**Mandeep Singh Bhar – 8989367**

**The presentation Question Response of rag code and Llama trigger - Team 5**

A white background with blue text

AI-generated content may be incorrect.

This is the numeric cutoff we compare the top similarity score against. Scores **≥ 0.30** are treated as strong; scores **< 0.30** are treated as medium/low and will route to the LLM for general queries.

A screenshot of a computer code

AI-generated content may be incorrect.

We compute a **similarity score** for the user’s query against our Faq.

If the top score is **below 0.30** (our “medium/low” band) **and** the intent is *general*, we **trigger the LLM** via call\_llm(user\_text). If the score is high, we return the answer directly from our knowledge store.

A screen shot of a computer code

AI-generated content may be incorrect.

This is the generation step. When the similarity is in the medium/low band, we call our LLM with the user’s query to produce the answer.

A close-up of a computer code

AI-generated content may be incorrect.

We use **meta-llama/llama-3-70b-instruct** through the OpenRouter API. This is what actually runs when the decision gate routes to the LLM on medium/low similarity.

**Concluion**

1. We calculate a similarity score for the query in the retrieval step of our RAG pipeline.
2. The score is compared against a fixed cutoff: SIMILARITY\_THRESHOLD = 0.30.
3. If the score is **below 0.30** and the intent is “general,” we the **generation** step of RAG by sending the query to the LLM.
4. The LLM used is **meta-llama/llama-3-70b-instruct**, invoked via call\_llm(user\_text) → \_post\_openrouter(...).
5. If the score is **0.30 or higher**, the answer comes directly from the retrieved content, skipping LLM generation.