RAG-Powered Smart Chatbot for Any Document

Overview

This project enables document understanding through a Retrieval-Augmented Generation (RAG) chatbot. It allows users to upload documents (PDF, DOCX, PPTX, TXT), which are processed, chunked, and embedded using Azure OpenAI's embedding models. The chunks are stored in ChromaDB and queried using semantic similarity to generate contextual responses using GPT-4o.

System Flow Structure

The following steps outline the processing pipeline:

- 1. Upload: User uploads documents via Streamlit UI (PDF, DOCX, PPTX, TXT).
- 2. Parsing: Extract and clean raw text from documents using unstructured loaders.
- 3. Chunking: Split the text into small manageable pieces
- 4. Embedding: Use Azure OpenAI's `text-embedding-3-small ` model to convert text chunks into high-dimensional vectors.
- 5. Storage: Save the vectors in a local ChromaDB database.
- 6. Search: Convert user query into embedding and find similar documents using vector similarity.
- 7. Response Generation: GPT-4o uses retrieved chunks as context to generate a grounded answer.
- 8. Response includes document source, page.

Azure OpenAI Embedding: text-embedding-3-small

- Efficient and lightweight embedding model.
- Converts text into 1536-dimensional vectors.
- Used for semantic similarity matching with user queries.
- Optimized for fast computation and small memory footprint.
-

GPT-4o: Advanced Chat Completion

- Multimodal and real-time generative model.
- Receives chunks and query context to produce relevant answers.
- Understands citations, formatting, and technical language.
- Excellent zero-shot reasoning and document summarization capabilities.

