Mini RAG System: YouTube Transcript Retrieval-Augmented Generation

Fitria Zusni Farida

Objectives

Retrieval-Augmented Generation (RAG) combines LLMs with an external knowledge source (e.g., YouTube transcripts), improving both factual correctness and domain relevance.

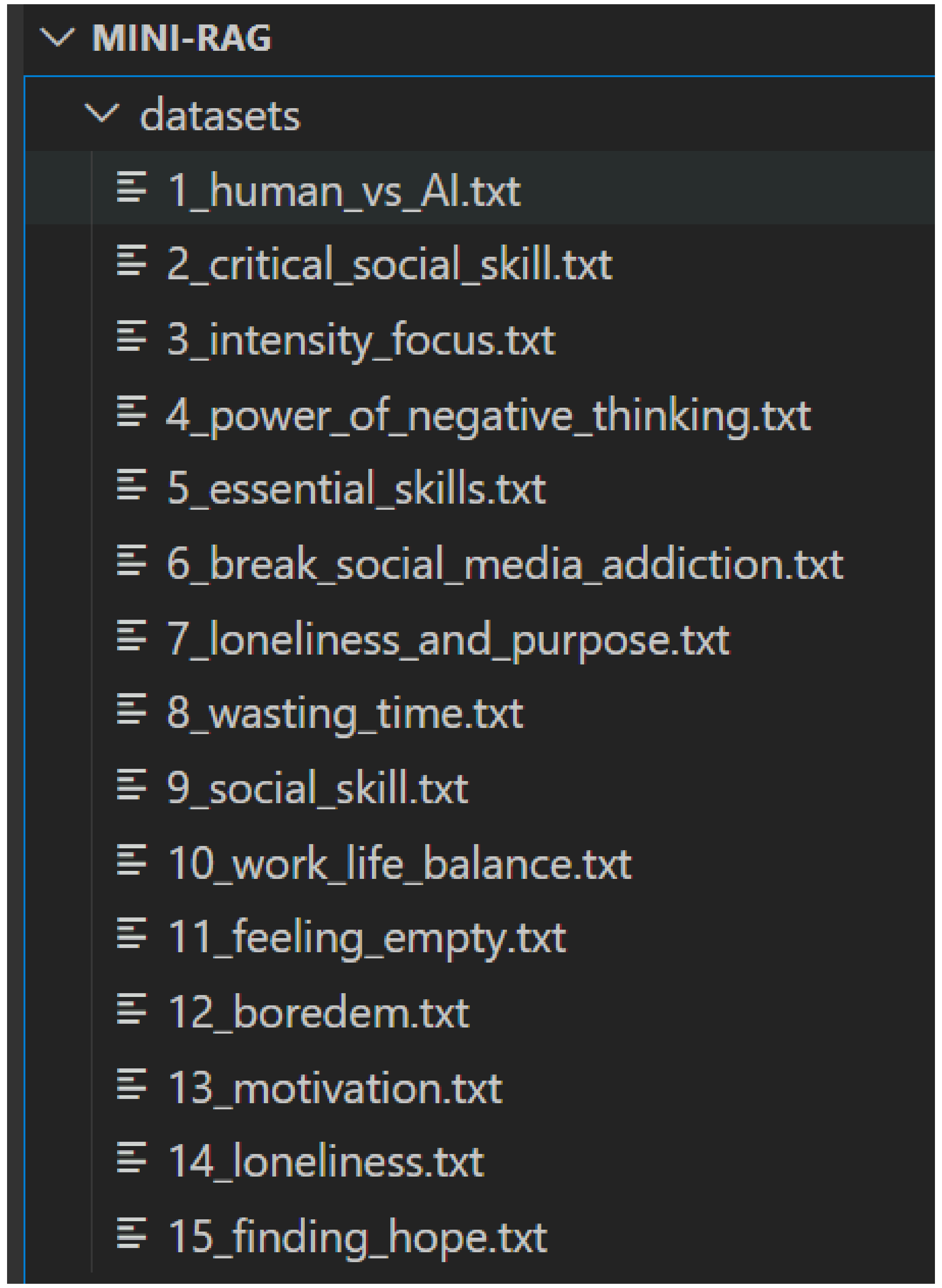
This project aims to simulate a mini RAG system that:

- Uses preprocessed YouTube transcripts as the knowledge base.
- Performs **semantic retrieval** using FAISS and Sentence Transformers.
- **Generates responses** using either a local GPT-2 model or the GPT-4 API.
- **Benchmarks and compares performance** using different encoder model (MiniLM).

