

## **Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions.**

Jocelyn Gravel MD<sup>1</sup>, Madeleine D'Amours-Gravel<sup>2</sup>, Esli Osmanlliu MD<sup>3</sup>

### **Affiliations:**

<sup>1</sup>Department of Pediatric Emergency Medicine, CHU Sainte-Justine, Université de Montréal, Montréal, Québec, Canada

<sup>2</sup>CEGEP de Saint-Hyacinthe, Saint-Hyacinthe, Québec, Canada

<sup>3</sup>Division of Pediatric Emergency Medicine, Montreal Children Hospital, McGill University, Montréal, Québec, Canada

**Corresponding author:** Jocelyn Gravel, Department of Pediatric Emergency Medicine, CHU Sainte-Justine, 3175 Chemin de la Côte-Sainte Catherine, Room A326 Montreal (Quebec) H3T 1C5 CANADA

Phone: 514-345-4931 #2559

Email: graveljocelyn@hotmail.com

**Financial Disclosure Statement:** The authors have no financial relationships relevant to this article to disclose. This study was conducted without financial support.

**Conflict of Interest:** The authors have no conflict of interest relevant to this article to disclose.

## Abstract

**Background:** ChatGPT have gained public notoriety and recently supported manuscript preparation. Our objective was to evaluate the quality of the answers and the references provided by ChatGPT for medical questions.

**Methods:** Three researchers asked ChatGPT a total of 20 medical questions and prompted it to provide the corresponding references. The responses were evaluated for quality of content by medical experts using a verbal numeric scale going from 0 to 100%. These experts were the corresponding author of the 20 articles from where the medical questions were derived. We planned to evaluate three references per response for their pertinence, but this was amended based on preliminary results showing that most references provided by ChatGPT were fabricated.

**Results:** ChatGPT provided responses varying between 53 and 244 words long and reported two to seven references per answer. Seventeen of the 20 invited raters provided feedback. The raters reported limited quality of the responses with a median score of 60% (1<sup>st</sup> and 3<sup>rd</sup> quartile: 50% and 85%). Additionally, they identified major (n=5) and minor (n=7) factual errors among the 17 evaluated responses. Of the 59 references evaluated, 41 (69%) were fabricated, though they appeared real. Most fabricated citations used names of authors with previous relevant publications, a title that seemed pertinent and a credible journal format.

**Interpretation:** When asked multiple medical questions, ChatGPT provided answers of limited quality for scientific publication. More importantly, ChatGPT provided deceptively real references. Users of ChatGPT should pay particular attention to the references provided before integration into medical manuscripts.

## Introduction

Large Language Models (LLM) constitute a branch of artificial intelligence (AI) at the intersection of linguistic and computer science(1). Trained on massive quantities of text-based data, LLMs have learned to interpret written input and produce language that is understandable to humans. LLMs incorporate several algorithms, including generative pre-trained transformers (GPT). This type of neural network architecture is useful in chatbots, rendering them particularly effective at simulating human conversations. On November 2022, San Francisco-based company OpenAI released a freely available version of ChatGPT, a LLM based on their proprietary GPT (GPT-3)(2). Since then, scientific articles have been written in part by ChatGPT and published(3-5). These publications were mainly to demonstrate the remarkable quality of the manuscripts written by ChatGPT. For example, a journal published an editorial for which the first five paragraphs were written by ChatGPT(4). In another publication, the researcher asked ChatGPT to discuss the potential impact of taking Rapamycin to increase longevity(5). The chatbot's output constituted most of the article. Finally, an article describing the potential use of ChatGPT for medical writing was mainly written by the chatbot(3). In these articles, the parts produced by the chatbot did not include any references. While being low, the number of publications discussing ChatGPT has exploded in the last weeks. It went from only 24 publications identified in Pubmed with the term "ChatGPT" on February 7<sup>th</sup>, 2023, to 92 publications on March 6<sup>th</sup>. These publications were mainly editorials or news from scientific journals praising the quality of the writing (1, 6-8) or questioning ethical aspects of using chatbot for scientific writing(9-12).

