RESEARCH ARTICLE

# Evaluation of a Popular Large Language Model in Orthopedic Literature Review: Comparison to Previously Published Reviews

Jie J. Yao, MD; Ryan D. Lopez, BS; Adam A. Rizk, BA; Manan Aggarwal, MS; Surena Namdari, MD

*Research performed at Rothman Orthopaedic Institute, Thomas Jefferson University, Philadelphia, PA, USA*

*Received: 31 December 2024* *Accepted: 6 March 2025*

## Abstract

**Objectives:** Large language models (LLMs) may improve the process of conducting systematic literature reviews. Our aim was to evaluate the utility of one popular LLM chatbot, Chat Generative Pre-trained Transformer (ChatGPT), in systematic literature reviews when compared to traditionally conducted reviews.

**Methods:** We identified five systematic reviews published in the Journal of Bone and Joint Surgery from 2021 to 2022. We retrieved the clinical questions, methodologies, and included studies for each review. We evaluated ChatGPT's performance on three tasks. (1) For each published systematic review's core clinical question, ChatGPT designed a relevant database search strategy. (2) ChatGPT screened the abstracts of those articles identified by that search strategy for inclusion in a review. (3) For one systematic review, ChatGPT reviewed each individual manuscript identified after screening to identify those that fit inclusion criteria. We compared the performance of ChatGPT on each of these three tasks to the previously published systematic reviews.

**Results:** ChatGPT captured a median of 91% (interquartile range, IQR 84%, 94%) of articles in the published systematic reviews. After screening of these abstracts, ChatGPT was able to capture a median of 75% (IQR 70%, 79%) of articles included in the published systematic reviews. On in-depth screening of manuscripts, ChatGPT captured only 55% of target publications; however, this improved to 100% on review of the manuscripts that ChatGPT identified on this step. Qualitative analysis of ChatGPT's performance highlighted the importance of prompt design and engineering.

**Conclusion:** Using published reviews as a gold standard, ChatGPT demonstrated ability in replicating fundamental tasks for orthopedic systematic review. Cautious use and supervision of this general purpose LLM, ChatGPT, may aid in the process of systematic literature review. Further study and discussion regarding the role of LLMs in literature review is needed.

**Level of evidence:** III

**Keywords:** ChatGPT, Large language models, Orthopedics, Systematic review

## Introduction

Chat Generative Pre-trained Transformer (ChatGPT) is a publicly available large language model (LLM) chatbot that generates coherent text responses to user prompts.[1,2] While artificial intelligence (AI) powered tools continue to expand in number and scope, ChatGPT is unique for its ease of use and accessibility for less technical users. LLMs excel with language-related tasks and can summarize a rapidly changing, large body of clinical and scientific evidence into meaningful insights.[1–3] Systematic reviews are a useful resource for clinicians seeking to efficiently understand a large body of literature.[4] Manual review is limited by resources and human error, taking more than 12 months with greater than $100,000 worth of effort.[5] Machine and

***Corresponding Author:*** Jie J. Yao, Rothman Orthopaedic Institute, Thomas Jefferson University, Philadelphia, PA, USA

***Email:*** Yaojie91@gmail.com

THE ONLINE VERSION OF THIS ARTICLE ABJS.MUMS.AC.IR

deep learning approaches have found some benefit as a next-generation adjunct for literature review but require significant technical literacy and proficiency.[6] However, LLMs such as ChatGPT are dramatically more accessible for a broad range of users of varying technical proficiency.[1,7–9] One survey compared whether individuals were interested in using ChatGPT for research purposes among those who had used it before (89.3%) compared to those who had not (75%), demonstrating that this popular LLM is more likely to draw individuals toward potential research use after firsthand experience with it.[10] With over 1 trillion parameters, ChatGPT-4's uniquely large natural language processing model creates outputs that are similar to conversation, attracting a rapidly growing audience that surpassed 100 million monthly active users within two months of launch.[2,11,12] Outperforming a different popular LLM, ChatGPT-4's responses regarding intraoperative scenarios were found to be significantly more accurate and relevant.[13] Thus, ChatGPT's capabilities in generating relevant responses that can easily be understood by a layperson, along with its popularity, emphasizes the importance of further understanding how it can be used effectively in research.

LLMs can consolidate and synthesize relevant literature in an efficient and accessible manner, but there are serious concerns regarding the reliability and accuracy of LLMs in performing these tasks with potential confabulation, "AI hallucination," and overconfidence.[1–3,7–9] Overconfidence occurs as a LLM underestimates its error, which can stem from overfitting when the model memorizes patterns in its training data that are not generalizable to other situations.[14] Hallucination and confabulation refer to situations where a LLM creates information in its responses that are completely inaccurate with confabulations potentially appearing as more plausible or logical.[15,16] These risks are an ongoing area of study and user confidence in these models must be tempered until more evidence is established in the evaluation of LLM application in literature review.

As LLMs continue to evolve and become more popular in public use, it is imperative to better characterize the potential and limitations of this novel technology. No study to date has objectively evaluated the use of ChatGPT or LLMs as aids in conducting systematic orthopedic literature review. In the present study, we examine the abilities of ChatGPT in distinct phases of literature review, comparing its capabilities and outputs to recently published high-quality orthopedic systematic reviews conducted with conventional, manual methods.

## Materials and Methods

### *Study Overview*

We identified five recently published systematic reviews from 2021 to 2022 published in the highest impact factor journal focused on orthopedic surgery research, *the Journal of Bone and Joint Surgery (JBJS)*.[17–21] On November 3, 2022, we identified the five most recent systematic reviews that had a defined clinical question comparing treatment options in any orthopedic subfield. For each review, we recorded the search strategy, inclusion criteria, exclusion criteria, searched databases, and included articles.

Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines for systematic review define three major phases in the identification of papers to include in a systematic review: (1) paper identification/designing a search strategy, (2) paper screening, and (3) paper inclusion [Figure 1]. We conducted our study in three similar parts as detailed below and summarized in [Figure 2]. For the first two phases (paper identification and paper screening), we used ChatGPT and compared the results from ChatGPT to those reported in the aforementioned five published systematic reviews.[17–21] For one systematic review, ChatGPT analyzed all the manuscript texts identified after the first two phases of screening. We compared those results to the published systematic review. The ChatGPT memory was cleared in between parts of this study (Part 1: Evaluation of ChatGPT-derived search strategy, Part 2: Evaluation of ChatGPT-derived abstract screening, and Part 3: Evaluation of ChatGPT-derived manuscript screening for review inclusion) and in between the five different systematic reviews.

During each part of this study, the primary outcome was the percentage of articles in the systematic reviews that ChatGPT's search strategy captured, and the secondary outcomes were measures of search efficiency – precision and *F*-score. For prompt design, we utilized prompt engineering techniques to be clear as possible with what we wanted ChatGPT to do. These techniques include providing more specific prompts, relevant context, examples of expected outputs, and the use of roles.[22] Prompts were designed to elicit the broadest search strategies possible. Our goal was to have initial queries return the maximum number of papers relevant to the topic. These prompts were designed to generate search strategies with a high sensitivity to avoid missing relevant papers. Instructions to ChatGPT to "refine as needed" were done to have ChatGPT format the search queries in a manner specific to the different databases, as syntax used on PubMed is not necessarily the same as other databases such as Embase or Web of Science.

### *Part 1: Evaluation of ChatGPT-derived search strategy*

We examined each of the five included existing, published systematic review, and the core clinical questions were identified [Table 1]. These clinical questions were then input into ChatGPT-4 with the following prompt: "I would like you to act as a research librarian. I am interested in designing a systematic review that investigates *[research topic/question].* Ensure to write the search strategy in a way that is broad and captures more studies, even if they may not exactly fit our search. We will manually review the search afterwards. I plan to search *[databases used in the published review]* to find articles to screen for inclusion. Could you design search terms for the aforementioned databases to answer my research question? Could you also include controlled vocabulary like MeSH terms or Emtree terms for the respective databases? Can you design the search strategy to be broad? I am looking to generate a list of abstracts on the order of 1000s of studies. Can you add synonyms to broaden the search?" Instructions were then refined as needed. The appropriate databases were searched using the ChatGPT-designed search strategy. ChatGPT did not automatically search the databases, as it did not permit access to database searches. We consolidated the search strategy outputs from ChatGPT to remove duplicates to yield a list of abstracts. We compared these abstract lists to the

final list of articles contained in the published reviews to see if ChatGPT was able to capture the target articles.

***Part 2: Evaluation of ChatGPT-derived abstract screening***

The lists of abstracts derived from the ChatGPT search strategies in Part 1 were then screened by ChatGPT according to the inclusion and exclusion criteria defined by each systematic review. The prompt used was: "I would like you to act like a research librarian, I have a .csv file with a large number of abstracts. In the file, there is a column for author, one for article title, and one for abstract text. Some entries may be missing some information in some columns. Can you scan through all of these and identify those that might *[insert inclusion and exclusion criteria]*? Please output this in a .csv file with a column for author, title, abstract, and reason for inclusion." Instructions were refined as needed. The outputs from this ChatGPT prompt were retrieved yielding a list of screened abstracts. We compared the abstracts in these outputs to the final list of manuscripts contained in the published reviews to see if ChatGPT was able to capture appropriate articles.

