Systematic Reviews

COMMENTARY

Open Access

Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation?

Riaz Qureshi[1,2*], Daniel Shaughnessy[1], Kayden A. R. Gill[2,3], Karen A. Robinson[2,4], Tianjing Li[1,2] and Eitan Agai[2]

Abstract

In this commentary, we discuss ChatGPT and our perspectives on its utility to systematic reviews (SRs) through the appropriateness and applicability of its responses to SR related prompts. The advancement of artificial intelligence (AI)-assisted technologies leave many wondering about the current capabilities, limitations, and opportunities for integration AI into scientific endeavors. Large language models (LLM)—such as ChatGPT, designed by OpenAI—have recently gained widespread attention with their ability to respond to various prompts in a natural-sounding way. Systematic reviews (SRs) utilize secondary data and often require many months and substantial financial resources to complete, making them attractive grounds for developing AI-assistive technologies. On February 6, 2023, PICO Portal developers hosted a webinar to explore ChatGPT's responses to tasks related to SR methodology. Our experience from exploring the responses of ChatGPT suggest that while ChatGPT and LLMs show some promise for aiding in SR-related tasks, the technology is in its infancy and needs much development for such applications. Furthermore, we advise that great caution should be taken by non-content experts in using these tools due to much of the output appearing, at a high level, to be valid, while much is erroneous and in need of active vetting.

Keywords Artificial intelligence, Large language models, Systematic review, Methodology

In this commentary, we discuss ChatGPT and our perspectives on its utility to systematic reviews through the appropriateness and applicability of its responses to systematic review tasks and prompts. ChatGPT is a large language model (LLM) and artificial intelligence (AI) system designed by OpenAI (https://openai.com/blog/chatgpt/) to interact with people in a natural and conversational way [1]. Standard machine-learning (ML) algorithms are trained to provide responses or to make classifications or predictions given a specific input using large sets of data that have already been, or are actively, categorized by users [2, 3]. Likewise, LLMs are trained to predict language and writing based on large datasets of written language, thereby learning contextual clues and what might be expected or predicted following a set of words (i.e., a prompt) [3–5]. For example, OpenAI's GPT-3.5 was trained on approximately 570 GB of text and is the original basis for ChatGPT [1, 6]. Depending on the prompt, LLMs can produce many different types of outputs. This has produced an explosion in the use of LLMs and ChatGPT, as well as creating controversy surrounding their applications [6].

Conducting a systematic review is a complex and arduous process that takes a great deal of expertise and time. It is not uncommon for reviews to take over 12 months to complete and require upwards of $100,000 in effort when considering the time spent on searching (by information

*Correspondence:
Riaz Qureshi
riaz.qureshi@cuanschutz.edu
[1] University of Colorado Anschutz Medical Campus, Aurora, CO, USA
[2] PICO Portal, New York, NY, USA
[3] University of Pittsburgh, Pittsburgh, PA, USA
[4] Johns Hopkins University, Baltimore, MD, USA

specialists), screening, data extraction, analysis, interpretation, and writing by methodologists and content experts [7–9]. There are areas where ML and AI have already been introduced in the systematic review process with great success [10–13]. Given the recent attention around LLMs, like ChatGPT, and the resource burden of conducting a systematic review, we wanted to explore and critique the responses of ChatGPT to systematic review tasks. On February 6, 2023, developers of PICO Portal—an AI-assisted systematic review platform [10]—hosted a webinar to demonstrate a variety of tasks and elicit feedback on the ChatGPT output.

We "tested" ChatGPT by asking it to complete systematic-review tasks with a focus on tasks relevant to interpretation of language and not test whether ChatGPT could perform a task that is more data-specific, such as data extraction [14]. Other biomedical uses for LLMs have been suggested, including for data and text mining, particularly of clinical records, and aiding in medical education and clinical decision making [15, 16]. Our intent was to see whether this kind of language model could be used by someone who may wish to plan a systematic review, further develop a review question, or get help in drafting the search or analysis methods. A detailed description of our experience and a link to the webinar recording can be found in the SUPPLEMENT.

