Enabling Large Language Models to Think Twice When Its Answer Is Unreliable: A Case Study In Cancer Screening

Minchong Wu†, Hongxiang Lin*, Xiaoqing Lyu*¶, Chenrui Zhang*, Sun Yu ‡,
*Wangxuan Institute of Computer Technology, Peking University, Beijing, China
†Mathematics and Computer Science, New York University, NY, 11201, United States
‡Instec Technology Co., Ltd.
mw4794@nyu.edu, linhongxiang@stu.pku.edu.cn, lvxiaoqing@pku.edu.cn, chenrui.zhang@pku.edu.cn, sunyu29@cnic.gt.cn

2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-3748-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/BIBM58861.2023.10385316

***Abstract*—To enhance the accuracy and reliability of the response of large language models (LLMs) in professional domain-specific question answering, this paper proposes an approach to enhance the capabilities of LLMs by prompting them to "think twice." This methodology encourages a second-pass processing, allowing for deeper analysis and more refined outputs. We developed a prototype-logic-inspired query optimizer (PQO) to enhance the response of LLM. It refines initial queries using few-shot learning and self-reflection, leading to more insightful questions. Our experimental results show that the proposed approach obtains empirically validated answers.**

***Index Terms*—Large language model, prompt optimization, query optimization, health care Q&A, Langchain**

I. Introduction

With the advancement of technology, LLM have become a cutting-edge technology in many application areas. From simple text generation and dialogue systems to complex analysis and interpretation tasks, LLMs have demonstrated its powerful capabilities in processing and understanding natural language. In real life, including in fields such as e-commerce, law, education, and healthcare, LLMs play an important role in assisting people with their work. However, in the professional application of LLMs, there still exists the issue of answering questions inaccurately, which hinders the practical application of LLMs. Take cancer screening in healthcare area for instance, Hana L.Haver et al. [1] found that the responses from the LLMs are not hundred percent correct. One of the example of the unreliable answer indicated by them is shown in Table 1.

To prompt LLMs to provide more reasonable answers, many researchers are considering how to optimize the original question, including:

- Optimizing prompts manually;
- Using multiquery [2] to automatically generate queries from multiple perspectives for a single query, overcoming the limitations of distance-based retrieval;
- Adopting the online tool (such as Promptperfect [3]) to optimize the prompts;
- Utilizing two interactive LLMs or agents via a dialogic error-correction mechanism to amend inaccuracies in response to a given response.

However, most of the existing methods could not obtain satisfactory performance so far. In fact, the professional application of LLM generally faces several new technical challenges, such as how to automatically optimize the prompt and enable LLM to obtain more satisfactory answers. In this paper, we propose a Prototype-logic-inspired Query Optimizer, PQO for short, aiming to enhance LLM by alternating the question formulations based on a self-reflection strategy. Specifically, PQO utilizes few-shot learning to obtain guidance for the initial inquiries and substitute them with more illuminating questions.

ChatGPT unreliable answer for question "How can I prevent breast cancer?"
• Maintain a Healthy Weight: Being overweight or obese, especially after menopause, can increase your risk of breast cancer. • Stay Physically Active: Regular physical activity can help you maintain a healthy weight and reduce your risk of breast cancer. • Limit Alcohol Consumption: The more alcohol you drink, the higher your risk of developing breast cancer. Limiting alcohol intake can help reduce this risk. • Eat a Healthy Diet: Focus on a diet rich in fruits, vegetables, and whole grains. Some studies suggest that diet may play a role in breast cancer risk, although this is not definitively proven. • Breast-feeding: Breast-feeding for a total of one year or more (combined for all children) reduces the risk of breast cancer. • Limit Postmenopausal Hormone Therapy: Combination hormone therapy for more than three to five years increases the risk of breast cancer. Discuss the risks and benefits of hormone therapy with your doctor. • Regular Screenings: Regular mammograms can help detect breast cancer early, possibly before it has spread. Discuss with your doctor when to start and how often to have a mammogram. • Avoid Exposure to Radiation and Environmental Pollution: Some research suggests a link between breast cancer and radiation exposure. Limit your exposure to avoidable radiation (like some medical imaging tests) when possible. • Consider Risk-Lowering Drugs or Surgery: For women at very high risk of breast cancer, medications or preventive surgery might be an option. It's important to talk to a healthcare professional for advice tailored to your personal health history and risk factors. Regular check-ups and discussing any changes in your breasts with your doctor are also crucial steps.

TABLE I: Example of an unreliable answer indicated by Hana L.Haver et al. [1]

II. Related Work

The substantive AI model [4] is a neural network structure with a vast number of parameters. This encompasses large visual models, extended language models, and comprehensive models. When few-shot examples are directly woven into their input prompts, LLM has an approach which enable termed in-context learning. These particular examples are instrumental in shaping how LLMs conform to a given task [5].

