JMIR Publications
The Leading eHealth Publisher

Hallucination Rates and Reference Accuracy in ChatGPT and Bard for Systematic Reviews: A Comparative Analysis

Mikaël Chelli, Jules Descamps, Vincent Lavoué, Christophe Trojani, Michel Azar, Marcel Deckert, Jean-Luc Raynier, Gilles Clowez, Pascal Boileau, Caroline Ruetsch-Chelli

Submitted to: Journal of Medical Internet Research
on: September 27, 2023

Table of Contents

Hallucination Rates and Reference Accuracy in ChatGPT and Bard for Systematic Reviews: A Comparative Analysis

Mikaël Chelli[1] MD, MSc; Jules Descamps[2] MD, MSc; Vincent Lavoué[1] MD; Christophe Trojani[1] MD, PhD; Michel Azar[1] MD; Marcel Deckert[3] PhD; Jean-Luc Raynier[1] MD, MSc; Gilles Clowez[1] MD; Pascal Boileau[1] MD, PhD; Caroline Ruetsch-Chelli[3, 4] MD, MSc

[1]Institute for Sports and Reconstructive Bone and Joint Surgery, Groupe Kantys Nice FR
[2]Hôpital Lariboisière Paris FR
[3]Université Côte d'Azur, INSERM, C3M, Team Microenvironment, Signalling and Cancer Nice FR
[4]Immunology department, Nice University Hospital, Côte d'Azur University Nice FR

Corresponding Author:
Jules Descamps MD, MSc
Hôpital Lariboisière
2 Rue Ambroise Paré
Paris
FR

Abstract

Background: Large Language Models (LLMs) have raised both interest and concern in the academic community. They offer potential for automating literature search and synthesis for systematic reviews but concerns regarding their reliability and the tendency to generate unsupported ('hallucinated') content persist.

Objective: To assess the performance of Large Language Models like ChatGPT and Bard to produce references in the context of scientific writing.

Methods: The performance of ChatGPT and Bard in replicating the results of human-conducted systematic reviews was assessed. Using systematic reviews pertaining to shoulder rotator cuff pathology, these LLMs were tested by providing the same inclusion criteria and comparing the results with original systematic review references. The study utilized three key performance metrics: Recall, Precision, and F1-score, alongside the hallucination rate. Articles were considered "hallucinated" if any two of the following information were wrong: title, first author, or year of publication.

Results: The LLMs could generate legitimate references, but also produced hallucinated articles at a rate between 28% to 91%. Although ChatGPT 4 demonstrated superior performance among the models tested, all failed significantly in adhering to the established eligibility criteria. Precision and recall ranged from 0% to 13.4% and 0% to 14.7% respectively, highlighting the limitations of these models in replicating human-conducted systematic reviews.

Conclusions: Given their current performance, it is not recommended for LLMs to be deployed as the primary or exclusive tool for conducting systematic reviews. Any references generated by such models warrant thorough validation by researchers. The high occurrence of hallucinations in LLMs highlights the necessity for refining their training and functionality before confidently employing them for rigorous academic purposes

(JMIR Preprints 27/09/2023:53164)
DOI: https://doi.org/10.2196/preprints.53164

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
   Please make my preprint PDF available to anyone at any time (recommended).
   Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
   Only make the preprint title and abstract visible.
✓ **No, I do not wish to publish my submitted manuscript as a preprint.**
2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Hallucination Rates and Reference Accuracy in ChatGPT and Bard for Systematic Reviews: A Comparative Analysis

Mikaël Chelli, MD, MSc (1), Jules Descamps, MD, MSc (2), Vincent Lavoué MD (1), Christophe Trojani, MD, PhD (1), Michel Azar, MD (1), Marcel Deckert, PhD (4), Jean-Luc Raynier, MD, MSc (1), Gilles Clowez, MD (1), Pascal Boileau, MD, PhD (1), Caroline Ruetsch-Chelli, MD, PhD ,4)

(1) Institute for Sports and Reconstructive Bone and Joint Surgery, Groupe Kantys, Nice (France)
(2) Orthopedic and traumatology unit, Hospital Lariboisière, APHP, Paris (France)
(3) Université Côte d'Azur, INSERM, C3M, Team Microenvironment, Signalling and Cancer, Nice, France

Corresponding author:
Mikael Chelli, MD
Institute for Sports and Reconstructive Bone and Joint Surgery, Groupe Kantys, Nice (France)
Co-founder of EasyMedStat
mikael.chelli@gmail.com
+33650542270

Abstract and keywords

Background. Large Language Models (LLMs) have raised both interest and concern in the academic community. They offer potential for automating literature search and synthesis for systematic reviews but concerns regarding their reliability and the tendency to generate unsupported ('hallucinated') content persist.
Objective. To assess the performance of Large Language Models like ChatGPT and Bard to produce references in the context of scientific writing.
Methods. The performance of ChatGPT and Bard in replicating the results of human-conducted systematic reviews was assessed. Using systematic reviews pertaining to shoulder rotator cuff pathology, these LLMs were tested by providing the same inclusion criteria and comparing the results with original systematic review references, serving as gold standards. The study utilized three key performance metrics: Recall, Precision, and F1-score, alongside the hallucination rate. Articles were considered “hallucinated” if any two of the following information were wrong: title, first author, or year of publication.
Results. Eleven systematic reviews across four fields yielded 33 prompts to LLMs (3 LLMs x 11 reviews), with 471 references analyzed. Precision rates for GPT-3.5, GPT-4, and Bard were 9.4%, 13.4%, and 0% respectively (P<.001). Recall rates were 11.9% for GPT-3.5 and 14.7% for GPT-4, with Bard failing to retrieve any relevant articles (P<.001). Hallucination rates stood at 39.6% for GPT-3.5, 28.6% for GPT-4, and 91.4% for Bard (P<.001). Further analysis of non-hallucinated articles retrieved by GPT models revealed significant differences in identifying various criteria, such as randomized studies, participant criteria, and intervention criteria. The study also noted the geographical and open-access biases in the articles retrieved by the LLMs.
Conclusions. Given their current performance, it is not recommended for LLMs to be deployed as the primary or exclusive tool for conducting systematic reviews. Any references generated by such models warrant thorough validation by researchers. The high occurrence of hallucinations in LLMs highlights the necessity for refining their training and functionality before confidently employing them for rigorous academic purposes.
Keywords: artificial intelligence; large language models; ChatGPT; Bard; rotator cuff

Introduction

The advent of artificial intelligence (AI) has led to significant advancements in various fields, including medical research. Large language models (LLMs), such as ChatGPT, could assist academic researchers in a variety of tasks, including writing scientific articles. These models have the potential to streamline the way researchers conduct literature searches, synthesize findings, and draft systematic reviews [1]. However, there is ongoing debate surrounding their reliability, ethical considerations, and appropriate use in academic publishing.

Recently, editorials and opinion articles have been published addressing the use of LLMs in the scientific community. One such example is an editorial in The Lancet Digital Health, which discusses the potential benefits and challenges of implementing AI in medical research [2]. As the application of LLMs like ChatGPT in research settings grows, concerns have arisen regarding their accuracy, the potential for generating misleading or false information, and the ethical implications of using AI-generated content without proper disclosure.

While it is known that ChatGPT can help researchers write articles [3–5], controversy exists about whether it should be used at all, whether its use should be disclosed, and if it should be listed as an author or not [6]. These debates raise important questions about the role of AI in scientific research and the potential consequences of using LLMs in generating systematic reviews and other research outputs [7].

In this study, we aim to address these concerns by systematically evaluating the reliability of ChatGPT in the context of searching for and synthesizing peer-reviewed literature for systematic reviews. We will compare its performance to that of traditional methods used by researchers, investigate the extent of the "hallucination" phenomenon, and discuss potential ethical and practical considerations for using ChatGPT in academic publishing. By providing evidence-based insights into the capabilities and limitations of LLMs in medical research, we hope to contribute to the ongoing debate about the role of AI in the research ecosystem and guide researchers in making informed decisions about using LLMs in their work.

