Chatbot for Multi-Format Documents

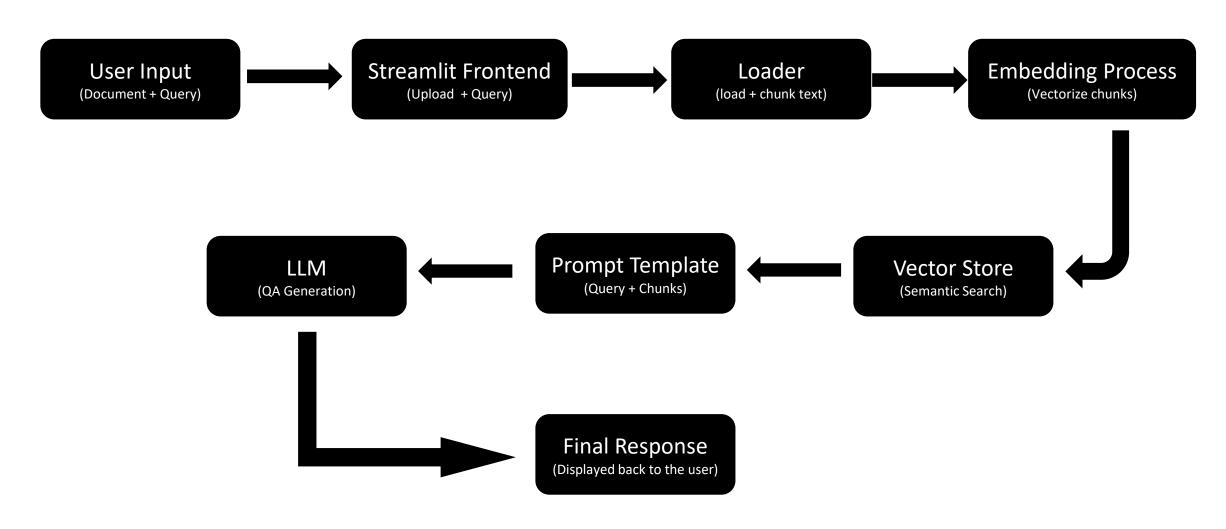
Architecture & Flow

A streamlined PDF chatbot using LangChain with HuggingFace embeddings, built for answering user queries from uploaded documents.

Architecture

Component	Role
Loader	LangChain loader to extract and chunk from uploaded documents (PDF, DOCX etc.)
Embedding Process	Applies HuggingFace embeddings to convert chunks into vectors
Vector Store	Uses FAISS (Facebook AI Similarity Search) to store and retrieve relevant chunks
Prompt Template	Combines query and retrieved chunks to guide the LLM
LLM	Uses DeepSeek Model via OpenRouter API with context-aware answers
Streamlit UI	For Frontend that sets for user interaction having page config, title, and description.