Dataset

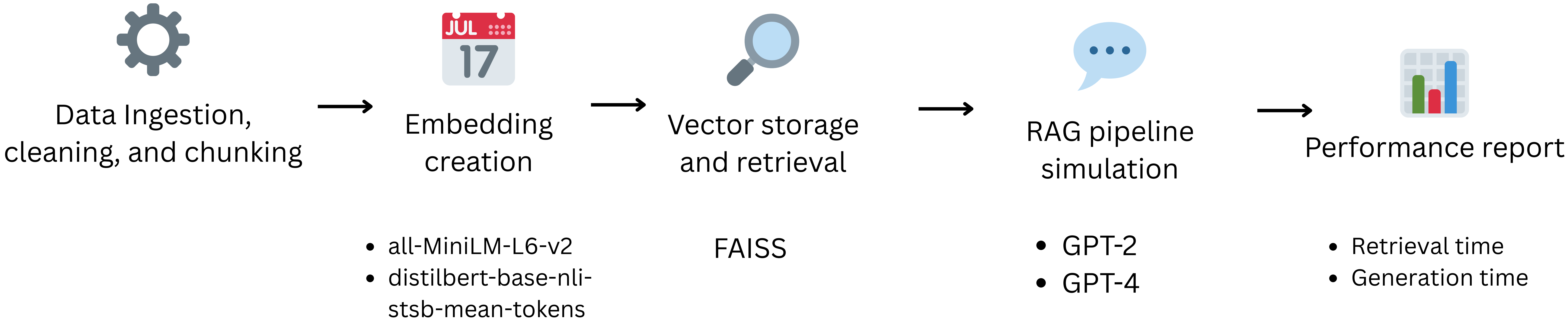
The dataset is YouTube 15 video's transcription taken from this youtube channel. The transcriptions are generated using this online generator.

- The video's length is around 4-10 minutes.
- The generated transcriptions don't contain punctuations, stop words, and special characters.



```
MINI-RAG
  datasets
    1_human_vs_AI.txt
    2_critical_social_skill.txt
    3_intensity_focus.txt
    4_power_of_negative_thinking.txt
    5_essential_skills.txt
    6_break_social_media_addiction.txt
    7_loneliness_and_purpose.txt
    8_wasting_time.txt
    9_social_skill.txt
    10_work_life_balance.txt
    11_feeling_empty.txt
    12_boredem.txt
    13_motivation.txt
    14_loneliness.txt
    15_finding_hope.txt
```

