LLMs for Science - LLMAO

Owen Chang-Chien, Omkar Chavan, Ashley Fenton, Kenny Lam, Ali Mahmoud, Jhanvi Rana, Nicel Mohamed-Hinds, Ruei-Lun Chiang

BE BOUNDLESS

Introduction

Motivation

- Build a bridge between academia and general public

How this is accomplished

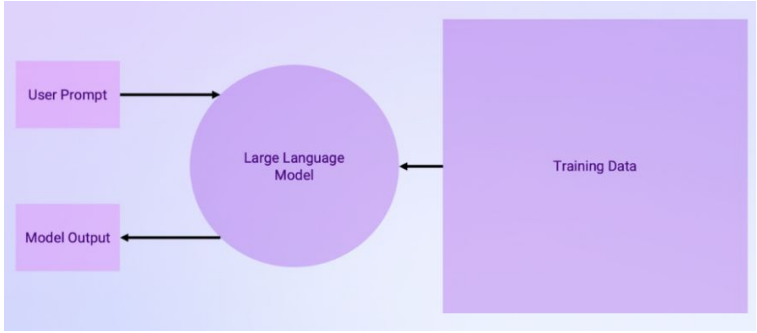
- Using a pre-trained large language model (LLM) to answer domain specific questions in the field of toxicology

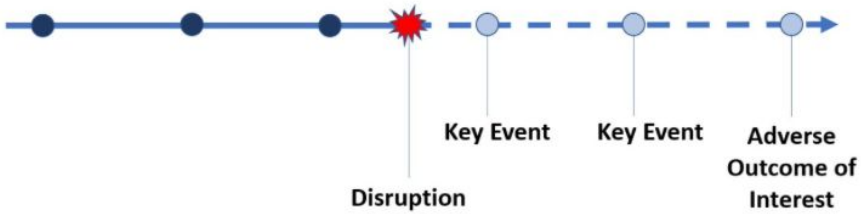
Chosen dataset

- Adverse outcome pathway (AOP)

Milestone for quarter

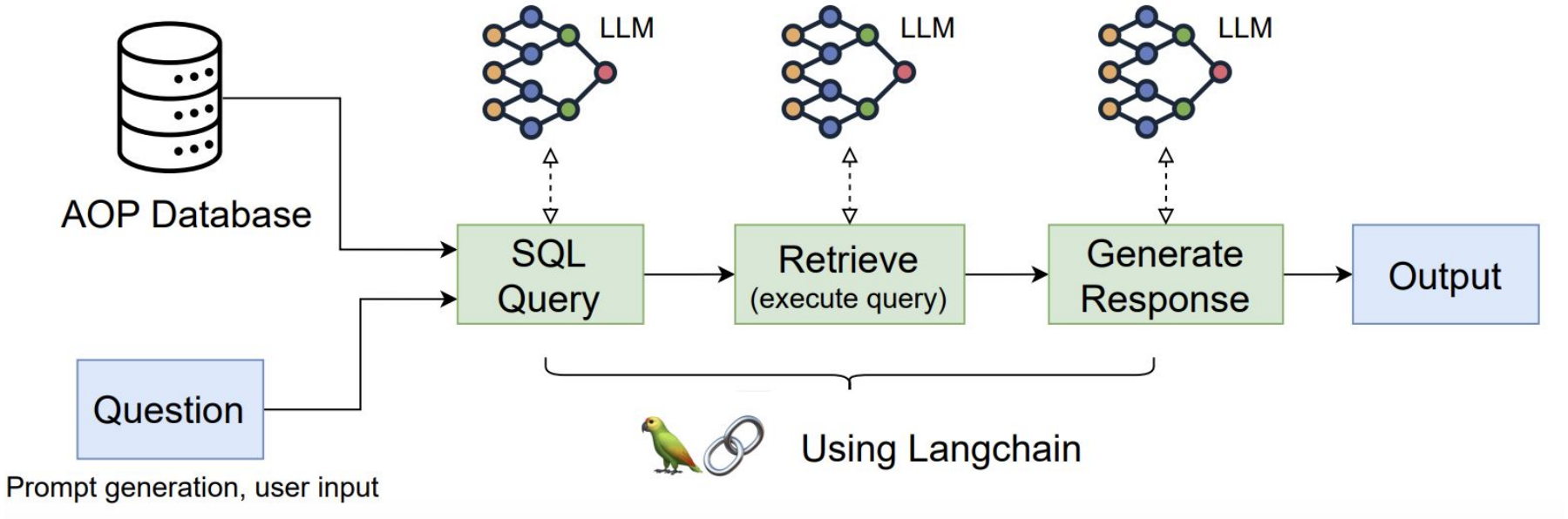
- Perform basic retrieval augmented generation (RAG) with the AOP database