While most articles suggest that the answers provided by ChatGPT are acceptable, there is no information about the sources of its knowledge and no reference is usually provided by ChatGPT, unless specifically requested. This limitation is acknowledged by OpenAI as the model was trained on multiple internet texts with a large number of sources(2). However, when conversing with ChatGPT, one can prompt it regarding the references backing its assertions.

To our knowledge, no study has evaluated the quality and appropriateness of the references provided by ChatGPT. This study aimed to evaluate the quality of the answers and corresponding references provided by ChatGPT, when responding to a wide range of medical questions. A secondary objective arose during the construction of the study as many citations provided by ChatGPT were not found. Because of this, we aimed to assess the validity of the suggested references.

## Methods

This was an experimental observational study conducted in February 2023 evaluating the quality of the responses provided by ChatGPT (Version 3.5. OpenAI Inc, San Francisco, CA, USA) to 20 medical questions from diverse fields.

The medical questions were identified by selecting five research articles published at the end of 2022 in four high-impact factor medical journals (BMJ(13-17), CMAJ(18-22), the Lancet(23-32) and NEJM). These 20 articles spanned different topics and fields. The

questions asked to ChatGPT were related to the primary objectives of the 20 studies. In most cases, it was the stated primary objective of the study. In a few instances where the objective was deemed too narrow to ensure a minimal breadth of references, we formulated the question in a broader context. For example, the primary objective “*to critically examine the leadership experiences of African Nova Scotian nurses in health care systems?*”(18), was changed to: *What are the leadership experiences of African nurses in the United States health care systems?*

In general, questions to ChatGPT started as “*What are...*” (*see example above*), without any word limit or other constraint. Following the answer by ChatGPT, a follow-up question asked: *Do you have references for this?* All references were counted, but only the first three were used for the analysis. To promote the external validity of the study and considering that ChatGPT is sensitive to previous chats, questions were asked by the three co-authors on different computers and ChatGPT accounts.

The primary outcomes were: 1. The appropriateness of the references; 2. The quality of the responses. We initially aimed to evaluate the appropriateness of the references by multiple raters using the following verbal numeric scale: *On a scale of 0 to 100% where 0 signifies that the references had no relationship with the study question and 100% is for the three most pertinent references for this topic, how would you rate the references provided?* Our initial plan was to provide the articles to the raters. However, given that we failed to find the first six articles, we modified this outcome to evaluate whether the reference existed. To verify this, we searched Pubmed using the title and authors. If unsuccessful, we then searched in the journal’s website. To better describe the references provided by ChatGPT, we evaluated whether the authors listed had previous publication in the field. We also looked at the title (*Is it a title that seems pertinent for the study question?*) and the journal (*Does the citation look a plausible article for this journal?*).

The quality of the response was measured using the following verbal numeric scale from 0 to 100%: *On a scale of 0 to 100% where 0 is equal to no answer, 100% to a perfect answer and 50% to the minimum acceptable answer how would you rate the answer provided to the question?* Also, raters were asked to report any factual error found in the ChatGPT answer, as well as any other relevant qualitative feedback.

To ensure domain expertise for the selected articles, we invited the corresponding author of each article to act as rater and determine the quality of the response. We contacted each corresponding authors by email, inviting them to provide feedback on the answer related to their respective study objective. When the corresponding author did not reply, we contacted other listed authors. Each corresponding author was contacted at least three times before saying it was unsuccessful. These authors were deemed content experts in their field of publication.

The primary analysis of this study determined the validity of the references by calculating the proportion of references that really existed among all evaluated references. We also reported the proportion of references that listed an author with previous publications in the field of interest. The other primary analysis pertained to the quality of ChatGPT responses, reported as the median and interquartile range on the verbal numeric scale. We also reported the proportion of responses that contained factual errors according to the

raters. Minor errors referred to erroneous details with limited impact to the quality of the overall response (e.g. an overoptimistic affirmation that is not supported by any data), whereas major errors consisted of flagrant mistakes that invalidated the response (e.g. wrong pathophysiological explanation).

We did not conduct the analysis using the verbal numeric scale for the appropriateness of the references due to the small number of real references.

