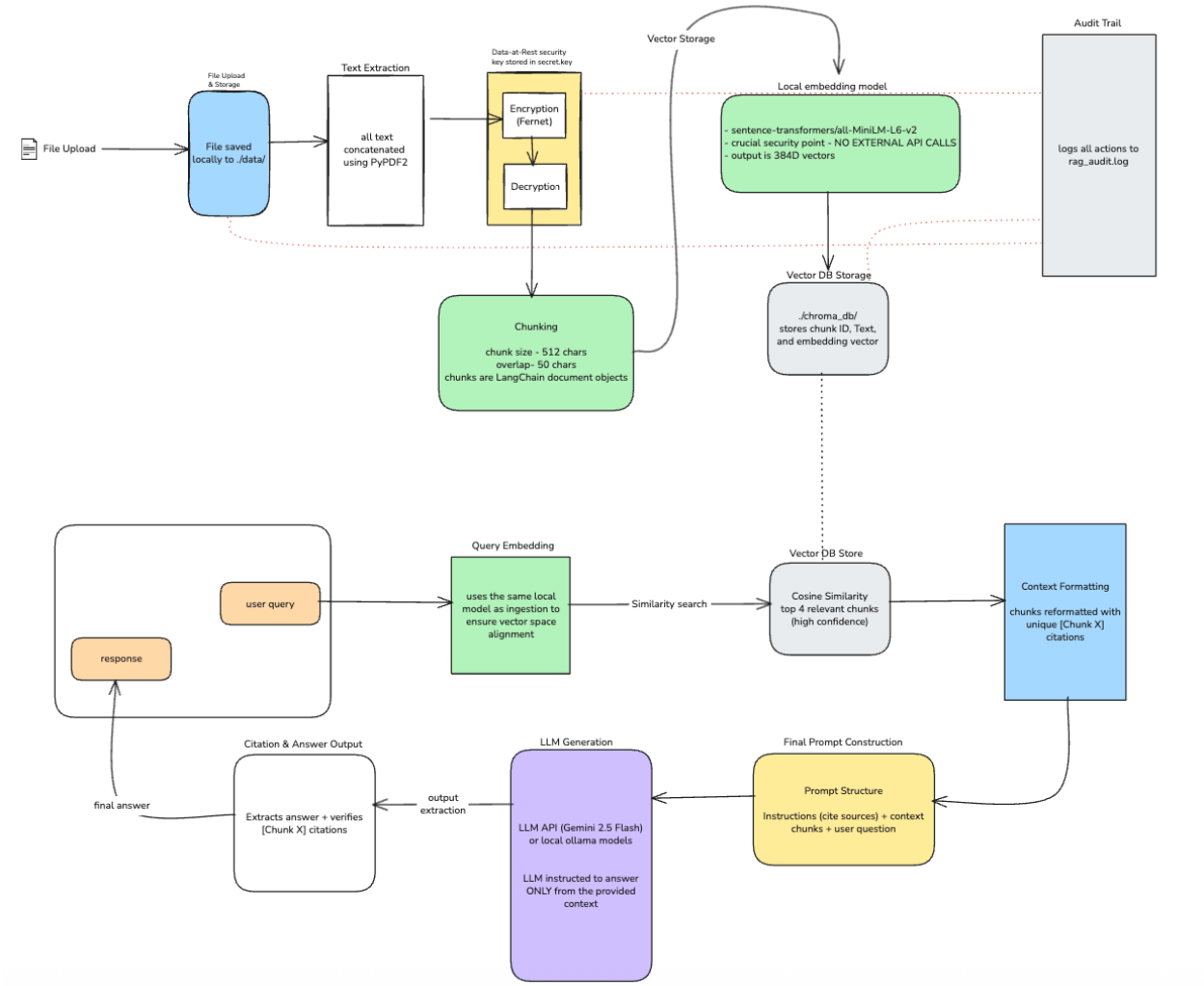
Offline-First RAG System with Quantized SLMs

Author: Santhi Maddipudi



Problem we are solving

Lawyers cannot upload case files to ChatGPT (HIPAA violations).

Doctors cannot use AI for patient records.

Privacy regulations block them.

Anyone with sensitive documents is stuck choosing between privacy and productivity.

Tech stack

llama.cpp, ChromaDB, Streamlit, Python, Sentence Transformers

Proof of work :

In progress