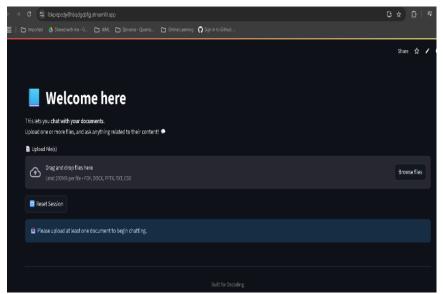
Flow



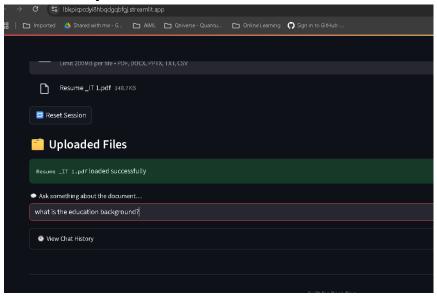
Screenshots

PDF

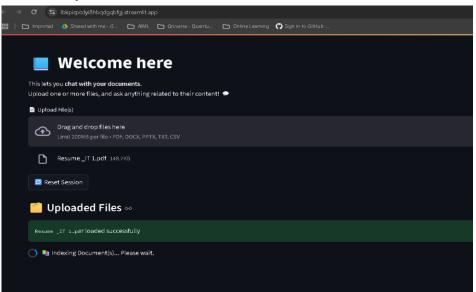
Home Page



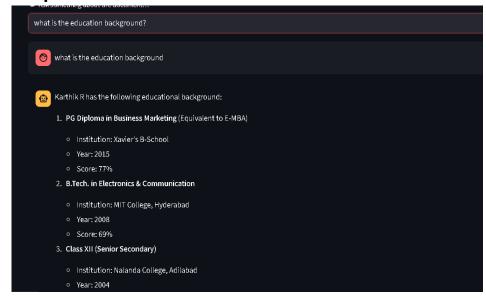
User Query



User Input



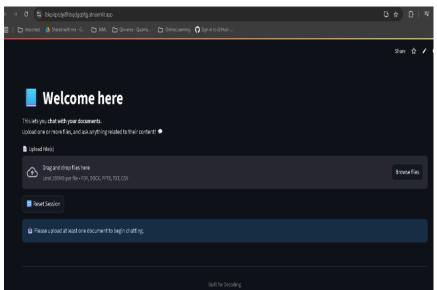
Response



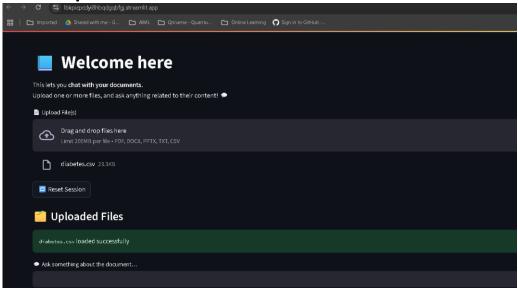
Screenshots

CSV

Home Page



User Input



User Query

Response





- 1. Features (Columns):
 - The first column could represent the number of pregnancies (integer values like 0, 1, 2, etc.).
 - The next few columns likely represent medical measurements such as glucose levels (e.g., 187, 146), blood pressure (e.g., 68, 82), skin thickness (e.g., 39, 0), insulin levels (e.g., 304, 0), and BMI (e.g., 37.7, 40.5).
 - o The column with values like 0.254, 0.962 might represent a diabetes pedigree function (a genetic risk score).
- The next column is likely age (e.g., 41, 31).
- The last column appears to be a binary outcome (0 or 1), possibly indicating the presence (1) or absence (0) of diabetes or another condition.
- 2. **Potential Source**: This resembles the **Pima Indians Diabetes Dataset**, a well-known dataset in machine learning for binary classification (predicting diabetes). The features align with common predictors like pregnancies, glucose, blood pressure, etc.
- 3. Missing/Zero Values:
 - o Some rows have zeros (e.g., insulin = 0), which may indicate missing/unknown values (common in raw datasets).

Summary:

This is likely a health dataset for diabetes prediction, where each row represents a patient's medical attributes, and the last column indicates whether they have diabetes (1) or not (0).

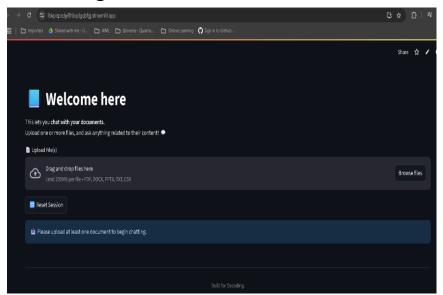
Would you like help analyzing or interpreting specific parts of the data?

Source: diabetes.csv

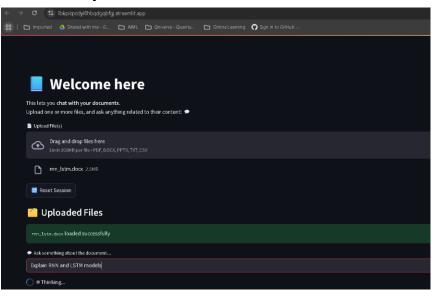
Screenshots

DOCX

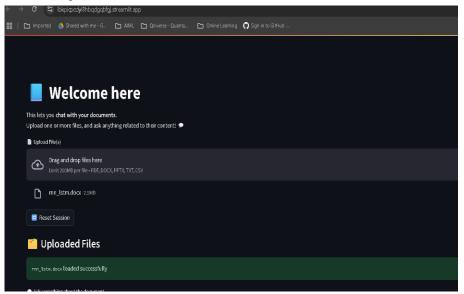
Home Page



User Query



User Input



Response