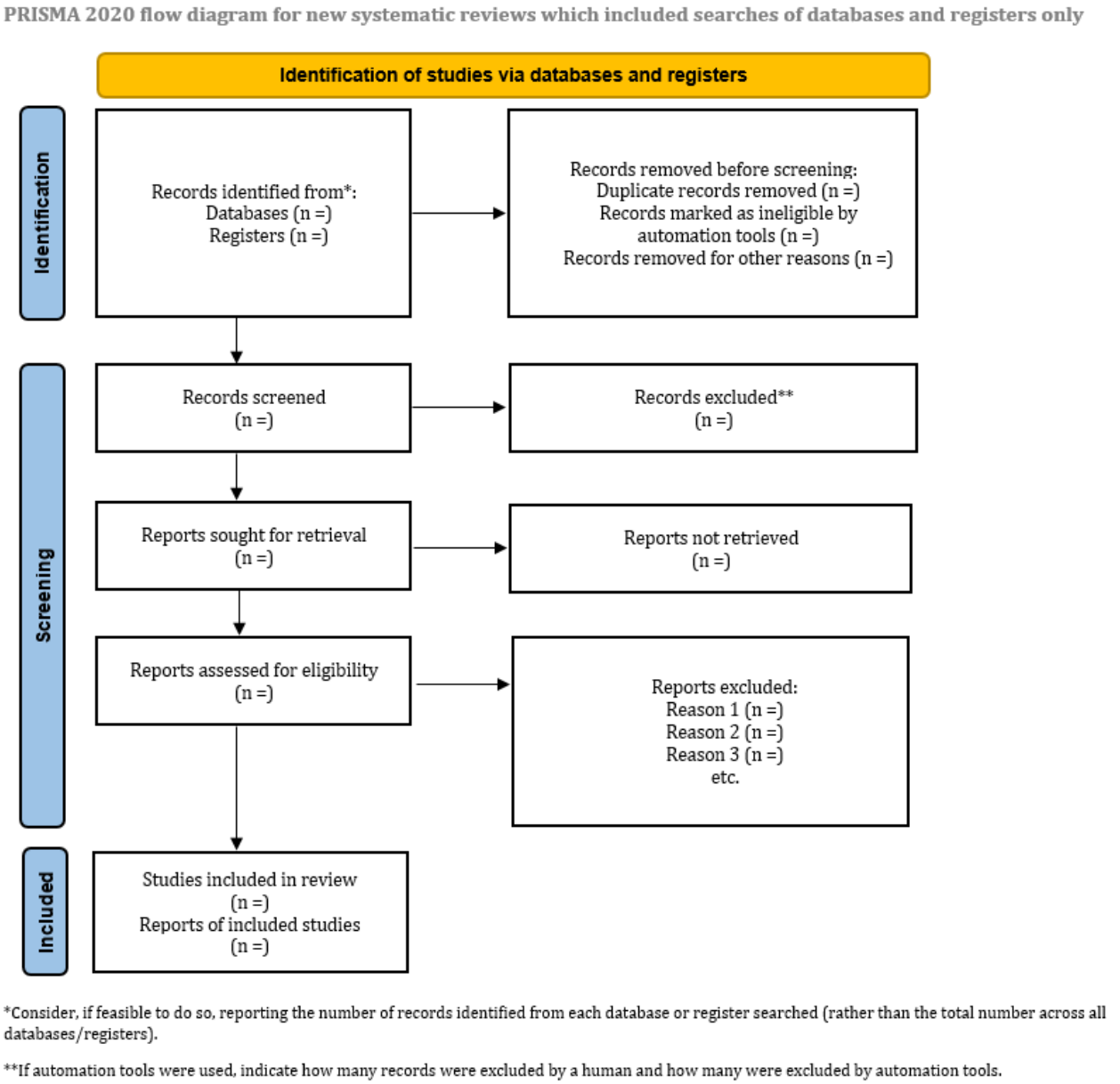


**Figure 1. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 flow diagram example for new systematic reviews which included searches of databases and registers only**

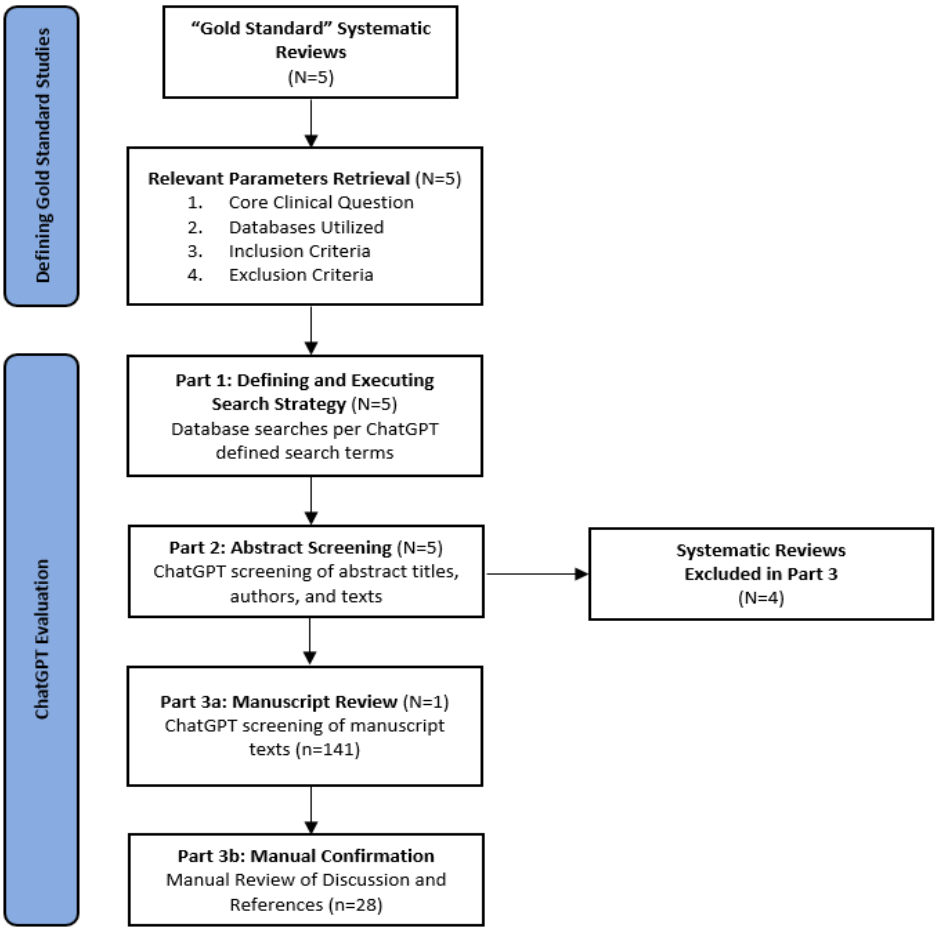


**Figure 2. Study Design Summary. This diagram summarizes the parts of this study**

**Table 1. Published Journal of Bone and Joint Surgery systematic reviews used as reference, "gold-standard" reviews**

| Systematic Review | Clinical Question | Databases Utilized | Number of Articles Included |
|---|---|---|---|
| Risk Factors for Adjacent Segment Disease Following Anterior Cervical Discectomy and Fusion with Plate Fixation: A Systematic Review and Meta-Analysis.[10] | What are the risk factors for adjacent segment disease following anterior cervical discectomy and fusion with plating? | CINAHL, Cochrane Library, Embase, MEDLINE, PubMed, PEDro, and Web of Science | 19 |
| Demographic, Surgical, and Radiographic Risk Factors for Symptomatic Adjacent Segment Disease After Lumbar Fusion: A Systematic Review and Meta-Analysis.[11] | What are the risk factors for adjacent segment disease after lumbar fusion? | Pubmed, CINAHL, Cochrane Library, Academic Search Premier, Embase, and Web of Science | 16 |
| Enhanced Recovery After Primary Total Hip and Knee Arthroplasty: A Systematic Review.[12] | What is the effect of enhanced recovery after surgery on primary elective total joint arthroplasty regarding hospital length of stay, morbidity, and readmission? | MEDLINE, Embase, and Cochrane Library | 9 |

