IC08 Leveraging LLM and RAG for Advanced Pharmaceutical Statistical Programming

A Novel Chatbot Framework Approach

Jaime Yan, Merck & Co., Inc.
Chao Su, Merck & Co., Inc.
Changhong Shi, Merck & Co., Inc.

Biography

Jaime Yan:

Jaime Yan is a statistical programmer with Merck Sharp & Dohme. He has seven years of experience working as a statistical programmer in the oncology and cardiovascular therapy area in the pharmaceutical industry. He has used SAS and R and Python for seven years to handle clinical trial data and prepare clinical reports.

Chao Su:

Chao Su has 14-year experience as statistical programmer. He has worked at different pharmaceutical companies and CROs. He is an expert in SAS and R, Currently, he is working in Merck Sharp & Dhome, based in new Jersey.

Changhong Shi:

Changhong Shi been a statistical programmer for 21 years. Before that, she worked as a software developer using Oracle, VB script and Java. Recently she acquired knowledge on R.

Currently, Changhong is a director in statistical programming in Merck Sharp & Dhome, based in new Jersey.

MERCK

Agenda

- Introduction
- Proposed framework
- Future work

MERCK

Abstract the three main tasks

As a statistical programmer, we:

1. *Review the document*
2. *Data explore(check the ADaM data)*
3. *Output the report*

MERCK

Target

Use LLM with Framework like LangChain/llamaIndex to optimize:

1. ***Documents review***
2. ***Data explore(check the ADaM data)***
3. ***Report generate***

MERCK

Former Process

Review the documents:

Contains:

- *CDISC standard(SDTM, ADaM IG)*
- *SAP & Protocol*
- *Company training material*

Self learning or workshop

Data explore(check the ADaM data):

- *Edit check(data quality)*
- *Request from custom(Stat/Clinical etc.)*

Based on requirement to prepare sas/r code. Take 1-2 days

Prepare code&Output the report:

- *Based on analysis data(ADaM datasets) and mock(ICH E3 guidance)*
- *Use company macro(most), user defined sas code*

Taken most of the time of the process

MERCK

Former Work to simplify the process

Data explore(check the ADaM data):

- *Edit check: P21 report*
- *Develop R shiny app*

Review the document:

- *Prepare training material*

Output the report:

- *Transfer from user defined code to standard macro.*
- *Provide interface to select the condition for macro(R shiny etc.)*
- *Make data machine readable(CDISC 360 project)*

Gaps in Using Pre-Trained LLM Directly

***Domain Knowledge**: Pre-trained LLMs often lack the nuanced understanding necessary for pharmaceutical tasks, such as interpreting complex regulatory standards.*

***Security Concerns**: Processing confidential data with LLMs poses risks, necessitating strict adherence to data protection laws like HIPAA.*

***Industry Optimization**: General-purpose LLMs may not meet the specific needs of pharmaceutical statistical programming, affecting performance and compliance.*

Exploring Potential Solutions

Helpful method:

Prompt engineering, let LLM better know user's request, not give extra info beside the LLM original training data.

Feasible method:

- *Advanced Retrieval-Augmented Generation (RAG): Improves LLM's contextual understanding, enabling precise information retrieval and response generation.(Doc and analysis)*
- *Open source LLM, deploy in server/cloud or run in local(security concern)*
- *SAS Macro Integration: Fine-tuning LLMs for seamless interaction with SAS macro libraries, streamlining the TLF generation process.(reporting)*

Why LLM + RAG + Fine tunning + Code interpreter could work

LLM:

- *LLM naturally good at process/organize and do summary for text.*
- *LLM has the ability to abstract the feature/rule/relationship by fine tuning.*

RAG:

- *Provide domain knowledge based on user prepared document to LLM.*
- *Would work as 'RAG code Model' by prepareing the train data to LLM.*

Fine tuning:

- *Link the private sas mcro library to LLM, go further, make anlysis and report process automation.*

Code interpreter:

- *Deal/run with the LLM provided code in desired local environment.*

Potential LLM solution

One Chatbot empowered by LLM:

Review the document:

- *RAG(optimized for pdf and support the other format)*

Outcome: Search key point and get summary by communicating with file

Data explore(check the ADaM data):

