RAG DATA CHALLENGE

BY ROY KREMER

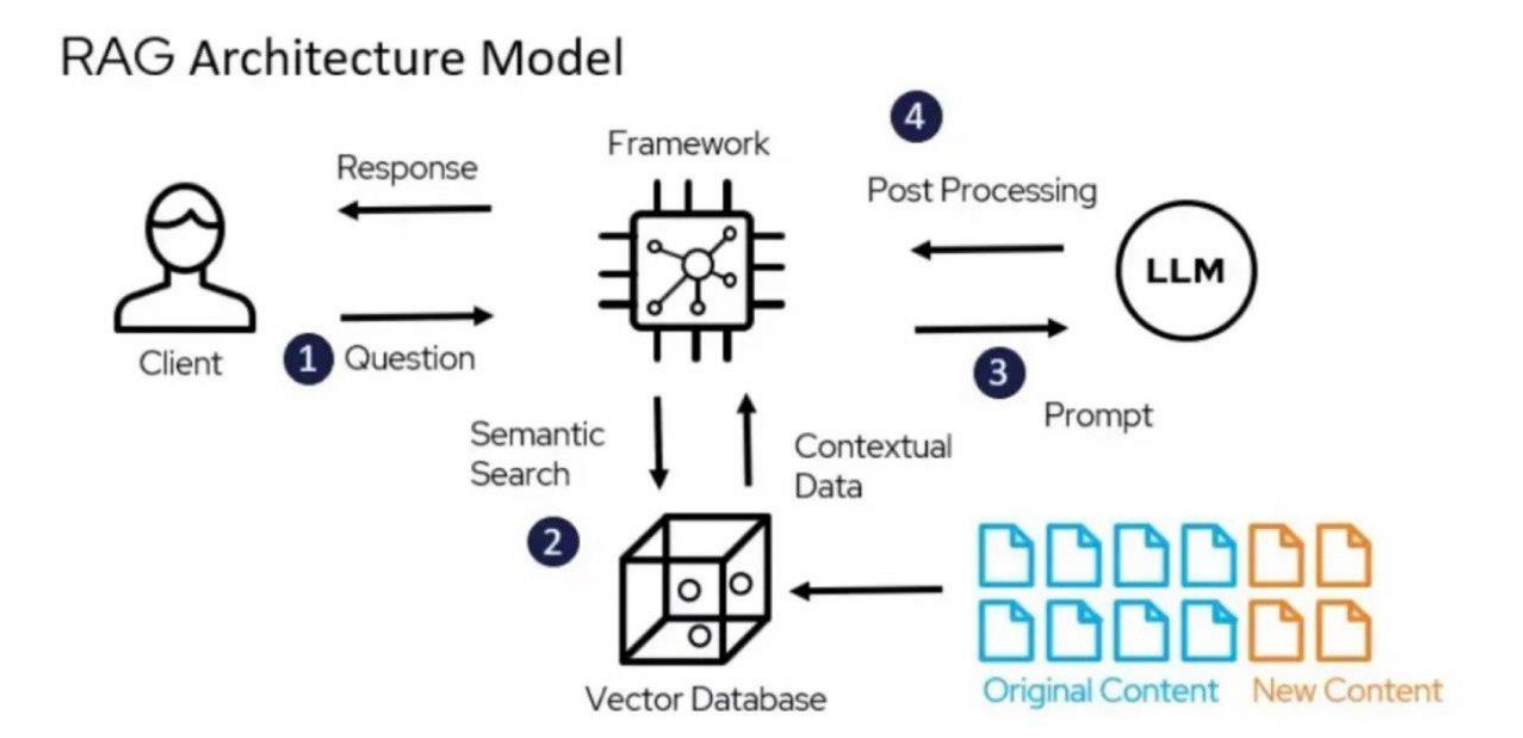
I received a data challenge with a two-day deadline, and the main tasks were:

1)create a Retrieval Augmented Generation (RAG) pipeline that can answer questions about verdicts.

2)evaluate its performance.

