The role of ChatGPT in developing systematic literature searches: an evidence summary

Veronica Parisi (a) and Anthea Sutton (b)
(a) Cruciform Hub, UCL Library Services, Library, Culture, Collections & Open Science (LCCOS), University College London (UCL), London, UK
(b) School of Medicine and Population Health, University of Sheffield, UK

Abstract
This evidence summary explores the potential and limitations of using ChatGPT for developing systematic literature searches. A systematic search identified the current peer-reviewed and grey literature. Studies were selected according to eligibility criteria. Included studies were analysed and synthesised narratively, focusing on the strengths, limitations, and recommendations for using ChatGPT to assist with the systematic literature searching process. Current literature is mostly opinion-driven, and there is limited published literature originating from the library and information profession. At present, limitations outweigh the strengths of ChatGPT for systematic literature searching, caution should be exercised, and human oversight is essential. More research is required, and information specialists and librarians are in a prime position to develop guidelines and share examples of best practice.

Key words: *artificial intelligence; information storage and retrieval; systematic reviews as topic; methods.*

Introduction

Health librarians and information specialists have long contributed to the conducting of systematic reviews for clinical decision making and evidence-based medicine (1). Typically, the information specialist role in a review team would be to design and conduct systematic searches across a range of information sources, including bibliographic databases, trial registries and grey literature (2). The rapid growth of scientific literature presents challenges for the information specialist, and has an impact on the quest for a comprehensive search. Large language models (LLMs) like ChatGPT, with their ability to process information and generate text, are attracting attention for their potential to revolutionise information retrieval (3, 4). Currently, there is a lack of guidance on how we, the health library and information community can harness this potential to aid our work, and little is known about the effectiveness of these AI tools in practice. The aim of this article is to identify and summarise the current research literature on using ChatGPT to develop systematic literature searches.

Methods

Eligibility criteria

Studies were included if they involved researchers or individuals engaged in the development of systematic literature searches using ChatGPT. All versions of ChatGPT utilised for these purposes were considered. Studies investigating the use of other AI tools for developing systematic literature searches were excluded. No restrictions were applied regarding the date, language, or study design.

Information sources

Searches were conducted across PubMed, Web of Science, arXiv, PROSPERO, Cochrane Library (CENTRAL), and Google Scholar from their inception to 1 May 2024. Additional sources included citation searching and relevant organisation websites to capture grey literature.

Search strategy

Tailored search strategies were devised for each database to ensure comprehensive coverage of relevant

Address for correspondence: Veronica Parisi, Cruciform Hub, UCL Library Services, Library, Culture, Collections & Open Science (LCCOS), University College London (UCL), Gower Street, London, WC1E 6BT, London, UK.
E-mail: v.parisi@ucl.ac.uk

Journal of EAHIL 2024; Vol. 20 (2): 30-34 doi: 10.32384/jeahil20623

literature. Given the novelty of ChatGPT, subject headings for this concept were unavailable, necessitating a strategy incorporating textwords, including synonyms and variations for both ChatGPT and literature search. For the complete PubMed search strategy, please refer to *Box 1*.

Box 1
PubMed search
("Chat Generative Pre-Trained Transformer" OR ChatGPT OR Chat-GPT) AND (Literature search* OR search strateg*)

Study selection

For the initial title and abstract screening, the total number of retrieved papers was divided equally among the review team using a randomly generated sample in Rayyan. Each reviewer independently screened their assigned portion of titles and abstracts. Following the initial screening, the full texts of the included papers were retrieved for further evaluation. Disagreements during the full-text screening phase were resolved through discussion and consensus was reached.

Results

The database searches retrieved 438 references, and a further 20 references were identified through website and citation searching. All references were imported into EndNote and 340 references were left after deduplication. Following title and abstract screening against the eligibility criteria, 24 references remained. Following full-text screening, a further 8 were excluded, leaving 16 included publications in this review. Two publications were merged as they contained the same information in two different formats (blog post and editorial), therefore for the purpose of this review we counted those as one publication. From this point onwards, we will summarise the findings relating to 15 publications. A PRISMA diagram illustrating the search and selection process can be found below (*Figure 1*).

