Internal Notes on the RAG Chatbot Project

These notes describe internal design decisions for the Retrieval-Augmented Generation (RAG) chatbot.

The backend is built using FastAPI and supports session-based conversations. Each user interaction includes a session_id so conversations do not mix.

RAG is implemented using vector embeddings generated from OpenAI's embedding models. Document chunks are stored in a FAISS index to enable semantic search.

Uploaded documents (PDF or TXT) are automatically indexed and immediately available to the chatbot without restarting the server.