📝 RAG PDF Chatbot – Project Report

📌 Overview

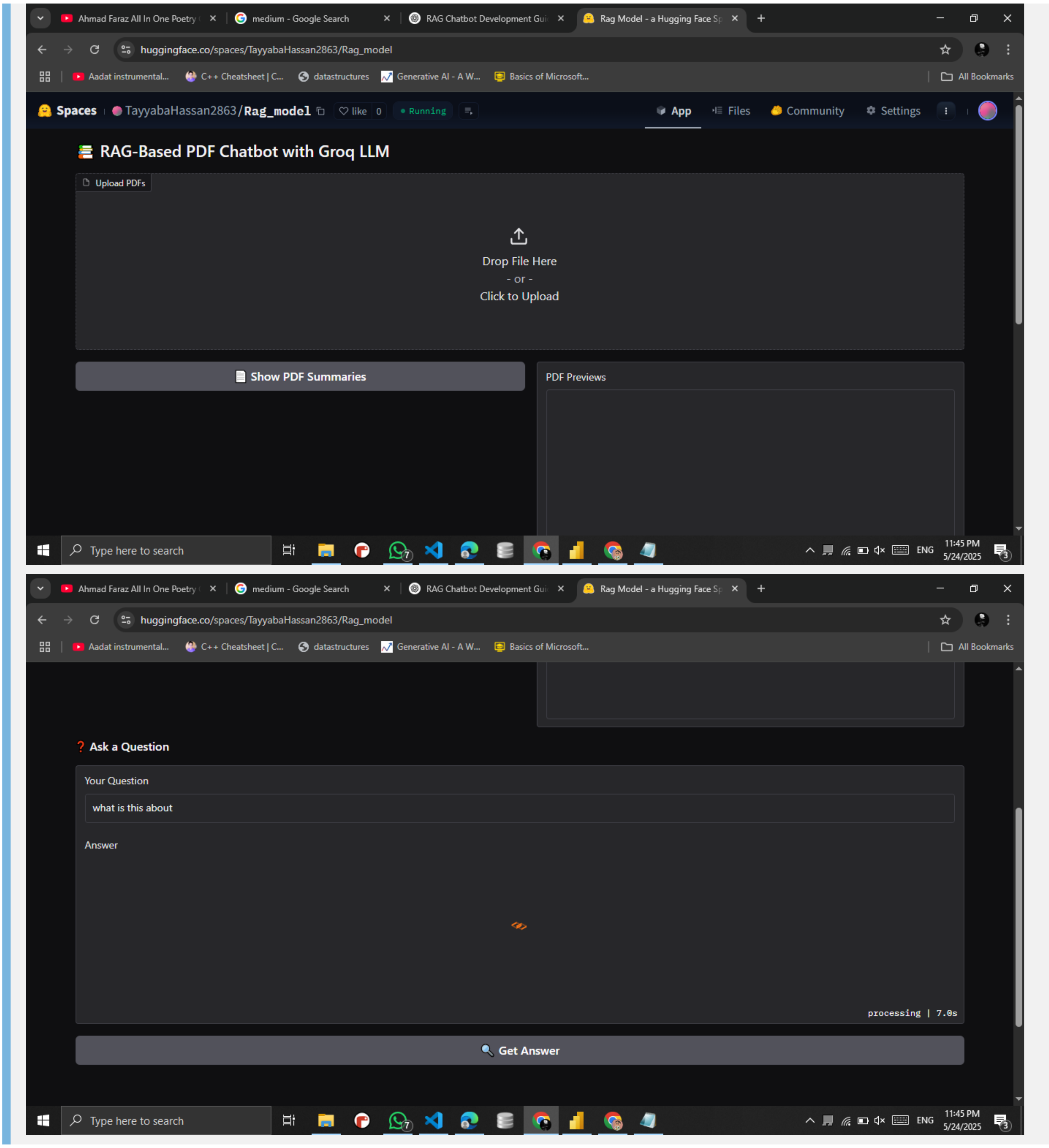
This project is a Retrieval-Augmented Generation (RAG) chatbot designed to answer user questions based on the content of multiple uploaded PDF documents. The app extracts text from each PDF, splits the text into overlapping chunks, generates vector embeddings using a `sentence-transformers` model, retrieves top relevant chunks based on cosine similarity, and sends them along with the user's question to a **Groq-powered LLM** . The LLM then returns a human-like answer. The app is built with Gradio and deployed on Hugging Face Spaces.

✨ Enhancements Added

1. **PDF Preview/Summary Feature**
   Before asking a question, users can click a button to view short previews (first ~500 characters) from each uploaded PDF, helping them understand the contents at a glance.

2. **Source Chunk Labels**
   Relevant chunks used to form the context are labeled (e.g., "Chunk #1") so that users know which part of the document their answer came from, improving transparency and trust.

🖼️ Screenshots



huggingface.co/spaces/TayyabaHassan2863/Rag_model

Spaces TayyabaHassan2863/Rag_model

RAG-Based PDF Chatbot with Groq LLM

Ask a Question

Your Question: what is this about

Answer

processing | 7.0s

🚧 Challenges Faced

- **Groq API Integration**
  Adjusting the prompt structure and headers to match Groq's API format required careful formatting to avoid errors and get proper responses.

- **Text Chunking**
  It was important to design a chunking method that preserves semantic meaning while staying within context length limits.

- **Deployment on Hugging Face Spaces**
  Managing external dependencies like `fitz` for PDF extraction and `sentence-transformers` for embeddings needed proper setup in `requirements.txt` and `apt.txt`.