RAG Evaluation & Hallucination Control (Day 15–16)

This document summarizes the design and implementation of a production-style Retrieval Augmented Generation (RAG) evaluation system built using FastAPI, FAISS, and a local LLM (Ollama). The focus is on correctness, faithfulness, and trust.

Day 15 – Core Evaluation Pipeline

• Implemented retrieval evaluation based on keyword overlap and source matching.
• Added faithfulness evaluation to detect hallucinations by measuring overlap between LLM answers and retrieved context.
• Enforced safe refusal behavior: if no relevant context is retrieved, the system responds with 'I don't know based on the provided context.'
• Secured evaluation endpoints using JWT-based authentication.

Day 16 – Document-Aware Evaluation

A major limitation of naive RAG evaluation is unfair failure when relevant documents are not available for a user. To solve this, document-aware evaluation was introduced.

• Each evaluation question declares required document sources (e.g., 'fastapi', 'rag').
• Evaluation is skipped when required documents are not present for the user.
• Skipped questions are explicitly reported instead of being marked as failures.
• This ensures evaluation metrics remain honest, explainable, and defensible.

Key Engineering Outcomes

• Prevented hallucinations by enforcing context-grounded answers only.
• Produced explainable evaluation metrics suitable for dashboards and audits.
• Designed a system that knows when not to answer, a critical property of trustworthy AI systems.

This project demonstrates production-grade thinking in AI backend engineering, covering evaluation, safety, observability, and user-scoped behavior.