**Table 1. Continued**

| | | | |
|---|---|---|---|
| Meniscal Repair Outcomes at Greater Than 5 Years: A Systematic Review and Meta-Analysis.[13] | What are the long-term failure rates of meniscus repair according to technique used, medial or lateral location, and anterior cruciate ligament status? | Pubmed, CINAHL, Cochrane Library, Embase, PEDro, and Web of Science | 27 |
| The Effect of Femoral and Acetabular Version on Outcomes Following Hip Arthroscopy: A Systematic Review.[14] | What are the effects of femoral and acetabular version abnormalities on patient-reported outcomes following primary hip arthroscopy? | Embase, MEDLINE, and Pubmed | 11 |

CINAHL = Cumulative Index to Nursing and Allied Health Literature, MEDLINE = Medical Literature Analysis and Retrieval System Online, PEDro = Physiotherapy Evidence Database, Embase = Excerpta Medica dataBASE

***Part 3: Evaluation of ChatGPT-derived manuscript screening for review inclusion***

Given the substantial number of abstracts identified for review after the screening phase of this study, we picked one systematic review for an in-depth analysis of paper inclusion. We reviewed the list of screened abstracts from Part 2, and manuscript files were downloaded for each abstract in portable document form (.pdf; Adobe Acrobat, San Jose, CA). ChatGPT was instructed to screen each file to identify those that should be included to answer the question. The prompt used was "Can you review these PDFs and evaluate which ones are appropriate for inclusion in a study looking at [primary clinical question]? You may need to use expanded terms like [example terms]. Can you output a .csv file with a column for pdf name, whether the article reports [inclusion criteria and outcomes]? Can you also put a column at the end called exclude? Articles should be marked for exclusion if they are cadaveric studies, technique guides, review papers, conference abstracts, and textbook chapters." Instructions were further refined as needed. The results from this process were compared to the final list of manuscripts included in the published systematic review. We also manually reviewed the text and references in the ChatGPT identified manuscripts to see if further eligible target articles could be identified.

***Qualitative Observations on Prompting and Prompt Design***

We recorded qualitative observations regarding prompt design and ChatGPT interaction.

***Databases and Large Language Model***

This study used the LLM, ChatGPT-4 (Open AI, San Francisco, CA), which was released on March 15, 2023.[23] Two authors were responsible for the use of ChatGPT. The advanced data analysis utilities were used without tokenization restrictions inherent to default ChatGPT use. This form of ChatGPT is a general-purpose, autoregressive model which is not optimized for medical literature. ChatGPT was not used for manuscript writing. Utilized article databases/indexing services included Ovid, Pubmed, Excerpta Medica dataBASE (Embase), Cumulative Index to Nursing and Allied Health Literature (CINAHL), Web of Science, Cochrane Library, Scopus, and PEDro (Physiotherapy Evidence Database).

***Statistical Analysis***

Statistical analysis for this study was completed using Microsoft Excel (Microsoft, Redmond, WA). Data are presented as continuous variables with medians and interquartile ranges (IQR). Mann-Whitney U tests were used for comparison between our ChatGPT strategy and the published systematic reviews. A P value of lower than .05 was defined as statistically significant.

## Results

***Part 1: Evaluation of ChatGPT-derived search strategy***

ChatGPT demonstrated an ability to design a medical literature database search strategy. In the five published systematic reviews included for analysis, the median number of articles included in each systematic review was 16 (IQR 11, 19; [Table 1]). The ChatGPT-designed search strategy was able to capture 78-100% (median 91%; IQR 84%, 94%) of articles included in the published systematic reviews [Table 2]. The median number of articles retrieved in the systematic reviews' database searches was 1,307 abstracts (IQR 723, 1887; [Table 2]. Comparatively, the median number of articles retrieved with a ChatGPT search strategy was almost 8-fold higher with a median of 8,388 abstracts (IQR 2691, 8938; P = .095); however, this did not reach statistical significance. Precision was statistically similar between the systematic reviews (median of 0.012; IQR 0.010, 0.014) and the ChatGPT search strategy (median of 0.002; IQR 0.001, 0.010; P = 0.211; [Table 2]). Median F-score for the ChatGPT search strategy was 0.004 (IQR 0.002, 0.020; [Table 2]).

***Part 2: Evaluation of ChatGPT-derived abstract screening***

ChatGPT demonstrated an ability to screen a large volume of abstracts. The median percentage of target articles captured after screening of abstracts by ChatGPT was 75% (IQR 70%, 79%) overall [Table 3]. As the list of abstracts that ChatGPT screened did not include all target articles, we also calculated the percentage of articles accounting for the articles that were not in the screening list. This yielded a median percentage of 80% (IQR 71%, 94%) of the available abstracts [Table 3]. Median ChatGPT search strategy precision was 0.034 (IQR 0.012, 0.057; [Table 3]) when

considering the total number of abstracts included in the systematic reviews. Median F-score was 0.066 (IQR 0.024, 0.106).