The open-sourced LLM models [6] can address issues of varying model sizes and complexity while maintaining good

¶: Corresponding author

979-8-3503-3748-8/23/$31.00 ©2023 IEEE

performance in tasks. Based on whether it's a few-shot or zero-shot learning mode, several LLM models have demonstrated more pronounced response results.

LLMs have been applied to many professional fields, such as finance [7], marine science [8], law [9], education [10]. In the healthcare area, Nori et al. [11] suggest that the excellent performance of ChatGPT-4 in tests highlights its potential in medical education, and it can assist doctors in improving their work efficiency. Kan Hatakeyama-Sato et al. [12] discover that ChatGPT has the capability to handle molecular data and even address specific predictions in chemical problems. Israt Jahan et al. [13] conclude that the zero-shot ChatGPT and fined-tuning BioGPT perform similarly across various biomedical tasks, with ChatGPT excelling when given more detailed prompts. The DRG-LLaMA [14] endows us with the awareness that with proper training and adjustments, LLM can effectively automate the classification of diagnosis-related groups in hospitals. Bertalan Mesko et al. [15] state that the involvement of LLMs in the analysis of medical images, videos, and pharmaceuticals has enhanced the interaction between patients and physicians, highlighting the potential for LLMs to facilitate better services through interaction between AI and medical experts in the future.

Overall, LLMs are involved in various fields, providing more convenient services to people. However, their shortcomings are still quite evident: Given the stringent requirements of LLMs for prompts, generic or non-specific prompts might not elicit optimal responses from the LLM. Prompts that cannot be optimized automatically may lead to many non-professionals receiving unsatisfactory responses.

III. Methodology

To address the problem of non-automated prompt optimization, thereby enabling people from non-professional fields to obtain more satisfactory answers and to save the labor involved in manually optimizing prompts, we design a self-refining approach, PQO (Prototype-logic-inspired Query Optimizer). This method allows LLMs to autonomously learn and evolve, generating optimized prompts with minimal manual intervention. We provide the LLM with a selection of structured examples as logical prototypes to guide its learning process, enabling it to independently refine and iteratively generate better prompts with the similar logic in prototypes.

In the context of LLM advancements, particularly in prompt engineering, we emphasize on the linguistics logics to guide the learning process of LLM. This involves a detailed examination of the linguistic frameworks and cognitive mechanisms that underlie effective prompt design. We explore how aspects like syntactic structure, semantic clarity, and contextual appropriateness impact LLM performance, aiming to improve its response accuracy and relevance. By integrating linguistic theory with practical LLM, we develop a systematic methodology for enhancing prompt engineering, ultimately ensuring that LLMs like ChatGPT produce high-quality, contextually appropriate prompts.

To provide more convenient and superior services across various fields with LLM, we proposed PQO as an optimized prompt model to maximize the advantages of LLM. The workflow of PQO is shown as in Figure 1. To achieve the

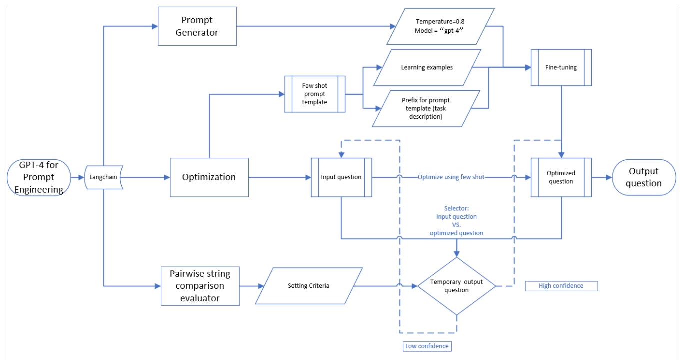

Fig. 1: Workflow of prototype-logic-inspired query optimizer (PQO) for LLM

objective of generating high-quality prompts, we explore a technological pathway by utilizing tools such as LangChain. It enables PQO to implement a series of algorithmic procedures to create, assess, and refine prompts, ensuring that ChatGPT's application is both precise and efficient. In this way, PQO ensures that the generated prompts are not only logically coherent in linguistic terms but also highly effective in practical interaction.

In the workflow of PQO, the optimization of prompts and the pairwise string comparison evaluator (PSCE) tool from Langchain are combined. Initially, PQO obtains a prompt generator through an OpenAI API. During optimization, the few-shot tool from Langchain is utilized to provide the model with examples of how to optimize prompts based on linguistic logic. The model learns from these examples and then uses them to refine the user's initial prompt, ultimately yielding an optimized prompt. The PSCE compares the original prompt with the optimized one and return a superior prompt. In the PSCE, linguistic logic is employed to establish criteria such as the reliability and clarity of prompts, to enable the PSCE to select a more effective prompt across various domains. To achieve more precise prompts, we add a loop into the combination of these two elements, setting the prompt generator's temperature parameter to a relatively high empirical value to generate more creative and varied prompts. The number of iterations can be altered based on user needs to ensure that the prompt generation time meets user expectations.This loop used to continually adjust the output of the prompts, ultimately delivering the most optimal prompt as the result.

