RAG Foundations – Goals

- Understand why RAG beats LLM-only for enterprise data.
- Learn chunking, embeddings, vector search, and prompt assembly.
- Hands-on: build a PDF/Text Q&A over a local vector DB (Chroma).

Naive RAG Pipeline

- 1) Ingest: chunk → embed → store in vector DB.
- 2) Query: embed query → retrieve top-k → assemble context.
- 3) Generate: prompt LLM with question + context → answer + citations.
- Key controls: chunk size/overlap, k, re-ranking, token budget.

Demo Architecture

- Local embeddings: sentence-transformers (all-MiniLM-L6-v2).
- Vector store: Chroma persistent volume.
- LLM: your self-hosted GPT/Azure OpenAI (chat/completions).

Evaluation & Guardrails

- Grounding: answer only from retrieved context.
- Citations: show source file and distance for transparency.
- Latency/cost: tune chunking, k, and model.