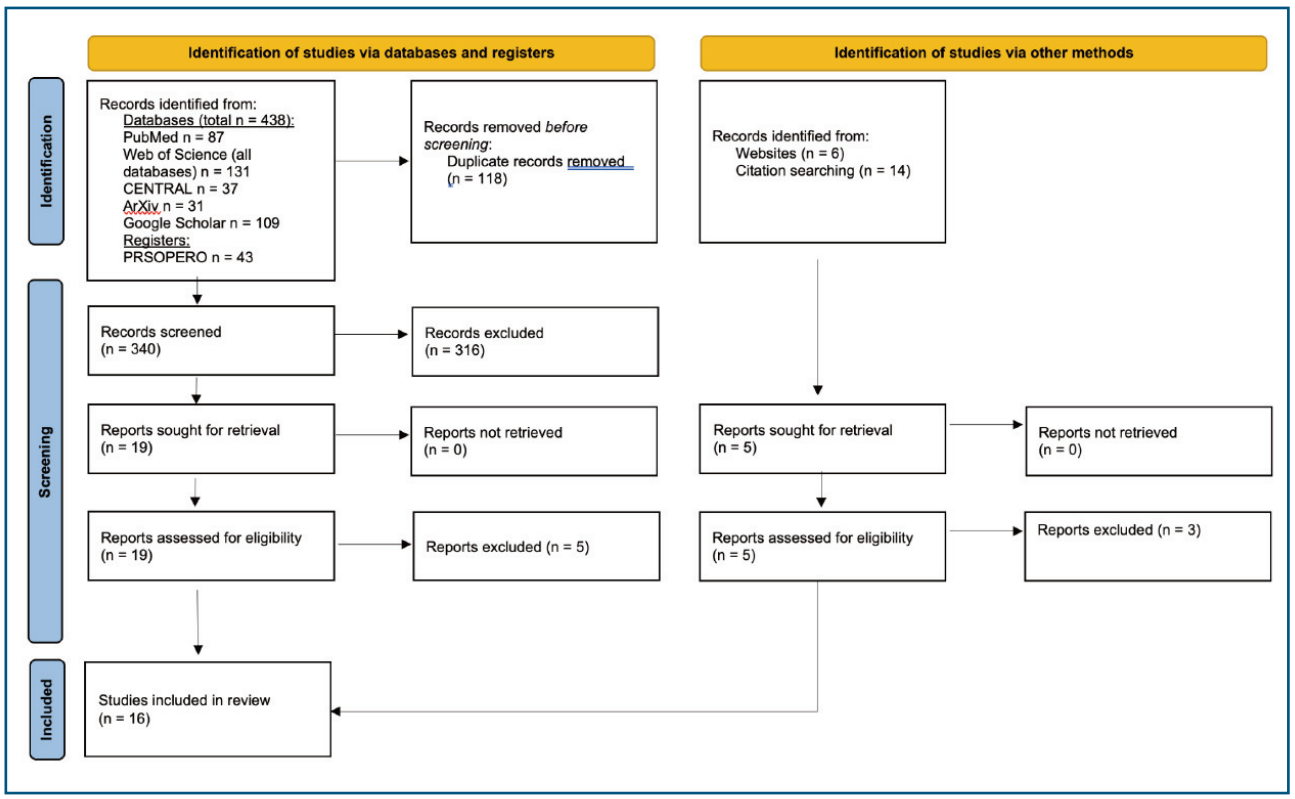

Fig. 1. *PRISMA flow diagram.*

Study characteristics

Over half of the publications included in this review are what we classified as "opinion pieces" (including letters, commentaries, editorials, blog posts). Four publications are articles in peer-reviewed journals (*Psychiatry Research, Journal of Clinical Medicine, JMIR Medical Informatics, Systems*) and two are published on preprint servers (e.g. pre-peer review). The majority of the publications are from the USA, with the remaining publications from Australia, Brazil, Czech Republic, and UK. We sought the author information roles from publications and checked for mentions of librarian/information specialist involvement. In all but three cases this was not reported. One single-authored publication was not by an information specialist role, and another publication stated that a librarian had been involved in validating the manual search strategy that was compared with ChatGPT. Only two publications were solely generated by library and information-based authors and these were our own on ChatGPT for systematic literature reviews and one on how ChatGPT and prompt engineering can be used in searching.

Strengths of ChatGPT for systematic literature searching

ChatGPT has the potential for generating search strategies, and there are some publications that suggest that this is possible, giving examples for PubMed/Medline and Scopus (5-10). In particular, Wang *et al.* (11) evaluate the precision and recall of ChatGPT's generated search strings and advise that these can lead to high search precision, to the expenses of recall. Some studies show that ChatGPT may also be helpful in translating search strategies from PubMed to Embase, Web of Science, Cochrane Library, and IEEE Xplore (5), and for ProQuest and Scopus databases (9). It is noted that ChatGPT could be a starting point for researchers unfamiliar with formulating search strategies (8, 12), but this would be at scoping stage, as the search strategies would need expert validation from a librarian/information specialist prior to utilising them for a systematic review. In the current literature, the limitations of ChatGPT appear to outweigh the strengths at this point.