Methods

Study Design

This study follows a sequential design, chosen for its ability to progressively build on each preceding phase, thus ensuring a comprehensive evaluation of the LLMs in the context of a systematic review. The process initiated with a systematic review search on PubMed, followed by retrieval of selected articles. Subsequently, the methodology of these articles served as inputs to the LLM which is tasked to search for articles using the same inclusion criteria as the systematic reviews. The final phase involves comparison of the LLM results with the systematic review references, which act as the ground truth, thus providing a robust evaluation of the LLMs' ability to replicate the results of human-conducted systematic reviews.
The ethical considerations of using AI, specifically LLMs, in research were carefully evaluated. All articles accessed by the LLMs were publicly available, and no proprietary or subscription-based sources were used without appropriate access rights. LLMs did not have access to any sensitive or private patient data. The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Systematic Review Search on PubMed

On the 27th of July 2023, a literature search was performed on PubMed to find literature published in the English language during the year 2020. The selected year aligns with ChatGPT's training cut-off point in September 2021, ensuring that the AI model has access to the comprehensive scope of literature for the given year. The focus was directed towards systematic reviews of randomized clinical trials pertaining to shoulder rotator cuff pathology. This prevalent condition spans multiple disciplines inclusive of surgery, anesthesiology, sports medicine, and physical therapy, thereby positioning it as an optimal candidate for this multidisciplinary appraisal. In addition, the collective clinical and scientific experience of the research team on the topic furnished a critical review of the references obtained from the PubMed search and from the LLMs [8–11].

An electronic search of PubMed was conducted using a combination of keywords, including “shoulder,” “rotator cuff,” and “randomized” (see Appendix 1). The search was restricted to articles published in 2020 and filtered to retrieve only systematic reviews and meta-analyses. Titles and abstracts were scrutinized, and articles indicating a systematic review of randomized studies on rotator cuff pathology were selected for further analysis.

Exclusion criteria were applied to eliminate articles that did not meet our study focus. Articles were excluded if they were not systematic reviews, if their primary concern did not pertain to rotator cuff pathology, if written in a language other than English, or if they included non-randomized clinical studies.

Two independent reviewers (MC and PB) screened titles, abstracts, and full texts retrieved by this query. Differences between reviewers were reconciled with a third reviewer (JD). To ensure the selection of relevant systematic reviews, the reviewers applied exclusion criteria that consisted of systematic reviews including non-randomized studies and articles that were not systematic reviews. The eligibility of the selected systematic reviews was further validated by assessing their adherence to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [12] . Additionally, we verified the registration status of these reviews in the PROSPERO database [13].

For each article referenced in the systematic reviews, we collected information on the article title, author list, country (based on the first author's affiliation with PubMed), journal name, journal date and issue, DOI, and open access status. We assessed the hypothesis that LLMs may favor publicly available articles in their results by using a broad definition of "open access." This definition included open access through the journal or any full-text PDF available on another server and accessible through a Google search (e.g., ResearchGate or university website).

Systematic Review on LLMs: Article Retrieval

For each new request, a fresh chatbot session was initiated to prevent any carryover effect from previous queries, ensuring the validity of the results. We prompted ChatGPT and BARD with a precise query to identify articles that could be included in the systematic review. The structure of the prompt consisted of a statement about the physician's and researcher's current work, followed by the inclusion criteria for the studies in the review (Figure 1). The criteria specified randomized controlled trials with specific participant criteria and interventions comparing two different treatments. LLMs were asked to provide references to randomized studies on the topic, excluding articles published after 2020 and systematic reviews or meta-analyses. To assess the impact of the prompt's specificity on the search results of LLMs, we tested two versions of the prompt for each request. One specifying the minimum number of articles to be found and the other without specifying this minimum number, thus providing us with an opportunity to assess if the presence or absence of a target number influences the LLMs' search results. The query that led to the largest number of results was retained for this study.

For each article provided by LLMs, we collected information on the existence or hallucination status of the article, authors' list, country (based on the first author's affiliation on PubMed), open access status, inclusion in the original systematic review, randomization status, participant criteria adherence, intervention criteria adherence, exclusion of systematic reviews (as requested in the prompt), and accuracy of the provided information (authors' list, journal, year and issue, title, and DOI). We also verified if the article was published before 2021, as requested in the prompt.

Articles were considered hallucinated if any two of the following information were wrong: title, first author, or year of publication. The hallucination rate was calculated to quantify the proportion of LLM-generated references that were irrelevant, incorrect, or unsupported by available literature, offering insights into the extent of spurious or inaccurate information production by the LLMs.
For non-comparative studies, the intervention criteria were considered adequate if at least one of the two interventions was studied in the proposed reference. For comparative studies, the intervention criteria were considered adequate if both interventions were studied in the proposed reference.

Comparison of LLMs Results

The sample size was determined based on an anticipated 10% rate of systematic review references overlooked by Large Language Models (LLMs), with an assumed power of 90% and an alpha of 0.05. This calculation yielded a requisite of 80 references for the comparison. The Pubmed search yielded eleven systematic reviews (Figure 2), each with an average of 9.9 references (SD: 6.6, range: 3 to 23). The evaluation of the Large Language Models (LLMs) was predicated on three widely utilized metrics: 1) Recall, representing the proportion of genuinely pertinent articles from the original systematic reviews accurately identified and retrieved by the LLMs, 2) Precision, quantifying the proportion of articles retrieved by the LLMs that are verifiably present in the original systematic reviews, and 3) F1-score, which serves as an aggregate metric encapsulating both the recall and precision values.

	Articles provided as an output by LLMs	Articles not provided as an output by LLMs
Articles cited by systematic reviews	True Positive (TP)	False Negative (FN)
Articles not cited by systematic reviews	False Positive (FP)	True Negative (TN)

$$Recall=\frac{TP}{TP+FN}$$

$$Precision=\frac{TP}{TP+FP}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

The LLMs incorporated in this study included ChatGPT 3.5 (*text-davinci-002-render-sha, July 19 version, OpenAI, San Francisco, USA*), ChatGPT 4 (*gpt-4-32k-0314, July 19 version, OpenAI, San Francisco, USA*) [14] and Bard (*PaLM version 2.0, released on July 13, 2023, Google AI, San Francisco, USA*). We conducted Chi-squared tests to compare each piece of information extracted from LLMs' responses, including authors' nationalities and the open-access status of the retrieved articles. The significance threshold used was P < .05. Statistical analysis was performed with EasyMedStat (www.easymedstat.com; version 3.24, France).

Use of Chat GPT for this manuscript

The entirety of this manuscript was composed by the authors of this research. As non-native English speakers, the authors utilized ChatGPT to refine the English language utilized in the manuscript [15]. Importantly, all modifications suggested by ChatGPT underwent meticulous evaluation and approval by the authors to ensure accuracy and clarity. ChatGPT was not used for bibliographic reference retrieval.

Results

Eleven systematic reviews were identified in four fields (Table 1): physiotherapy (3 articles), sports medicine (3 articles), orthopedic surgery (3 articles) and anesthesiology (2 articles), leading to 33 prompts to LLMs (3 tested LLMs x 11 systematic reviews). LLM prompts returned references in 32 out of 33 cases: Bard did not return any result for the systematic review about “subacromial analgesia via continuous infusion catheter”. In most cases, the number of references returned by LLMs was greater or equal to that of original articles (Table 1). Overall, 471 references were included in this study and analyzed.
Articles identified by LLMs were present in the original systematic reviews (precision) in 9.4%, 13.4% and 0% of cases for GPT-3.5, GPT-4 and Bard (*P*<.001). Conversely, 11.9% of articles from the systematic reviews (recall) were retrieved by GPT-3.5 and 14.7% by GPT-4. No article from the systematic reviews were retrieved by Bard *(P*<.001) (Table 2).
The hallucination rates were respectively 39.6%, 28.6% and 91.4% for GPT-3.5, GPT-4 and Bard (*P*<.001).
When analyzing the articles retrieved by GPT that were not hallucinated (n=84 for GPT-3.5 and n=85 for GPT-4 respectively), the following criteria were successfully identified (Figure 3): randomized studies (39.3% vs 49.4%; *P*=.242), participant criteria (57.1% vs 67.1%; *P*=.242), intervention criteria (69.1% vs 84.7%; *P*=.026), not a systematic review (81.0% vs 77.7%; *P*=.734) and published before 2021 (100% vs 100%; *P*>.99). Nine articles retrieved by Bard were not hallucinated. This limited sample was not appropriate for further inferential statistics.
Regarding the same non-hallucinated articles retrieved by GPT, the following bibliographic information were considered accurate (Figure 4): authors list (86.9% vs 87.1%; *P*>.99), journal title (96.4% vs 100%; *P*=.121), date and issue (84.5% vs 95.3%; *P*=.023), article title (98.8% vs 98.8%; *P*>.99) and DOI (15.9% vs 20.2%; *P*=.595).
Open access articles were selected in 27.5% of original systematic reviews, 38.1% of GPT-3.5

articles and 36.5% of GPT-4 articles (P=.237). Articles from American (USA) authors were selected in 16.5% of original systematic reviews, 44.1% of GPT-3.5 articles and 32.9% of GPT-4 articles (P<.001).

