Milestone 3

1. Approach

For this milestone, we used prompt engineering as the model enhancement approach for both models used. We utilized **retrieval-augmented generation (RAG)** to supplement the LLM with external knowledge from a domain-specific dataset. The LLMs we chose were the following:

- **Gemma 2 9B-IT**: An instruction-tuned LLM from Google designed for high-performance conversational tasks.
- **LLaMA 3 8B-8192**: Meta's state-of-the-art 8B parameter model with extended context window, enabling deeper multi-turn context retention.

The chatbot consists of the following components that were created using **langchain**:

- **User Interface Layer:** Accepts user input and displays model output
- **Memory Module:** Maintains conversation history over multiple turns.
- **Retrieval Module:** Uses vector similarity search to retrieve top-k relevant documents (for RAG).
- **LLM Interface:** Handles the API calls to the LLM and formats prompts accordingly.

The documents that were given to the model were extracted from the dataset we have and they were chunked and embedded using **sentence-transformers/ all-MiniLM-L6-v2 embedding model** and added to a vector database the one we used is **FAISS**. On each turn, the top-k relevant chunks are retrieved and passed along with the user query and conversation history to the LLM.

The system prompt that was given to models was:

"You are a helpful assistant."
"You are given a context and a question."
"You should answer the question based on the context."
"If the context does not contain the answer, say 'I don't know'."
"Dont answer the question with a question."
"Be specific and concise."

2. Dataset

We got the dataset from **HuggingFace** **NarrativeQA** which consists of question answer pairs and their context. The scope of the dataset was different movies and e-books.

We took about 1000 random rows and fed their context to the vector database that the LLM looks into and takes the most semantically close context to the query asked of him.

3. Results

For the Results we created a csv file that show the evaluation of both models used, and has the **F1** and **rougeL** scores for both,

The results show good answers for the questions however the evaluation metrics are sometimes inaccurate as it can be tricky to match sentences that are correct to the ground truth, so sometimes even though the answer is correct we get low values for the evaluation metrics.

The results were conducted by testing on 30 random questions from the dataset. The model **Gemma 2 9B-IT** gives better results in the evaluation metrics, but this is only due to the fact that Gemma 2 9B-IT can be very straight forward so its answers get better ranking in the evaluation metrics used. The avg F1 score for llama was 0.1 while it was 0.21 for Gemma, The avg rogueL for llama was 0.161 while for Gemma it was 0.324

This is an example on the chat history working: