Enterprise RAG System - Design Notes

Purpose:

This document explains why the system uses chunking, overlapping windows, LanceDB, local LLMs, and confidence-based abstention.

Chunking and Overlap:

Documents are split into chunks to fit LLM context limits and to ensure efficient retrieval. Overlap is used so that important context that spans boundaries is not lost. This significantly improves retrieval grounding.

Vector Database (LanceDB):

LanceDB is used as the vector store because it supports fast similarity search, metadata filtering, and versioned document storage. Each chunk is stored with metadata like doc_key, version, and (for PDFs) page number.

Local LLM via Ollama:

Instead of relying on paid APIs, the system uses a locally hosted large language model via Ollama. This allows offline execution, improved privacy, and zero inference cost.

Confidence and Abstention:

The system does not always answer. It calculates top1 similarity, average top-K similarity, and top1 minus top2 margin to estimate retrieval confidence. If confidence is low, the system abstains to prevent hallucinations.

Intended Usage:

This RAG system is meant to answer technical and architectural questions about enterprise systems, documents, and internal knowledge bases, while ensuring answers stay grounded in actual content.