Discussion

The most important finding of the present study is that using large language models like ChatGPT and Bard to conduct systematic reviews for a common condition such as rotator cuff disease can generate misleading or 'hallucinated' references, exceeding a 25% rate.

This concern has been broached in previous literature [16–19], but our study provides an experimental design to probe the matter more deeply. OpenAI, the developer of ChatGPT, acknowledges this issue, stating that their model “occasionally generates plausible but incorrect or nonsensical responses” [20]. As LLMs increasingly assist academic researchers in producing scientific literature, this phenomenon warrants careful scrutiny.

When comparing the three models tested, ChatGPT 4 was the most efficient to retrieve non-hallucinated references, while ChatGPT 3.5 produced almost 40% of non-existing references. Bard, however, appears ill-suited for conducting systematic reviews in the selected areas, with over 90% of the references failing to correlate with legitimate articles. Bard seemed to have a try-and-repeat approach, providing multiples versions of hallucinated articles with close titles and journal name (Figure 5).

Despite this, LLMs typically encouraged users to conduct their own systematic reviews, recognizing the necessity of human involvement. However, in none of our queries, did the LLMs ask to verify the authenticity of the produced citations. Nonetheless, the convincing verisimilitude of the references generated by LLMs presents a risk for incautious researchers, potentially undermining the quality of scientific bibliographies if improperly employed (Figures 5 and 6). Moreover, the efficiency of LLMs in retrieving original articles from systematic reviews ranged from negligible to modest (0%-14%), emphasizing that researchers should not overly rely on these tools for systematic reviews. Nevertheless, in numerous instances, both ChatGPT and Bard “encouraged [users] to conduct their own research” (Figure 5), a suggestion that appears crucial considering the findings of this study.

It could be expected that LLMs were not able to retrieve the same references as authors of systematic reviews. However, our study also reveals that LLMs, despite being provided with the same eligibility criteria as those in the original systematic reviews, were not able to consistently apply them. For instance, the criterion of "randomized study" was adhered to in only 40% to 50% of non-hallucinated articles generated by ChatGPT, even when the term "randomized" appeared in the title or abstract of the articles from the original systematic reviews. The same finding was observed for the “not a systematic review” criterion which was not respected in about 20% of cases, while the publicly available information of the produced articles clearly states the nature of these studies.

These discrepancies could potentially stem from the underlying statistical nature of these LLMs, which predict subsequent text (tokens) based on a model reinforced by human feedback [21]. However, as human supervision does not extend to validating the accuracy of LLM outputs, especially in specialized fields like medicine, inaccuracies can prevail.

In the case of non-hallucinated articles, however, ChatGPT demonstrated significant efficiency in retrieving accurate bibliographic information like the exact article title, the authors list, and the journal title.

Potential biases in LLMs due to training on biased datasets and the risk of perpetuating stereotypes have been highlighted [2]. Our findings suggest that American authors were more frequently represented in ChatGPT references. However, further investigation across diverse medical fields is warranted to ascertain whether these LLMs may introduce such biases

definitively.
This investigation, by virtue of its specific and circumscribed parameters, comes with several inherent limitations. The scope of the study was exclusively focused on systematic reviews related to shoulder rotator cuff pathology. Consequently, it must be recognized that the findings might not be universally applicable across diverse medical specialties or disciplines. The examination was also restricted to three large language models, specifically ChatGPT 3.5, ChatGPT 4, and Bard. The landscape of available language models is vast and continually evolving, and it is conceivable that different models might yield divergent results. In addition, the field lacks established guidelines for leveraging LLMs to optimize accuracy. Notwithstanding rigorous attempts to devise specific, comprehensive prompts, it remains plausible that alternative queries could generate more precise outcomes. This fact underscores the multifaceted nature of the challenge and the need for further research in this domain.
The choice of prompt plays a crucial role in determining the output generated by LLMs. During the exploratory phase of our study, various prompt versions were tested. While our study did not focus on identifying the optimal prompts, several techniques employed in our prompts appeared to enhance output quality: specifying a minimum number of articles ("a minimum of 9 articles"), utilizing bullet points to delineate criteria such as "type of studies," "participants," and "interventions," and explicitly instructing to "Exclude systematic reviews and meta-analyses." Introducing prompts by specifying the researcher's profession provides additional context, aligning with recommendations from LLM providers. Lastly, enforcing a specific reference style format facilitated the retrieval of vital information, including authors' names, journal titles, publication dates, and DOIs when available.
Our decision not to provide the initial PubMed results list to LLMs for assessing paper eligibility was deliberate, aimed at preserving study integrity and interpretability. While providing the list might enhance LLM accuracy, it introduces bias by guiding models towards replicating the provided set rather than autonomously identifying relevant studies. Our study design, though sacrificing some precision, ensures that LLM results reflect genuine capabilities in navigating scientific literature independently.

Large language models present a highly efficient instrument that may aid academics in the drafting of research articles. However, upon analyzing the findings of this study, it becomes imperative to emphasize that the bibliographic references proposed by the AI are not intrinsically trustworthy. These citations necessitate human validation, focusing on the authors, the title, and the subject matter.
We thereby deduce that, in the context of GPT iterations, user verification is indispensable for preserving the scientific integrity and relevance of the output. A statement or a scholarly usage guideline should be prominently featured before the tool is utilized or should be integrated into the software itself, to outline its lack of liability for any inaccuracies in the citation of articles. This is paramount as such errors could potentially mislead a considerable number of users. We also propose that the application of GPT-based chatbots for tasks such as spelling correction, proofreading, or text restructuring ought to be explicitly mentioned within the Materials and Methods section of academic writings.

Conclusion

ChatGPT and Bard exhibit the capacity to generate convincingly authentic references for systematic reviews but also yield hallucinated articles in 28% to 91% of cases. Among the models tested, ChatGPT 4 displayed superior performance in generating legitimate and relevant references but, like the other models, largely failed to respect the established eligibility criteria. Given their current state, LLMs such as ChatGPT and Bard should not be utilized as the sole or primary means for conducting systematic reviews of literature and it is crucial that references generated by these tools undergo rigorous validation by the authors of scientific articles.

Acknowledgments

Authors' Contributions: MC, PB screened titles, abstracts, and full texts retrieved by PubMed search. Differences between reviewers were reconciled with a third reviewer (JD).
CRC designed the study, drafted, and critically reviewed the manuscript.
MC, JD, VL, CT, MA, MD, JLR, GC, PB, CRC confirm that they had full access to all the data in the study and accept responsibility to submit for publication.
MC, JD, VL, CT, MA, MD, JLR, GC, VL, PB, CRC contributed commented on, revised, and approved the final version to data acquisition.

Role of Funding Source: None

Conflict of Interest Statements: The authors have nothing to declare.

Ethics Committee approval: Not required

Figures and Tables

Table 1. Systematic Reviews Included in the Study and the Count of Articles Retrieved by Original Authors and Large Language Models.
Table 2. Evaluative metrics of the assessed large language models.
Figure 1. Captured screenshots demonstrating a prompt to large language models.
Figure 2. Flow diagram of included systematic reviews.
Figure 3. Efficiency of the tested large language models in complying to inclusion and exclusion criteria. At the exception of the "Article exists" criteria, hallucinated articles were excluded from this analysis.
Figure 4. Efficiency of the tested large language models in generating accurate bibliographic information of the retrieved articles. Hallucinated articles were excluded from this analysis.
Figure 5. Instances of hallucinated articles. Four out of five article titles commence with "Tranexamic acid for the prevention of bleeding in arthroscopic" and were allegedly published in the journal "Arthroscopy".
Figure 6. Instance of a hallucinated reference. "A" presents the output of a large language model. "B" and "C" showcase authentic articles with similarities in title and author list, potentially serving as original data for large language model reference generation.

References

1. Lund BD, Wang T. Chatting about ChatGPT: how may AI and GPT impact academia and libraries? Library Hi Tech News Emerald Publishing Limited; 2023 Jan 1;40(3):26–29. doi: 10.1108/LHTN-01-2023-0009

2. Health TLD. ChatGPT: friend or foe? The Lancet Digital Health Elsevier; 2023 Mar 1;5(3):e102. PMID:36754723

3. Else H. Abstracts written by ChatGPT fool scientists. Nature 2023 Jan 12;613(7944):423–423. doi: 10.1038/d41586-023-00056-7

4. Biswas S. ChatGPT and the Future of Medical Writing. Radiology 2023 Apr;307(2):e223312. PMID:36728748

5. Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? Critical Care 2023 Feb 25;27(1):75. doi: 10.1186/s13054-023-04380-2

6. Stokel-Walker C. ChatGPT listed as author on research papers: many scientists disapprove. Nature 2023 Jan;613(7945):620–621. PMID:36653617

7. Zheng H, Zhan H. ChatGPT in Scientific Writing: A Cautionary Tale. The American Journal of Medicine Elsevier; 2023 Mar 9;0(0). PMID:36906169

8. Azar M, Van der Meijden O, Pireau N, Chelli M, Gonzalez J-F, Boileau P. Arthroscopic Revision Cuff Repair: Do Tendons have a Second Chance to Heal? Journal of Shoulder and Elbow Surgery 2022 Jun;S1058274622004888. doi: 10.1016/j.jse.2022.04.024

9. Boileau P, Andreani O, Schramm M, Baba M, Barret H, Chelli M. The Effect of Tendon Delamination on Rotator Cuff Healing. Am J Sports Med 2019 Apr;47(5):1074–1081. PMID:30943087

10. Muccioli C, Chelli M, Caudal A, Andreani O, Elhor H, Gauci M-O, Boileau P. Rotator cuff integrity and shoulder function after intra-medullary humerus nailing. Orthopaedics & Traumatology: Surgery & Research 2019 Dec;S1877056819303755. doi: 10.1016/j.otsr.2019.11.004

11. Boileau P, Baqué F, Valerio L, Ahrens P, Chuinard C, Trojani C. Isolated arthroscopic biceps tenotomy or tenodesis improves symptoms in patients with massive irreparable rotator cuff tears. J Bone Joint Surg Am 2007 Apr;89(4):747–757. PMID:17403796

12. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JPA, Clarke M, Devereaux PJ, Kleijnen J, Moher D. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 2009 Dec 4;339(jul21 1):b2700–b2700. doi: 10.1136/bmj.b2700

13. Booth A, Clarke M, Dooley G, Ghersi D, Moher D, Petticrew M, Stewart L. The nuts and bolts of PROSPERO: an international prospective register of systematic reviews. Syst Rev 2012 Feb 9;1:2. PMID:22587842

14. OpenAI. GPT-4 Technical Report. arXiv; 2023. doi: 10.48550/arXiv.2303.08774

15. Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. The Lancet Digital Health Elsevier; 2023 Apr 27;0(0). PMID:37120418

16. Hillier M. Why does ChatGPT generate fake references? TECHE. 2023. Available from: https://teche.mq.edu.au/2023/02/why-does-chatgpt-generate-fake-references/ [accessed May 17, 2023]

17. Gravel J, D'Amours-Gravel M, Osmanlliu E. Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions. medRxiv; 2023. p. 2023.03.16.23286914. doi: 10.1101/2023.03.16.23286914

18. Day T. A Preliminary Investigation of Fake Peer-Reviewed Citations and References Generated by ChatGPT. The Professional Geographer Routledge; 2023 Apr 12;0(0):1–4. doi: 10.1080/00330124.2023.2190373

19. Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. Cureus 2023 Feb;15(2):e35179. PMID:36811129

20. OpenAI. Introducing ChatGPT. 2022. Available from: https://openai.com/blog/chatgpt [accessed May 18, 2023]

21. Stiennon N, Ouyang L, Wu J, Ziegler DM, Lowe R, Voss C, Radford A, Amodei D, Christiano P. Learning to summarize from human feedback. arXiv; 2022. doi: 10.48550/arXiv.2009.01325

22. Lähdeoja T, Karjalainen T, Jokihaara J, Salamh P, Kavaja L, Agarwal A, Winters M, Buchbinder R, Guyatt G, Vandvik PO, Ardern CL. Subacromial decompression surgery for adults with shoulder pain: a systematic review with meta-analysis. Br J Sports Med 2020 Jun;54(11):665–673. PMID:30647053

23. Catapano M, Zhang K, Mittal N, Sangha H, Onishi K, de Sa D. Effectiveness of Dextrose Prolotherapy for Rotator Cuff Tendinopathy: A Systematic Review. PM R 2020 Mar;12(3):288–300. PMID:31642203

24. Gutiérrez-Espinoza H, Araya-Quintanilla F, Cereceda-Muriel C, Álvarez-Bueno C, Martínez-Vizcaíno V, Cavero-Redondo I. Effect of supervised physiotherapy versus home exercise program in patients with subacromial impingement syndrome: A systematic review and meta-analysis. Physical Therapy in Sport 2020 Jan 1;41:34–42. doi: 10.1016/j.ptsp.2019.11.003

25. Chen X, Jones IA, Togashi R, Park C, Vangsness CT. Use of Platelet-Rich Plasma for the Improvement of Pain and Function in Rotator Cuff Tears: A Systematic Review and Meta-analysis With Bias Assessment. Am J Sports Med 2020 Jul;48(8):2028–2041. PMID:31743037

26. An VVG, Farey JE, Karunaratne S, Smithers CJ, Petchell JF. Subacromial analgesia via continuous infusion catheter vs. placebo following arthroscopic shoulder surgery: a systematic review and meta-analysis of randomized trials. J Shoulder Elbow Surg 2020 Mar;29(3):471–482. PMID:32067710

27. Craig RS, Goodier H, Singh JA, Hopewell S, Rees JL. Shoulder replacement surgery for osteoarthritis and rotator cuff tear arthropathy. Cochrane Database Syst Rev 2020 Apr 21;4(4):CD012879. PMID:32315453

28. Naunton J, Street G, Littlewood C, Haines T, Malliaras P. Effectiveness of progressive and resisted and non-progressive or non-resisted exercise in rotator cuff related shoulder pain: a systematic review and meta-analysis of randomized controlled trials. Clin Rehabil 2020 Sep;34(9):1198–1216. PMID:32571081

29. Malliaras P, Johnston R, Street G, Littlewood C, Bennell K, Haines T, Buchbinder R. The Efficacy of Higher Versus Lower Dose Exercise in Rotator Cuff Tendinopathy: A Systematic Review of Randomized Controlled Trials. Arch Phys Med Rehabil 2020 Oct;101(10):1822–1834. PMID:32682937

30. Simpson M, Pizzari T, Cook T, Wildman S, Lewis J. Effectiveness of non-surgical interventions for rotator cuff calcific tendinopathy: A systematic review. J Rehabil Med 2020 Oct 31;52(10):jrm00119. PMID:32830280

31. Belk JW, McCarty EC, Houck DA, Dragoo JL, Savoie FH, Thon SG. Tranexamic Acid Use in Knee and Shoulder Arthroscopy Leads to Improved Outcomes and Fewer Hemarthrosis-Related Complications: A Systematic Review of Level I and II Studies. Arthroscopy 2021 Apr;37(4):1323–1333. PMID:33278534

32. Belk JW, Kraeutler MJ, Houck DA, Chrisman AN, Scillia AJ, McCarty EC. Biceps tenodesis versus tenotomy: a systematic review and meta-analysis of level I randomized controlled trials. J Shoulder Elbow Surg 2021 May;30(5):951–960. PMID:33373685

Tables

Table 1. Systematic Reviews Included in the Study and the Count of Articles Retrieved by Original Authors and Large Language Models.

