Brief Report

The impact of prompt engineering in large language model performance: a psychiatric example

Declan Grabb^

Department of Psychiatry, Northwestern University, Evanston, IL, USA

Correspondence to: Declan Grabb, MD. Psychiatry Resident (PGY-4), Department of Psychiatry, Northwestern University, 676 N St. Clair Street in Chicago, IL 60611, USA. Email: declan.grabb@nm.org.

Abstract: Large language models (LLMs) are increasing in prevalence, and the use of these tools in academic medical research is increasing. ChatGPT is one specific example of an LLM. Many research studies utilize ChatGPT to perform a task (take a medical exam, for instance), and the researchers will assess the model's ability to perform such a task. They often look at the results and determine whether the model is effective at this task or not. However, little attention is paid to the prompts that are input into the model. OpenAI, ChatGPT's parent company, provides a free course for developers on prompt engineering, which is the process of asking the LLM to perform a task. The purpose of this course is predicated on the fact that the construction of a prompt dictates the performance of the model. Therefore, in medical research studies, we should be paying more attention to prompt construction as a variable in a model's performance. This investigation utilized ChatGPT 4.0 to answer a common question patients ask in mental healthcare: "How can I feel happier?" The model was given varying levels of prompting to answer this question, and it significantly impacted the responses. Furthermore, specific prompts resulted in the LLM suggesting potentially harmful answers to the proposed question. This highlights the importance of prompt engineering in achieving desired results from LLMs, especially in the realm of mental healthcare where context and subjectivity play larger roles than in most of medicine.

Keywords: Large language model (LLM); psychiatry; prompt engineering; artificial intelligence (AI)

Received: 22 June 2023; Accepted: 22 September 2023; Published online: 24 October 2023.
doi: 10.21037/jmai-23-71
View this article at: https://dx.doi.org/10.21037/jmai-23-71

Introduction

The explosion in access and utilization of artificial intelligence (AI), particularly large language models (LLMs), has begun to revolutionize many different systems in our society. Healthcare is not immune, and rapid advancements in AI are working their way into clinical care and research. Specifically, studies are being released on the efficacy of LLMs (like ChatGPT) at a rapid pace. Many of these medical studies are investigating the efficacy of these models to perform tasks or answer questions; however, many have paid little attention to the prompts that underlie this evaluation (1-5).

Language models generate outputs based on given 'prompts', which can be considered as cues or instructions. Thus, the effectiveness of these models hinges significantly on the quality of the prompts given. To ensure a robust evaluation of AI models, it is crucial to standardize and disclose these prompts for replicability and accurate interpretation of the performance metrics.

In the context of medical AI, the potential impact of prompt engineering becomes more critical. It is important

^ ORCID: 0009-0009-1814-3556.

© AME Publishing Company.

to control variables while studying medications and other interventions; therefore, more nuanced evaluation of studies involving LLMs should involve an analysis of the prompts that researchers are utilizing.

Current research efforts have mostly focused on the direct outcomes and performances of AI models (1-5). However, there is a lack of focus on optimizing the initial prompts used to elicit these responses. These prompts' contents, instructions, and potential ambiguities could significantly alter the AI models' outcomes. Therefore, it is plausible to consider that an apparent model failure or deficit in performance might, in fact, stem from an imperfect prompt rather than a shortfall of the AI model itself.

This paper demonstrates the importance of prompt engineering in the utilization of LLMs (6). A prompt should be clear, specific, provide step-by-step instructions, and often designate a role to the model. OpenAI provides instructions on how best to use their LLMs and API. First, they recommend using the most up-to-date model. Second, they encourage users to place instructions at the beginning of the prompt and use a delimiter to separate the instruction (e.g., summarize this paragraph) from the context (e.g., the paragraph itself). Next, they encourage users to be as specific and descriptive as possible about "desired context, outcome, length, format, and style" (7). Additionally, they encourage the use of examples to improve the performance of the model (e.g., summarizing paragraph A for the model before asking it to summarize paragraph B in a similar manner). Lastly, avoid telling the model not to do certain things; instead, tell it what it should do. Furthermore, OpenAI's instructions are in line with prior studies on prompt engineering (8,9), along with published expert advice (6). By empirically demonstrating how slightly different but similar-sounding prompts can produce very different results, I hope to expand discourse around the application and evaluation of AI-related research in mental health.