- *RAG + SAS kernel*

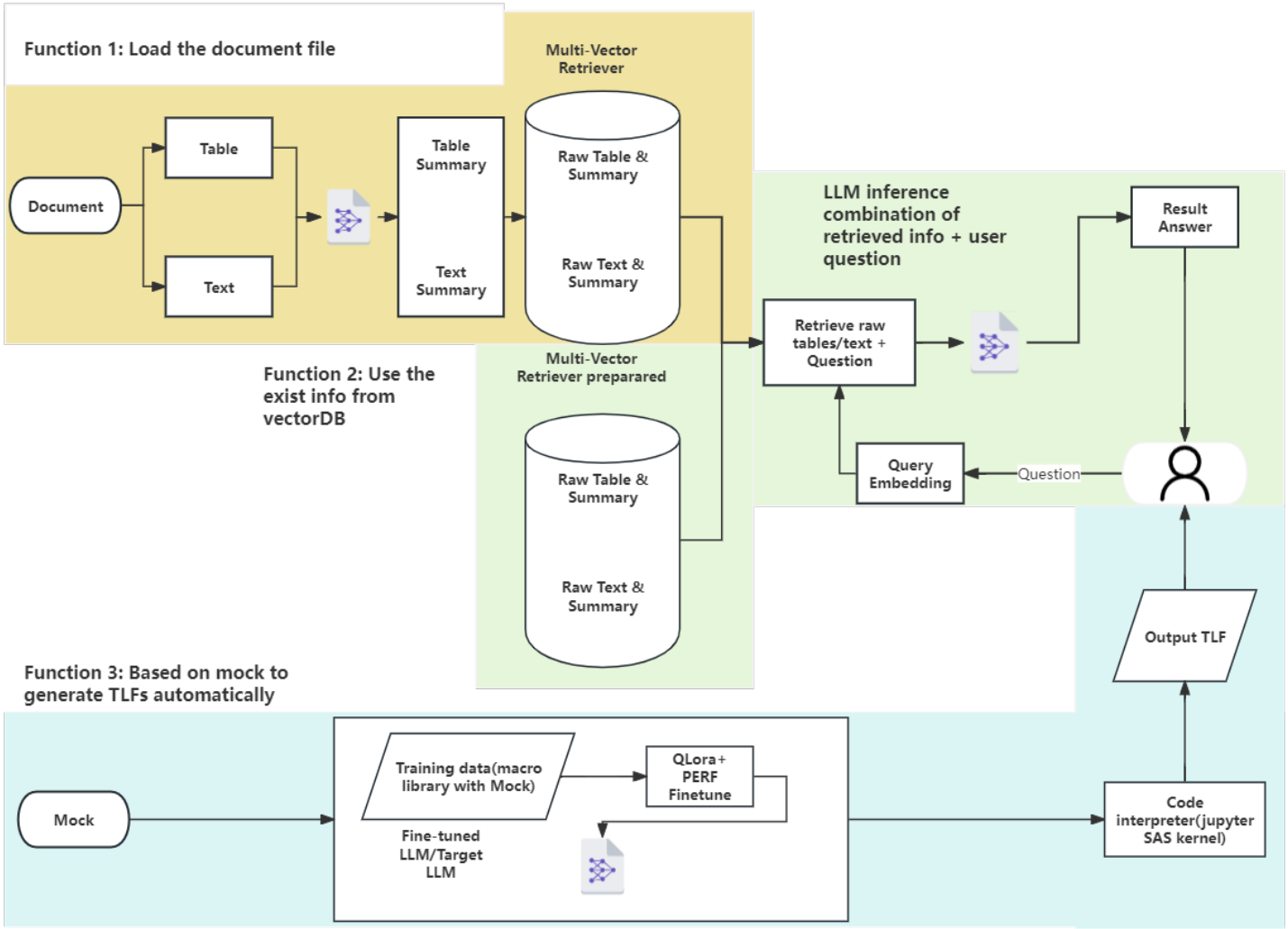
Outcome: Deal with most of the task immediately by ask question(chat using nature language)

Output the report:

- *Well-designed SAS reporting macro system + Fine-tuned LLM*

Outcome: End to end automation report generator (from mock to report immediately)