We had no prespecified idea of the median scores that would be obtained for the responses. It was estimated that the evaluation of at least 12 questions would provide a great variety of study subjects and allows to demonstrate the general quality of the responses/references. Also, this would lead to at least 30 references to evaluate. Based on this, and the premise that at least 60% of the invited authors would agree to help us, we invited 20 authors to rate 20 study questions in order to have at least 12 evaluations.

We did not seek institutional review board approval given that all data were publicly available, and no participants were involved. The study was completed without financial support. The manuscript is an honest, accurate, and transparent account of the study being reported; no important aspects of the study have been omitted. Any discrepancies from the study as originally planned was explained (change in the outcome regarding the validity of the references).

## **Results**

Each of the 20 study questions were asked to ChatGPT by a member of the study team (JG n= 10, MDG n=5 and EO n=5) and received a response that varied in length between 53 and 309 words. When prompted to provide its references, ChatGPT provided two to seven references. A total of 59 references were included in the primary analysis.

When searching for the references suggested by ChatGPT, we noted that while they looked credible (see figure 1), most of them were fabricated by ChatGPT. Indeed, 56/59 (95%) references contained authors with previous publications on a related topic found in Pubmed or were from recognized organizations (ex: CDC, FDA, etc.) (see table 1). Also, all titles seemed appropriate because it was related to the study question. In reality, 41/59 (69%) references were fabricated. Among the 18 real references, 11 were titles of real published articles (including three with minor citation errors and five with major citation errors), five were existing websites and two were books (Figure 2). The remaining reference (n=41) did not exist. Of those, 29 (71%) of the fabricated articles were reportedly published in a known medical journal, website or manuscript repository (ex: CDC or MedRxiv) using an appropriate format of citation (i.e.: they reported a year, volume number and page that were coherent with the journal). However, the reported volume and page range pertained to an unrelated article. All responses and references can be found on the web appendix.

Of the 20 corresponding authors, 17 (0.85) agreed to evaluate the responses. The median score they provided was 60% (1<sup>st</sup> and 3<sup>rd</sup> quartiles 50% and 85%). Raters identified major factual error in five (0.29) responses. For example, a rater identified that the mechanism of action of antipsychotic described by ChatGPT was incorrect. Another noted that

ChatGPT overestimated the global burden of mortality associated with Shigella infections by a factor of 10. Minor errors were identified in seven (0.41) other responses.

### **Interpretation**

This study demonstrated that over two-thirds of the references provided by ChatGPT to a diverse set of medical questions were fabricated, though most seemed deceptively real. Moreover, domain experts identified major factual errors in a quarter of the responses. These findings are alarming, given that trustworthiness is a pillar of scientific communication,

A previous study evaluated answers provided by ChatGPT using a scientific methodology in January 2023(33). This study demonstrated that ChatGPT's score (60%) was lower than that of Korean medical students (90%) in a single parasitology examination. Another study compared results of ChatGPT to two other LLMs on the United States Medical Licensing Examination Step 1 and Step 2 exams using two question banks. With a mean score of 60%, they concluded that ChatGPT outperformed the other chatbots and “achieved the equivalent of a passing score for a third-year medical student”(34). A recent comment published in Nature reported that “ChatGPT fabricated a convincing response that contained several factual errors”(35). To our knowledge, this is the first study evaluating the quality of the references provided by ChatGPT. However, a recent article described a manuscript written by ChatGPT using mock data(36). When asked to conduct a literature search, ChatGPT suggested 9 references. The author of the article acknowledged that: “*Interestingly, at least some of these references that ChatGPT suggested do not exist in the form that is presented here*”. We looked for these 9 articles and, none exist. Reference inaccuracies do not constitute a novelty in medical publishing, but usually relate to minor mistakes(37-40). For example, Browne noted that most of the errors in referencing among papers submitted to any of five radiology journals were related to a failure to follow author submission guidelines(37). Another study reported that 15% of articles published in a single journal had an error but it was most commonly a spelling or punctuation error(38). Finally, two studies reported that approximately 15% of articles had errors, and up to 4.5% consisted of major errors. However, these errors did not relate to purely fabricated references.