INTRODUCING TO RAG

- RAG Explained (8-minute video):
  RAG Explained (youtube.com)



CREATING RAG PIPELINE

- After searching the internet, I chose to use the Haystack library. I followed a guide that is suited to the size of my project and my level of experience in the field.
- loaded these two databases and stored them in a vector database:

1. About 1,000 US-SupremeCourtVerdicts: https://huggingface.co/datasets/macadeliccc/US-SupremeCourtVerdicts
2. About 500k U.S. states supreme court verdicts (1845-2024): https://huggingface.co/datasets/macadeliccc/US-SupremeCourtVerdicts

- connected to the OpenAI API and set the temperature to get less random results.

CREATING PIPELINE

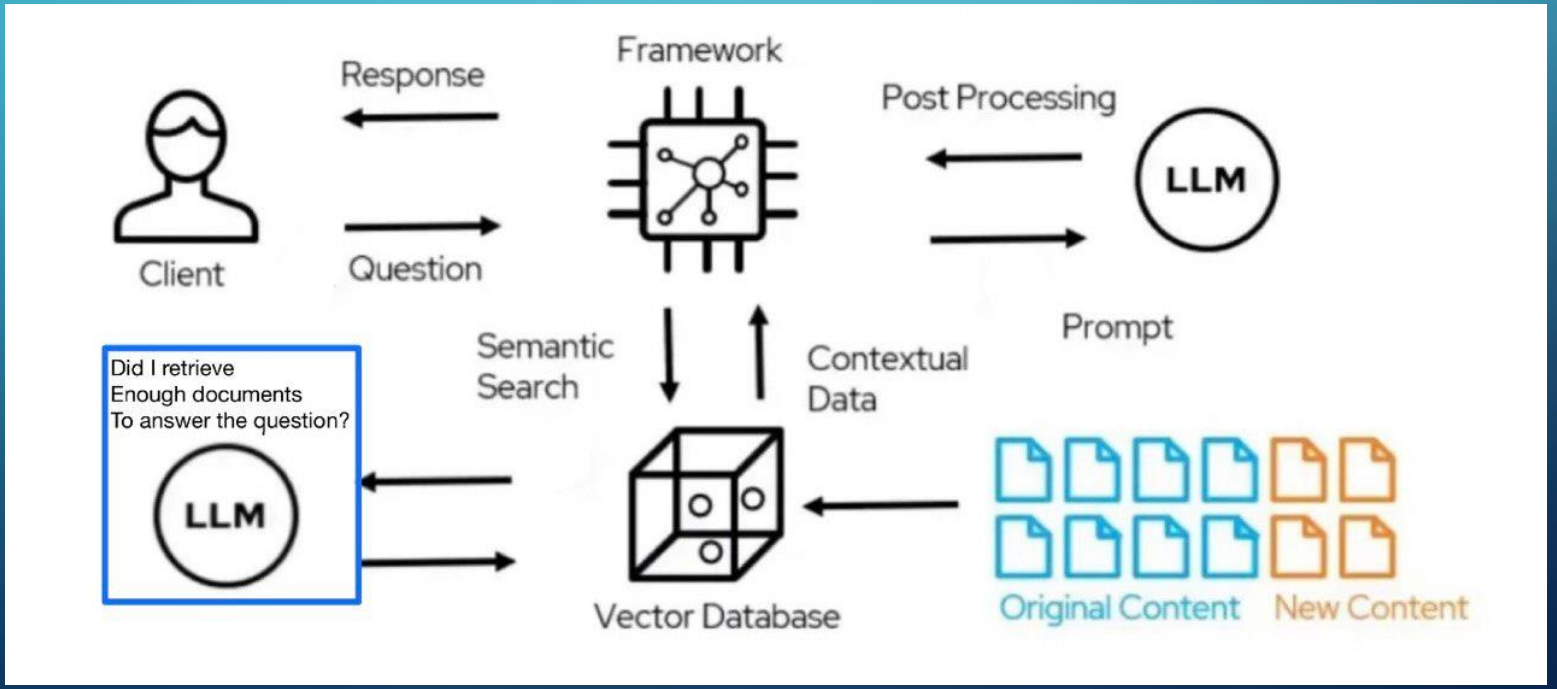
- After I managed to run the program, I encountered the first issue: when I sent all the documents, I downloaded to the LLM, the API refused to accept prompts of that size.

