Project Info

Project Name : LLM-Fusion

Description : Fusion of chat and document summarization using LLMs

Problem Statement: LLM-Fusion

In an era where users are overwhelmed with vast volumes of digital documents and simultaneously seek personalized AI-driven interactions, there is a growing demand for a unified solution that can handle both natural language conversation and intelligent document summarization — all while maintaining data privacy and functioning without cloud dependencies.

Existing solutions often:

- Rely on cloud-based LLMs, risking data exposure.
- Require multiple disconnected tools for chat and document processing.
- Lack flexibility for offline or on-premise environments.
- Are not user-friendly for non-technical users.

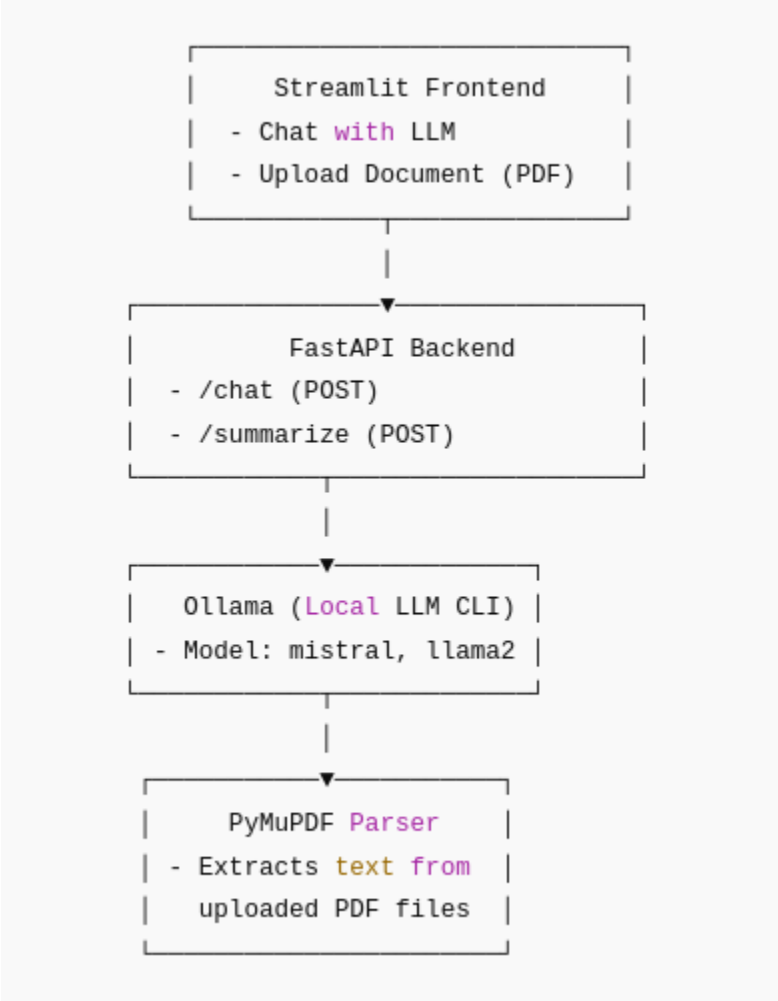
LLM-Fusion aims to address this gap by providing a lightweight, offline-capable application that:

- Uses local LLMs (via Ollama) for both chatbot interaction and document summarization.
- Offers a simple, unified Streamlit-based UI and a modular FastAPI backend.
- Enables users to upload documents and instantly receive summaries.

- Allows conversational queries powered by the same local model.
- Ensures full data control by avoiding cloud APIs or external services.

Solution Architecture – *LLM-Fusion*

The solution is built as a modular system combining a user-friendly frontend, a REST-based backend API, and a local LLM (via Ollama) for both conversation and summarization. It ensures offline capability, data privacy, and seamless integration.



Key Tech Stack

Frontend : Streamlit

Backend : FastAPI

LLM Engine : Ollama CLI

Document Parsing : PyMuPDF (fitz)

HTTP Client : requests

Language : Python 3.10+

Runner : Uvicorn

Local Model : Mistral (via Ollama)

Code repository (GitHub)

https://github.com/VijilaVijayanVS/LLM_ChatBot