4-28-2023

Can ChatGPT Accurately Answer a PICOT Question? Assessing AI Respones to a Clinical Question

Candise Branum
Gonzaga University, branum@gonzaga.edu

Martin Schiavenato
Gonzaga University, schiavenato@gonzaga.edu

Can ChatGPT Accurately Answer a PICOT Question? Assessing Artificial Intelligence Response to a Clinical Question

ABSTRACT

Background: ChatGPT, an artificial intelligence (AI) text generator trained to predict correct words, can provide answers to questions but has shown mixed results in answering medical questions.

Purpose: To assess the reliability and accuracy of ChatGPT in providing answers to a complex clinical question.

Methods: A PICOT formatted question was queried, along with a request for references. Full-text articles were reviewed in order to verify the accuracy of the evidence summary provided by the chatbot.

Results: ChatGPT was unable to provide a certifiable response to a PICOT question. The references cited as evidence included incorrect journal information, and many study details summarized by ChatGPT proved to be patently false, including providing fabricated data.

Conclusions: ChatGPT provides answers that appear legitimate but may be factually incorrect. The system is not transparent in how it gathers data to answer questions and sometimes fabricates information that looks plausible, making it an unreliable tool for clinical questions.

Keywords: Natural Language Processing, Artificial Intelligence, Information Literacy, Information Storage and Retrieval, Machine Learning, ChatGPT

INTRODUCTION

Natural language processing (NLP) is a branch of artificial intelligence (AI) that involves the understanding and generation of language.[1] Built on OpenAI's GPT-3.5 language model, ChatGPT (Chat Generative Pre-trained Transformer) is an NLP model released in November, 2022 which has captured the public's attention.[2] The chatbot interacts conversationally to generate essays,[3] answer clinical questions,[4] and has even passed medical licensing examinations.[5]

The power of AI deep learning techniques lies in its ability to predict a correct answer. ChatGPT is revolutionary because it is trained not only to provide an accurate prediction (in this case, predicting words), it is also trained to predict what answers humans would prefer. That is, the model optimizes for what humans expect in an answer.[6] This is an important caveat that may very well explain not only the model's success so far, but also its reception, since ChatGPT attempts to predict answers that feel right to humans.

In evaluating its potential clinical applicability, Sarraju and colleagues asked ChatGPT 25 questions on the topic of cardiovascular disease; the model appropriately answered 21 questions, or performed at 84% accuracy.[4] Although this is an impressive performance, especially considering that ChatGPT provides answers near instantly and in a conversational dialogue, the researchers found that some of the inappropriate responses provided potentially harmful advice. Their conclusion is that although the current version of the model is clearly not for medical use, there is a potential use for the technology in patient education.[4]

The mnemonic PICOT (Population, Intervention, Comparator, Outcome and Time) is an often-used tool to ask clinical, evidence-based practice (EBP) questions. Its format facilitates the selection of key terms used in searching bibliographic databases like PubMed and the Cumulated Index to Nursing and Allied Health Literature (CINAHL).[7] In principle, this structured format should generate better answers from the model. However, the model is trained on massive amounts of data from the internet, the surface web,[6] but not necessarily on password protected scientific databases, or the deep web where databases like PubMed and CINAHL reside. This raises the question of how accurate ChatGPT is in answering a clinical query. Previously, Wang and colleagues[8] explored ChatGPT's effectiveness in building a boolean query for use in a systematic review with mixed findings. Here, we set out to qualitatively evaluate ChatGPT's ability to accurately answer a structured PICOT question.

METHODS

In order to assess ChatGPT's reliability in helping to answer a complex PICOT question, we (1) identified a question for query and pasted it into a new ChatGPT session, (2) asked follow up questions to identify references, and (3) evaluated the references to verify that the answer provided by ChatGPT was accurate. This study was conducted in March of 2023, using ChatGPT 3.5.

The conversation was initiated by inputting a PICOT question directly into the chatbox, without any additional prompts or context: "*In African Americans with hypertension, is telemonitoring blood pressure effective in reducing blood pressure within 12 months of initiation?*" The initial response to this question included summaries from three key studies; although no citations were provided, the chatbot did provide vague references to the studies, such as a journal title and year of publication, and study type (i.e., "A systematic review and meta-analysis published in the Journal of Hypertension in 2021 included…"). This prompted the researchers to submit a follow up question, requesting the full citations. We then attempted to locate these articles and if full-text could be obtained, we would move on to the next step of verifying that the summary provided by ChatGPT was accurate.

RESULTS

Upon first glance, the answer provided by ChatGPT appeared promising. The response, written in academic tone, stated that there is evidence indicating that telemonitoring blood pressure is effective in reducing systolic blood pressure. Key findings from three published studies were highlighted in support of this statement.

The first step in verifying the information provided by ChatGPT was to locate the references. Although ChatGPT summarized key findings from only three studies, when prompted to provide references, four articles were cited as evidence. A quick search of the DOI numbers proved that these citations were all incorrect. The citations appeared to be an amalgamation of unrelated articles, with article titles and authors not matching the journal information or DOI number. Extracting the article titles from these citations, three of the four citations were located in PubMed. The last citation appeared to be completely fabricated.