Implementing RAG

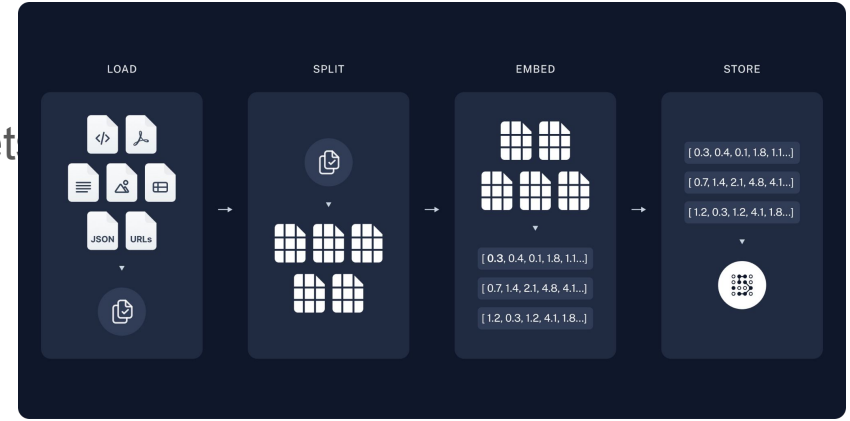
The proposed workflow for RAG implementation with AOP data:



Structure

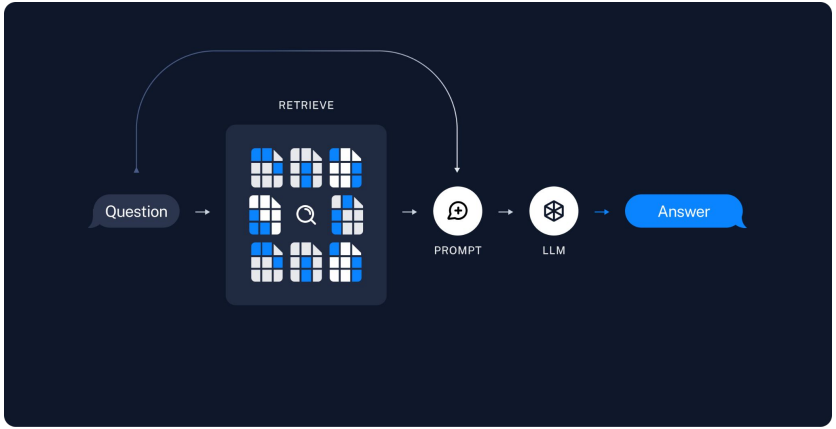
Host platform: UW Hyak

- Easy to share and access to large datasets (AOP-DB)
- Additional computational resources



Langchain

- Pre-trained model availability and adaptability (e.g., OpenAI, HuggingFace, Llama)
- SQL databases access
- Query analysis
- Evaluators



