Prompt engineering techniques to enhance Large Language Models' performance in introductory physics

Giulia POLVERINI, Bor GREGORCIC

Department of Physics and Astronomy, Uppsala University, Box 516, 75120, Uppsala, Sweden

Abstract. Prompt engineering has increasingly garnered attention with the widespread use of AI-based chatbots over the past year. The formulation of prompts highly impacts the output of chatbots, which rely on Large Language Models and thus generate text that is a statistically good fit with both its training data and the users' prompt. Through examples from introductory physics, this study shows how selected prompt techniques can enhance the performance of chatbots like ChatGPT. In our investigation, we observed that upon using two specific prompt engineering techniques, the chatbot's responses improved both in the rate of correctness and quality of the argumentation.

Introduction

In the rapidly evolving landscape of artificial intelligence, large language models (LLMs) have emerged as transformative tools. LLMs-based chatbots' proficiency in processing and generating text has unlocked new potential for education in different disciplines, including physics. However, their use presents challenges and limitations arising from both their inherent operational mechanisms and users' proficiency in crafting prompts [1]. While we cannot alter the fundamental working principles of LLMs, the domain of prompt engineering offers a powerful way to enhance their utility. In doing this, mastering the subtleties of effectively communicating with LLM-based chatbots becomes crucial [2]. While there are no fixed rules, there exist strategic approaches and techniques that we can use to direct LLM-based chatbots towards more productive output generation.

In this study, we show how selected prompt strategies enhance the performance of LLMs in physics-related tasks, by both increasing the likelihood of a correct response and improving the quality of the reasoning in the responses. By doing so, we aim to provide some practical guidance for a deliberate and effective integration of LLMs into physics education.

Theoretical framework and research question

Existing research in physics education has been exploring the potential uses of LLM-based chatbots and framing their potential roles within the educational process in various ways. A notable framework, first proposed by Robert Taylor, can summarise these roles into three distinct functions: a tutor, a tutee, and a tool [3]. The effectiveness of chatbots in each of these roles is intrinsically linked to their performance, which in turn is highly influenced by the nature of the interaction (*i.e.* how they are prompted). Recognising this, our study seeks to offer the physics education research community insights into how to prompt efficiently. We aim to address the following research question: *How can prompt engineering strategies affect LLM-based chatbots' performance in physics education?*

Methods and findings

We initiated our study by assessing ChatGPT-4's proficiency in solving conceptual physics tasks, designed to evaluate the fundamental understanding of phenomena rather than complex mathematical abilities. Our initial findings using a selection of such questions revealed that the chatbot provided incorrect answers in approximately half of the attempts, and that even correct responses often lacked robust argumentation.

By analysing prompt engineering techniques, we identified two families of prompting strategies that improved ChatGPT's responses.

The first one consists of providing the chatbot with a contextualisation. In this approach, we tailored our prompts to include specific contextual information relevant to the physics problems at hand. For example, we included details such as the specific domain of physics the question referred to and the level of required explanation. This technique aims to guide the chatbot in focusing on the appropriate physics principles and reasoning methods [4].

The second family of strategies is Chain-of-thought (CoT), which allows LLM-based chatbots to methodically lay out the reasoning before reaching a conclusion. Consequently, the provided steps for solving a task are appended to the prompt, enabling the LLM to build a chain of argumentation that is more likely to be logical and coherent [5]. A CoT technique called Zero-Shot is particularly simple to use and effective, since it encourages the chatbot to articulate its reasoning process in a step-by-step fashion by simply instructing it to do so [6].

Applying these strategies significantly improved the chatbot's initial performance in terms of the likelihood of providing a relevant and correct response. Furthermore, when using CoT, the reasoning steps that the chatbot provided were more detailed and logically well-structured. Our findings suggest that the examined prompt engineering techniques have the potential to impact the usability of LLM-based chatbots in physics education in the three roles proposed by Taylor.

Conclusion

This study underscores the significant potential of prompt engineering in harnessing the capabilities of Large Language Models-based chatbots for educational purposes, particularly in the domain of introductory physics. Our investigation reveals that, while LLMs have some inherent limitations due to their reliance on training data and statistical language modelling, strategic use of selected prompt engineering techniques can significantly improve their output quality. Educators and developers can use these insights to effectively tailor interactions with chatbots, potentially leading to better learning outcomes and more engaging educational experiences.

This study contributes to the growing body of knowledge on the practical application of LLMs in education, offering a path forward for future research and applications.

References

[1] G. Polverini and B. Gregorcic, How Understanding Large Language Models Can Inform the Use of ChatGPT in Physics Education, *Eur. J. Phys.* **45** (2024) 02570.

[2] J. D. Zamfirescu-Pereira, R. Y. Wong, B. Hartmann, Q. Yang, Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts, in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (ACM, Hamburg Germany, 2023), pp. 1–21.

[3] R. Taylor, *The Computer in the School: Tutor, Tool, Tutee*, Teachers College Press, Totowa, NJ, 1980.

[4] T. B. Brown et al., *Language Models Are Few-Shot Learners*, arXiv:2005.14165.

[5] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*, arXiv:2201.11903.

[6] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, *Large Language Models Are Zero-Shot Reasoners*, arXiv:2205.11916.