System Architecture Overview

This system is designed as a local AI assistant using Retrieval-Augmented Generation (RAG).

Key Components:

- FastAPI backend
- LangChain for orchestration
- Ollama for running local LLMs
- Chroma for vector storage

Flow:

1. User asks a question
2. Relevant documents are retrieved
3. Context is injected into the prompt
4. The local LLM generates an answer

Design Principles:

- Privacy-first (no cloud calls)
- Modular architecture
- Fast local inference