PROBLEM :

- After overcoming this problem and managing to retrieve the top K most relevant documents, I wondered how many documents I should pass for a query to provide enough information to answer the question while not overwhelming it with too much information (i.e., finding the optimal K for each query). I thought that for 'easy' queries, it wouldn't need much information, whereas for more difficult ones, it would need more.

MY SOLUTION

- I solved the problem as follows: Before sending the prompt and asking for an answer from the LLM, I sent the question and the retrieved documents and asked the LLM if it had enough information to answer the question. If the answer was negative, I would retrieve additional documents to ensure that the answer provided was accurate and up-to-date for our database.



PROBLEMS WITH THE PROMPT

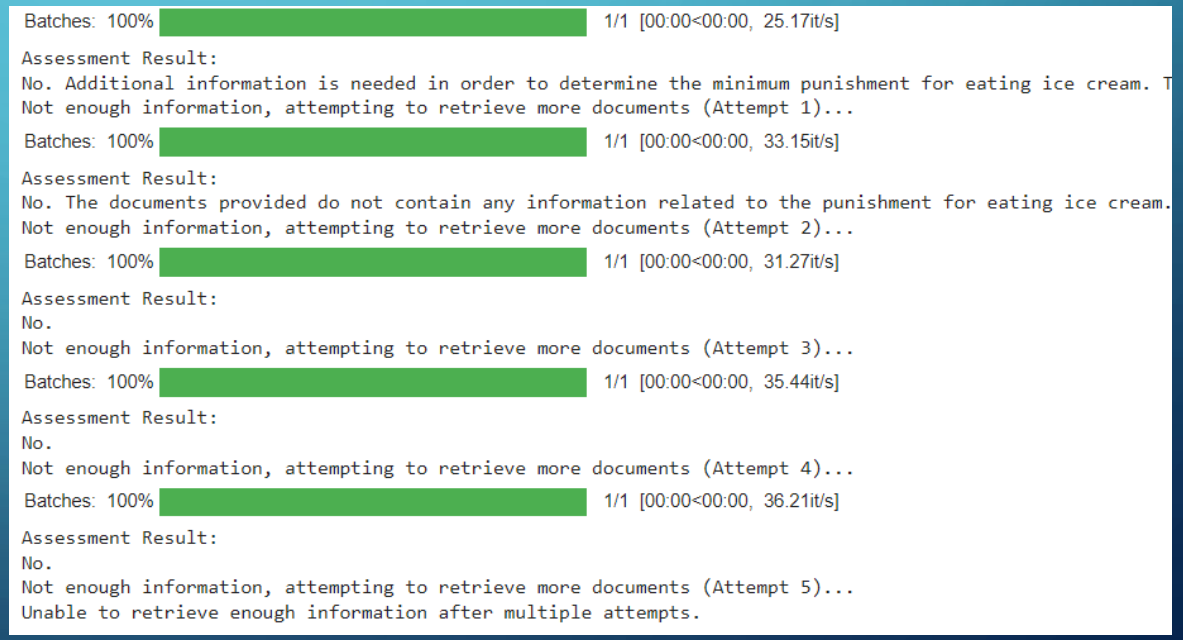
- Initially, I formulated the prompt for the first model inaccurately, and it returned partial answers. This is what it looked like:
- query = "What is the maximum and minimum punishment range for murder?"

Answer:
The maximum punishment for murder varies by state and may include the death penalty or life imprisonment without the possibility of parole. The minimum punishment for murder is typically a term of imprisonment, which also varies by jurisdiction and the circumstances of the crime.