Methodology



Data Ingestion, cleaning, and chunking →

Embedding creation →

- all-MiniLM-L6-v2
- distilbert-base-nli-stsb-mean-tokens

Vector storage and retrieval →

FAISS

RAG pipeline simulation →

- GPT-2
- GPT-4

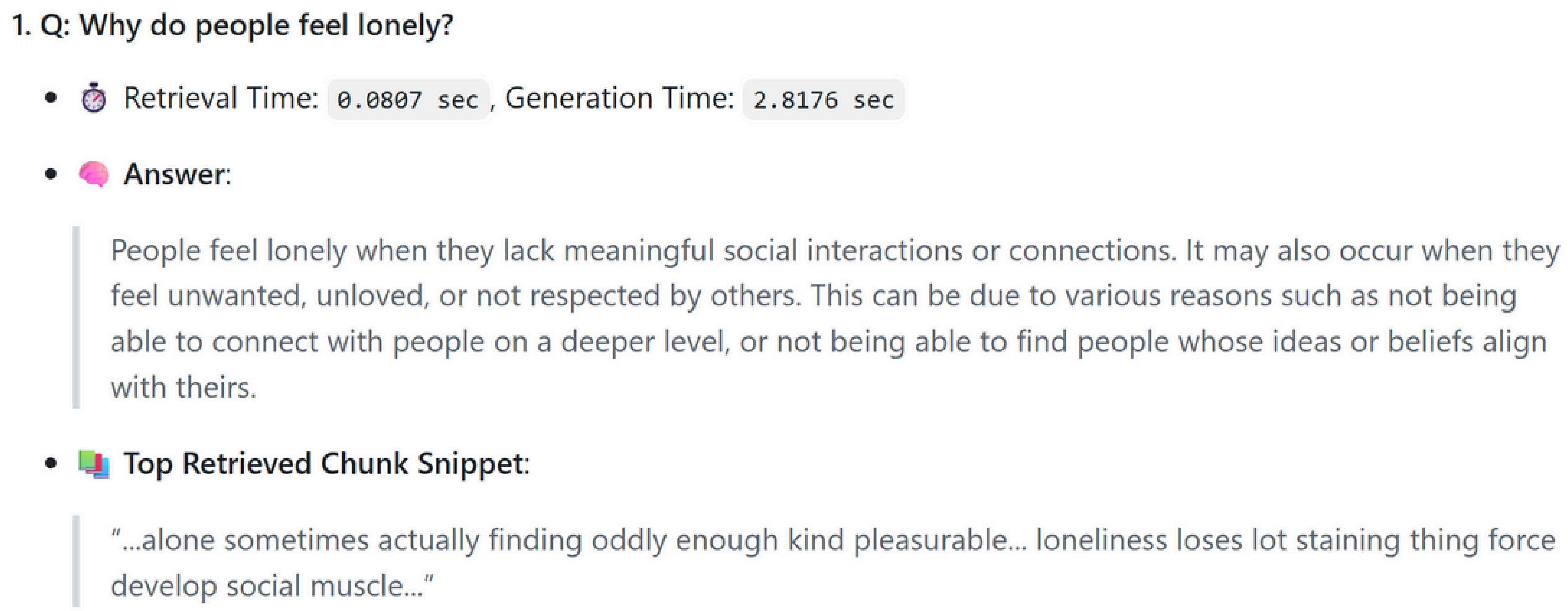
Performance report

- Retrieval time
- Generation time

Results

Embedding model: MiniLM

Generation model: GPT-4

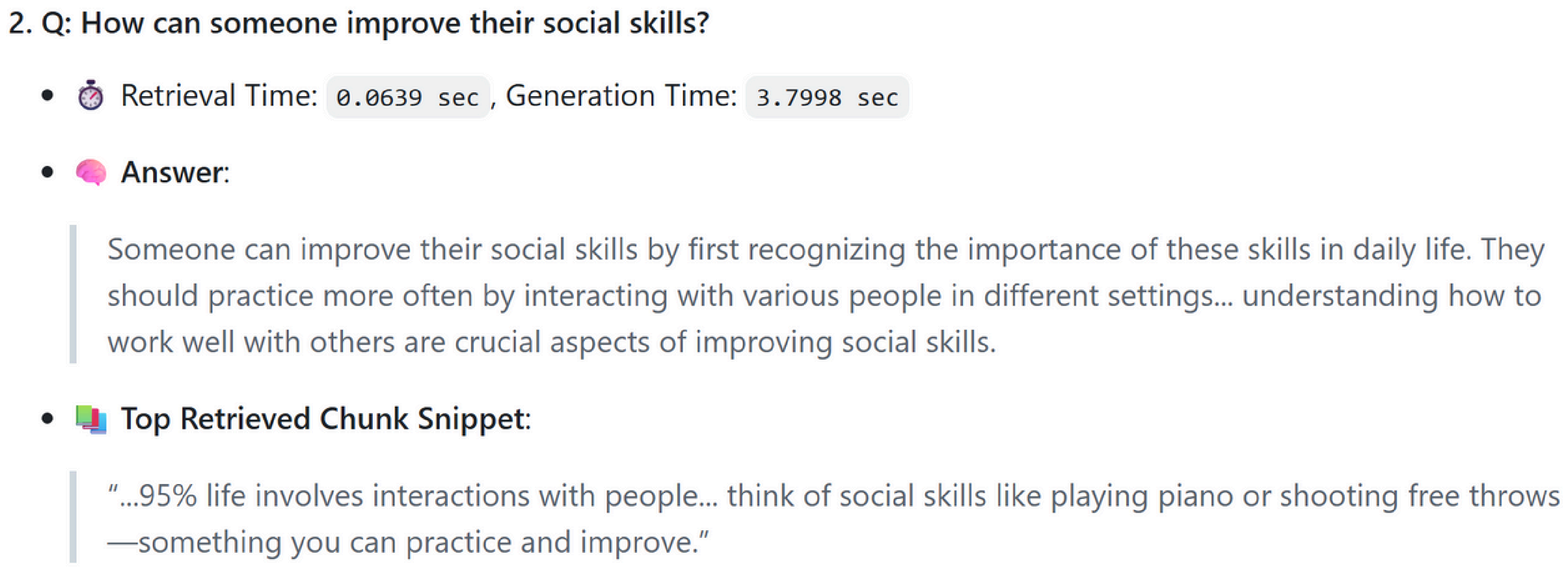


1. **Q: Why do people feel lonely?**
   - ⏱️ Retrieval Time: `0.0807 sec`, Generation Time: `2.8176 sec`
   - 🧠 **Answer**:

     > People feel lonely when they lack meaningful social interactions or connections. It may also occur when they feel unwanted, unloved, or not respected by others. This can be due to various reasons such as not being able to connect with people on a deeper level, or not being able to find people whose ideas or beliefs align with theirs.