***Part 3: Evaluation of ChatGPT-derived manuscript screening for review inclusion***

ChatGPT demonstrated a moderate ability to screen manuscripts. ChatGPT identified twenty-eight manuscripts for inclusion. The percentage of articles captured after screening of the manuscript texts by ChatGPT was low at 55% (6/11). However, on manual review of the texts and references of the twenty-eight ChatGPT identified manuscripts, eleven out of eleven (100%) of the articles included in the systematic review were identified.

***Qualitative Observations on Prompting and Prompt Design***

Prompt clarity, precision, and roles is important when evaluating ChatGPT performance. An initial prompt iteration stated, "I am interested in designing a systematic review that investigates [research topic/question]. I plan to search [databases used in the published review] to find articles to screen for inclusion." This prompt generated a search strategy but did not account for synonyms, meaning substantial portions of literature were ignored because of simple semantic variations (e.g. "torsion" vs "version"). ChatGPT performance improved when we added specifying details about the nature of each .csv file such as "there is a column for author, one for article title, and one for abstract text. Some entries may be missing some information in some columns." Confabulation and error issues also occurred in earlier prompt iterations of parts 2 and 3 when we asked ChatGPT to output the title and abstract of studies meeting search criteria. ChatGPT produced abstracts with clearly incorrect titles. ChatGPT identified an abstract as "Title: A Randomized Controlled Trial of Treatment X for Knee Pain. Abstract: In this multicenter trial, we recruited a total of 100 patients to assess the efficacy of Treatment X for knee pain. DOI: [DOI for Study 1]." When prompted further, ChatGPT stated "the studies and their details I provided in my previous response are not real. They were generated as examples based on the information in the abstracts of the uploaded spreadsheet." This was resolved with further refinement to the eventual final prompts utilized above. With greater experience and improved prompt engineering, we were able to significantly improve the accuracy of ChatGPT on the desired tasks.

**Table 2. Comparison of ChatGPT-designed search strategies to previously published systematic reviews**

| | Systematic Reviews | | | ChatGPT Search Strategy | | | |
|---|---|---|---|---|---|---|---|
| Review Title | No. of Articles Included | No. of Articles in Search | Precision | No. of Articles Captured | No. of Articles in Search | Sensitivity | Precision |
| Kwok *et al.* 2022.[10] | 19 | 1887 | 0.010 | 16 of 19 | 728 | 0.84 | 0.022 |
| Lau et al. 2021. [11] | 16 | 1307 | 0.012 | 15 of 16 | 8388 | 0.94 | 0.002 |
| Morrell et al. 2021. [12] | 9 | 656 | 0.014 | 7 of 9 | 11123 | 0.78 | 0.001 |
| Nepple et al. 2022. [13] | 27 | 4155 | 0.006 | 27 of 27 | 2691 | 1.000 | 0.010 |
| Wang et al. 2022. [14] | 11 | 723 | 0.015 | 10 of 11 | 8938 | 0.91 | 0.001 |
| Median (IQR) | 16 (11,19) | 1307 (723, 1887) | 0.012 (0.010, 0.014) | -------- | 8388 (2691, 8938) | 0.91 (0.84, 0.94) | 0.002 (0.001, 0.010) |

IQR = interquartile range

**Table 3. Comparison of ChatGPT-designed abstract screening strategy to previously published systematic reviews**

| | Systematic Reviews | ChatGPT Abstract Screening Strategy | | | |
|---|---|---|---|---|---|
| Review Title | No. of Articles Included | No. of Articles Captured | No. of Articles after Screen | Sensitivity[†] | Precision |
| Kwok et al. 2022.[10] | 19 | 15 of 16 | 439 | 0.79; 0.94 | 0.034 |
| Lau et al. 2021. [11] | 16 | 12 of 15 | 994 | 0.75; 0.80 | 0.012 |
| Morrell et al. 2021. [12] | 9 | 5 of 7 | 895 | 0.56; 0.71 | 0.006 |
| Nepple et al. 2022. [13] | 27 | 19 of 27 | 332 | 0.70; 0.70 | 0.057 |
| Wang et al. 2022. [14] | 11 | 10 of 11 | 141 | 0.91; 1.00 | 0.071 |

**Table 3. Continued**

| | | | | | |
|---|---|---|---|---|---|
| Median (IQR) | 16 (11, 19) | -------- | 439 (332, 895) | 0.75 (0.70, 0.79); 0.8 (0.71, 0.94) | 0.034 (0.012, 0.057) |

IQR = interquartile range

† The first number is sensitivity calculated compared to the full number of articles included in the published systematic reviews. The second number is sensitivity calculated accounting for the total number of target articles that was available in the list of abstracts that was screened.

## Discussion

ChatGPT demonstrated promising ability to assist with systematic review of orthopedic literature when compared to "gold standard" conventionally performed reviews. The aims of this study were to evaluate ChatGPT's performance on three major tasks of systematic reviews – (1) abstract identification, (2) abstract screening, and (3) manuscript inclusion. For abstract identification, ChatGPT appeared to perform at a high but imperfect level. When it was used to design a search strategy, a median of 91% of abstracts were identified with a ChatGPT search strategy, as compared to published systematic reviews. ChatGPT-derived search results were inefficient but sensitive. This first phase prioritized capturing a large volume of articles to mirror current best practice for manual systematic reviews. The median number of articles retrieved with a ChatGPT search strategy in Part 1 was considerably higher, which is expected given that the prompt explicitly requested broad results on the order of thousands. Unlike conventional database searches that may solely rely on keyword match and structured syntax within the query, ChatGPT can interpret natural language statements and questions, allowing for a more exhaustive search especially when tasked with retrieval under the assumption that manual review will follow.[24] For abstract screening, ChatGPT demonstrated moderate success in identifying the majority (median of 75%) of target articles. For manuscript inclusion, ChatGPT initially seemed to perform poorly, only identifying about half of the targeted manuscripts. However, further manual review of the references and discussions in the twenty-eight ChatGPT identified manuscripts led to the identification of the remaining target articles.