Architecture Diagram

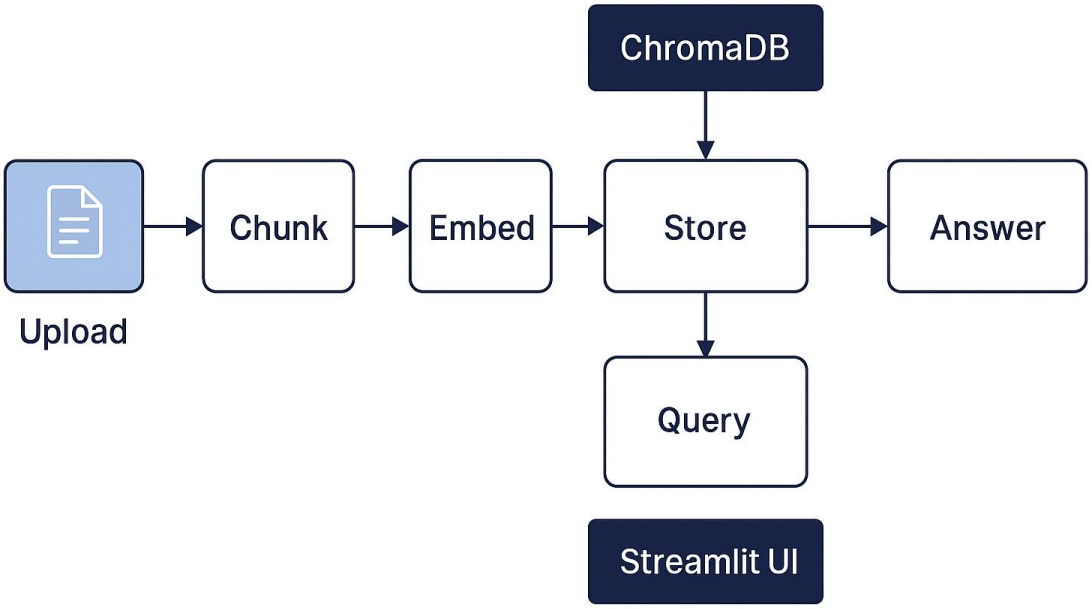


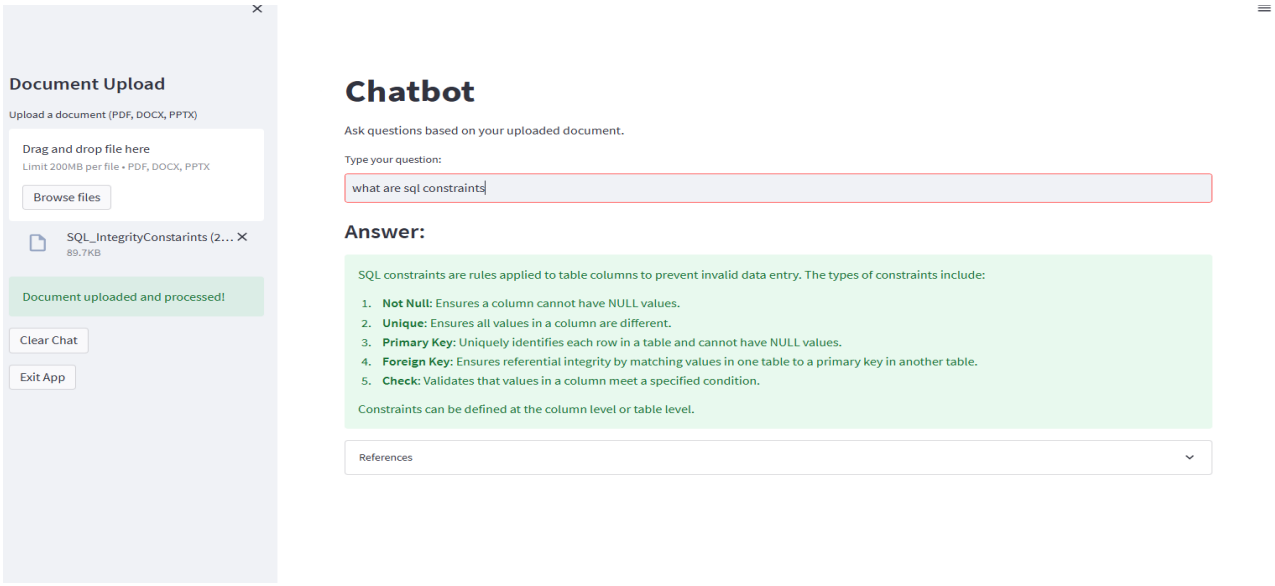
Figure: End-to-end architecture of the RAG chatbot system.

