LATENCY ANALYSIS

Step 1: OpenAI Embedding Generation — Estimated Time: ~0.5 – 2 seconds

- When a user asks a question (e.g., *"Who is Jamie?"*), the system first converts that text into a vector representation (embedding).
- This is done by sending a request to **OpenAI's API**, specifically using the `"text-embedding-3-small"` model.
- The request travels from your server to OpenAI's servers (likely in the US), is processed, and the embedding is returned.
- This API call introduces:
    - Network travel time.
    - OpenAI's internal processing time.
    - Network return time.

Note: Even with a fast connection, this step typically takes between **0.5 – 2 seconds**.

Step 2: Qdrant Vector Search (Remote) — Estimated Time: ~1 – 2 seconds

- Once the embedding is generated, your system performs a **similarity search** against the knowledge base stored in **Qdrant**.
- Your current setup uses **Qdrant Cloud**, meaning this involves:
    - Sending the embedding to the Qdrant Cloud API.
    - Qdrant searching its vector index for the most relevant knowledge chunks.

    - Receiving the matching documents and similarity scores back.
- Even if Qdrant is efficient, network travel time adds to the delay.
- Depending on:
    - Region proximity between your backend and Qdrant.
    - Qdrant search configuration (e.g., `k` values).
    - Overall load on Qdrant Cloud.

The typical time for this search step is around **1 to 2 seconds**.

Step 3: Post-Processing — Estimated Time: ~0.5 seconds

- After receiving results from Qdrant:
    - The system filters out duplicate results (using your `_seen_results` logic).
    - It formats the chunks into readable context for the LLM to generate a reply.
    - It logs information like latency and source IDs.
- This string manipulation and light filtering adds minimal overhead, typically around **0.5 seconds**.

Total Estimated Time per Query: ~2 to 5 seconds

Step		Estimated Time
OpenAI Embedding Generation	----------	~0.5 to 2 sec
Qdrant Vector Search	----------	~1 to 2 sec
Post-processing	----------	~0.5 sec
Total	----------	**~2 to 5 sec**