AI Notes Assistant - Build Guide

Overview

Build a system that answers questions from your own notes using embeddings and vector search (RAG).

Steps

1. Collect notes (PDF/Text). 2. Chunk text. 3. Create embeddings. 4. Store in vector DB. 5. Retrieve relevant chunks and answer with LLM.

Tech Stack

Python, LangChain, OpenAI/Any LLM, FAISS/Chroma, FastAPI/Streamlit, PyMuPDF for PDFs.

Chunking

Split text into 300-1000 token chunks with overlap 50-200 tokens.

Embedding

Use models like text-embedding-3-small/large or any embeddings compatible model.

Vector DB

FAISS (local), Chroma (local/cloud), Pinecone (cloud).

Prompt

Use retrieved context + user query: 'Answer using the context only.'

Deployment

Wrap with API (FastAPI) or UI (Streamlit). Deploy on Render, Hugging Face Spaces or Railway.

Evaluation

Check accuracy, latency, hallucination rate. Use metrics: hit rate, MRR.