Systematic review	Field	PRISMA guidelines	PROSPERO registration	Number of articles in the original article	Number of articles returned by GPT 3.5	Number of articles returned by GPT 4	Number of articles returned by Bard
Lähdeoja T [22]	Surgery	Yes	Yes	9	10	10	9
Catapano [23]	Sports medicine	Yes	No	5	7	5	5
Gutiérrez-Espinoza [24]	Physiotherapy	Yes	Yes	7	15	7	7
Chen [25]	Sports medicine	Yes	No	18	18	18	16
An [26]	Anesthesiology	Yes	Yes	9	9	9	0
Craig [27]	Surgery	Yes	No	23	22	22	23
Naunton [28]	Physiotherapy	Yes	Yes	7	10	8	7
Malliaras [29]	Physiotherapy	Yes	Yes	3	5	7	5
Simpson [30]	Sports medicine	Yes	Yes	18	18	18	23
Belk [31]	Anesthesiology	Yes	No	5	10	8	5
Belk [32]	Surgery	Yes	No	5	15	7	5
Total		11/11	6/11	109	139	119	104

Table 2. Evaluative metrics of the assessed large language models

Metric	ChatGPT 3.5	ChatGPT 4	Bard
True positive (TP)	13	16	0
False positive (FP)	126	103	104
False negative (FN)	96	93	109
Recall $\frac{TP}{TP+FN}$	11·9%	13·4%	0%
Precision $\frac{TP}{TP+FP}$	9·4%	14·7%	0%

