

1

LLMs for Science - LLMAO

Qianqian Cheng, Qianqian Cheng, Ashley Fenton, Henry Lam, Ali Mahdavi, Daniel Rana, Hani Mahdavi, Hani, Ravi Lavi Cheng

BE BOUNDLESS

W

2

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

Diagram showing the flow from Question to Answer, involving LLM, AOP Database, and Retrieval Augmented Generation (RAG).

3

Implementing RAG

The proposed workflow for RAG implementation with AOP data:

Diagram showing the workflow: Question → Prompt generation, user input → SQL Query → Retrieve (Relevant text) → Generate (Response) → Output. The workflow is powered by LLM and Langchain.

4

Structure

Host platforms: 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

Diagram showing the structure of the RAG system, including the AOP Database, LLM, and Langchain.

5

Demo

Question: "List two chemicals which can be found within the database"

Response:

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)]

6

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 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

7

Conclusion

This quarter

- Basic-RAG with OpenAI models

Next quarter

- Complex RAG w/ self-hosted LLMs

Diagram showing the conclusion of the project, comparing the current state (Basic-RAG with OpenAI models) to the next quarter (Complex RAG w/ self-hosted LLMs).

🗨️

Omkar

⏪

⏩