Limitations of ChatGPT for systematic literature searching

While large language models like ChatGPT hold promise for various tasks, their application in systematic reviews currently faces significant limitations. These limitations hinder ChatGPT's current ability to generate comprehensive and reliable search strategies, a crucial step in the systematic review process.

The most often stated limitation in the current literature is ChatGPT's tendency to "hallucinate". Whilst in theory, ChatGPT can generate a search string, it struggles with database specific syntax and fabricates index terms such as MeSH headings (5, 8, 11-13), and is unable to execute the search once created (6). In some cases, ChatGPT states itself that it does not support database searching (8, 14). One publication raised the inability of ChatGPT to incorporate established search filters (for example to identify randomised controlled trials) in the search strings it produces (7).

ChatGPT has limited access to real-time data. The free version of ChatGPT, ChatGPT 3.5 (although in early May OpenAI has made available a free version of ChatGPT-4 on a limited basis), currently has a cut-off date of 2021, therefore concerns about currency are expressed in the literature. Alshami (6) emphasises the model's reliance on user prompts, which can be subjective and introduce bias. However, a manual search strategy is also subject to human input. Of more concern is the length of the prompts required and the iterative process, demonstrating that ChatGPT is unlikely to save time for the experienced information specialist. There is also a lack of transparency in prompts, and inconsistencies. Studies by Guimaraes, Qureshi, and Wang (7, 8, 11) raise concerns about inconsistent outputs, reporting different responses to the same research question at different times. This is because answers in such LLMs are non-deterministic, which may affect the reproducibility and transparency of searches.

Some publications attempt to validate ChatGPT against human-generated search strategies (14-18). However, it is not clear whether the manually generated searches have been validated or appraised for efficiency, so they may not be a reliable benchmark to use.

Recommendations from the literature

There are several important considerations to be considered when using ChatGPT for literature searches. While ChatGPT can assist in developing search strategies, relying solely on it is not recommended (7, 10, 19), and some authors suggest that traditional search methods and expert reviews are essential to ensure thoroughness and comprehensiveness (8).

Given the risks associated with hallucinations and inaccurate information, some authors even recommend against using ChatGPT at all for literature searches (14, 17).

To help mitigate these risks, it is essential to verify ChatGPT-assisted searches for accuracy and relevance. Human oversight is necessary to cross-check the validity of the information generated by ChatGPT (9, 13, 18).

Implementing a structured framework, such as the one proposed by Alshami (6), may help integrate ChatGPT into the workflow with predefined protocols for human oversight, verification, and periodic reassessment of ChatGPT-generated outputs.

In terms of search strategy formulation, Boolean query development, as recommended by Wang (11), involve extensive refinement to ensure precision and comprehensiveness. Expert intervention is necessary to tailor the queries to the specific requirements of the systematic review and to optimise the search results (5).

Limitations and strengths of this study

This paper has certain limitations that should be acknowledged. Firstly, due to time constraints it has been developed at pace, which may have led to a less systematic and comprehensive exploration of the topic. To this end, it should also be noted that this manuscript has been augmented with the use of ChatGPT for summarisation and proofreading purposes. Secondly, the research question addressed in this study has been kept narrow by design, focusing specifically on the use of ChatGPT for developing search strategies for systematic literature searching. While this is a relevant area of study, it excludes broader discussions on ChatGPT's capabilities such as its use in creating literature reviews, aiding in clinical decision making, generating references, as well as the ethical and legal implications of using ChatGPT in education and research. All of these topics may offer invaluable intersections to help enrich the current discourse.

This study presents several strengths as well. In addition to traditionally structured searches, the adoption of iterative and purposive searching contributed to the identification of more sparse and unsystematic literature. As it has been discussed, "opinion pieces" represent more than half of the body of evidence on this topic, offering invaluable insights for our research. Another strength of this paper is that it has been devised and developed by information specialists/librarians, bringing a breadth of knowledge and expertise in the field of systematic literature searching. More significantly, this study has helped identify a research gap, which is the paucity of literature from information specialists/librarians on using ChatGPT for literature searches. In this regard, our review found that most studies included were not conducted by professionals in this field, despite their expertise in search strategies and systematic searching. This gap underscores the need for further research and contributions from information specialists/librarians, who are ideally positioned to provide insights and develop best practices in this area.

Call for action and conclusion

To address this gap, we would like to make a call for action and encourage more research that involves information specialists and librarians. The expertise brought by these professionals can significantly contribute to the development of a more informed and judicious use of ChatGPT for literature search processes. To help achieve this, opportunities for funding could be sought nationally, internationally or at institutional level to support the development of research initiatives to explore this topic.