The importance of proper referencing is undeniable. As suggested by Glick, “you are what you cite”(41). The quality and breadth of the references provided demonstrates that the researchers have done a complete literature review and are knowledgeable about the topic. This process enables the integration of findings in the context of previous work, a fundamental aspect of medical research advancement(42). It limits the risk of biases. Failing to provide references is one thing but creating fake references would be considered fraudulent for researchers. When asked for references, ChatGPT made very appealing suggestions. It blended authors with a good research track to an interesting title in addition to a relevant journal like if the chatbot wanted to put the best of everything in a single reference. Some titles seemed to be the perfect article for our question. For example, for the question “*What is the impact of haloperidol in intensive care unit patients with delirium?*”, ChatGPT suggested the following inexistent title as reference: “*Haloperidol in critically ill patients with delirium: a randomized, placebo-controlled*

*trial*". In most cases, it suggested authors that published multiple scientific articles on the subject of interest.

This study highlights an important shortcoming of LLMs, namely their risk of being "confidently wrong". While such models have now reached astounding performance in simulating human conversations, the next stages in their development must improve information validity and responsible deployment. Though a series of disclaimers upon user registration highlight some of the limitations of ChatGPT(2), scientists considering the use of this tool to support manuscript preparation must be aware of its limitations. To promote responsible responses, chatbot developers should adapt the reward models that guide the algorithm's output, so that they learn to optimize the validity of the information and references provided. OpenAI uses a supervised learning framework described as reinforcement learning from human feedback(2), which should be amenable to such modifications.

These findings can help orient the use of LLMs by scientists in ways that are safe, contribute to their wellbeing and that of their teams, while targeting feasible applications(43). This includes the partial automation of tedious tasks, such as the initial steps in knowledge mapping and synthesis. This would free up time to critically appraise the information presented, thoroughly verify the corresponding sources, and orient subsequent searches. Future LLM iterations may also support manuscript formatting according to journal requirements, and facilitate knowledge translation to diverse audiences (e.g., translation, summarization and adjustment of readability). On the long term, use of LLM might improve knowledge translation by accelerating manuscript redaction and improve the quality of the writing. However, it could also threaten the scientific validity of manuscript if it incorporates inaccurate information and is misused(43). Researchers using ChatGPT may be misled by false information because clear, seemingly coherent and stylistically appealing references can conceal poor content quality. Journals should consider clear guidelines regarding the allowed uses and reporting guidelines when tools such as ChatGPT are used. In the future, such tools may be re-designed to support human reviewers and publishers when appraising submitted articles and their corresponding bibliography. This use case highlights the enormous potential and alarming pitfalls of LLM integration in scientific writing. The inherent risk of overconfidence among LLMs also emphasizes the need for "humans in the loop" as a key element for their responsible implementation.

There are limitations to this study. First, we always asked the same question regarding references and did not specify to ChatGPT to limit references to published articles. However, when trying this *a posteriori*, we obtained similar answers. Second, the information provided was accurate as of February 2023 and may improve in the following months/years. For example, there were two questions related to the COVID 19 pandemic. ChatGPT being constructed based on data published before sept 2021, it may have been difficult to find sufficient knowledge for adequate answers. Also, ChatGPT has been launched publicly in November 2022, with the stated objective of iterative cycles of learning and improvement. It is probable that LLM will improve and learn to provide responses exempt from factual error. Finally, we used multiple raters to measure the quality of the responses. We preferred to have experts in each specific field than to have experts in scoring.

## **Conclusion**

ChatGPT proposes undeniable progress. This study should alert the scientific community to be careful about the important risks of relying on its references because it is assembling very convincing, yet often fabricated citations. To be useful for medical editing, chatbots should embrace the values of the scientific community such as integrity and completeness. Considering the speed of improvement of LLMs, we are hopeful that future versions of ChatGPT will suggest more accurate responses when asked to provide references.

## **Acknowledgements**

The study team would like to acknowledge the contribution of Drs Changhai Ding, Areef Ishani, Yazdan Yazdanpanah, Nina Christine Andersen-Ranberg, Marc Rothenberg, Evan S Dellon, Ken Parhar, Jason Weatherald, Alberto Ruano Raviña, John Cleland, John H Krystal, Luciana Vercoza Viana, Kevin Ituka, Miriam Sander, Amanda Roberts, Fan Wang and all many other authors who agreed to rate the quality of the responses for this study.



