Peer Review File

Article information: https://dx.doi.org/10.21037/jmai-23-71

Reviewer A

There is a long history related to how queries are constructed when performing meta-analysis. It's a fundamental requirement that the query used to search the literature for relevant studies is shown, including any exclusions (e.g,. English language only) and the date when the query was produced. This is necessary for reproducibility. Why would it be any different when using AI for research purposes?

There is a more fundamental problem with the study. It is not sufficient to do a single query and generate reliable conclusions. It is known that asking the same question to a LLM in separate conversations will generate different responses, often conflicting. You need to repeat each of the queries multiple times and evaluate the range of responses within the same formulation of the question at least for contradictory recommendations. Failing to do this has been a consistent issue with studies looking at LLM performance.

While the article is focused on research applications, it is unrealistic to expect that individuals will undertake training in AI prompt construction. That speaks to increasing issues related to the ethics of providing these models for uneducated use. There is a large danger that people will be harmed if they take the advice provided at face value. Other ethical concerns relate to the responsibility of the companies providing these tools on the consequences of individuals experiencing harm as a result of their products.

My Response: Thank you for these comments. I have updated the study to indicate when the model was accessed. The table also includes the exact prompt asked of the model, and it includes the model's response verbatim. It indicates that the model was accessed on 8/20/23 only, as this is the day that I began anew with four simple questions and asked the model each question five times. This gave me the ability to better evaluate the range of responses within the same formulation of the question, as indicated by your comment.

Reviewer B

1. A number of previous studies on prompt engineering have been reported and should be reviewed. The author's argument might be quite obvious and lack novelty.

My Response: Thank you for encouraging me to look through this literature. I provided some of these studies as references in addition to more specific guidance from OpenAI on best methods for utilizing their models. These strategies may be well-understood in computer science circles, but a lot of healthcare related studies do not address prompt engineering as a concept (i.e. they do not even reference the fact that their task may not require it). Furthermore, as LLM’s are employed in the field of mental health (a more nuanced, complex situation than a test question), prompt engineering will increase in relevance.

2. Although the papers cited by the authors evaluated ChatGPT performance by asking questions with definite answers, such as national exams, is it possible that the nature of the study did not require advanced prompt engineering?

My Response: I think this is a good point, and it made me scrutinize the cited papers even more. I think source (1) is a good example. They categorized incorrect answers by type: logical error, information error, and statistical error. They describe the logical error as “The response adequately found the pertinent information, but did not properly convert the information into an answer,” and they describe information error as “ChatGPT either did not identify a key piece of information, whether present in the question stem or external that would be considered expected knowledge.” Given that the model committed a fair amount of both types of errors, I would like them to have specified an expert role, asked it to think step-by-step, or provided clear instructions on how to approach a multiple choice question. I think that could have improved the model’s performance, and I think future studies should make sure to do just that.

3. The questions asked in the ChatGPT are unclear, and we cannot evaluate which prompts and the responses were appropriate. Does the subject in the question have a history of depression, but is not currently depressed? Why is the subject asking questions about how to be happier, not how to improve depressive symptoms?

My Response: Thank you for this question! The questions, responses, and date accessed are all listed in Table 1. Patients use more natural language to discuss their moods and feelings (e.g. “happier” instead of “fewer depressive symptoms”). Providers attempt to mirror this language, and happiness is more than the absence of depressive symptoms. Many patients are unhappy without having depressive symptoms (trauma, anxiety, and more). Asking a tool how to minimize depressive symptoms would not represent best how most individuals would use this tool. The purpose of this study is not necessarily to demonstrate that one prompt yielded more "appropriate" suggestions (except when it provided harmful advice, which I highlighted); rather, it is to emphasize that -- in mental health particularly -- there are many "correct" answers and tweaking the prompt will generate very different suggestions.

4. It is recommended to introduce general prompt engineering techniques and those utilized in this study.

My response: Thank you for this suggestion. As mentioned in my response to (1), I included general prompt engineering techniques specific to OpenAI’s models. I added portions to my methods section indicating those utilized in this study.

5. Please add the limitations of this study in the discussion section.

My response: Thank you, I have added a limitation section in my discussion.

6. When will the research period begin and end?

My Response: The research period began on June 20 and will end at the end of August. For research purposes, though, ChatGPT 4.0 was only accessed on August 20, 2023. (I repeated study and assessed all prompts on the same day to remove as many variables as possible.)

7. What languages were used on ChatGPT?

My response: The study was conducted in English, utilizing the conversation window on ChatGPT.