes the reaction mixture warm. Reactions in which heat is  
released along with the formation of products are called exothermic  
chemical reactions.

Other examples of exothermic reactions are –

(i)

Burning of natural gas




(1.17)

(ii)

Do you

-----

 **CONTEXT USED FOR ANSWERING:** This is the **concatenation of the top 3 chunks** provided to the chat model as context.

The context acts as a **knowledge base**, from which GPT-4o must generate its answer.

The system prompt instructs GPT-4o to:

- Only use the provided context, Avoid guessing or hallucinating and Provide citations if relevant

=====

always be balanced.

n

In a combination reaction two or more substances combine to form a new single substance.

n

Decomposition reactions are opposite to combination reactions. In a decomposition reaction, a single substance decomposes to give two or more substances.

n

Reactions in which heat is given out along with the products are called exothermic reactions.

n

Reactions in which energy is absorbed are known as endothermic reactions.

n

When an element displaces another element from its compound in the reactions we have studied in our previous classes. Whenever a chemical change occurs, we can say that a chemical reaction has taken place. You may perhaps be wondering as to what is actually meant by a chemical reaction. How do we come to know that a chemical reaction has taken place? Let us perform some activities to find the answer to these questions.

Figure 1.1

Burning of a magnesium ribbon in air and collection of magnesium oxide in a watch-glass

Activity 1.1

Activity 1.1

Activity 1.1

Activity 1.1



## Activity 1.1

### CAUTION:

say that when two or more substances (elements or compounds) combine to form a single product, the reactions are called combination reactions.

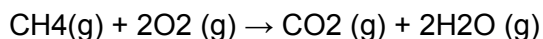
In Activity 1.4, we also observed that a large amount of heat is evolved.

This makes the reaction mixture warm. Reactions in which heat is released along with the formation of products are called exothermic chemical reactions.

Other examples of exothermic reactions are –


(i)

Burning of natural gas



(1.17)

=====

 **GPT-4o FULL RESPONSE:** This is the **answer generated by the Azure OpenAI GPT-4o chat model**, printed and shown in the UI.

- It uses the system prompt to ensure factual correctness and citation.
- The response might include:
  - Bullet points (for structured answers)
  - Page references (if present in the text)
  - A disclaimer like "I don't know" if no relevant info is found

=====

A reaction, specifically a chemical reaction, occurs when a chemical change takes place. It involves the transformation of one or more substances (reactants) into new substances (products). Characteristics of a chemical reaction can be identified through certain indicators such as heat production, color change, gas evolution, etc.

(Source: Context provided)

=====