1. Kitamura FC. ChatGPT Is Shaping the Future of Medical Writing but Still Requires Human Judgment. *Radiology*. 2023:230171.
2. ChatGPT: Optimizing language models for dialogue. : OpenAI; [updated November 30, 2022. Available from: <https://openai.com/blog/chatgpt/>.
3. Biswas S. ChatGPT and the Future of Medical Writing. *Radiology*. 2023:223312.
4. O'Connor S, ChatGpt. Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse? *Nurse Educ Pract*. 2023;66:103537.
5. Chat GPTGP-tT, Zhavoronkov A. Rapamycin in the context of Pascal's Wager: generative pre-trained transformer perspective. *Oncoscience*. 2022;9:82-4.
6. Else H. Abstracts written by ChatGPT fool scientists. *Nature*. 2023;613(7944):423.
7. Gao CAH, F.M.; Markov, N.S.; Dyer, E.C.; Ramesh, S.; Luo, Y.; Pearson, A.T. Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers 2023 [Available from: <https://www.biorxiv.org/content/10.1101/2022.12.23.521610v1>.
8. Cahan P, Treutlein B. A conversation with ChatGPT on the role of computational systems biology in stem cell research. *Stem Cell Reports*. 2023;18(1):1-2.
9. Thorp HH. ChatGPT is fun, but not an author. *Science*. 2023;379(6630):313.
10. Stokel-Walker C. ChatGPT listed as author on research papers: many scientists disapprove. *Nature*. 2023;613(7945):620-1.
11. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. *Nature*. 2023;613(7945):612.
12. Looi MK. Sixty seconds on . . . ChatGPT. *BMJ*. 2023;380:205.
13. Weatherald J, Parhar KKS, Al Duhailib Z, Chu DK, Granholm A, Solverson K, et al. Efficacy of awake prone positioning in patients with covid-19 related hypoxemic respiratory failure: systematic review and meta-analysis of randomized trials. *BMJ*. 2022;379:e071966.
14. Xie J, Wang M, Long Z, Ning H, Li J, Cao Y, et al. Global burden of type 2 diabetes in adolescents and young adults, 1990-2019: systematic analysis of the Global Burden of Disease Study 2019. *BMJ*. 2022;379:e072385.
15. Santer M, Muller I, Becque T, Stuart B, Hooper J, Steele M, et al. Eczema Care Online behavioural interventions to support self-care for children and young people: two independent, pragmatic, randomised controlled trials. *BMJ*. 2022;379:e072007.
16. McIlroy DR, Shotwell MS, Lopez MG, Vaughn MT, Olsen JS, Hennessy C, et al. Oxygen administration during surgery and postoperative organ injury: observational cohort study. *BMJ*. 2022;379:e070941.
17. Candal-Pedreira C, Ross JS, Ruano-Ravina A, Egilman DS, Fernandez E, Perez-Rios M. Retracted papers originating from paper mills: cross sectional study. *BMJ*. 2022;379:e071517.
18. Jefferies K, Martin-Misener R, Murphy GT, Gahagan J, Bernard WT. African Nova Scotian nurses' perceptions and experiences of leadership: a qualitative study informed by Black feminist theory. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*. 2022;194(42):E1437-E47.
19. Zhu Z, Huang JY, Ruan G, Cao P, Chen S, Zhang Y, et al. Metformin use and associated risk of total joint replacement in patients with type 2 diabetes: a population-

based matched cohort study. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*. 2022;194(49):E1672-E84.

20. Rudoler D, Peterson S, Stock D, Taylor C, Wilton D, Blackie D, et al. Changes over time in patient visits and continuity of care among graduating cohorts of family physicians in 4 Canadian provinces. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*. 2022;194(48):E1639-E46.

21. Skowronski DM, Kaweski SE, Irvine MA, Kim S, Chuang ESY, Sabaiduc S, et al. Serial cross-sectional estimation of vaccine-and infection-induced SARS-CoV-2 seroprevalence in British Columbia, Canada. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*. 2022;194(47):E1599-E609.

