Assignment: RAG Pipeline for Job Data Retrieval

Dataset:

Link: LF Jobs

- Total jobs: 1000
- Unique companies: 145
- Unique job categories: 7
- 9 Columns and 1000 rows

Dataset Overview

Feature Name	Description
ID	Unique identifier for each job listing (format: LF####)
Job Category	Job category classification
Job Title	Job title/position name
Company Name	Name of the hiring company
Publication Date	Date when the job was posted (ISO 8601 format)
Job Location	Geographic location(s) where the job is basedBuild a classification and clustering algorithm to predict the diabetic class/cluster for a given dataset.
Job Level	Experience level required (e.g., Mid Level, Senior Level, Internship)
Tags	Additional tags for the job
Job Description	Full job description including responsibilities, requirements, and benefits (HTML formatted)

Requirements:

1. **RAG Pipeline Components**
    - **Preprocessing**: Clean and chunk job descriptions.
    - **Embeddings**: Generate embeddings (e.g., via Gemini, Cohere, or Hugging Face).
    - **Vector Store**: Store embeddings in Vector Database.
    - **Retriever**: Fetch top-N relevant chunks based on queries.
    - **LLM Integration**: Combine query + retrieved chunks to generate enriched responses.
2. **API Endpoint**
    - **Tech Stack:** Python (FastAPI preferred)
    - **Endpoint:** POST api/query

Expectations:

1. Fully functional RAG that returns relevant job listings for queries.
2. Modular, well-organized, and clearly documented code.
3. Thoughtful prompt design for accurate and concise LLM output.

Optional Enhancements:

1. **Hybrid Search** (combine vector + keyword search for better precision).
2. **Reranker Model** (use a cross-encoder or other reranking approach to improve retrieved result order).

Documentation:

Include a **Document (Google Docs)** that covers:

1. High-level architecture and engineering decisions and reasoning behind each decision.
2. Setup and installation instructions.
3. Example usage: requests & expected responses.
4. Any assumptions made during development.
5. Drawbacks and future enhancements.

Submission:

1. Please upload your complete codebase and all related files including Documentation Report to a GitHub repository. Share the link to the repository upon completion.