We found that ChatGPT could complete some systematic review tasks well, while others had clear room for improvement:

- In formulating a structured review question, creating eligibility criteria, and screening titles for relevance, ChatGPT's output suggested the interpretation and contextualization of the prompt was appropriate. We felt the proposed criteria and selected articles could serve as a starting point for refinement depending on the complexity of the question.
- Having a ChatGPT generated PubMed search strategy, or an initial version, would be helpful to those who may not have access to an informationist in their resources. However, the proposed search strategy was unusable with multiple issues, including fabricating controlled vocabulary, that would not be apparent without expertise in search construction.
- ChatGPT is able to produce code in various programming languages and was able to create an outline of code for conducting a meta-analysis in Python and R. However, as with the search strategy, there were coding errors that required troubleshooting from a user with methodologic expertise.
- Synthesis and summary of multiple studies is a challenge but ultimately the most essential product of a systematic review. The time required to pick relevant information and create a summary is substantial, but there is potential for tools like ChatGPT to help begin these processes for reviewers. We found promise in the ability of the system to identify and summarize relevant information from a set of three abstracts. However, there were errors that suggest the technology is not yet ready for such a task.

In its current form, ChatGPT presents as an "uncanny valley" in research and information sciences: from a distance, the output mimics and passes as authentic; however, on closer inspection, it becomes apparent that it is not expertly formed material based on a depth of understanding of the systematic review process. A particularly strong limitation of the system is the lack of referencing appropriate and verifiable sources when asked for factual information. When we asked for references, we could not verify what it presented to us. This is a common occurrence as LLMs are designed to build a response using predictions and not by looking through literature to find real sources [17]. Indeed, when asked where it finds information and to search bibliographic databases, ChatGPT responds only that it cannot conduct any real literature retrieval.

With the model's current capabilities, we anticipate that anyone attempting to use ChatGPT for providing verifiable and content/context-specific research will find that the recipient must have expertise in the subject matter. Unfortunately, this pre-requisite defeats the purpose of having an "intelligent" automation help with the tasks. It should be noted that other LLMs are being developed and entering the public domain, so ChatGPT may not perfectly reflect all LLMs. On March 15, 2023, GPT-4.0 was released for testing among OpenAI's paid subscribers [18]. This new model purports to be more powerful than GPT-3.5 and better at recognizing and producing language and contextual cues in writing [18]. It should be noted that some minor testing of GPT-4.0 with similar questions showed a mild improvement in the summarization of three abstracts, but no additional improvements in systematic review task completion as far as we could discern. Additionally, there may be other systematic review tasks and use cases that we have not conceived and may elicit a more trustworthy and usable response from ChatGPT or other such systems. Furthermore, for better or worse, since generating a response in an LLM is not deterministic, the response will not be identical each time the same question is asked. We expect that there will be further advancements in the capabilities of these systems.

We know that the broader scientific community has concerns with the use of LLMs in research, and from our experience, we believe the systematic review community shares the same sentiments. Those in attendance during our demonstration posted many comments about ChatGPT and its application in education and research, primarily echoing concerns with the use of the technology and a large number of questions about its capabilities, limitations, internal processes, and output. Comments reflecting the uncertainty and hesitancy to use LLMs were also common, alongside the risks with non-expert use, as it was apparent that there was a requirement for content expertise in the various tasks. Many attendees posted links to resources and other tools to help perform the systematic review tasks we explored. There were also some comments on potential applications and areas for developing LLMs in the field of evidence syntheses and general positive and negative reactions from people about the utility of AI systems and LLMs in science. It is clear that discussion of the potential applications, challenges, and risks with integrating these technologies into the systematic review process need to happen and should take place in large, public forums. One group that is working towards addressing some of the questions and methodological issues such integration brings is the International Collaboration for the Automation of Systematic Reviews (ICASR) [19].

Despite the challenges with its use in systematic review tasks, ChatGPT was able to contextualize our questions and formulate responses that fit what we requested, which is encouraging for future development. In particular, we believe more attention should be given to accurately creating search strategies, as these utilize logic and rules (e.g., Boolean operators) and structured language in a way that should be more easily trainable than conversational language. Likewise, strengthening the ability of LLMs to take sections of text and identify relevant information for summary purposes could provide starting points for researchers or high-level summaries of results when time is limited (e.g., in a pandemic with hundreds of articles published daily that could contain relevant information to inform guidelines). Additionally, as the text writing and editing capabilities improve, the potential utility increases for these systems in polishing drafts of systematic reviews for authors who need help revising their writing [20].

In conclusion, we believe ChatGPT and other LLMs hold promise in being integrated into systematic reviews, but they are not yet able to be used with confidence in any way. We encourage others to attempt similar exploration and testing to understand the current limitations and capacity of ChatGPT and LLMs in the context of evidence synthesis.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13643-023-02243-z.