F1-score $\frac{TP}{TP+\frac{1}{2}(FN+FP)}$	10·5%	14·0%	0%

Figures

Figure 1 - Prompt example

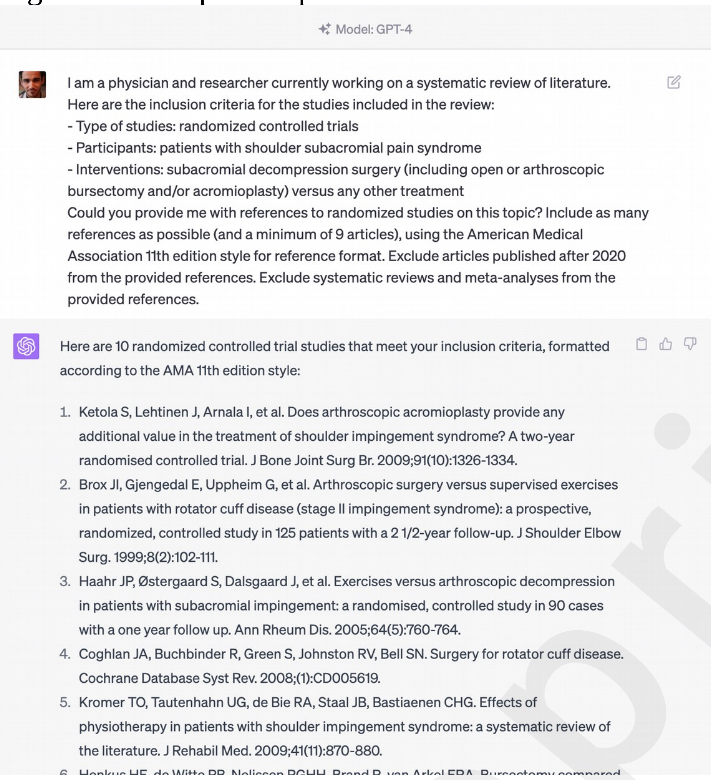


Figure 2 – Flowchart

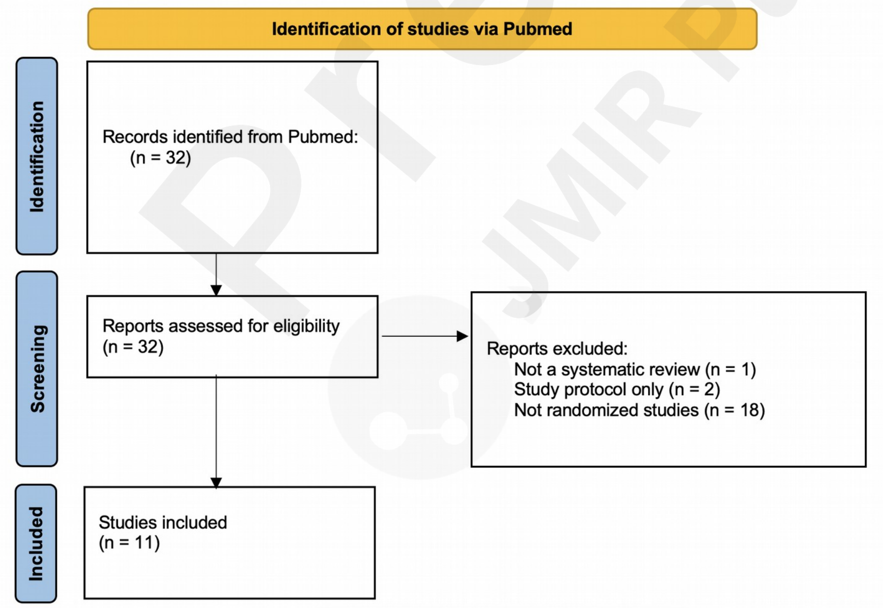


Figure 3 – Criteria

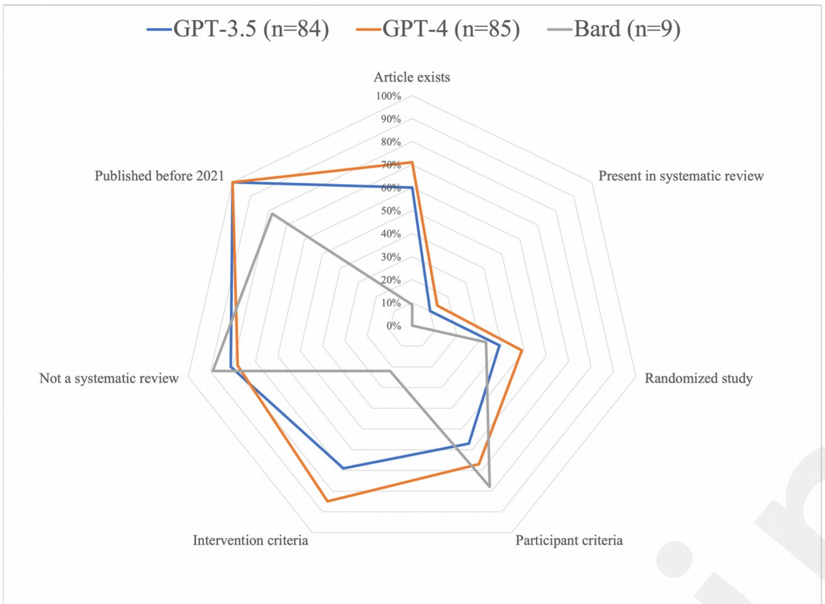


Figure 4 - Exact data

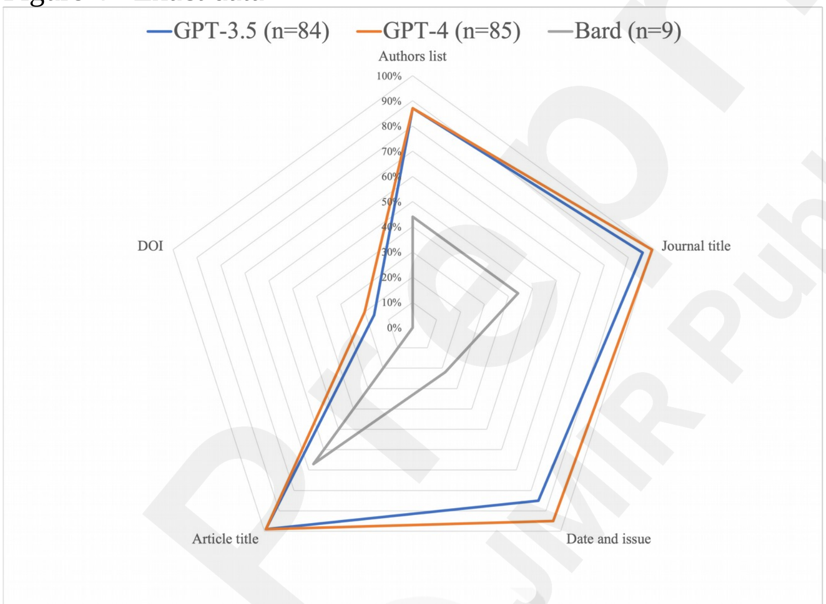


Figure 5 - Obviously hallucinated

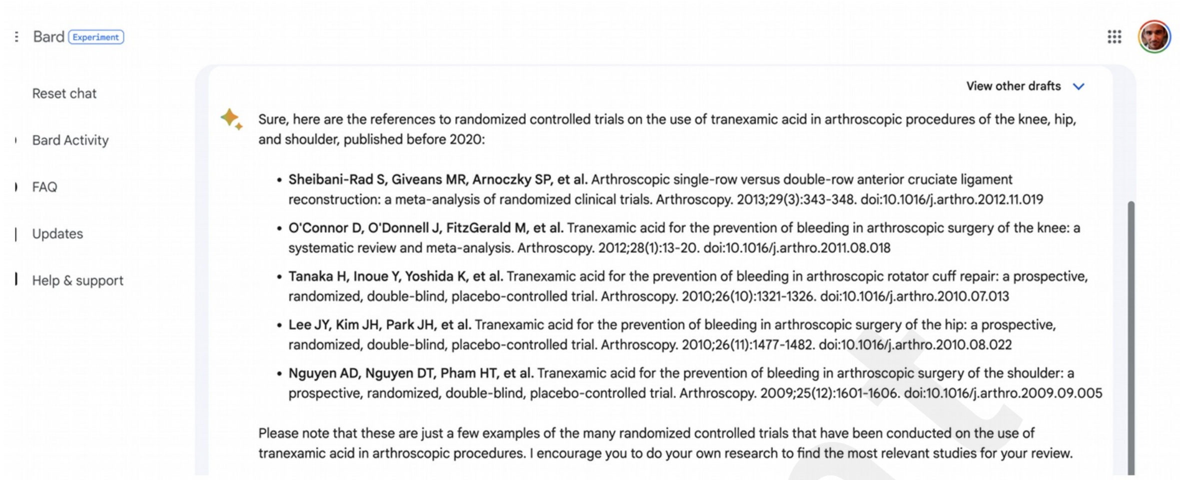
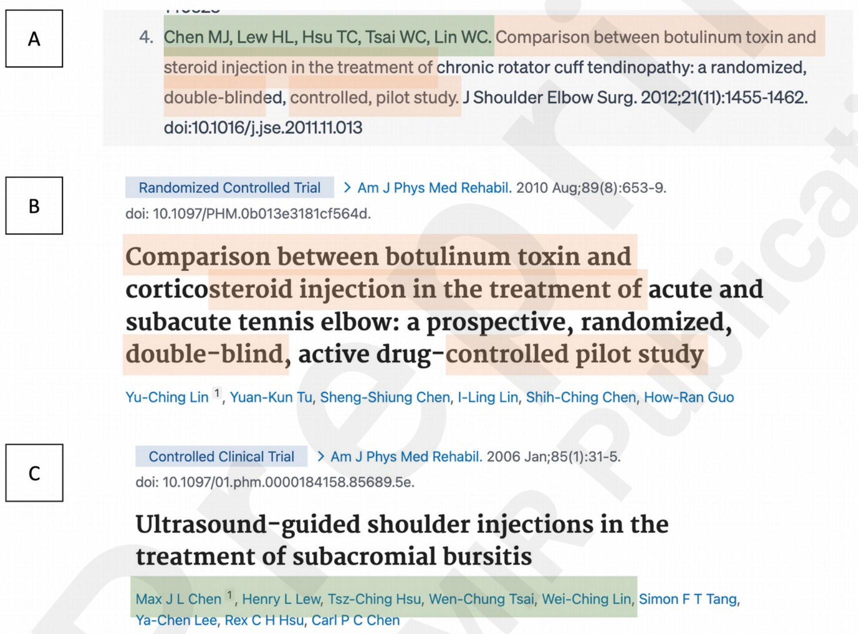


Figure 6 - Hallucinated Reference Example



Appendix

Search strategy

(("shoulder"[MeSH Terms] OR "shoulder"[All Fields] OR "shoulders"[All Fields] OR "shoulder s"[All Fields]) AND ("rotator cuff"[MeSH Terms] OR ("rotator"[All Fields] AND "cuff"[All Fields]) OR "rotator cuff"[All Fields]) AND ("random allocation"[MeSH Terms] OR ("random"[All Fields] AND "allocation"[All Fields]) OR "random allocation"[All Fields] OR "randomization"[All Fields] OR "randomized"[All Fields] OR "random"[All Fields] OR "randomisation"[All Fields] OR "randomisations"[All Fields] OR "randomise"[All Fields] OR "randomised"[All Fields] OR "randomising"[All Fields] OR "randomizations"[All Fields] OR "randomize"[All Fields] OR "randomizes"[All Fields] OR "randomizing"[All Fields] OR "randomness"[All Fields] OR "randoms"[All Fields])) AND ((systematicreview[Filter]) AND (2020:2020[pdat]))

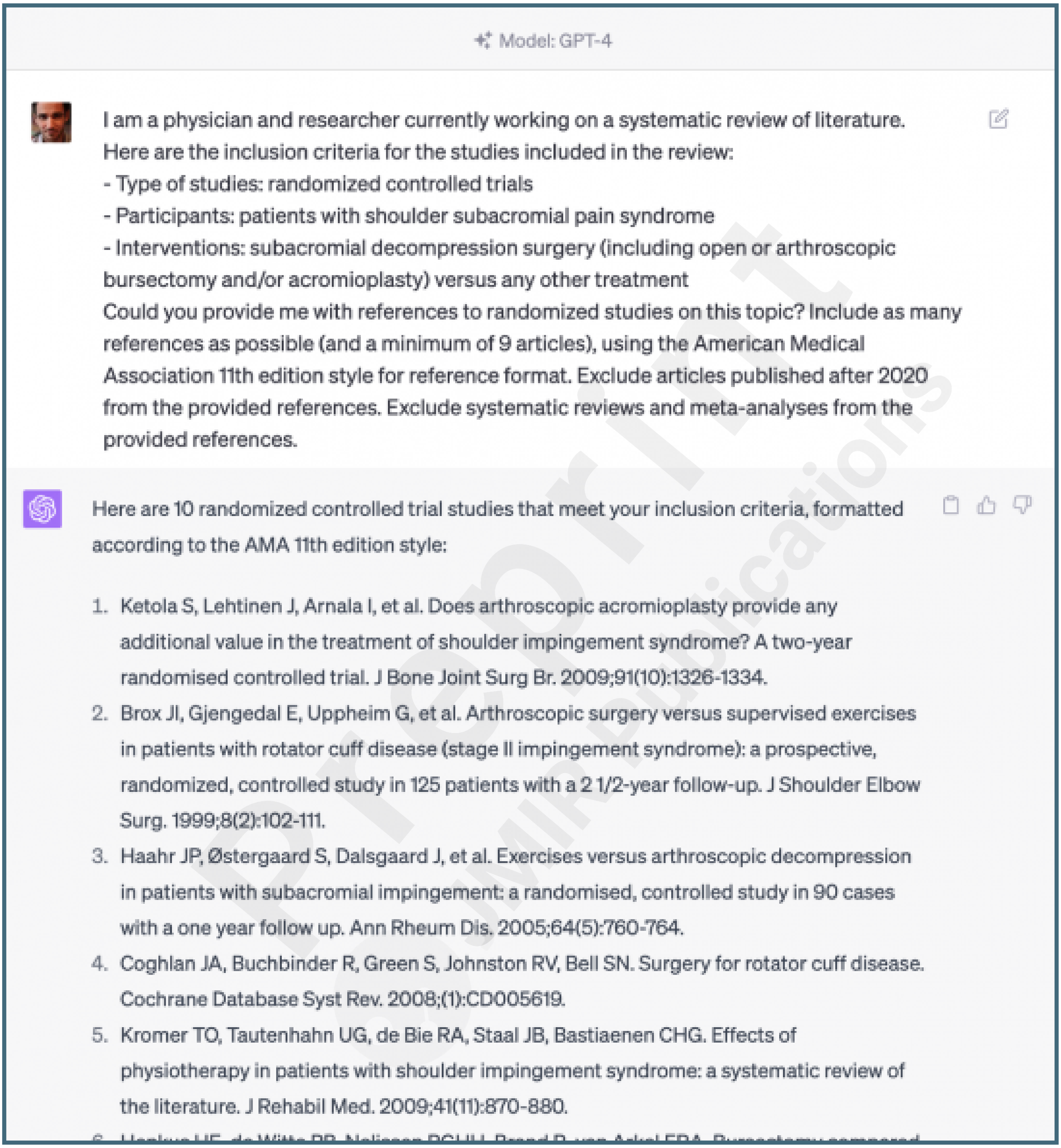
Supplementary Files

Untitled.
URL: http://asset.jmir.pub/assets/0f8da4adbbd97b55012cd53d9d2b6176.docx

Untitled.
URL: http://asset.jmir.pub/assets/3fe2b48e9dae33c22a1bc98f2e681b5f.docx

Figures

Prompt example.