22. Naveed Z, Li J, Spencer M, Wilton J, Naus M, Garcia HAV, et al. Observed versus expected rates of myocarditis after SARS-CoV-2 vaccination: a population-based cohort study. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*. 2022;194(45):E1529-E36.

23. Kalra PR, Cleland JGF, Petrie MC, Thomson EA, Kalra PA, Squire IB, et al. Intravenous ferric derisomaltose in patients with heart failure and iron deficiency in the UK (IRONMAN): an investigator-initiated, prospective, randomised, open-label, blinded-endpoint trial. *Lancet*. 2022;400(10369):2199-209.

24. Krystal JH, Kane JM, Correll CU, Walling DP, Leoni M, Duvvuri S, et al. Emraclidine, a novel positive allosteric modulator of cholinergic M4 receptors, for the treatment of schizophrenia: a two-part, randomised, double-blind, placebo-controlled, phase 1b trial. *Lancet*. 2022;400(10369):2210-20.

25. Collaborators GBDAR. Global mortality associated with 33 bacterial pathogens in 2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet*. 2022;400(10369):2221-48.

26. Sheikh J, Allotey J, Kew T, Fernandez-Felix BM, Zamora J, Khalil A, et al. Effects of race and ethnicity on perinatal outcomes in high-income and upper-middle-income countries: an individual participant data meta-analysis of 2 198 655 pregnancies. *Lancet*. 2022;400(10368):2049-62.

27. Kramer CK, Leitao CB, Viana LV. The impact of urbanisation on the cardiometabolic health of Indigenous Brazilian peoples: a systematic review and meta-analysis, and data from the Brazilian Health registry. *Lancet*. 2022;400(10368):2074-83.

28. Ishani A, Cushman WC, Leatherman SM, Lew RA, Woods P, Glassman PA, et al. Chlorthalidone vs. Hydrochlorothiazide for Hypertension-Cardiovascular Events. *The New England journal of medicine*. 2022;387(26):2401-10.

29. Team PS, Kieh M, Richert L, Beavogui AH, Grund B, Leigh B, et al. Randomized Trial of Vaccines for Zaire Ebola Virus Disease. *The New England journal of medicine*. 2022;387(26):2411-24.

30. Andersen-Ranberg NC, Poulsen LM, Perner A, Wetterslev J, Estrup S, Hastbacka J, et al. Haloperidol for the Treatment of Delirium in ICU Patients. *The New England journal of medicine*. 2022;387(26):2425-35.

31. Farber A, Menard MT, Conte MS, Kaufman JA, Powell RJ, Choudhry NK, et al. Surgery or Endovascular Therapy for Chronic Limb-Threatening Ischemia. *The New England journal of medicine*. 2022;387(25):2305-16.

32. Dellon ES, Rothenberg ME, Collins MH, Hirano I, Chehade M, Bredenoord AJ, et al. Dupilumab in Adults and Adolescents with Eosinophilic Esophagitis. *The New England journal of medicine*. 2022;387(25):2317-30.
33. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. *J Educ Eval Health Prof*. 2023;20:1.
34. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med Educ*. 2023;9:e45312.
35. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. *Nature*. 2023;614(7947):224-6.
36. Macdonald C, Adeloye D, Sheikh A, Rudan I. Can ChatGPT draft a research article? An example of population-level vaccine effectiveness analysis. *J Glob Health*. 2023;13:01003.
37. Browne RF, Logan PM, Lee MJ, Torreggiani WC. The accuracy of references in manuscripts submitted for publication. *Can Assoc Radiol J*. 2004;55(3):170-3.
38. O'Connor AE, Lukin W, Eriksson L, O'Connor C. Improvement in the accuracy of references in the journal *Emergency Medicine Australasia*. *Emerg Med Australas*. 2013;25(1):64-7.
39. Montenegro TS, Hines K, Partyka PP, Harrop J. Reference accuracy in spine surgery. *J Neurosurg Spine*. 2020:1-5.
40. Al-Benna S, Rajgarhia P, Ahmed S, Sheikh Z. Accuracy of references in burns journals. *Burns*. 2009;35(5):677-80.
41. Glick M. You are what you cite: the role of references in scientific publishing. *J Am Dent Assoc*. 2007;138(1):12, 4.
42. Gasparyan AY, Yessirkepov M, Voronov AA, Gerasimov AN, Kostyukova EI, Kitas GD. Preserving the Integrity of Citations and References by All Stakeholders of Science Communication. *J Korean Med Sci*. 2015;30(11):1545-52.
43. Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. *NPJ Digit Med*. 2021;4(1):93.

