Mandatory Explanations – RAG System

1. Chunk Size Selection

A chunk size of **300–500 characters** was chosen to balance retrieval accuracy and context completeness. Smaller chunks improve semantic similarity and reduce topic overlap, while also keeping retrieved context within LLM token limits. Chunks smaller than this often lose meaning, and larger chunks reduce retrieval precision.

2. Retrieval Failure Case Observed

A retrieval failure was observed when a single chunk contained multiple related but distinct topics. For example, when asking *“What is RAG?”*, the retrieved chunk also included unrelated FastAPI information. This resulted in a lower similarity score even though the correct explanation was present. Despite this, the LLM generated a correct answer because the relevant context was available.

3. Metric Tracked

Latency: Measures the total time taken from receiving a user question to returning an answer. This metric helps evaluate system responsiveness and user experience.

Similarity Score: Tracks semantic similarity between the user query and retrieved document chunks, helping analyze and tune retrieval quality.