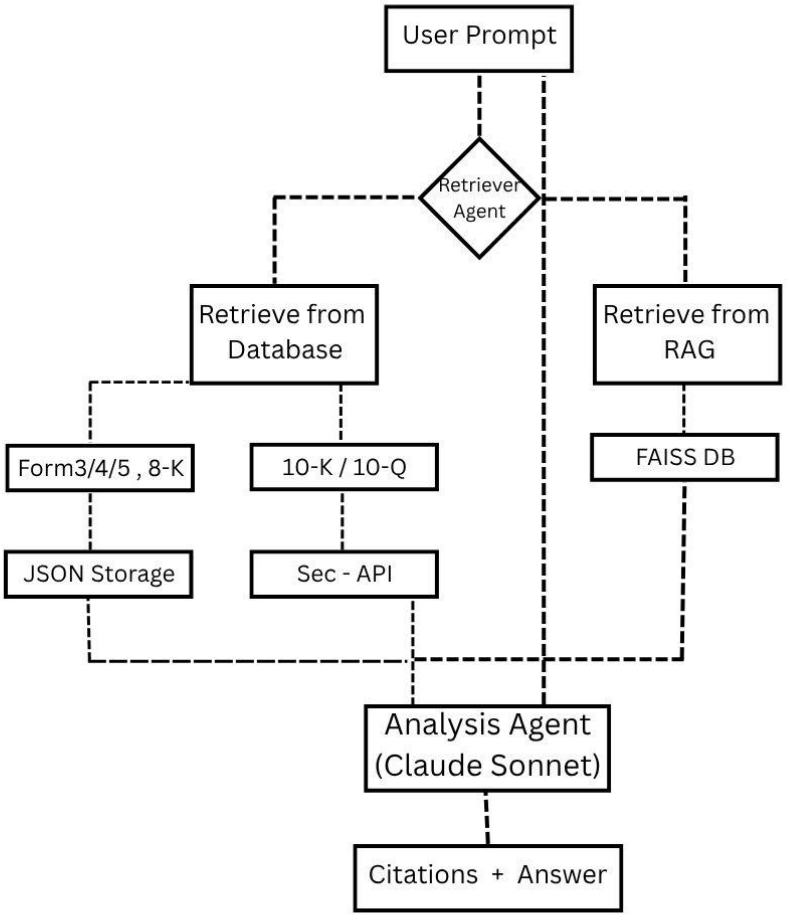
Submission - Sree Snehan

Local System: M1 Max 64 GB Ram



Methodology:

1. Identified the important and standard columns from each of the forms and extracted them into a JSON structure directly from SEC EDGAR website under each section. Eg : Derivative table , non Derivative table and reporting owner among Form 3/4/5.
2. Retriever agent based on Claude Haiku is used to parse the prompt to identify relevant context to be retrieved. It could be based directly on multi dimensional query or semantic based. So the retriever agent decides where the context to bring along with parameters.
3. Since 10-K and 10-Q filings could not parsed into items with automation , sec api was used to retrieve those contents.

4. Individual forms were chunked separately into vectors and stored in FAISS by all-MiniLM-L6-v2 transformer model (about 9.6 GB model size).
5. These were compiled together to form the context for the analysis agent which is handled by CoT Prompting with Citations generations handled by the Claude Sonnet 4.

Key Decision and Intuition:

1. The simplest method to implement citations is by referencing the meta data of chunk , the latest improvements , the **local LongCite 9B** , which is finetuned for this purpose to generate Citations inline with the text was used initially to have the **best citation performance**. But since CoT reasoning and deep analysis was required , the model was not chosed due to the performance limitations. And hence Claude was chosen as it had the best balance between performance and citation generation. It was able to properly cite over multiple CoT interventions without losing quality in Citations , which was observed in LongCite model as a limitation.

https://medium.com/@techsachin/longcite-training-llm-models-to-generate-fine-grained-citations-in-long-context-question-answering-0575cc9ca4b2

https://arxiv.org/pdf/2410.11217

2. Instead of fixed sequential data retriever , since the LLM can read and understand context , the quality and accuracy of retrieved context is improved drastically.
3. There were not a one size fits all parsing method for 10-K and 10-Q FORMS , hence had to resort to sec-api for direct fetching whenever LLM requests. I consulted multiple forms and methods , but was not able to satisfactorily parse all the sections of these forms.

https://highdemandskills.com/using-regular-expressions-to-search-sec-10k-filings/

4. RAG vectoriser was chosen to the largest possible size for enhancing performance.

Limitations and Improvement Areas:

1. Instead of accessing the context once , the loop can be modified to retrieve the context on request of the Analyse agent based on its requirements while analysing in detail.
2. Similar to Sec API , need to identify a standard parsing method to categorise the items in 10-k and 10-q.

3. There are only 3500 forms for 6 companies in total , the scale is limited in this implementation.
4. RAG retrieval performance has to be improved despite having a large model.
5. Local models like LongCITE have to be modified for Deep Analyse use case of ours to have additional benefits and better cost efficiency.