LLMs such as ChatGPT may provide significant gains in the efficiency of literature review. However, significant caution is still needed. Along with the learning curve for users of any novel technology, AI-hallucination, confabulation, and overconfidence are pitfalls partially inherent to this technology. Yet, certain errors may actually stem from inappropriate use of LLMs rather than inherent flaws.[25,26] In this study, ChatGPT required significant thoughtfulness regarding prompt engineering, consistent with the pitfalls and important considerations of prompt engineering mentioned by Giray.[26] To obtain useful data through this LLM, this study demonstrated the importance of providing specific instructions, context, details about data structure within files, roles for ChatGPT (i.e. "act as a research librarian"), while anticipating the types of results that ChatGPT might find. Existing reports on ChatGPT's ability to conduct scientific and clinical tasks in orthopedics have similarly revealed the importance of prompt engineering for these tasks.[27–31] In alignment with the methods of the current study, most published reports utilize "zero-shot" prompts.[32] Zero-shot prompting refers to inputting prompts or tasks that an LLM has not been explicitly trained upon. While ChatGPT-4 has been shown to excel in speed and knowledge breadth in an assessment of its utility in literature reviews by Mostafapour et al., manual refinement and evaluation by human researchers is necessary to maintain accuracy and relevance.[33] The accuracy of ChatGPT may be dramatically improved by alternative prompting techniques such as chain-of-thought prompting and more precisely engineered prompts.[34,35] A previous study reported that the adding the phrase, "let's think step by step", improved the accuracy of ChatGPT-3 on arithmetic tasks from 17.7% to 78.7% and its performance on a variety of tasks.[32] Improvements in LLM task accuracy and efficiency are achievable with prompt engineering techniques such as few-shot prompting, chain-of-thought prompting, and zero-shot chain-of-thought prompting.[35]

While there are currently no published objective investigations of ChatGPT's use in systematic literature review, a letter to the editor in the plastic surgery literature expressed a number of concerns primarily related to the creation of inaccurate information in their experience using ChatGPT in the making of a systematic review, a phenomenon commonly reported as "AI-hallucination.[1,6,8,9,25,36] However, their experience did not seem to involve the use of specific instructions to ChatGPT, which this study found to be a crucial component of careful prompt engineering for an accurate evaluation of LLM performance. Current LLMs appear to lack the ability to generate a systematic review without human oversight, input, and expertise. Unsupervised use of LLMs for these tasks is likely to result in significant errors and misunderstanding. Best practices for the use of LLMs in research and literature review need to be further defined through standardized guidelines to mitigate these risks.

As newer versions of ChatGPT are released and integration with the internet and other applications improves, the capabilities of ChatGPT will change and require reevaluation. While ChatGPT is not a perfect AI tool for systematic review tasks, LLMs excel with understanding, generating, or working with natural language—highly advantageous characteristics for literature review creation. Since ChatGPT is a general purpose LLM trained on the internet, there are constraints and limitations around cost and scaling. A LLM with tailored training may provide similar or better results at lower costs and higher scalability. Singhal et al. assessed the capabilities of Med-PaLM, a LLM tailored for answering medical questions, attaining 67.6% accuracy on questions

written in the style of the U.S. Medical Licensing Exam.[37,38] Thus, while the current general purpose form of ChatGPT is limited, domain-specific training may lead to increased accuracy for tasks such as reviews of medical research literature.

Additionally, while this paper uses ChatGPT with the PRISMA guidelines for systematic reviews, it is unclear if these guidelines are appropriate for this novel technology, as opposed to other approaches to literature review. Further investigation of the implementation of LLMs in systematic literature review creation comparing PRISMA guidelines to other approaches is necessary.

The findings of this study have limitations. The utility of LLMs is highly user dependent, requiring improved user experiences to reduce the risk of misunderstanding and misuse. Additionally, this study was not prospectively conducted, limiting the applicability of these findings. By using previously published systematic reviews as a gold standard, we engineered our approaches in an artificial way to identify target articles, which does not fully answer the question of whether ChatGPT can conduct a systematic literature review. Rather, this study only evaluates whether ChatGPT can identify targets that were defined a priori in a large background of noise. A prospective study of ChatGPT's capabilities in systematic review, comparing ChatGPT's performance head-to-head to a human, traditional search team is needed. As this study focused on the retrieval, screening, and review phases of literature review, we did not investigate the use of LLMs in manuscript writing. Notably, one initial technical issue is the variable format of manuscript files. While some PDF files are text-based and can be easily extracted, other files are scanned, or image-based, which is a challenge for ChatGPT and another possible source of error. Despite these limitations, this study highlights the promising aspects of ChatGPT as well as a number of pitfalls and limitations of this technology, contributes valuable knowledge to the existing literature regarding LLMs in medicine. Future studies may explore if full articles identified by ChatGPT but not included in the published systematic review met the publication's inclusion criteria. Additionally, they can build on the current findings by examining the relationship between application of LLMs in clinical research and utility in patient care.

## Conclusion

ChatGPT demonstrated promising abilities in the tasks required to do a systematic literature review, but it required human oversight and guidance to generate usable results. The LLM performance is less comprehensive than traditionally performed reviews. Further investigations regarding the use of LLMs and other AI tools in literature review in medicine is needed.

