Multi-PDF Q&A Chatbot using LLaMA and Gradio

This app allows users to upload multiple PDFs and ask questions based on their content. It uses the **GROQ API (LLaMA 3-8B)** as the LLM backend, with **sentence-transformers** for semantic search. The chatbot remembers the last few interactions to provide context-aware answers.

Overview

This application provides:

- Multi-PDF upload and intelligent text extraction
- Sentence chunking and embedding using MiniLM
- Semantic search for relevant content
- Context-aware querying using LLM
- Conversational memory to enhance user experience
- PDF export of the last LLM response
- Clean Gradio interface with improved styling

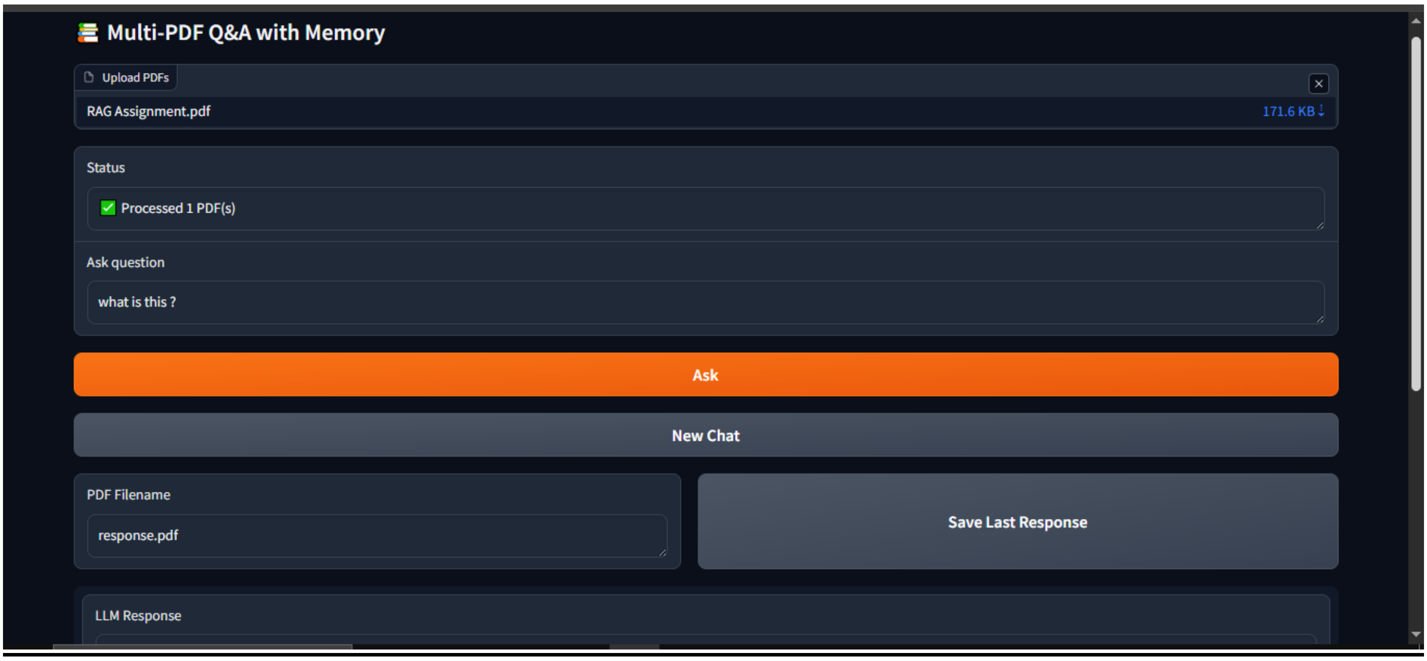
Enhancements Added

- Support for multiple PDF uploads
- Conversation memory to track recent Q&A
- Relevant document chunk retrieval based on semantic similarity
- Save last LLM response as a downloadable PDF
- Improved UI with scrollable response display and styling

