Historical Figures Chatbot – Project Documentation

This document explains the design and working of the Historical Figures Chatbot project. The chatbot answers user questions based strictly on the content of a provided PDF document using a Retrieval-Augmented Generation (RAG) approach.

1. LangSmith Configuration

LangSmith tracing is enabled using environment variables. This allows monitoring and debugging of LangChain workflows during execution.

2. PDF Ingestion

The chatbot loads historical_figures.pdf using PyPDFLoader. This PDF is the only knowledge source used to answer questions.

3. Text Splitting

The PDF content is split into overlapping chunks to improve retrieval accuracy.

4. Embeddings and Vector Store

Text chunks are converted into vector embeddings using Ollama’s granite-embedding model and stored in ChromaDB for similarity search.

5. Language Model

A local Ollama LLM (llama3) generates answers using retrieved context.

6. RetrievalQA Chain

LangChain’s RetrievalQA combines retrieval and generation, ensuring answers come only from the document.

7. Chat History

In-memory chat history stores the conversation and can be cleared using the Clear History button.

8. User Interface

The Gradio UI includes a banner image, chatbot window, input box, Submit and Clear History buttons, and a footer showing tracing status.

Conclusion

This project demonstrates a complete PDF-based chatbot using LangChain, Ollama, ChromaDB, and Gradio with a clean and professional interface.