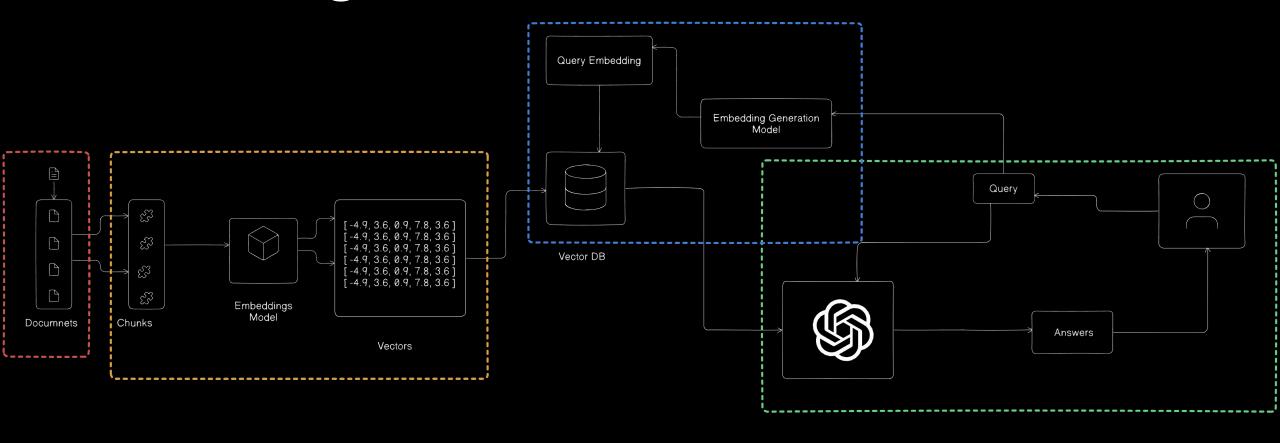
Retrieval Augmented Generation



RAG Working Diagram

PDF Processing and Query System



User initiates the process by uploading PDF files.

PDFs Stored Permanently

The backend server stores the PDFs for future use.

Vector Embeddings Generated

Vector embeddings are created using HuggingFace models.

User Submits Query

User enters a query through the frontend interface.

Chunks Passed to LLM

The retrieved chunks are sent to an LLM for processing.

Answer Displayed in Chat UI

The generated answer is displayed to the user.

























PDFs Sent to Backend

The uploaded PDFs are transmitted to the backend server.

Text Splitter Breaks Content

The PDF content is divided into smaller chunks.

Embeddings Stored in ChromaDB

The embeddings are stored in a persistent vector store.

Relevant Chunks Retrieved

The system retrieves the most relevant chunks from ChromaDB.

Context-Aware Answer Generated

The LLM generates an answer based on the context.

Chat History Stored

The conversation history is saved for future reference.

RAG-based PDF Chatbot Cycle

