Enterprise RAG Platform Overview

The Enterprise RAG platform lets teams upload confidential documents, automatically chunk them with
sentence-aware windows, store embeddings via pgvector, and serve a single FastAPI endpoint that answers questions with grounded citations.

Idempotent ingestion ensures re-uploads are skipped by hashing the PDF bytes, while OCR fallback (pdf2image + pytesseract) extracts text from scans when needed.

Prompt versioning lives in the database so operators can promote new system prompts without redeploying, and token usage is logged per query for cost tracking.

The REST API exposes /documents, /query, /metrics, /query-logs, and prompt management endpoints,
accompanied by a Vue dashboard for uploads, chat, and observability.