In addition to this, the creation of a special interest group (SIG) across EAHIL, which focuses on the use of AI tools in literature searching could provide a platform for information specialists and librarians to collaborate, share knowledge, and advance the field. This SIG could organise conferences, workshops, and publications to disseminate findings and best practices, thereby contributing to the advancement of research in AI-assisted literature searches.

In conclusion, while this paper presents initial findings on the use of ChatGPT for developing search strategies for systematic literature searching, it also underscores the need for broader research. By involving information specialists and librarians, the academic and research communities can enhance their knowledge and understanding of literature searching and its applications within the context of AI. Future research, supported by appropriate funding and collaborative efforts, may be crucial in addressing the current gap and advancing the field.

Submitted on invitation.
Accepted on 9 June 2024.

REFERENCES

1. Spencer AJ, Eldredge JD. Roles for librarians in systematic reviews: a scoping review. J Med Libr Assoc. 2018;106(1):46-56.
2. Cooper C, Booth A, Varley-Campbell J, Britten N, Garside R. Defining the process to literature searching in systematic reviews: a literature review of guidance and supporting studies. BMC Med Res Methodol. 2018;18(1):85.
3. Huang Y, Huang J. Exploring ChatGPT for next-generation information retrieval: opportunities and challenges 2024 February 01, 2024:[arXiv:2402.11203 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2024arXiv240211203H
4. Jin Q, Leaman R, Lu Z. PubMed and beyond: biomedical literature search in the age of artificial intelligence. EBioMedicine. 2024;100:104988.
5. Alaniz L, Vu C, Pfaff MJ. The utility of artificial intelligence for systematic reviews and Boolean query formulation and translation. Plastic and Reconstructive Surgery-Global Open. 2023;11(10):e5339.
6. Alshami A, Elsayed M, Ali E, Eltoukhy AEE, Zayed T. Harnessing the power of ChatGPT for automating systematic review process: methodology, case study, limitations, and future directions. Systems. 2023;11(7).
7. Guimarães NS, Joviano-Santos JV, Reis MG, Chaves RRM. Development of search strategies for systematic reviews in health using ChatGPT: a critical analysis. J Transl Med. 2024;22(1):1.
8. Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? Syst Rev. 2023;12(1):72.
9. Nguyen-Trung K, Saeri AK, Kaufman S. Applying ChatGPT and AI-powered tools to accelerate evidence reviews [online]. OSF Preprints. 2023. [2 May 2024]. Available from: osf.io/pcrqf
10. Schopow N, Osterhoff G, Baur D. Applications of the natural language processing tool ChatGPT in clinical practice: comparative study and augmented systematic review. JMIR Medical Informatics. 2023;11:e48933.
11. Wang S, Scells H, Koopman B, Zuccon G, Acm, editors. Can ChatGPT write a good Boolean query for systematic review literature search? 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR); 2023 Jul 23-27; Taipei, Taiwan 2023.
12. Wood H. Using ChatGPT as a Knowledge Specialist. Health Education England, Knowledge and Library Services [online]. 2023. [2 May 2024]. Available from: https://library.hee.nhs.uk/about/blogs/using-chatgpt-as-a-knowledge-specialist
13. Parisi V, Sutton A. How can AI help you with your systematic literature review? Reflections from a two-day seminar. Notion [online]. 2024. [5 February 2024]. Available from: https://elated-broker-fc3.notion.site/How-can-AI-help-you-with-your-systematic-literature-review-Reflections-from-a-two-day-seminar-0c45dc93910145afaedfb0fceed273e3
14. Blum M. ChatGPT produces fabricated references and falsehoods when used for scientific literature search. J Card Fail. 2023;29(9):1332-4.
15. Corti C, Castellano G, Curigliano G. Exploring the utility and limitations of ChatGPT in scientific literature searches. ESMO Real World Data and Digital Oncology. 2023;1.
16. Haman M, Školník M. Using ChatGPT to conduct a literature review. Accountability in research. 2023:1-3.
17. McGowan A, Gui Y, Dobbs M, Shuster S, Cotter M, Selloni A, et al. ChatGPT and Bard exhibit spontaneous citation fabrication during psychiatry literature search. Psychiatry Res. 2023;326:115334.
18. Suppadungsuk S, Thongprayoon C, Krisanapan P, Tangpanithandee S, Garcia Valencia O, Miao J, et al. Examining the validity of ChatGPT in identifying relevant nephrology literature: findings and implications. J Clin Med. 2023;12(17).
19. Wood H. Prompt engineering: adventures with ChatGPT, Bing, and Bard. Health Education England, Knowledge and Library Services [online]. 2023. [2 May 2024]. Available from: https://library.hee.nhs.uk/about/blogs/prompt-engineering--adventures-with-chatgpt-bing-and-bard