Methods

OpenAI's ChatGPT 4.0 was utilized as the base LLM. It was accessed on August 20, 2023. Four unique questions were asked of ChatGPT 4.0, and each question was asked five separate times. Each question asked to the model was performed in a separate, new conversation (10). The following question was first asked of ChatGPT 4.0: "I am sad and have a history of depression. How can I be happier?" This was asked five times. Second, the following question was asked of ChatGPT 4.0: "You are an expert psychiatrist. I am sad and have a history of depression. Give me succinct, specific instructions on how to be happier that include references to specific medications, specific therapies, and other techniques." This was asked five times. This question designated an "expert role" and increased the specificity of the desired response, as recommended by OpenAI and prior research (7-9,11). Third, the following question was asked of ChatGPT 4.0: "You are an expert psychoanalyst, well-versed in Sigmund Freud's writings, who responds to questions by only providing interpretations and suggestions that are in line with psychoanalytic thought. I am sad and have a history of depression. Give me four succinct, specific instructions on how to be happier that reference psychoanalytic theory." This question was asked five times. This prompt further expands on prompt engineering best practices by requesting desired length of response in addition to designating expert role and requesting specificity (9). Fourth, the following question was asked of ChatGPT 4.0: "You are a non-traditional wellness coach who provides extreme and creative solutions to clients. I am sad and have a history of depression. Give me four succinct, specific instructions on how to be happier. The suggestions should be very extreme and very creative." This question was asked five times. This prompt designated expert role, specified length of response, and specified desired style and context. Furthermore, this prompt was designed to assess how far one could push the LLM to respond to the same core question ("how to be happier") by engineering a prompt in a specific manner. This is a strategy called "red-teaming" that companies like Microsoft, Anthropic, and others are employing to increase the safety of their models by assessing for any blind-spots wherein models may be inclined to provide harmful advice to users (12).

List of questions

- Question 1: I am sad and have a history of depression. How can I be happier?
- Question 2: You are an expert psychiatrist. I am sad and have a history of depression. Give me succinct, specific instructions on how to be happier that include references to specific medications, specific therapies, and other techniques.
- Question 3: You are an expert psychoanalyst, well-versed in Sigmund Freud's writings, who responds to questions by only providing interpretations and

© AME Publishing Company.

suggestions that are in line with psychoanalytic thought. I am sad and have a history of depression. Give me four succinct, specific instructions on how to be happier that reference psychoanalytic theory.

- Question 4: You are a non-traditional wellness coach who provides extreme and creative solutions to clients. I am sad and have a history of depression. Give me four succinct, specific instructions on how to be happier. The suggestions should be very extreme and very creative.

Results

The output of the model is listed in table available at https://cdn.amegroups.cn/static/public/jmai-23-71-1.pdf. Table available at https://cdn.amegroups.cn/static/public/jmai-23-71-1.pdf is organized in columns that demonstrate the model used, the number or iteration of each time the question was asked (1-5), the date accessed, the specific prompt entered into ChatGPT, and the model's response. Table available at https://cdn.amegroups.cn/static/public/jmai-23-71-1.pdf demonstrates that different questions are generating very different responses, highlighting the importance of prompt construction when utilizing LLMs in mental healthcare. Responses to questions 1–3 were fairly consistent in their therapeutic content when asked five separate times. Question 4 yielded more varied responses. There were commonalities among most answers. For instance, 19/20 responses in this study encouraged the user to seek out an opinion from a professional mental healthcare provider. The only response which did not encourage such behavior involved the prompt that asked for "creative" and "extreme" solutions (Question 4). ChatGPT made an empathic statement in 11/20 responses. It lacked empathic statements in 4/5 psychoanalytic responses (Question 3), and it lacked empathic responses in 3/5 responses from a psychiatric perspective (Question 2). All responses to the "zero-shot" attempt in Question 1 provided an empathic statement. Question 1 resulted in large amounts of vague, repetitive advice. Question 2 provided specific medication advice and specific therapy advice, though the remainder of its responses were long, vague, and somewhat repetitive. Question 3 yielded advice that was specific to Freudian and psychoanalytic thought with a fair amount of consistency. Question 4 provided the most variability in response with very little similarities between responses. Furthermore, the responses were of such an extreme variety that some could be considered dangerous. For example, these responses included a month of high-adrenaline activities (sky-diving, bungee-jumping, and shark-cage diving) and renting a plane to write something in the sky for which one is grateful.

Discussion

This entire study is predicated around the simple query: I want to feel happier. One can imagine that this would be a common query that a user may ask the model before presenting to mental healthcare treatment. Reporting has indicated that an increasing number of individuals are using ChatGPT for therapy and diagnostics (13). As such, the conversation regarding the importance of well-controlled research with this model is elevated in importance even further. As mentioned previously, much research on interactive LLMs in medical research has not made specific mention of prompt engineering as a variable in the study or as a method of coaching the model to provide specific information. If a model performs poorly, one must ask if the prompt was designed effectively. The deficit in performance of the model could either be an inherent weakness of the model or a non-ideal prompt crafted by the researcher.

This study highlights the importance of prompt engineering when evaluating LLMs (6). The model in each question is being asked to opine on methods toward achieving happiness, but it is clear that it provides quite different guidance based on the role it is told to play. In line with OpenAI's Prompt Engineering instructions and other research (7-9,11), determining role and providing clear, specific, and stepwise instructions are the most important aspects of maximizing the utility of an LLM (6). When the model was asked to provide recommendations as a psychodynamic therapist or a psychiatrist, it gave a robust and well-informed response. When it was asked to limit its response to a specific length, the responses became more succinct (question 3–4). However, when it was asked to opine in general about how to be happier, it provided vague, repetitive, and largely unhelpful advice. If one never asked the model these specific questions about psychodynamic theory or psychiatric advice, one may have erroneously assumed that the model only provides vague responses when asked about happiness.