You can see that the first model answered 'yes' and sent the question for analysis. The model was only able to answer half of the question because I wasn't specific enough in the initial prompt, and this issue was corrected.

EVALUATE

- After addressing the prompt issue (which resulted in partial answers), I started to understand whether what I was doing was logical. I asked a question unrelated to the topic ('difficult') and wanted to see if the model would decide whether it had enough information to answer the question
- query = "What is the minimum punishment for eating ice cream?“
- The model did not find enough documents on the topic even after retrieving 13 documents and decided that it could not answer the question given the information.



- These are the metrics for the documents retrieved in relation to the question.
- It can be seen that the similarity of the most similar document is 0.36.

```
# Example usage
query = "What is the minimum punishment for eating ice cream?"
generate_answer(query, top_k=5)
```

```
Batches: 100% 1/1 [00:00<00:00, 25.47it/s]
Similarity Scores for Retrieved Documents:
Batches: 100% 1/1 [00:00<00:00, 10.29it/s]
Document 1:
Similarity Score: 0.36
Batches: 100% 1/1 [00:00<00:00, 8.79it/s]
Document 2:
Similarity Score: 0.35
Batches: 100% 1/1 [00:00<00:00, 28.19it/s]
Document 3:
Similarity Score: 0.34
Batches: 100% 1/1 [00:00<00:00, 25.29it/s]
Document 4:
Similarity Score: 0.34
Batches: 100% 1/1 [00:00<00:00, 8.82it/s]
Document 5:
Similarity Score: 0.33
```

- Then I asked an 'easy' question and wanted to see how many documents the model would agree to use to provide an answer.query = "What is the difference between the punishment of consuming drugs for personal use and selling drugs?"
- ANSWER with 3 docs :
  The punishment for consuming drugs for personal use is typically less severe than the punishment for selling drugs. Consuming drugs for personal use is often considered a misdemeanor, while selling drugs is generally considered a felony punishable by a longer prison sentence.
- We received a good and accurate answer after retrieving 3 documents!

The similarity metrics of the documents to this query are: (the highest being 0.48)

- In this part, I forced it to provide an answer even if it did not find enough documents.

- query = "If I stole but expressed remorse, is there a difference in punishment because I expressed remorse? What was the punishment if I didn't express remorse?"

- Answer with 3 docs

  Based on the facts provided, no information was provided about Perkins expressing remorse for the murder he confessed to committing. Therefore, there is no basis to determine if there would be a difference in punishment based on remorse. The punishment for Perkins confessing to the murder while being held in jail was not specified.

  It was unable to answer even with the 3 most similar documents.

  Answer with 13 docs

  In cases involving theft or other criminal acts, the expression of remorse may be a mitigating factor that could potentially lead to a lesser punishment. However, the actual impact of expressing remorse on the punishment can vary based on the specific circumstances of the case and the discretion of the judge or jury.

- It provided a reasonable answer.