Model: GPT-4

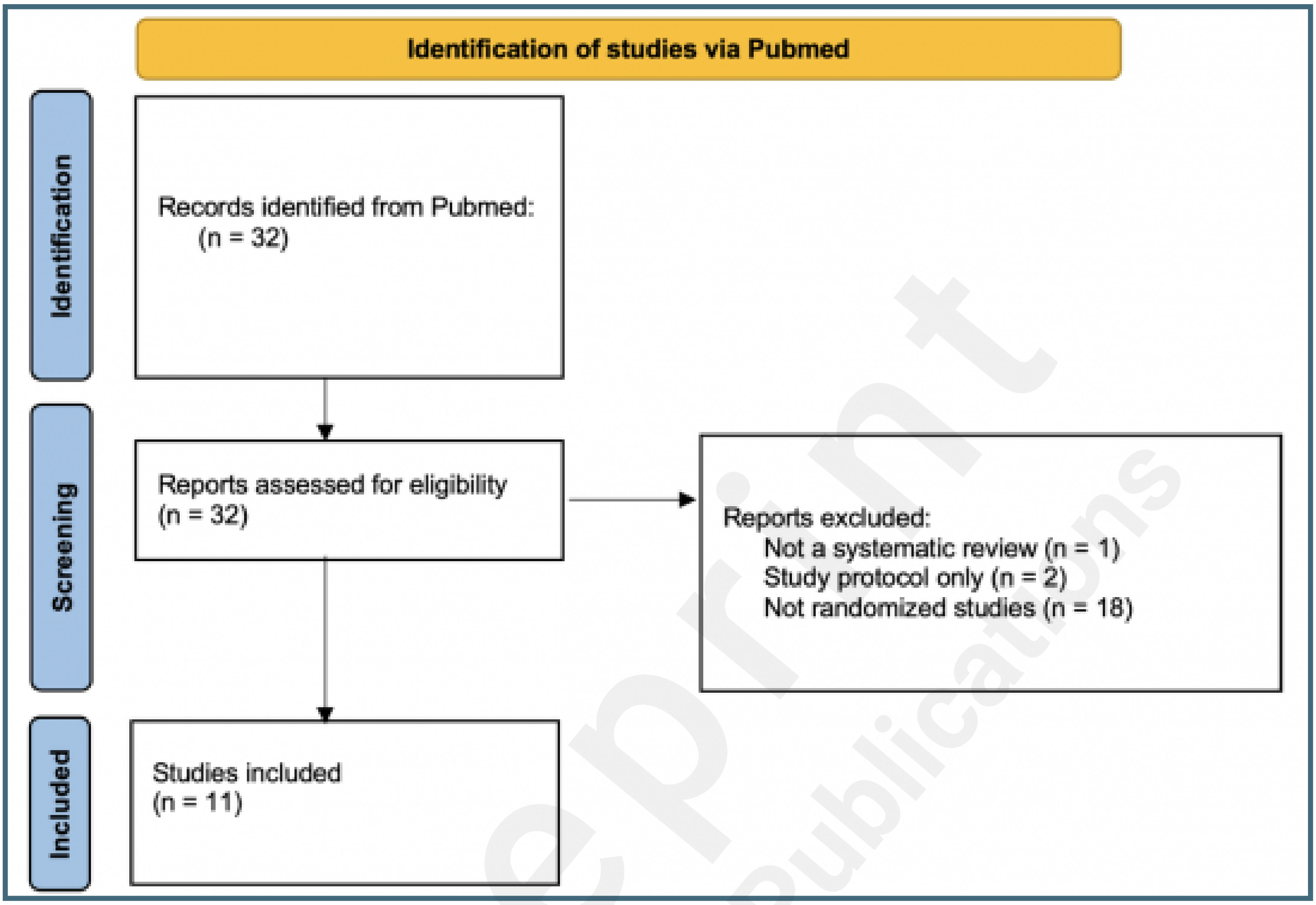
I am a physician and researcher currently working on a systematic review of literature.
Here are the inclusion criteria for the studies included in the review:
- Type of studies: randomized controlled trials
- Participants: patients with shoulder subacromial pain syndrome
- Interventions: subacromial decompression surgery (including open or arthroscopic bursectomy and/or acromioplasty) versus any other treatment

Could you provide me with references to randomized studies on this topic? Include as many references as possible (and a minimum of 9 articles), using the American Medical Association 11th edition style for reference format. Exclude articles published after 2020 from the provided references. Exclude systematic reviews and meta-analyses from the provided references.

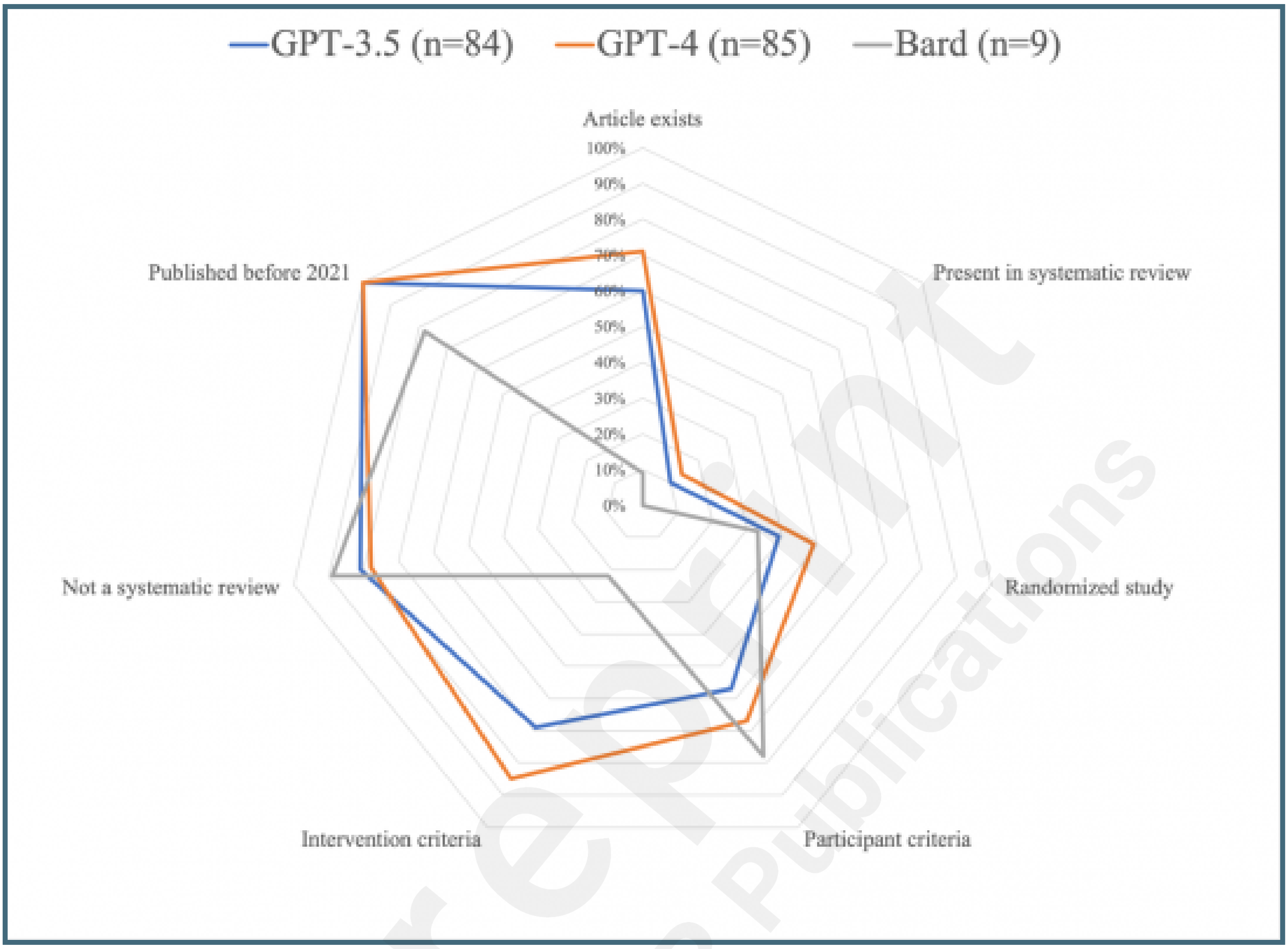
Here are 10 randomized controlled trial studies that meet your inclusion criteria, formatted according to the AMA 11th edition style:

1. Ketola S, Lehtinen J, Arnala I, et al. Does arthroscopic acromioplasty provide any additional value in the treatment of shoulder impingement syndrome? A two-year randomised controlled trial. J Bone Joint Surg Br. 2009;91(10):1326-1334.
2. Brox JI, Gjengedal E, Uppheim G, et al. Arthroscopic surgery versus supervised exercises in patients with rotator cuff disease (stage II impingement syndrome): a prospective, randomized, controlled study in 125 patients with a 2 1/2-year follow-up. J Shoulder Elbow Surg. 1999;8(2):102-111.
3. Haahr JP, Østergaard S, Dalsgaard J, et al. Exercises versus arthroscopic decompression in patients with subacromial impingement: a randomised, controlled study in 90 cases with a one year follow up. Ann Rheum Dis. 2005;64(5):760-764.
4. Coghlan JA, Buchbinder R, Green S, Johnston RV, Bell SN. Surgery for rotator cuff disease. Cochrane Database Syst Rev. 2008;(1):CD005619.
5. Kromer TO, Tautenhahn UG, de Bie RA, Staal JB, Bastiaenen CHG. Effects of physiotherapy in patients with shoulder impingement syndrome: a systematic review of the literature. J Rehabil Med. 2009;41(11):870-880.

