Warning: Artificial intelligence chatbots can generate inaccurate medical and scientific information and references

Catherine L. Clelland1, Stuart Moss2, James D. Clelland3*

1Department of Psychiatry and Taub Institute, Columbia University Irving Medical Center, New York, NY 10032, USA
2Health Services Library, The Nathan Kline Institute for Psychiatric Research, Orangeburg, NY 10962, USA
3Clinical Research Department, The Nathan Kline Institute for Psychiatric Research, Orangeburg, NY 10962, USA

*Correspondence: James D. Clelland, Clinical Research Department, The Nathan Kline Institute for Psychiatric Research, 140 Old Orangeburg Road, Orangeburg, NY 10962, USA. James.Clelland@nki.rfmh.org

Academic Editor: Hari Prasad Devkota, Kumamoto University, Japan

Received: April 28, 2023 Accepted: August 3, 2023 Published: January 10, 2024

Cite this article: Clelland CL, Moss S, Clelland JD. Warning: Artificial intelligence chatbots can generate inaccurate medical and scientific information and references. Explor Digit Health Technol. 2024;2:1–6. https://doi.org/10.37349/edht.2024.00006

Abstract
The use of generative artificial intelligence (AI) chatbots, such as ChatGPT and YouChat, has increased enormously since their release in late 2022. Concerns have been raised over the potential of chatbots to facilitate cheating in education settings, including essay writing and exams. In addition, multiple publishers have updated their editorial policies to prohibit chatbot authorship on publications. This article highlights another potentially concerning issue; the strong propensity of chatbots in response to queries requesting medical and scientific information and its underlying references, to generate plausible looking but inaccurate responses, with the chatbots also generating nonexistent citations. As an example, a number of queries were generated and, using two popular chatbots, demonstrated that both generated inaccurate outputs. The authors thus urge extreme caution, because unwitting application of inconsistent and potentially inaccurate medical information could have adverse outcomes.

Keywords
ChatGPT, YouChat, artificial intelligence, inaccurate information

Since late 2022, the use of generative artificial intelligence (AI) chatbots including ChatGPT (https://chat.openai.com/chat) and YouChat (https://you.com) has exploded, with ChatGPT becoming the fastest growing consumer application in history, reaching 100 million monthly active users in January 2023 following its release only two months prior [1]. The number of AI chatbots is rapidly expanding, with more recent examples including Claude (https://claude.ai/login), character.ai (https://beta.character.ai/), Perplexity (https://www.perplexity.ai/), and Bard (https://bard.google.com/). There has been much concern expressed over the potential misuse of generative AI in essay writing and exams in kindergarten through 12th grade (K-12), college, and even US medical licensing exams [2, 3]. Many publishers have also now updated their editorial policies to prohibit chatbot authorship on publications, with the Science...
journals recently updating their editorial policies to clarify that text and figures generated by AI tools cannot be used in manuscripts and an AI cannot be an author [4], and the Nature portfolio clarifying that an AI tool cannot be an author [5]. Indeed, the Lancet groups’ guide for authors specifically documents that AI or AI-assisted technology should not be used when producing scientific insights, analyzing and interpreting data, or drawing scientific conclusions. Nevertheless, given the descriptive power of chatbots, and their interface similarity to actual search engines such as the National Center for Biotechnology Information (NCBI)’s PubMed, their use for medical and scientific literature searches seems inevitable.