Challenges Faced

Challenge	Solution
Secure API key handling	Used Hugging Face Secrets to manage the GROQ API key
Chunking large documents	Implemented sentence-based chunking with a size limit
Managing context for LLM	Limited memory to recent interactions for performance
Latency on large files	Applied chunk and embedding limits to ensure responsiveness

Screen Shots of running Model



Multi-PDF Q&A with Memory

Upload PDFs

RAG Assignment.pdf 171.6 KB

Status

✅ Processed 1 PDF(s)

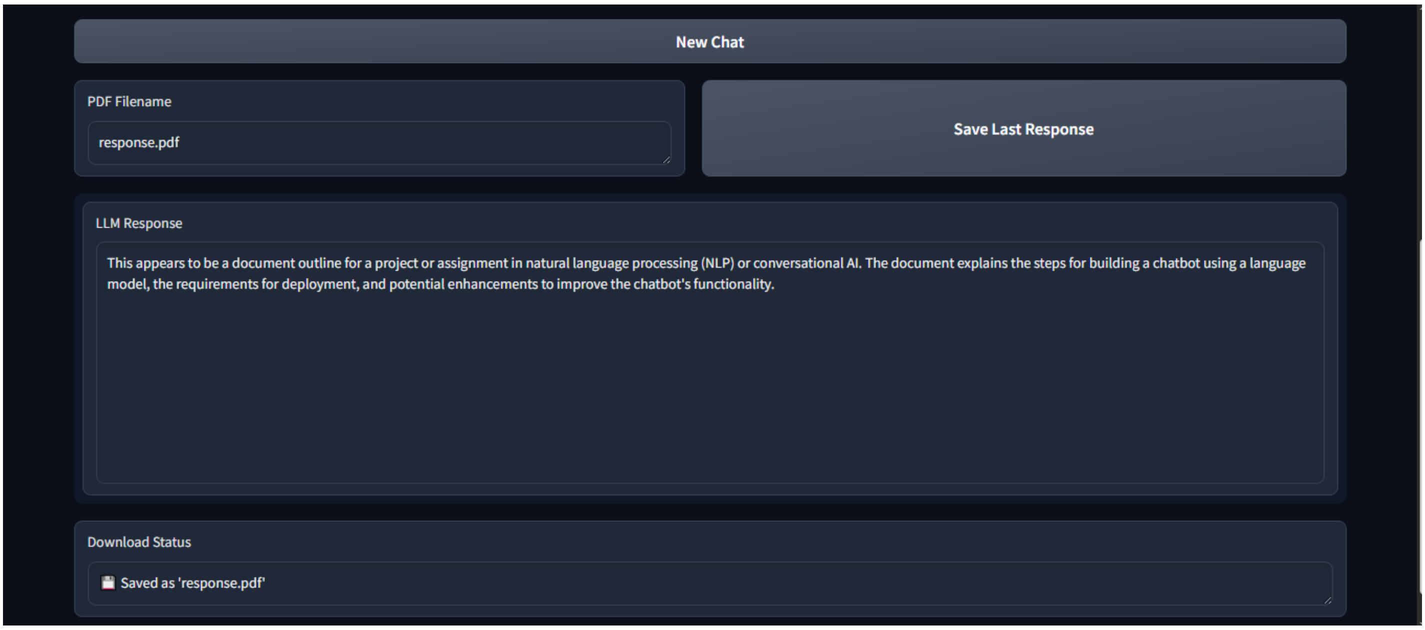
Ask question

what is this ?

PDF Filename

response.pdf

LLM Response



This appears to be a document outline for a project or assignment in natural language processing (NLP) or conversational AI. The document explains the steps for building a chatbot using a language model, the requirements for deployment, and potential enhancements to improve the chatbot's functionality.

Saved as 'response.pdf'

Tech Stack

- Gradio
- PyMuPDF
- Sentence Transformers
- scikit-learn
- FPDF
- httpx
- GROQ API (LLaMA 3-8B)