IV. Experiments

In order to verify the effectiveness of the PQO, an unreliable prompt indicated by Hana L.Haver et al. [1] is used in our experiments. We use the generator to generate the LLM responses for both the input question and the optimized question. The comparison result is shown in Figure 2.

Original question:	**Generated question:**
where can I get screened for breast cancer?	Your task is to provide information on where a person can get screened for breast cancer. In your response, please include a variety of options such as hospitals, clinics, and mobile screening units. You may also discuss the process of making an appointment, the types of tests used in breast cancer screening, and any factors that might influence where someone chooses to get screened. Please provide accurate and practical information that will help individuals seeking breast cancer screening services.
ChatGPT answer: If you're looking for places to get screened for breast cancer, consider the following general steps: • Primary Care Physician: Your first point of contact should be your primary care physician or general practitioner. They can provide guidance on recommended screenings based on your age, family history, and other risk factors. • Local Hospitals: Most hospitals offer mammography services. Contact your local hospital's radiology department to inquire. • Radiology Centers: Many radiology centers offer mammograms and other imaging tests. You can search online for accredited centers near you. • Breast Cancer Clinics: Some areas have specialized clinics that focus exclusively on breast health. These clinics often offer mammograms and may also provide consultations with specialists. • Community Health Centers: Some community health centers provide low-cost or even free mammograms, especially during Breast Cancer Awareness Month in October. • Mobile Mammography Units: In some regions, mobile mammography units travel to various locations, making it more convenient for people to get screened. • Insurance Provider: Check with your health insurance provider. They may have a list of approved facilities where you can get a mammogram covered by your plan. • Online Directories: Websites like the American College of Radiology or the National Breast Cancer Foundation often have directories of mammography providers. • Events: Look for local events, especially in October (Breast Cancer Awareness Month), where free or discounted screenings might be provided. Remember, the recommended age and frequency for mammograms can vary based on individual risk factors and prevailing medical guidelines. It's essential to consult with a healthcare professional about when and how often you should be screened. If you provide your specific location (e.g., city or country), I might be able to give more tailored advice or resources.	ChatGPT answer: Breast cancer screening is an essential health practice for early detection and management. Here's a comprehensive guide on where to get screened, how to schedule an appointment, what to expect during the screening, and some helpful tips: • **1. Places for Breast Cancer Screening** – Hospitals: Many hospitals have dedicated breast cancer screening centers. They often offer the latest screening technologies such as mammograms. – Clinics and Health Centers: Local clinics, health centers, and even some doctor's offices provide breast cancer screening services. These might be more accessible for routine screenings. – Mobile Screening Units: In many areas, mobile screening units travel to communities to provide easier access to screening services. They are especially beneficial in rural or underserved areas. – Community Health Events: Occasionally, free or low-cost breast cancer screenings are offered at community health events or fairs. – Specialized Radiology Centers: These centers focus solely on diagnostic imaging and often have advanced equipment for breast cancer screening. • **2. Scheduling an Appointment** – Referral from a Doctor: In some cases, you might need a referral from your primary care physician. – Direct Contact: For most clinics and hospitals, you can schedule an appointment directly by calling their radiology or mammography department. – Online Booking: Some facilities offer online appointment scheduling through their websites. – Mobile Units and Community Events: Keep an eye on local news or community boards for announcements regarding mobile unit schedules or events offering screenings. • **3. What to Expect at the Screening** – Preparation: Wear comfortable clothing, preferably a two-piece outfit, as you will need to undress from the waist up. – During the Screening: A typical screening involves a mammogram, which is an X-ray of the breast. The process requires compressing the breast for a few seconds to get a clear image. – Duration: The actual mammogram takes only a few minutes, but the whole appointment may last about 30 minutes. – Comfort: While mammograms can be uncomfortable, they shouldn't be painful. Inform the technician if you feel pain. • **4. Important Advice or Tips** – Timing: Schedule your mammogram when your breasts are least likely to be tender, usually a week after your menstrual period. – Previous Records: If you've had mammograms at another facility, bring those records to your new screening appointment. – Post-Screening: After the screening, it's normal to wait for the results. If there are any concerns, the healthcare provider will contact you for further steps. – Regular Screening: Follow the recommended screening guidelines for your age and risk factors. Most guidelines suggest starting at age 40 and then every 1 to 2 years. • **Conclusion**Breast cancer screening is a critical part of women's health. Knowing where to go, how to schedule an appointment, and what to expect can make the process less daunting. Always consult with your healthcare provider for personalized advice and follow their recommendations for regular screenings.