Key Features

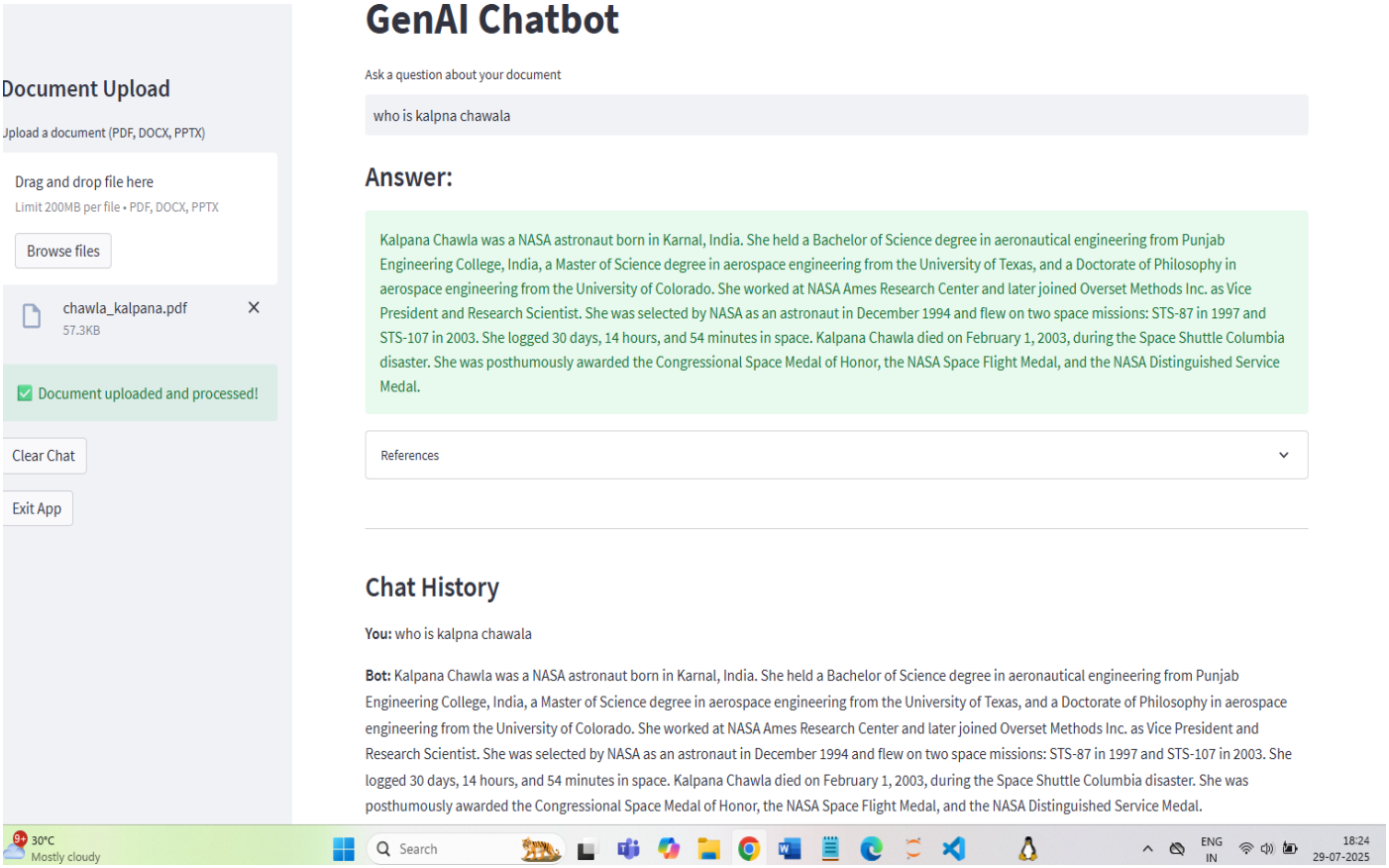
- Streamlit UI with file upload and reset/clear options.
- Supports PDFs, Word, PowerPoint, and plain text files.
- Local ChromaDB vector store for efficient retrieval.
- Chunked embedding for context-preserving answers.
- GPT-4o driven responses grounded in original content.

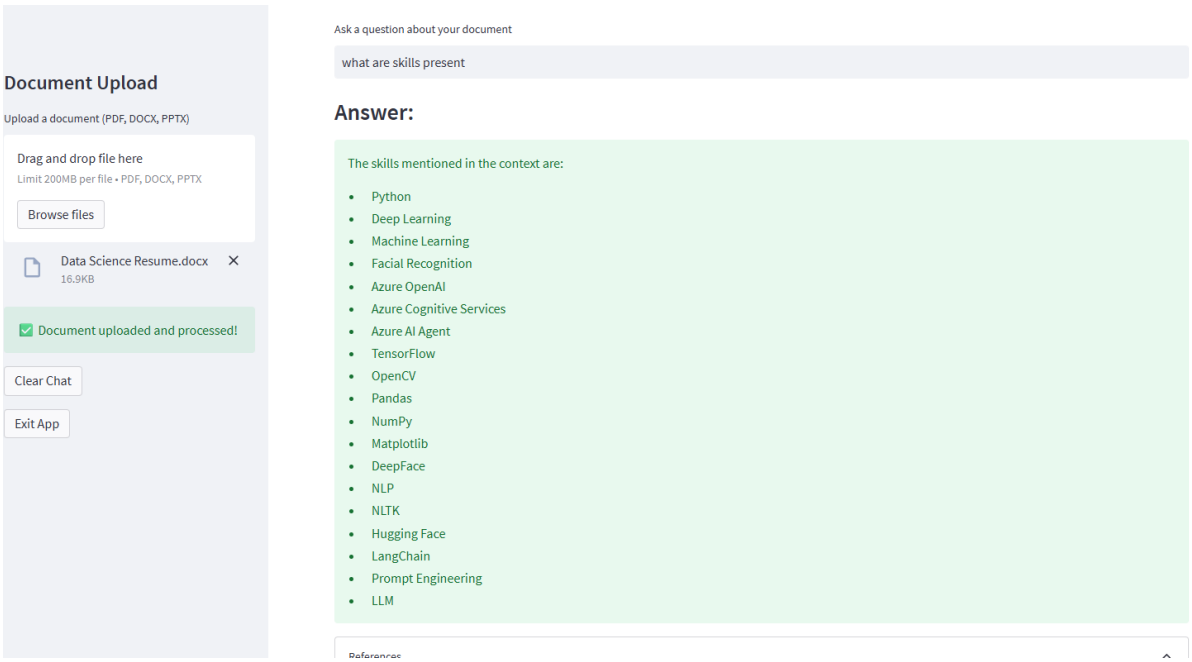
Please refer to the snippet below from the chatbot project.