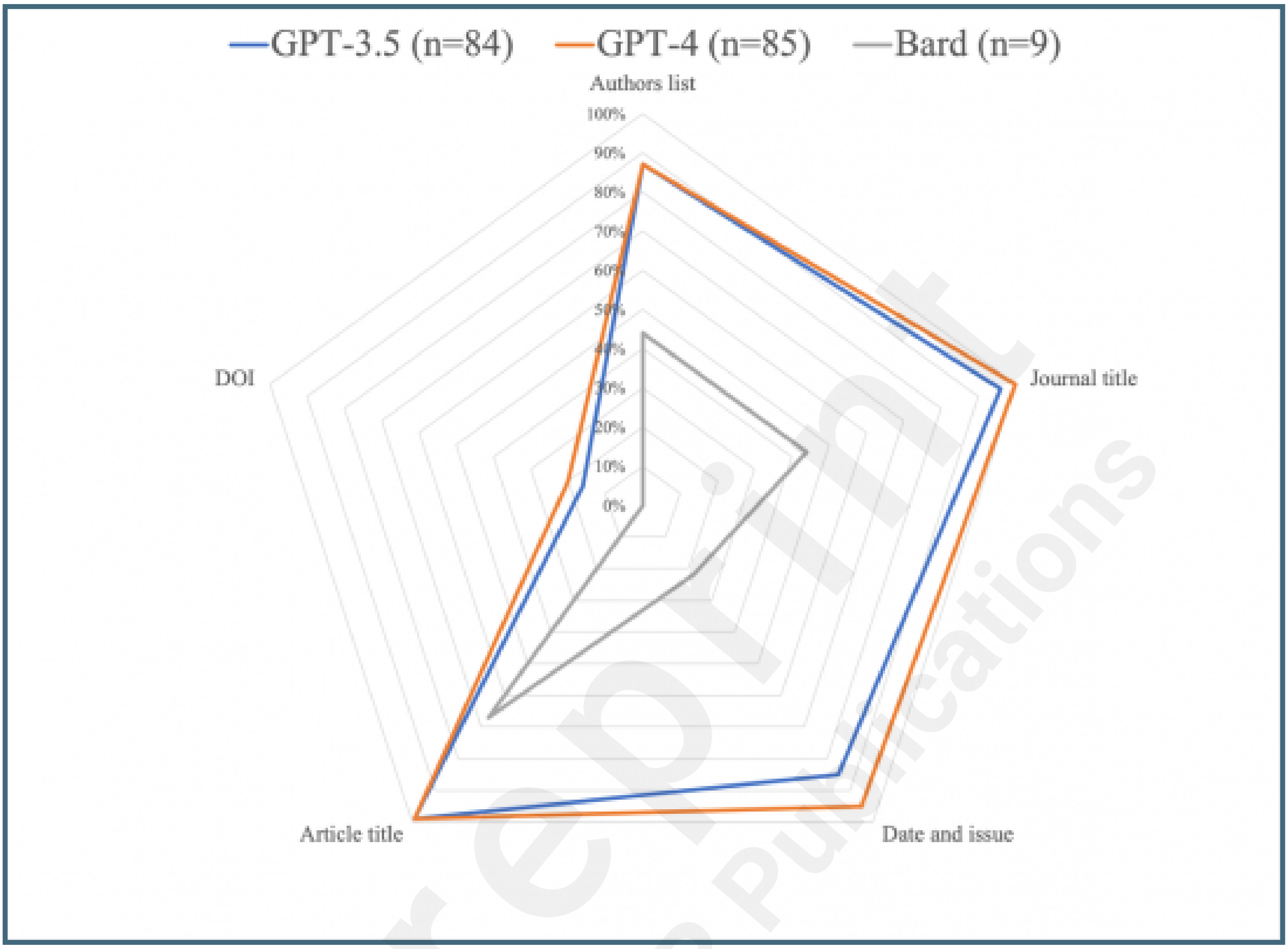
Flowchart.


Identification of studies via Pubmed
Identification
Screening
Included
Records identified from Pubmed:
(n = 32)
Reports assessed for eligibility
(n = 32)
Reports excluded:
Not a systematic review (n = 1)
Study protocol only (n = 2)
Not randomized studies (n = 18)
Studies included
(n = 11)

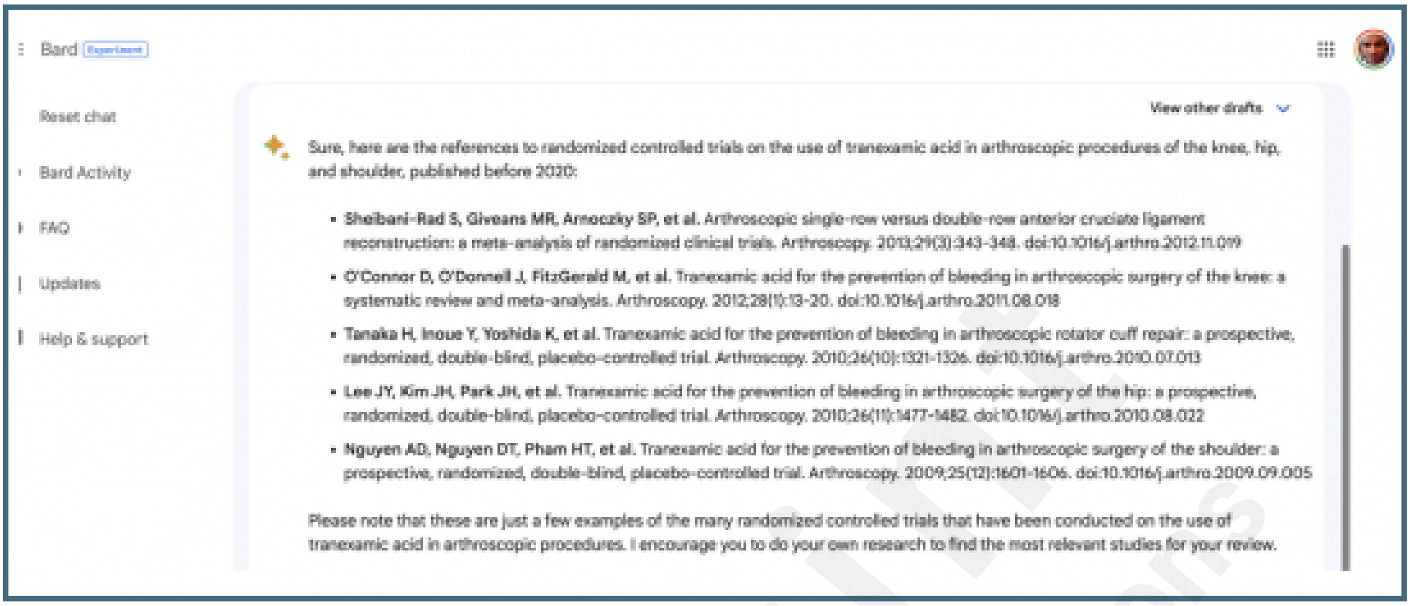
Criteria.


GPT-3.5 (n=84)
GPT-4 (n=85)
Bard (n=9)
Article exists
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Present in systematic review
Randomized study
Participant criteria
Intervention criteria
Not a systematic review
Published before 2021

Exact data.


GPT-3.5 (n=84)
GPT-4 (n=85)
Bard (n=9)
Authors list
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Journal title
Date and issue
Article title
DOI

Obviously hallucinated.



Hallucinated Reference Example.