   - 📚 **Top Retrieved Chunk Snippet**:

     > "...alone sometimes actually finding oddly enough kind pleasurable... loneliness loses lot staining thing force develop social muscle..."

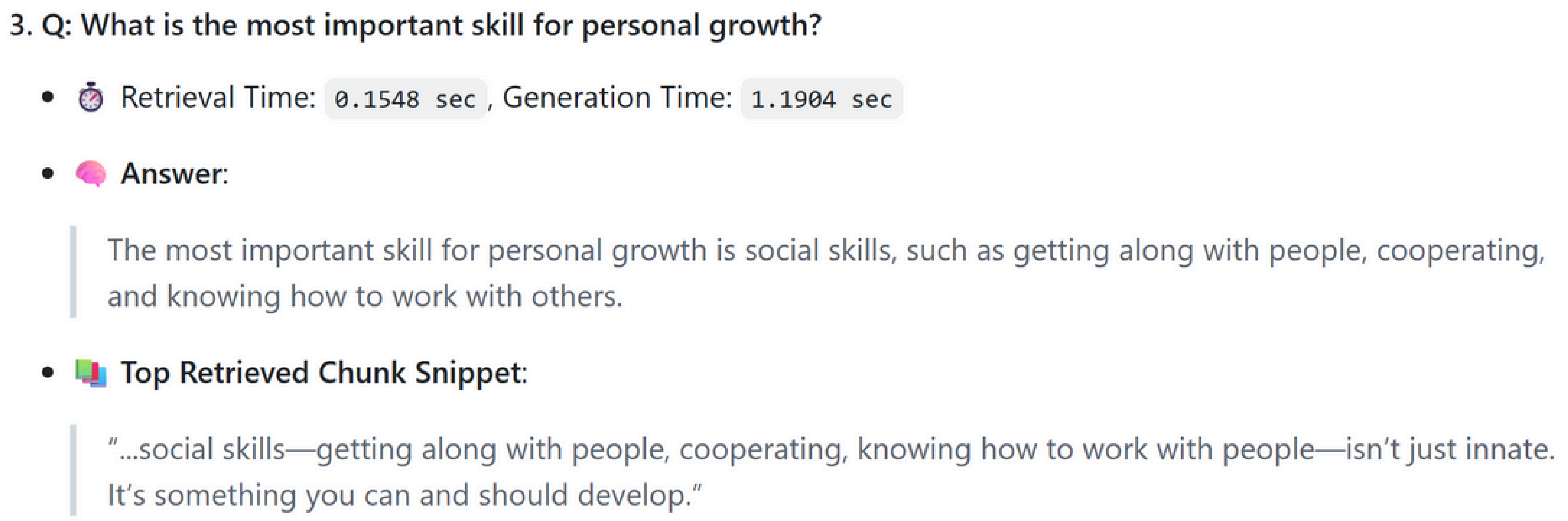


2. **Q: How can someone improve their social skills?**
   - ⏱️ Retrieval Time: `0.0639 sec`, Generation Time: `3.7998 sec`
   - 🧠 **Answer**:

     > Someone can improve their social skills by first recognizing the importance of these skills in daily life. They should practice more often by interacting with various people in different settings... understanding how to work well with others are crucial aspects of improving social skills.

   - 📚 **Top Retrieved Chunk Snippet**:

     > "...95% life involves interactions with people... think of social skills like playing piano or shooting free throws —something you can practice and improve."



3. **Q: What is the most important skill for personal growth?**
   - ⏱️ Retrieval Time: `0.1548 sec`, Generation Time: `1.1904 sec`
   - 🧠 **Answer**:

     > The most important skill for personal growth is social skills, such as getting along with people, cooperating, and knowing how to work with others.

   - 📚 **Top Retrieved Chunk Snippet**:

     > "...social skills—getting along with people, cooperating, knowing how to work with people—isn't just innate. It's something you can and should develop."

Conclusion

Insights:

- GPT-4 with retrieved context shows coherence and deeper understanding across questions compared with GPT-2.
- Retrieval latency remained low (mostly under 0.2s) using FAISS with sentence embeddings (all-MiniLM-L6-v2).
- Using overlapping token windows (e.g., 75 tokens with 50 overlap) preserved semantic continuity across chunks and improved retrieval relevance.

Challenges:

- Chunking the texts without knowing the expressiveness of the text.
- Selecting which pretrained model is suitable for the given task and datasets.
- Measuring the quality of the generated response without human involvement.

Recommendation:

- Switch to faster embedding models.
- Replace GPT-2 with a stronger local model (e.g., mistral or gemma) for offline.
- Implement contextual memory for handling multi-turn conversations more robustly.
- Improve banchmarking method to evaluate the quality of the response.