Fig. 2: Different answers from a same LLM between the original question and the generated question via optimized prompts

Here are the analysis of the reason why the generated question generate a better response from LLM:

- Comprehensiveness: The answer for generated question not only lists screening locations but also explains the scheduling process and what to expect, offering a more complete guide compared to the first answer's general list of locations.
- Process Clarity: The answer for generated question provides a step-by-step breakdown of the entire screening process, which is more user-friendly compared to the first answer's broader and less detailed suggestions.
- Practical Tips: The answer for generated question offers actionable advice like scheduling around menstrual cycles and carrying previous records, which is more practically useful compared to the first answer's general information.
- Structure: The answer for generated question possesses clearer structure and logic, with each main point broken down into specific sub-points for the user to follow. In comparison, responses to original questions might be more difficult for users to understand, due to their broad structure.

In short, the results from our experiments demonstrate that our proposed model yields practical results.

V. Conclusion

By automatically optimizing the prompt, it's possible to further improve the response quality of ChatGPT. Looking ahead, there's a wealth of possibilities to explore. Among them is the potential to automate follow-up questions to yield more detailed results. Additionally, assessing the confidence levels of the responses emerges as a future direction. This metric could serve as a valuable indicator for the reliability and trustworthiness of the generated answers, guiding further refinements to the optimization process. To improve the learning capabilities of LLMs, researchers can incorporate extensive specialized knowledge, including scientific papers and knowledge graphs, to facilitate the generation of more expert-level responses by LLMs. Furthermore, LLMs require understanding diverse multimodal information, such as medical images and videos, to enrich their capability to yield more refined answers.

VI. Acknowledgement

This work was supported by National Key R&D Program of China (No. 2021ZD0113301).

References

[1] H. L. Haver, E. B. Ambinder, M. Bahl, E. T. Oluyemi, J. Jeudy, and P. H. Yi, "Appropriateness of breast cancer prevention and screening recommendations provided by chatgpt," *Radiology*, vol. 307, no. 4, p. e230424, 2023.

[2] Langchain, https://python.langchain.com/docs/get_started/introduction, 2023.

[3] PromptPerfect, https://promptperfect.jina.ai/home, 2023.

[4] Y. Wang, Y. Pan, M. Yan, Z. Su, and T. H. Luan, "A survey on chatgpt: Ai-generated contents, challenges, and solutions," *arXiv preprint arXiv:2305.18339*, 2023.

[5] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, B. Yin, and X. Hu, "Harnessing the power of llms in practice: A survey on chatgpt and beyond," *arXiv preprint arXiv:2304.13712*, 2023.

[6] K. Gao, S. He, Z. He, J. Lin, Q. Pei, J. Shao, and W. Zhang, "Examining user-friendly and open-sourced large gpt models: A survey on language, multimodal, and scientific gpt models," *arXiv preprint arXiv:2308.14149*, 2023.

[7] Y. Yang, Y. Tang, and K. Y. Tam, "Investlm: A large language model for investment using financial domain instruction tuning," *arXiv preprint arXiv:2309.13064*, 2023.

[8] Z. Bi, N. Zhang, Y. Xue, Y. Ou, D. Ji, G. Zheng, and H. Chen, "Oceangpt: A large language model for ocean science tasks," *arXiv preprint arXiv:2310.02031*, 2023.

[9] H.-T. Nguyen, "A brief report on lawgpt 1.0: A virtual legal assistant based on gpt-3," *arXiv preprint arXiv:2302.05729*, 2023.

[10] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying large language models and knowledge graphs: A roadmap," *arXiv preprint arXiv:2306.08302*, 2023.

[11] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of gpt-4 on medical challenge problems," *arXiv preprint arXiv:2303.13375*, 2023.

[12] K. Hatakeyama-Sato, N. Yamane, Y. Igarashi, Y. Nabae, and T. Hayakawa, "Prompt engineering of gpt-4 for chemical research: what can/cannot be done?" *Science and Technology of Advanced Materials: Methods Volume 3, Issue 1*, 2023.

[13] I. Jahan, M. T. R. Laskar, C. Peng, and J. Huang, "Evaluation of chatgpt on biomedical tasks: A zero-shot comparison with fine-tuned generative transformers," *arXiv preprint arXiv:2306.04504*, 2023.

[14] H. Wang, C. Gao, C. Dantona, B. Hull, and J. Sun, "Drg-llama: Tuning llama model to predict diagnosis-related group for hospitalized patients," *arXiv preprint arXiv:2309.12625*, 2023.

[15] B. Meskó, "The impact of multimodal large language models on health care's future," *Journal of Medical Internet Research*, vol. 25, p. e52865, 2023.