Additional file 1: SUPPLEMENT. Detailed description of our experience and a link to the webinar recording.

Acknowledgements

We would like to thank the webinar attendees for their feedback and discussion of ChatGPT's responses to our prompts.

Authors' contributions

RQ and EA conceptualized the commentary and led the webinar discussed within. RQ drafted the manuscript. RQ, DS, and KG read the chat transcripts to identify themes. All authors critically revised the manuscript for content and approved the final version for publication.

Funding

Not applicable.

Availability of data and materials

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

RQ, KG, TL, and KR provide consultative services to PICO Portal. EA is the founder and owner of PICO Portal. PICO Portal itself has no affiliation with ChatGPT. DS has no relevant interests.

Received: 21 March 2023 Accepted: 20 April 2023
Published online: 29 April 2023

References

1. OpenAI. ChatGPT: optimizing language models for dialogue. OpenAI. Published 2023. Accessed 6 Feb 2023. https://openai.com/blog/chatgpt/.
2. Ray S. A quick review of machine learning algorithms. In: Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, COMITCon 2019. IEEE; 2019:35–39.
3. Mahesh B. Machine learning algorithms - a review. Int J Sci Res. 2018;18(8):381–6. https://doi.org/10.21275/ART20203995.
4. Drenik G. Large language models will define artificial intelligence. Forbes. Published online 2023. https://www.forbes.com/sites/garydrenik/2023/01/11/large-language-models-will-define-artificial-intelligence/?sh=698337a9b60f.
5. Wiggers K. The emerging types of language models and why they matter. TechCrunch. Published online 2022. https://techcrunch.com/2022/04/28/the-emerging-types-of-language-models-and-why-they-matter/.
6. Shen Y, Heacock L, Elias J, Hentel K, Reig B, Shih G, Moy L. ChatGPT and other large language models are double-edged swords. Radiology. 2023;1. https://doi.org/10.1148/radiol.230163.
7. Michelson M, Reuter K. The significant cost of systematic reviews and meta-analyses: a call for greater involvement of machine learning to assess the promise of clinical trials. Contemp Clin Trials Commun. 2019;16:100443. https://doi.org/10.1016/j.conctc.2019.100443.
8. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using

data from the PROSPERO registry. BMJ Open. 2017;7(2):e012545. https://doi.org/10.1136/bmjopen-2016-012545.
9. Bullers K, Howard AM, Hanson A, Kearns WD, Orriola JJ, Polo RL, Sakmar KA. It takes longer than you think: librarian time spent on systematic review tasks. J Med Libr Assoc. 2018;106(2):198–207. https://doi.org/10.5195/jmla.2018.323.
10. PICO Portal. Introducing PICO Portal. Published 2023. Accessed 10 Feb 2023. https://picoportal.org.
11. DistillerSR. DistillerSR smarter reviews: trusted evidence. DistillerSR. Published 2023. Accessed 10 Feb 2023. www.distillersr.com.
12. Covidence. Covidence - better systematic review management. Covidence. Published 2023. Accessed 10 Feb 2023. https://www.covidence.org/.
13. Rayyan. Rayyan - Intelligent Systematic Review. Faster Systematic Reviews. Published 2023. www.rayyan.ai.
14. RobotReviewer. RobotReviewer - automating evidence synthesis. RobotReviewer. Published 2023. www.robotreviewer.net.
15. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS Digit Heal. 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198.
16. Lewis P, Ott M, Du J, Stoyanov V. Pretrained language models for biomedical and clinical tasks: understanding and extending the state-of-the-art. In: Proceedings Of the 3rd Clinical Natural Language Processing Workshop. 2020. p. 146–57. https://doi.org/10.18653/v1/2020.clinicalnlp-1.17.
17. Smerdon D. @dsmerdon. Twitter. Published 2023. Accessed 10 Feb 2023. https://twitter.com/dsmerdon/status/1618816703923912704?lang=en.
18. Chen R. GPT-4. OpenAI. Published 2023. Accessed 4 Mar 2023. https://openai.com/research/gpt-4.
19. ICASR. International Collaboration for the Automation of Systematic Reviews. 2023. Accessed 4 Mar 2023. https://icasr.github.io/.
20. Staiman A. Guest Post — Academic Publishers Are Missing the Point on ChatGPT ChatGPT as Author. The Scholarly Kitchen. Published online March 2023. https://scholarlykitchen.sspnet.org/2023/03/31/guest-post-academic-publishers-are-missing-the-point-on-chatgpt/.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.