Home Task: Threat Hunting RAG System

Background

Build a Retrieval-Augmented Generation (RAG) system that can hunt for phishing threats across an email dataset. This system should demonstrate your ability to work with embeddings, vector databases, and large language models.

Task Requirements

Core Functionality

Build a threat hunting chatbot that can:

1. **Data Preparation**
   - Generate a synthetic dataset of 100+ emails using AI agents or Faker library
   - Include mix of legitimate and phishing emails
   - Extract and structure relevant metadata (sender, subject, body, timestamps)
   - Generate embeddings for semantic search
2. **Intelligent Query Processing**
   - Support natural language queries like:
     - "Show me emails with urgent payment requests from new senders"
     - "Find emails with suspicious attachment names"
     - "Identify emails that impersonate executives"
   - Implement both keyword and semantic search capabilities
3. **Threat Analysis & Reasoning**
   - Return ranked results with confidence scores
   - Provide clear explanations for why each email was flagged
   - Support iterative refinement of searches based on findings

Technical Requirements

1. **Architecture Design**
   - Create a Mermaid graph showing your RAG pipeline architecture
2. **Implementation**
   - Working code that processes queries and returns results
   - Ensure query response time is reasonable
   - Include at least 10 example queries demonstrating the system's capabilities

Deliverables

1. **Code Repository** containing:
    - Complete implementation with clear README
    - Requirements.txt / package.json
    - Sample .env file for API keys
    - Dataset generation script
    - Mermaid architecture diagram
    - Examples of queries and their outputs

Submission

- GitHub repository (public or provide access)