Figure 1 screenshot of an example of responses provided by ChatGPT

GR Is metformin use associated with a reduced risk of total knee replacement or total hip replacement among patients with type 2 diabetes mellitus ?

There is limited evidence to suggest that metformin use may be associated with a reduced risk of total knee or total hip replacement among patients with type 2 diabetes mellitus. However, more research is needed to establish a definite relationship between metformin use and the risk of joint replacement in this patient population.

GR Do you have references for this ?

Yes, here are a few references to support the statement:

1. Kim JH, Kim J, Kim HK, Kim JH, Cho JH, Kim JP. Metformin Use and Reduced Risk of Total Knee Replacement and Total Hip Replacement in Patients with Type 2 Diabetes Mellitus: A Nationwide Population-Based Study. *Diabetes Metab J*. 2019 Dec;43(6):470-475. doi: 10.4093/dmj.2018.0129.
2. Lu Y, Huang Y, Li L, et al. The Protective Effect of Metformin on Joint Replacements in Patients with Type 2 Diabetes Mellitus: A Systematic Review and Meta-Analysis. *J Diabetes Res*. 2016;2016:6247936. Published 2016 Jun 10. doi:10.1155/2016/6247936
3. Lee YH, Kim BJ, Kim K, et al. Metformin Use and the Risk of Joint Replacements in Patients with Type 2 Diabetes Mellitus: A Nationwide Retrospective Cohort Study. *PLoS One*. 2015;10(8):e0136320. Published 2015 Aug 28. doi:10.1371/journal.pone.0136320

Figure 2 Distribution of the references provided by ChatGPT (n=59)

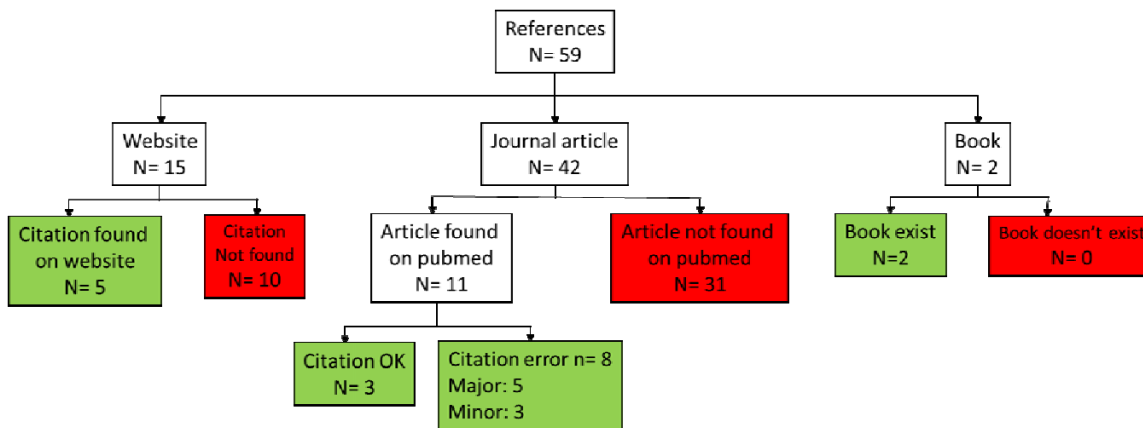


Table 1 Characteristics of the references provided by ChatGPT (n=59)

	N (%)
Type of reference format	
• Journal	• 42 (71)
• Website	• 15 (25)
• Book	• 2 (3)
One or more author has published on the topic	• 56 (95)
The title seems appropriate	• 59 (100)
Journal/website exists and format adequate	• 48 (81)
Title found on Pumed	• 11* (19)
Article found on the journal/website	• 8 (13)

\* Three articles with minor citation errors (volume, and pages) and five with fatal citation errors (wrong year, volume, and pages; or wrong authors)