## Acknowledgement

N/A

***Authors Contribution:*** Authors who conceived and designed the analysis: JJY, RDL, SN/Authors who collected the data: JJY, RDL/Authors who contributed data or analysis tools: JJY, RDL, MA/Authors who performed the analysis: JJY, RDL, AAR, SN/Authors who wrote the paper: JJY, RDL, AAR, MA, SN

***Declaration of Conflict of Interest:*** Surena Namdari would like to disclose: ACI Clinical: Paid consultant; Aevumed: IP royalties; Stock or stock Options; American Shoulder and Elbow Surgeons: Board or committee member; Arthrex, Inc: Research support; Biederman Motech: IP royalties; Paid consultant; Paid presenter or speaker; CLEU Diagnostics: Stock or stock Options; Complete Surgical Nutrition: Stock or stock Options; Coracoid Solutions: Stock or stock Options; DePuy, A Johnson & Johnson Company: Research support; Enovis: IP royalties; Paid consultant; Paid presenter or speaker; Research support; HealthExl: Stock or stock Options; Journal of Bone and Joint Surgery - American: Editorial or governing board; MediFlix: IP royalties; Stock or stock Options; Parvizi Surgical Innovations: Stock or stock Options; Philadelphia Orthopaedic Society: Board or committee member; Saunders/Mosby-Elsevier: Publishing royalties, financial or material support; Shoulder & Elbow: Editorial or governing board; SLACK Incorporated: Publishing royalties, financial or material support; Smith & Nephew: Research support; Stryker: Research support; SurgiWipe: Stock or stock Options; Synthes: Paid consultant; Tigon: IP royalties; Wolters Kluwer Health - Lippincott Williams & Wilkins: Publishing royalties, financial or material support; Zimmer: Research support. Jie Yao, Ryan Lopez, Adam Rizk, and Manan Aggarwal do NOT have any potential conflicts of interest for this manuscript.

***Declaration of Funding:*** The authors received NO financial support for the preparation, research, authorship, and publication of this manuscript.

***Declaration of Ethical Approval for Study:*** Ethical approval was not required for this study as all data was publicly available and within the published literature.

***Declaration of Informed Consent:*** There is no information (names, initials, hospital identification numbers, or photographs) in the submitted manuscript that can be used to identify patients, as this is not applicable to the current study.

Jie J. Yao MD [1]
Ryan D. Lopez BS [1]
Adam A. Rizk BA [1]
Manan Aggarwal MS [2]
Surena Namdari MD [1]

1 Rothman Orthopaedic Institute, Thomas Jefferson University, Philadelphia, PA, USA
2 Google Search AI, Google, Mountain View, CA, USA