A review of the existing three studies found that the article summaries provided by ChatGPT were not reliable, containing incorrect information about the study details and methodology, as well as fabricating data. The first study[9] retrieved was introduced by ChatGPT as a systematic review and meta-analysis involving a total of 4,375 participants, but this was not the case; rather, this was a position paper that included a brief discussion of multiple systematic reviews. The second study[10] was reported by ChatGPT to be an intervention administered by pharmacists, but in fact, this was a nurse-led intervention. In addition to false methodology details, this article was out of date; ChatGPT stated that it was from 2018, but it was actually published in 2007. The final article[11] focused on hypertensive patients in Ghana, and two key factors from our PICOT question were missing: Not only did the study not include our population of interest (African Americans), but telemonitoring was not one of the primary interventions.

Two of the studies reviewed were categorized by ChatGPT as being randomized controlled trials of "450 African American patients with uncontrolled hypertension," though this was not the case for either article. Most concerning, all three article summaries provided statistical data that suggested a correlation between telemonitoring blood pressure and a reduction in systolic blood pressure (SBP). In reviewing the articles, not only were the study methodologies misrepresented, but the data provided in all three summaries proved to be factually incorrect.

DISCUSSION

ChatGPT was unable to provide a verifiable response to a PICOT formatted clinical question. The references cited as evidence were either full of mistakes or completely fabricated, making it difficult to follow the research trail and verify the answers for accuracy, and when articles were finally located, it was discovered that many of the study details represented by ChatGPT were simply not true. The chatbot's responses stated that there was evidence supporting the use of telemonitoring in reducing blood pressure, but the statistical data provided as evidence to support these claims all appeared to be self-generated.

Although ChatGPT displays a convincing understanding of how to answer a PICOT formatted clinical question, upon a careful review, the chatbot is unable to support its conclusions. The model responds in a clear and convincing dialogue but fabricates evidence to support its claims, making it impossible to verify any information it provides. ChatGPT's lack of transparency in where it gets its data introduces inscrutable bias. We do not currently know much about the data set that is used by OpenAI, and thus we cannot identify potential biases in how the data is retrieved and interpreted. Another major drawback of ChatGPT is that it provides answers that are stylistically written to give the appearance of legitimacy, while being factually incorrect. OpenAI directly acknowledges this, noting that "ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers."[2] At its core, ChatGPT is a text generator, and its penchant for fabricating sources make it an unreliable tool for clinical questions, especially when it comes to providing evidence-based answers. AI will no doubt continue to make a profound impact in our world, but it is important to acknowledge that in answering a PICOT question, ChatGPT sacrifices accuracy for expediency, and verifiability for appealingness. This is a sobering thought for its implications both to education and patient care. We must remind ourselves that as of now, AI tools have not yet mastered the ability to replace human operated database searches in conducting evidence based research.

REFERENCES

1. Lauriola I, Lavelli A, Aiolli F. An introduction to deep learning in natural language processing: Models, techniques, and tools. *Neurocomputing*. 2022;470:443-456. doi:10.1016/j.neucom.2021.05.103

2. Introducing ChatGPT. OpenAI. Published November 30, 2022. Accessed March 15, 2023. https://openai.com/blog/chatgpt

3. Stokel-Walker C. AI bot ChatGPT writes smart essays - should professors worry? *Nature*. Published online December 9, 2022. doi:10.1038/d41586-022-04397-7

4. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *JAMA*. 2023;329(10):842-844. doi:10.1001/jama.2023.1044

5. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198

6. Montti R. What is ChatGPT and how can you use it? Search Engine Journal. Published December 26, 2022. Accessed April 11, 2023. https://www.searchenginejournal.com/what-is-chatgpt/473664/

7. Schiavenato M, Chu F. PICO: What it is and what it is not. *Nurse Education in Practice*. 2021;56:103194. doi:10.1016/j.nepr.2021.103194

8. Wang S, Scells H, Koopman B, Zuccon G. Can ChatGPT write a good boolean query for systematic review literature search? Published online February 9, 2023. doi:10.48550/arXiv.2302.03495

9. Omboni S, McManus RJ, Bosworth HB, et al. Evidence and recommendations on the use of telemedicine for the management of arterial hypertension: An international expert position paper. *Hypertension*. 2020;76(5):1368-1383. doi:10.1161/HYPERTENSIONAHA.120.15873

10. Artinian NT, Flack JM, Nordstrom CK, et al. Effects of nurse-managed telemonitoring on blood pressure at 12-month follow-up among urban African Americans. *Nurs Res*. 2007;56(5):312-322. doi:10.1097/01.NNR.0000289501.45284.6e

11. Ogedegbe G, Plange-Rhule J, Gyamfi J, et al. A cluster-randomized trial of task shifting and blood pressure control in Ghana: Study protocol. *Implementation Science*. 2014;9(1):73. doi:10.1186/1748-5908-9-73