Encouraging LLM Thought Improvements for Medical Diagnosis Consistency

Myung Jin Kim, YeongHyeon Park

SK Planet Co., Ltd.

ISBI 2025
2025 IEEE International Symposium on Biomedical Imaging
April 14-17, 2025 | Houston, TX, USA

Poster No. P03-150

ABSTRACT

The development of the Large Language Model (LLM) has recently stimulated active research in the field of medical. However, LLMs are difficult to use for medical diagnosis because the limitation of the generative model makes it difficult to produce consistent answers. This creates a problem of reliability in performing diagnostics. This study, we propose a method of using prompts and way to call LLM to encourage the generation of consistent response. We evaluate performance of consistency using the American Academy of Orthopedic Surgeons (AAOS) osteoarthritis (OA) evidence-based guidelines. We compared the consistency of the findings with guidelines across different evidence levels using LLM judge.

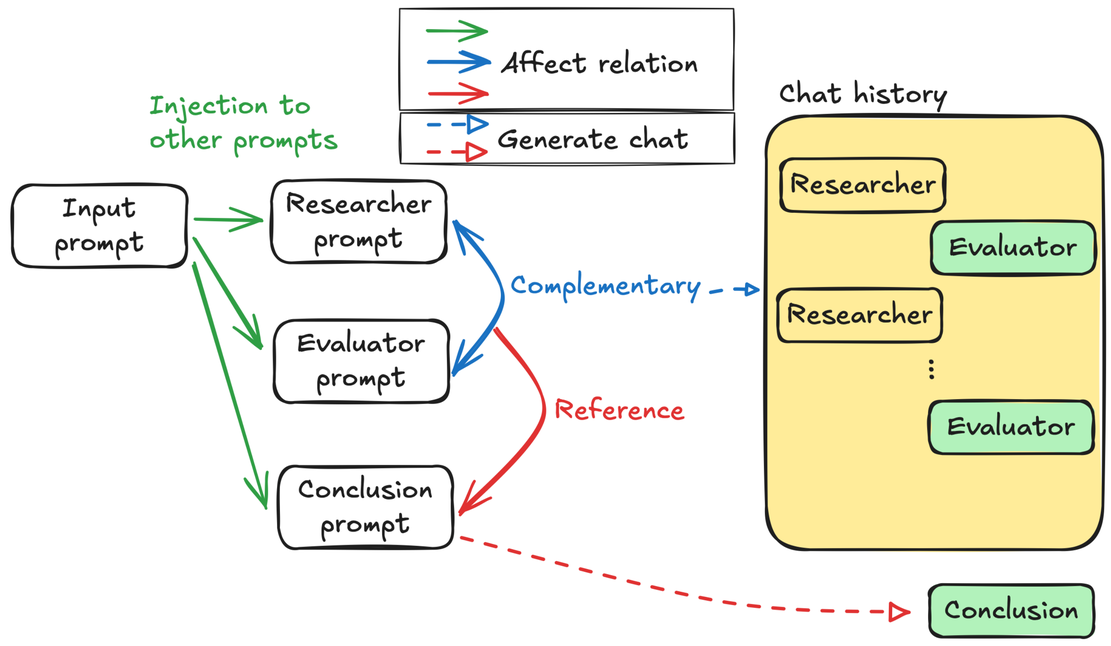

Fig. 1. Expressing the correlation of each prompt, the researcher and evaluator complement each other for a few turns, and then combine their dialogue in a conclusion to form a final conclusion.

INTRODUCTION

Recently, there have been many different use cases for LLM in medical. However, especially in the field of diagnostics, LLM is limited by its lack of consistency for some complex diagnoses[1].

This study, we propose a new method for maintaining consistency by increasing the accuracy of the problem. The LLM's logic skills allow you to delve deeper into the answer to a problem by designing a "thought experiment" and drawing conclusions, then identifying logical problems with the experiment and providing advices. Then, through all the dialogue in the course, we suggest how to reach a final conclusion.

METHOD AND EXPERIMENT

Each prompt has the same structure as Fig.1. All additional prompts contain input prompts to help you understand the purpose of input. The researcher and the evaluator continue the dialogue. Here we continue with the sixth round of the conversation to modify logical errors about experiments. The experiments used DeepSeek-R1-14B, and tried to query each advice 5 times as previous research. After that, when the getting result about consistency rate, we use LLM judge. Because of the recent LLM's performance of judgement is similar to human evaluation result[2].

The final consistency rate is 64.7% for each advice evaluation.

CONCLUSION

The experimental results from the limited parameter model were 2.9% higher than the previous research. That seems to be within the margin of error, this may mean that LLM are able to derive results through logical thinking, and that they are capable of deeper logical thinking under their own guidance. Also, it does not have to search for the best way to reach a good conclusion. However, current experiments with limited data and models can weaken reliability.

The next step is to run experiments on a wider range of models to accumulate more reliable data, and to use prompt optimization to draw clearer conclusions.

REFERENCES

[1] Li Wang, Xi Chen, Xiang Wen Deng, Hao Wen, Ming Ke You, Wei Zhi Liu, Qi Li, and Jian Li, "Prompt engineer- ing in consistency and reliability with the evidence-based guideline for llms," npj Digital Medicine, vol. 7, no. 1, pp. 41, 2024.

[2] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu, "G-eval: Nlg evaluation using gpt-4 with better human alignment," arXiv preprint arXiv:2303.16634, 2023.

This work was supported by SK Planet Co., Ltd., Korea.