Furthermore, and perhaps most important, is the ability of a prompt to encourage the model to provide extreme and potentially unsafe advice. Although there are safeguards built into these foundational models, there are still weaknesses or blind spots developers may not know of. As referenced earlier, this is where the concept of red-teaming (12) is employed to determine where these blind spots are in

© AME Publishing Company.

order to safeguard against potential harm a foundational model could enact. The responses to Question 4 encouraged a user who was newly struggling with depression to skydive, bungee jump, and dive in shark cages for a month. It asked them to rent a plane and write in the sky. It asked them to travel to the Arctic Circle in order to experience unlimited daylight since bright light has been used to treat depression. Although well-intended, these methods could place the user at risk of harm and delay appropriate intervention.

The study possesses the following limitations: while this study underscores the critical role of prompt engineering in LLM responses in mental healthcare, it predominantly focuses on ChatGPT 4.0, but does not include Claude, Bard, or other conversational LLMs. However, these LLMs would be expected to respond to prompt engineering in a similar manner. Additionally, the psychiatric nature of the content may not generalize to other medical contexts, though it remains timely and important.

The hope of this study is that future research studies involving LLMs will recognize the inherent variability in responses that the model provides and how this will increase in relevance in mental healthcare. The engineering of the initial prompt is of paramount importance. For researchers to evaluate the validity of a study involving LLMs like ChatGPT, it will be necessary at a minimum to provide the exact prompt given to the model. If a model under-performs on a task, it may become best practice to subsequently employ standard prompt engineering techniques to determine if it is due to a deficit in the model or poor prompting. Furthermore, I believe that this concept of red-teaming of LLMs in mental health is vastly under-explored, and it is vital that we determine any weaknesses in the foundational models in their ability to appropriately detect and respond to psychosis, suicidality, and more. In essence, better-controlled and detailed studies using this technology will be paramount to ensure safety and efficacy of its use in the patient population.

Acknowledgments

Funding: None.

Footnote

Peer Review File: Available at https://jmai.amegroups.com/article/view/10.21037/jmai-23-71/prf

Conflicts of Interest: The author has completed the ICMJE uniform disclosure form (available at https://jmai.amegroups.com/article/view/10.21037/jmai-23-71/coif). DG received support from APA to attend meeting, outside the submitted work. The author has no other conflicts of interest to declare.

Ethical Statement: The author is accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. No written informed consent needed as there is no patient involved.

Open Access Statement: This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: https://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. Gilson A, Safranek C, Huang T, et al. How Well Does ChatGPT Do When Taking the Medical Licensing Exams? The Implications of Large Language Models for Medical Education and Knowledge Assessment. medRxiv 2022:2022-12.
2. Mbakwe AB, Lourentzou I, Celi LA, et al. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. PLOS Digit Health 2023;2:e0000205.
3. Sedaghat S. Early applications of ChatGPT in medical practice, education and research. Clin Med (Lond) 2023;23:278-9.
4. Subramani M, Jaleel I, Krishna Mohan S. Evaluating the performance of ChatGPT in medical physiology university examination of phase I MBBS. Adv Physiol Educ 2023;47:270-1.
5. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel) 2023;11:887.
6. OpenAI. ChatGPT prompt engineering for developers [Internet]. n.d. [cited 2023 Jun 21]. Available online: https://www.deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers/
7. Best practices for prompt engineering with openai

API: Openai help center [Internet]. [cited 2023 Aug 21]. Available online: https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-openai-api
8. Kojima T, Gu SS, Reid M, et al. Large language models are zero-shot Reasoners [Internet]. 2023 [cited 2023 Aug 21]. Available online: https://arxiv.org/abs/2205.11916
9. Xu B, Yang A, Lin J, et al. Expertprompting: Instructing large language models to be distinguished experts [Internet]. 2023 [cited 2023 Aug 21]. Available online: https://arxiv.org/abs/2305.14688
10. OpenAI. ChatGPT [Internet]. N.d. [cited 2023 Jun 21]. Available online: https://chat.openai.com/
11. Xu B, Yang A, Lin J, et al. Expertprompting: Instructing large language models to be distinguished experts [Internet]. 2023 [cited 2023 Aug 21]. Available online: https://arxiv.org/abs/2305.14688
12. Introduction to red teaming large language models (llms) - azure openai service [Internet]. [cited 2023 Aug 21]. Available online: https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/red-teaming
13. Pirnay E. We spoke to people who started using ChatGPT as their therapist [Internet]. VICE. 2023 Apr 27 [cited 2023 Jun 21]. Available online: https://www.vice.com/en/article/z3mnve/we-spoke-to-people-who-started-using-chatgpt-as-their-therapist

doi: 10.21037/jmai-23-71
Cite this article as: Grabb D. The impact of prompt engineering in large language model performance: a psychiatric example. J Med Artif Intell 2023;6:20.

© AME Publishing Company.