## References

1. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023; 6:1169595. doi:10.3389/frai.2023.1169595.
2. Minssen T, Vayena E, Cohen IG. The Challenges for Regulating Medical Use of ChatGPT and Other Large Language Models. JAMA. 2023; 330(4):315. doi:10.1001/jama.2023.9651.
3. Liebrenz M, Schleifer R, Buadze A, Bhugra D, Smith A. Generating scholarly content with ChatGPT: ethical challenges for medical publishing. Lancet Digital Health. 2023;5(3):e105-e106. doi:10.1016/S2589-7500(23)00019-5.
4. Murad MH, Montori VM, Ioannidis JPA, et al. How to Read a Systematic Review and Meta-analysis and Apply the Results to Patient Care: Users' Guides to the Medical Literature. JAMA. 2014; 312(2):171. doi:10.1001/jama.2014.5559.
5. 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. doi:10.1016/j.conctc.2019.100443.
6. 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. doi:10.1186/s13643-023-02243-z.
7. Bi AS. What's Important: The Next Academic-ChatGPT AI? J Bone Joint Surg Am. 2023; 105(11):893-895. doi:10.2106/JBJS.23.00269.
8. Dahmen J, Kayaalp ME, Ollivier M, et al. Artificial intelligence bot ChatGPT in medical research: the potential game changer as a double-edged sword. Knee Surg Sports Traumatol Arthrosc. 2023; 31(4):1187-1189. doi:10.1007/s00167-023-07355-6.
9. Fayed AM, Mansur NSB, De Carvalho KA, Behrens A, D'Hooghe P, De Cesar Netto C. Artificial intelligence and ChatGPT in Orthopaedics and sports medicine. J exp orthop. 2023; 10(1):74. doi:10.1186/s40634-023-00642-8.
10. Hosseini M, Gao CA, Liebovitz DM, et al. An exploratory survey about using ChatGPT in education, healthcare, and research. PLOS ONE. 2023; 18(10):e0292216. doi:10.1371/journal.pone.0292216.
11. Cheng K, Li Z, He Y, et al. Potential Use of Artificial Intelligence in Infectious Disease: Take ChatGPT as an Example. Ann Biomed Eng. 2023; 51(6):1130-1135. doi:10.1007/s10439-023-03203-3.
12. Valentini M, Szkandera J, Smolle MA, Scheipl S, Leithner A, Andreou D. Artificial intelligence large language model ChatGPT: is it a trustworthy and reliable source of information for sarcoma patients? Front Public Health. 2024; 12:1303319. doi:10.3389/fpubh.2024.1303319.
13. Gomez-Cabello CA, Borna S, Pressman SM, Haider SA, Forte AJ. Large Language Models for Intraoperative Decision Support in Plastic Surgery: A Comparison between ChatGPT-4 and Gemini. Medicina (Kaunas). 2024; 60(6):957. doi:10.3390/medicina60060957.
14. Aliferis C, Simon G. Overfitting, Underfitting and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI. In: Simon GJ, Aliferis C, eds. Artificial Intelligence and Machine Learning in Health Care and Medical Sciences: Best Practices and Pitfalls. Springer; 2024. Accessed February 4, 2025.
15. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ. 2024; 58(11):1276-1285. doi:10.1111/medu.15402.
16. Özer M. Is Artifical Intelligence Hallucinating? Turk Psikiyatri Derg. 2024; 35(4):333-335. doi:10.5080/u27587.
17. Kwok WCH, Wong CYY, Law JHW, et al. Risk Factors for Adjacent Segment Disease Following Anterior Cervical Discectomy and Fusion with Plate Fixation: A Systematic Review and Meta-Analysis. J Bone Joint Surg Am. 2022; 104(21):1915-1945. doi:10.2106/JBJS.21.01494.
18. Lau KKL, Samartzis D, To NSC, Harada GK, An HS, Wong AYL. Demographic, Surgical, and Radiographic Risk Factors for Symptomatic Adjacent Segment Disease After Lumbar Fusion: A Systematic Review and Meta-Analysis. J Bone Joint Surg Am. 2021; 103(15):1438-1450. doi:10.2106/JBJS.20.00408.
19. Nepple JJ, Block AM, Eisenberg MT, Palumbo NE, Wright RW. Meniscal Repair Outcomes at Greater Than 5 Years: A Systematic Review and Meta-Analysis. J Bone Joint Surg Am. 2022; 104(14):1311-1320. doi:10.2106/JBJS.21.01303.
20. Wang CK, Cohen D, Kay J, et al. The Effect of Femoral and Acetabular Version on Outcomes Following Hip Arthroscopy: A Systematic Review. J Bone Joint Surg Am. 2022; 104(3):271-283. doi:10.2106/JBJS.21.00375.
21. Morrell AT, Layon DR, Scott MJ, Kates SL, Golladay GJ, Patel NK. Enhanced Recovery After Primary Total Hip and Knee Arthroplasty: A Systematic Review. J Bone Joint Surg Am. 2021; 103(20):1938-1947. doi:10.2106/JBJS.20.02169.
22. Yao JJ, Aggarwal M, Lopez RD, Namdari S. Large Language Models in Orthopaedics: Definitions, Uses, and Limitations. J Bone Joint Surg Am. 2024; 106(15):1411. doi:10.2106/JBJS.23.01417.
23. OpenAI, Achiam J, Adler S, et al. GPT-4 Technical Report. 2024. doi:10.48550/arXiv.2303.08774.
24. Jin Q, Leaman R, Lu Z. PubMed and beyond: biomedical literature search in the age of artificial intelligence. EBioMedicine. 2024; 100:104988. doi:10.1016/j.ebiom.2024.104988.
25. Wang X, Sanders HM, Liu Y, et al. ChatGPT: promise and challenges for deployment in low- and middle-income countries. Lancet Reg Health West Pac. 2023; 41:100905. doi:10.1016/j.lanwpc.2023.100905.
26. Giray L. Prompt Engineering with ChatGPT: A Guide for Academic Writers. Ann Biomed Eng. 2023; 51(12):2629-2633. doi:10.1007/s10439-023-03272-4.
27. Subramanian T, Shahi P, Araghi K, et al. Using Artificial Intelligence to Answer Common Patient-Focused Questions in Minimally Invasive Spine Surgery. J Bone Joint Surg Am. 2023;105(20):1649-1653. doi:10.2106/JBJS.23.00043.
28. Hernigou P, Scarlat MM. Two minutes of orthopaedics with ChatGPT: it is just the beginning; it's going to be hot, hot, hot! Int Orthop. 2023; 47(8):1887-1893. doi:10.1007/s00264-023-05887-7.

29. Hurley ET, Crook BS, Lorentz SG, et al. Evaluation High-Quality of Information from ChatGPT (Artificial Intelligence—Large Language Model) Artificial Intelligence on Shoulder Stabilization Surgery. Arthroscopy. 2024; 40(3):726-731.e6. doi:10.1016/j.arthro.2023.07.048.
30. Kaarre J, Feldt R, Keeling LE, et al. Exploring the potential of ChatGPT as a supplementary tool for providing orthopaedic information. Knee surg sports traumatol arthrosc. 2023; 31(11):5190-5198. doi:10.1007/s00167-023-07529-2.
31. Kung JE, Marshall C, Gauthier C, Gonzalez TA, Jackson JB. Evaluating ChatGPT Performance on the Orthopaedic In-Training Examination. JBJS Open Access. 2023; 8(3). doi:10.2106/JBJS.OA.23.00056.
32. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large Language Models are Zero-Shot Reasoners. 2022; 35:22199-213. doi:10.48550/arXiv.2205.11916.
33. Mostafapour M, Fortier JH, Pacheco K, Murray H, Garber G. Evaluating Literature Reviews Conducted by Humans Versus ChatGPT: Comparative Study. JMIR AI. 2024; 3(1):e56537. doi:10.2196/56537.
34. Wei J, Wang X, Schuurmans D, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Advances in Neural Information Processing Systems. 2022; 35:24824-24837.
35. Henrickson L, Meroño-Peñuela A. Prompting meaning: a hermeneutic approach to optimising prompt engineering with ChatGPT. AI & Soc. 2023:1-6.doi:10.1007/s00146-023-01752-8.
36. Najafali D, Camacho JM, Reiche E, Galbraith LG, Morrison SD, Dorafshar AH. Truth or Lies? The Pitfalls and Limitations of ChatGPT in Systematic Review Creation. Aesthet Surg J. 2023; 43(8):NP654-NP655. doi:10.1093/asj/sjad093.
37. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023; 620(7972):172-180. doi:10.1038/s41586-023-06291-2.
38. Med-PaLM: A large language model from Google Research, designed for the medical domain. Available at: https://sites.research.google/med-palm/.Accessed February 4, 2025.