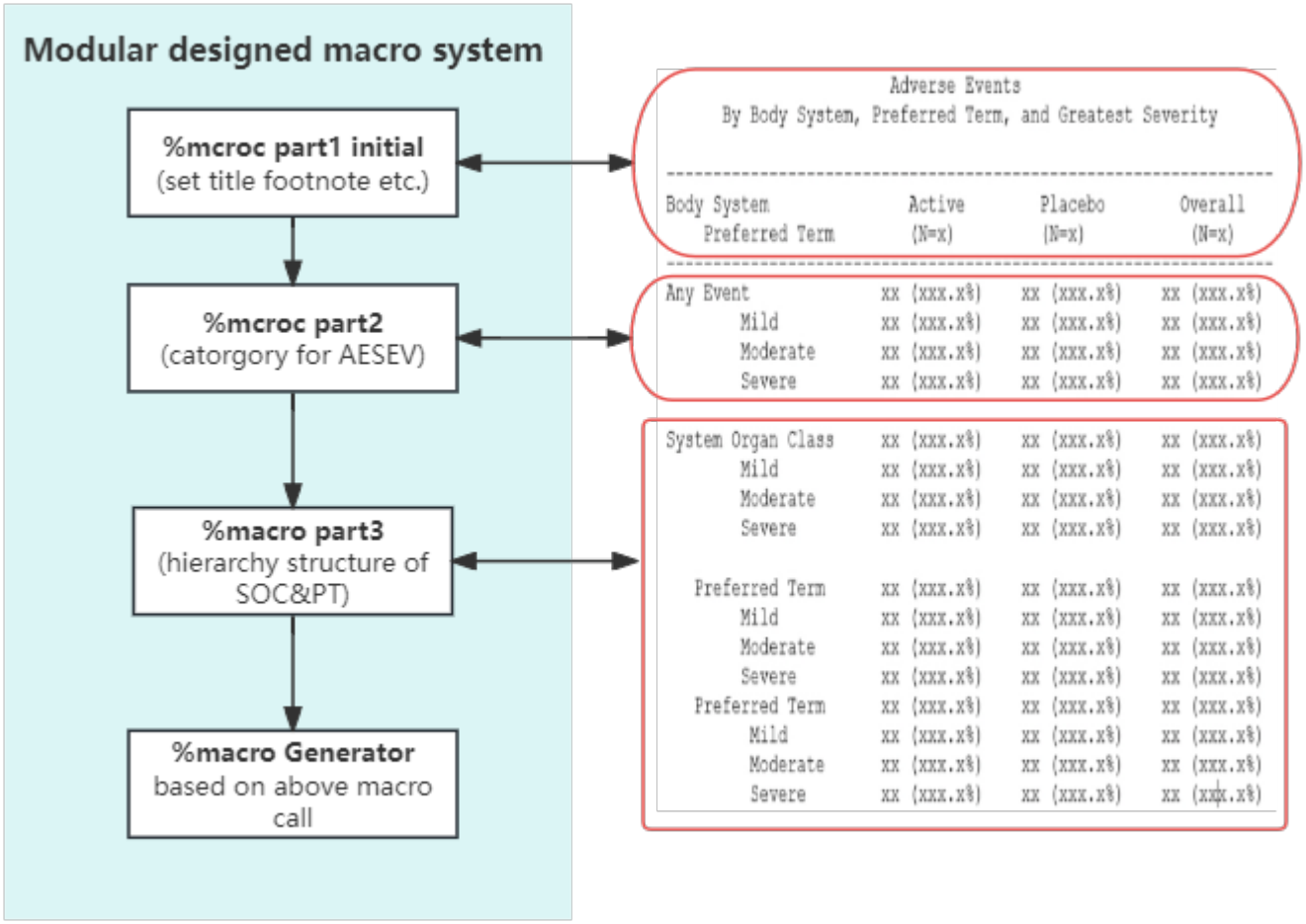
Proposed Framework for the Chatbot



MERCK

Design the reporting macro fit to LLM

Macro Library Compatibility: Adjusts macro libraries to be more compatible with LLM architectures, facilitating dynamic report generation.