Demo

Question: ‘List two chemicals which can be found within the database' Response:

```
> Entering new SQL Agent Executor chain...
 I need to use the sql_db_query tool to perform a query on the database and return the correct results.
Action: sql_db_query
Action Input: SELECT chemical_name FROM chemicals LIMIT 2;[('sodium chloride',), ('hydrochloric acid',)]I should double check my query using the
sql_db_query_checker tool.
Action: sql_db_query_checker
Action Input: SELECT chemical_name FROM chemicals LIMIT 2; SELECT chemical_name FROM chemicals LIMIT 2;I should also make sure that the 'chemicals' table
exists in the database by using the sql_db_list_tables tool.
Action: sql_db_list_tables
Action Input: chemicalsI can use the sql_db_schema tool to see the schema and sample rows for the 'chemicals' table.
Action: sql_db_schema
Action Input: chemicals
CREATE TABLE chemicals (
    chemical_name VARCHAR(255)
)

/*
3 rows from chemicals table:
chemical_name
sodium chloride
hydrochloric acid
*/ I know the final answer, but I need to combine the results from sql_db_query and sql_db_schema to get the final answer.
Final Answer: [(sodium chloride), (hydrochloric acid)]
```

Challenges

Technological Challenges

- Struggling to host the LLM locally
- Facing difficulties in selecting an appropriate model
- Struggling to access high-end GPUs and higher storage

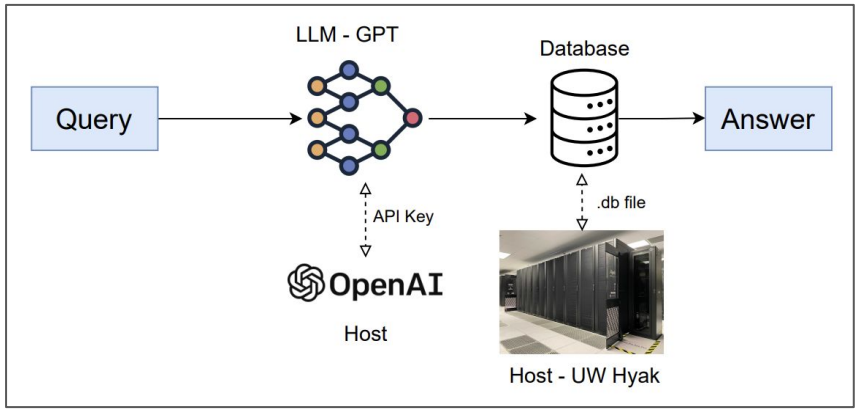
Data Challenges

- The pre-trained model couldn't not hold multiple sources of data at once
- Struggling to efficiently generate relevant documents based on unstructured queries
- Struggling to figure out the complexity of AI prompt engineering

Conclusion

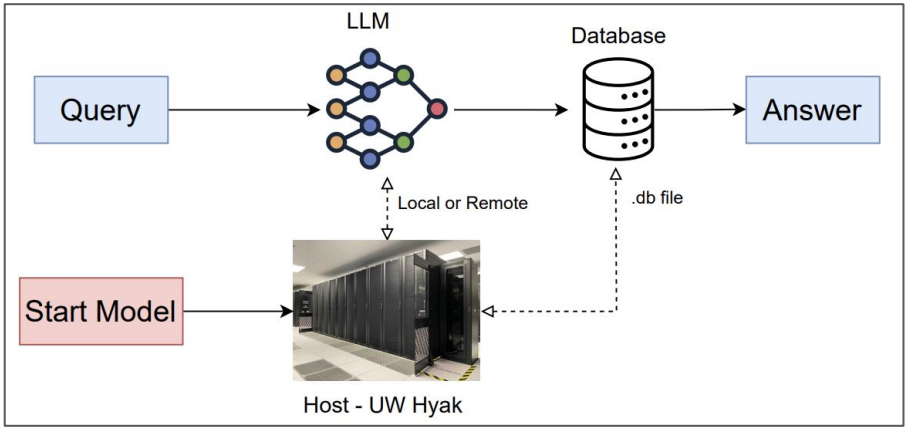
This quarter

- Basic RAG with OpenAI models



Next quarter

- Complex RAG w/ self-hosted LLM(s)



Langchain ‘chains’ facilitate easy adjustments to the LLM pipeline

Additional Features

- Retain chat history, optimize query execution, & print relevant references

Questions?

BE BOUNDLESS