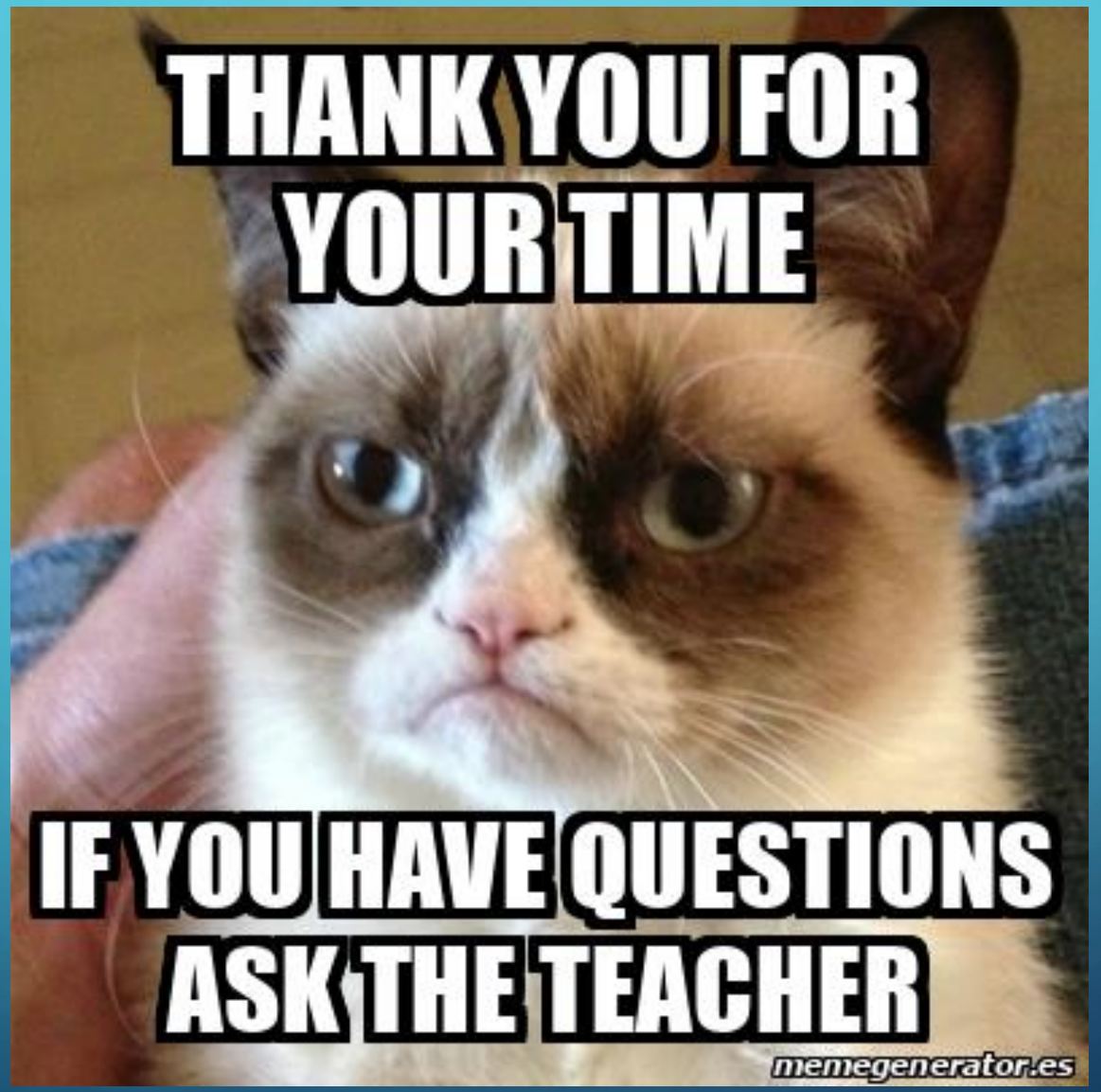
- query = "If I committed murder but repented, is there a difference in punishment because I repented? What was the punishment if I didn't repent?“

- Answer with 13 docs
  Based on the information provided, there does not appear to be a difference in punishment based on repentance after committing murder. The individuals who committed murder and were found guilty were sentenced based on the severity of their crimes and the laws in place, regardless of whether they showed remorse or not. The punishment for murder in the cases mentioned varied from life imprisonment to the death penalty, depending on the circumstances and the laws of the respective states where the crimes were committed. Repentance or lack thereof did not seem to have a direct impact on the severity of the punishment handed down by the courts.

- A reasonable answer. In this case, I tried to retrieve results using cosine similarity instead of BM25. When I attempted to get an answer with cosine similarity, the retrieved documents were too large, and I couldn’t get an answer from the language model due to the length of the prompt (this happened with several queries).

What I would improve if I had more time:

- Evaluate the model's performance more professionally.
- Improve the prompt and provide more precise guidance, possibly by creating specific prompts for each query.
- Test with different language models and compare the responses.
- Find alternative solutions for the issue of the number of documents returned.


THANK YOU FOR YOUR TIME
IF YOU HAVE QUESTIONS ASK THE TEACHER
memegenerator.es