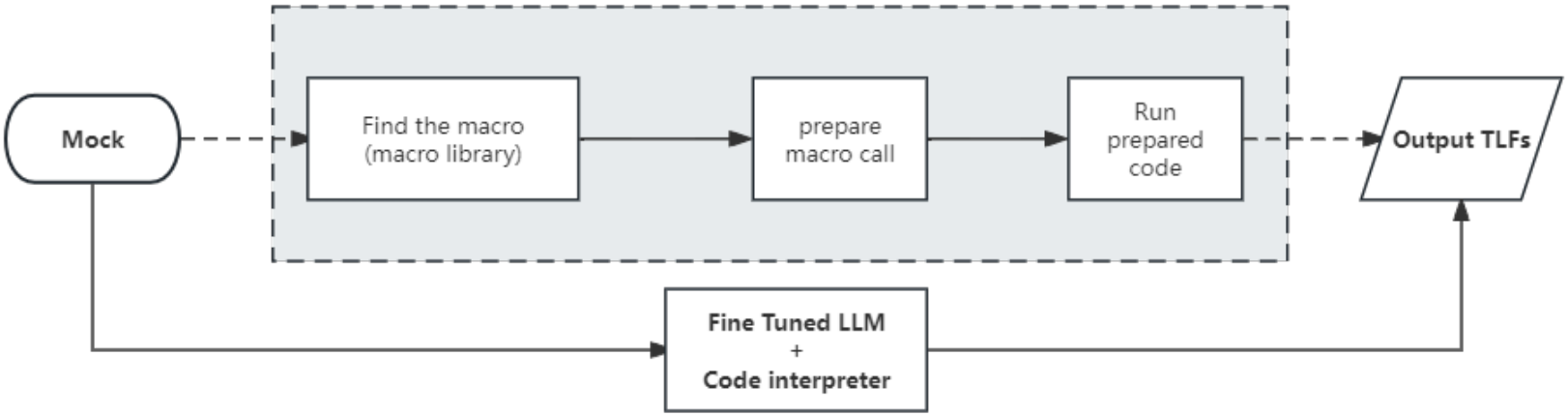


MERCK

End to end TLFs generator process

Fine-Tuning Process: Focuses on teaching LLM to recognize and replicate patterns between mock templates and SAS macro calls, improving TLF generation accuracy.

Fine-Tuned LLM give ready to use macro call based on input mock, then code interpreter(SAS kernel) will run the code and generate the report.



MERCK

Innovation point

Integrate the framework and technic in LLM filed for statistical programming usage.

Give the solution to solve the three main task in statistical programming:

1. *Use RAG for statistical programming related documents, optimize the pdf embedding.*
2. *'Train' RAG by providing the dataset using CDISC standard and SDTM/ADaM data structure/scheme.*
3. *Test and indicate which kind of SAS report macro library would fit to LLM and achieve end to end report generating automatically.*

In summary, the proposed framework and solution is cost efficient and would cover most of routine daily work of statistical programming.

MERCK

Future Work

- *Automation the validation process by combine well designed qc sas macro system with LLM and check the result with human involve.*
- *Implement new advanced technic and framework in LLM area from latest paper.*
- *Adopt in local environment by using distillation model.*
- *Cover some task in early phase(PK/PD, simulation etc.).*
- *Seeks to foster further development of these solutions in an open, collaborative environment across the industry, ultimately enhancing the efficacy and reliability of statistical programming in clinical research*

MERCK

Reference

- *Langchain-ai. Langchain. GitHub, 2024,* https://github.com/langchain-ai/langchain.
- *Ollama. Ollama. GitHub, 2023,* https://github.com/ollama/ollama.
- *Run-llama. Llama_index. GitHub, 2024,* https://github.com/run-llama/llama_index.
- *PromtEngineer. LocalGPT. GitHub, 2023,* https://github.com/PromtEngineer/localGPT.
- Vanna-ai. *Vanna*. GitHub, 2023, https://github.com/vanna-ai/vanna.

MERCK

Contact

Name:	*Jaime Yan* *Chao Su* *Changhong Shi*
Affiliation:	*Merck*
Contact Number:	*732-594-6459*
E-mail:	*mingyu.yan1@merck.com* *chao.su@merck.com* *changhong.shi@merck.com*
Web:	*www.merck.com*

MERCK

Q&A

MERCK