This article was written to highlight a continuing issue and urges the use of caution because research using AI chatbots, queries requesting medical information and its underlying references using ChatGPT and YouChat, can often generate plausible looking but inaccurate responses, with the chatbots also generating nonexistent citations. For example, in April 2023 we generated a simple question, asking two AI chatbots (ChatGPT and YouChat): “Describe useful treatments for hyperprolinemia type II. Provide full citations.”. In one iteration of the query response, ChatGPT generated very plausible looking information and citations (see Figure 1A). In the summary response, two treatments were noted with references included. The first was high-dose pyridoxine (vitamin B6) supplementation (up to 200 mg/kg per day), which was reported by Pérez-Dueñas et al., to be effective in reducing plasma proline levels and improving neurological symptoms in some hyperprolinemia type II (HPII) patents. While this does initially appear to be genuine, because pyridoxine is a routine therapy for vitamin B6 dependent epilepsies (which includes HPII) [6], the European Journal of Pediatric Neurology citation does not seem to exist. Specifically, the 2010 volume 14, number 6 issue does not include pages 465–468, the web-link provided (https://doi.org/10.1016/j.ejpn.2010.02.003) links to a paper reporting “school performance in a cohort of children with CNS inflammatory demyelination”, no manuscripts with the citation title are retrieved during searches with either PubMed or Google search engines, and neither the first nor the last author have any citations retrieved in PubMed using the search term “hyperprolinemia” or “pyrroline-5-carboxylate reductase”, although they have co-authored manuscripts on other childhood neurological disorders. To our knowledge, no studies have reported reduced proline levels or improvement of non-seizure symptoms following pyridoxine supplementation in HPII. Moreover, and of particular concern, the dose reported by ChatGPT seems relevant for infant intravenous delivery but is above the recommended dosage for long-term oral dosing [6, 7]. Thus, this query response is vague in details, is not actually supported by data, and is even potentially dangerous given the potential for incorrect dosing. Similarly, the second potential treatment (betaine supplementation) includes a nonexistent citation, and we could find no reports supporting betaine supplementation for HPII, although betaine (also known as trimethylglycine) is routinely used for lowering homocysteine levels in pyridoxine non-responsive cystathionine beta-synthase deficient homocystinuria (HCU), and Forny P, the first author of the second AI referenced source (Figure 1A), has publications in this field (for example, [8]).

When provided with the same query, YouChat generated the response shown in Figure 1B. Once again vitamin B6 was suggested as a potential treatment that reduces proline levels. The first citation references a National Institutes of Health (NIH) website, but the web-link is broken and does not retrieve an actual page. The NIH website (https://rarediseases.info.nih.gov) is valid and includes HPII, but the information provided does not include any data supporting vitamin B6 treatment for HPII. The second reported citation retrieves an unrelated paper (https://pubmed.ncbi.nlm.nih.gov/9643291/), while the third source is actually a broken web-link from the European Peptide Society, rather than the European Society of Pediatric Neurology.

To test whether this was a single inaccurate event due to a broad question being asked, a very specific query was generated regarding the effectiveness of proline as an inhibitor of glutamic acid dehydrogenase. We asked ChatGPT and YouChat: “What is the Ki of proline for mammalian glutamic acid dehydrogenase? List citations.”. Again, ChatGPT generated plausible looking information and citations, with proline inhibition constants (Ki) of 2.8 mmol/L, 0.46 mmol/L, and 0.48 mmol/L, along with citations. YouChat stated that: “The Ki of proline for mammalian glutamic acid dehydrogenase (GAD) is approximately 0.02–0.1 mmol/L.” and generated 3 citations (Figure 1C and 1D, respectively). Again, none of the generated
Figure 1. Medical and scientific questions related to hyperprolinemia. A. ChatGPT response to “Describe useful treatments for hyperprolinemia type II. Provide full citations.”; B. YouChat response to “Describe useful treatments for hyperprolinemia type II. Provide full citations.”; C. ChatGPT response to “What is the Ki of proline for mammalian glutamic acid dehydrogenase? List citations.”; D. YouChat response to “What is the Ki of proline for mammalian glutamic acid dehydrogenase? List citations.”

citations exist. The only citation known to us that addresses the test question, states a Ki of 7.84 mmol/L [9], and was not returned by any of the AI queries. Different iterations of the same query generated different outputs, but each time the information was not genuine.