I uploaded a Pptx and asked, "what are SQL Constraints ?" The chatbot responded with relevant information extracted from the file.



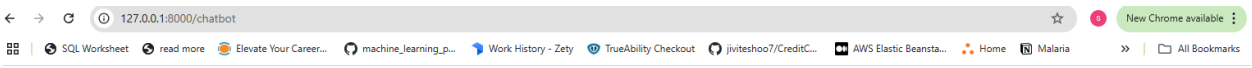
I uploaded a PDF and asked, "Who is Kalpna?" The chatbot responded with relevant information extracted from the file. And also references from a file .





- chatbot project was built using **FastAPI** for handling API endpoints like document upload and chat-based querying. FastAPI was chosen for its speed, simplicity, and automatic documentation support.

- The APIs are fully documented with Swagger UI (available at /docs) These tools make it easy to **test**, **integrate**, and **extend** the chatbot capabilities with minimal setup. The backend supports file uploads (PDF, DOCX, PPTX, TXT), processes them into vector embeddings, and responds to natural language questions.



GenAI RAG Chatbot 0.1.0 OAS 3.1

/openapi.json

Upload documents and chat using GPT + ChromaDB.

default

POST /upload/ Upload File

POST /chat/ Chat With Bot

Schemas

Body_chat_with_bot_chat__post > Expand all object

Body_upload_file_upload__post > Expand all object

HTTPValidationError > Expand all object

ValidationError > Expand all object

Curl

```
curl -X 'POST' \
  'http://127.0.0.1:8000/chat/' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/x-www-form-urlencoded' \
  -d 'query=who%20is%20kalpna'
```

Request URL

```
http://127.0.0.1:8000/chat/
```

Server response

Code	Details

200

Response body


```
{
  "answer": "Kalpana Chawla was a NASA astronaut born in Karnal, India. She held a Bachelor of Science degree in aeronautical engineering from Punjab Engineering College, India, a Master of Science degree in aerospace engineering from the University of Texas, and a Doctorate of Philosophy in aerospace engineering from the University of Colorado. She worked at NASA Ames Research Center and later joined Overset Methods Inc. as Vice President and Research Scientist. Selected by NASA in December 1994, she flew on two space shuttle missions, STS-87 in 1997 and STS-107 in 2003, logging a total of 30 days, 14 hours, and 54 minutes in space. She perished on February 1, 2003, when Space Shuttle Columbia and her crew were lost during re-entry. She was posthumously awarded the Congressional Space Medal of Honor, the NASA Space Flight Medal, and the NASA Distinguished Service Medal.",
  "documents": [
    {
      "id": null,
      "metadata": {
        "sourcemodified": "D:20140821135000",
        "producer": "Adobe PDF Library 15.0",
        "moddate": "2017-05-03T15:48:16-05:00",
        "page": 0,
        "total_pages": 1,
        "creationdate": "2017-05-03T15:48:16-05:00",
        "keywords": "",
        "company": "NASA/JSC",
        "creator": "Acrobat PDFMaker 17 for Word",
        "comments": "",
        "source": "data\\chawla_kalpana.pdf",
        "subject": "",
        "author": "FCOD(ASTRONAUT OFFICE)",
        "title": "",
        "page_label": "1"
      },
      "page_content": "KALPANA CHAWLA (PH.D.) \nNASA ASTRONAUT (DECEASED) \n \nPERSONAL DATA:  Born in Karnal, India.  Died on February 1, 2003 over the southern \nUnited Stat
```


Download