Interestingly, more recent queries of ChatGPT (July 2023) using the same HPII search initially produced the following response: “I apologize, but as of my last knowledge update in September 2021, there isn’t an established treatment for hyperprolinemia type II. It’s important to note that medical advancements and research are continually progressing, and new treatments may have been developed since then. I recommend consulting a healthcare professional or referring to recent medical literature for the most up-to-date information on potential treatments for hyperprolinemia type II.”. However, re-querying later the same day produced a response that provided information on HPII and potential medical treatments, but with no supporting citations. In July 2023, ChatGPT searches of the question “Describe useful treatments for scurvy. Provide full citations.” (Figure 2) and “Describe useful treatments for DiGeorge syndrome. Provide full citations.” continued to provide non-relevant, inaccurate, or nonexistent supporting citations (output not shown).

Thus, as convincing as chatbots may appear, we would warn professionals and the public that, in terms of medical or scientific inquiry, they should be regarded as currently unreliable, particularly as their output is very inconsistent, and can closely mimic the language style prevalent in the literature and thus upon first read the information seems genuine and relevant to the topic queried. Even though chatbot disclaimers exist (ChatGTP: “May occasionally generate incorrect information.”; YouChat: “This product is in beta and its accuracy may be limited.”), and chatbots can be trained to realize their limitations and refuse to respond
Scurvy is a condition caused by a deficiency of vitamin C (ascorbic acid) and can lead to symptoms such as fatigue, weakness, joint pain, and bleeding gums. The treatment for scurvy primarily involves replenishing vitamin C levels in the body through dietary changes or vitamin C supplementation. Here are some useful treatments for scurvy along with their citations:

1. Vitamin C Supplementation:
   Vitamin C supplements are a common and effective treatment for scurvy. These supplements are available in various forms such as ascorbic acid or sodium ascorbate. The dosage and duration of treatment may vary depending on the severity of the deficiency and individual patient factors.


2. Dietary Changes:
   Increasing the intake of vitamin C-rich foods in the diet can help treat and prevent scurvy. Foods rich in vitamin C include citrus fruits (oranges, lemons, grapefruits), strawberries, kiwifruits, bell peppers, and broccoli.


3. Intravenous Vitamin C:
   In severe cases of scurvy or when oral supplementation is not feasible, intravenous administration of vitamin C may be necessary. Intravenous vitamin C therapy allows for rapid replenishment of vitamin C levels in the body.


4. Medical Supervision:
   It is essential for individuals with scurvy to be under medical supervision to monitor their progress and ensure appropriate treatment. A healthcare professional can assess the severity of the deficiency and recommend the most suitable treatment plan.


Remember that while these treatments are effective for scurvy, prevention is always the best approach. Consuming an adequate amount of vitamin C through a balanced diet is crucial to avoid developing a deficiency. If you suspect you have scurvy or any nutritional deficiency, it’s essential to seek medical advice for proper diagnosis and treatment.
self-medicating with dietary supplements, chatbots may have a harmful impact, and we propose that responses to medical queries should come with strong warnings about the potential to generate inaccurate information, or at the very minimum it seems prudent that all AI-generated information and references be verified by the user.

**Abbreviations**
AI: artificial intelligence  
HPII: hyperprolinemia type II

**Declarations**
**Author contributions**
JDC and CLC: Conceptualization, Investigation, Writing—original draft, Writing—review & editing. SM: Writing—review & editing. All authors read and approved the submitted version.

**Conflicts of interest**
The authors declare that they have no conflicts of interest.

**Ethical approval**
Not applicable.

**Consent to participate**
Not applicable.

**Consent to publication**
Not applicable.

**Availability of data and materials**
Not applicable.

**Funding**
Not applicable.

**Copyright**
© The Author(s) 2024.

**References**
3. Varanasi L. GPT-4 can ace the bar, but it only has a decent chance of passing the CFA exams. Here’s a list of difficult exams the ChatGPT and GPT-4 have passed. [Internet]. Insider Inc; 2023 [cited 2023 Feb 19]. Available from: https://www.businessinsider.com/list-here-are-the-exams-chatgpt-has-passed-so-far-2023-1?op=1#us-medical-licensing-exam-2

