

High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content

Review began 05/11/2023
Review ended 05/16/2023
Published 05/19/2023

© Copyright 2023
Bhattacharyya et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Mehul Bhattacharyya¹, Valerie M. Miller², Debjani Bhattacharyya³, Larry E. Miller¹

1. Clinical Research, Miller Scientific, Johnson City, USA 2. Leadership, University of the Cumberlands, Williamsburg, USA 3. Education, University of Massachusetts Lowell, Lowell, USA

Corresponding author: Larry E. Miller, larry@millerscientific.com

Abstract

Background

The availability of large language models such as Chat Generative Pre-trained Transformer (ChatGPT, OpenAI) has enabled individuals from diverse backgrounds to access medical information. However, concerns exist about the accuracy of ChatGPT responses and the references used to generate medical content.

Methods

This observational study investigated the authenticity and accuracy of references in medical articles generated by ChatGPT. ChatGPT-3.5 generated 30 short medical papers, each with at least three references, based on standardized prompts encompassing various topics and therapeutic areas. Reference authenticity and accuracy were verified by searching Medline, Google Scholar, and the Directory of Open Access Journals. The authenticity and accuracy of individual ChatGPT-generated reference elements were also determined.

Results

Overall, 115 references were generated by ChatGPT, with a mean of 3.8 ± 1.1 per paper. Among these references, 47% were fabricated, 46% were authentic but inaccurate, and only 7% were authentic and accurate. The likelihood of fabricated references significantly differed based on prompt variations; yet the frequency of authentic and accurate references remained low in all cases. Among the seven components evaluated for each reference, an incorrect PMID number was most common, listed in 93% of papers. Incorrect volume (64%), page numbers (64%), and year of publication (60%) were the next most frequent errors. The mean number of inaccurate components was 4.3 ± 2.8 out of seven per reference.

Conclusions

The findings of this study emphasize the need for caution when seeking medical information on ChatGPT since most of the references provided were found to be fabricated or inaccurate. Individuals are advised to verify medical information from reliable sources and avoid relying solely on artificial intelligence-generated content.

Categories: Medical Education, Public Health, Healthcare Technology

Keywords: references, machine learning, large language model, chatgpt, artificial intelligence

Introduction

Large language models (LLMs) are sophisticated artificial intelligence (AI) systems that are capable of understanding and responding to prompts in a manner resembling human communication. These models are trained on massive amounts of data in order to recognize statistical relationships between words, allowing them to generate almost instantaneous responses to even the most complex questions. LLMs are routinely utilized across numerous applications such as text translation, content and product recommendation systems, and virtual assistants. Some common LLMs include Bidirectional Encoder Representations from Transformers (BERT), Language Model for Dialogue Applications (LaMDA), and Chat Generative Pre-trained Transformer (ChatGPT) [1].

The widespread availability of ChatGPT (OpenAI) [2] has enabled users from various backgrounds to receive answers to a wide variety of questions, including those related to medical topics. However, the accuracy of ChatGPT responses is a topic of ongoing debate. Providing accurate medical content is critically important, especially when used for medical education by the general public. The credibility of this medical content partly relies on the sources used to provide this information. To the extent that source referencing may be missing or incorrect, the integrity of the associated content may be compromised.

How to cite this article

Bhattacharyya M, Miller V M, Bhattacharyya D, et al. (May 19, 2023) High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content. Cureus 15(5): e39238. DOI 10.7759/cureus.39238

ChatGPT was released in November 2022, and related peer-reviewed literature has only recently started to emerge. An interesting characteristic of AI is that it tends to “hallucinate”, a phenomenon where nonsensical or inaccurate content is generated [3]. While several studies have evaluated the accuracy of ChatGPT medical content [4,5], the accuracy of the citations within this content remains to be determined. Athaluri et al. [6] reported that 16% of the references cited within ChatGPT-generated medical content were fabricated. In order to extend this previous research, the current study investigated the proportion of fabricated and inaccurate references in biomedical articles created by ChatGPT by examining the impact of prompt variations and inaccuracies within individual reference elements.

Materials And Methods

On April 12, 2023, ChatGPT-3.5 was prompted to generate 30 unique short papers on various biomedical topics. The prompts were developed to summarize a specific topic within common therapeutic areas intended for a general audience. The standardized prompt is shown in Table 1.

Write a short paper intended for a general audience that summarizes A within the field of B. Include at least 3 references to papers that are indexed in Medline. Provide in-text citations and a complete reference list, including the PMID number for each.

The prompt variations included:

A:

preventative strategies

recent advances

healthcare disparities

B:

oncology

immunology

cardiology

orthopedics

neurology

infectious disease

gastroenterology

dermatology

pulmonology


mental health

TABLE 1: Standardized prompt with variations used to develop 30 ChatGPT-generated biomedical papers with references.

ChatGPT: Chat Generative Pre-trained Transformer.

Each prompt was issued in a new chat window to avoid duplicating prior responses. Figure 1 provides an example prompt and response generated within ChatGPT.

L Write a short paper intended for a general audience that summarizes prevention strategies within the field of dermatology. Include at least 3 references to papers that are indexed in Medline. Provide in-text citations and a complete reference list, including the PMID number for each.

 Dermatology is the branch of medicine that deals with the diagnosis and treatment of skin, hair, and nail disorders. Skin disorders, in particular, are highly prevalent, and their incidence is increasing worldwide. Therefore, preventing skin disorders is crucial to maintaining skin health and preventing associated complications. This paper will summarize some of the prevention strategies within the field of dermatology.

Sun protection is one of the most important prevention strategies in dermatology. Prolonged sun exposure is the primary cause of skin cancer and skin aging. Therefore, using sunscreens, wearing protective clothing, and avoiding the sun during peak hours are essential in preventing skin damage. A study by Chen et al. (2021) found that sunscreen use significantly reduced the incidence of melanoma and squamous cell carcinoma, indicating the importance of sun protection.

Another important prevention strategy is early detection and treatment of skin disorders. Regular skin exams by a dermatologist can help detect and treat skin cancer and other skin disorders early. In a study by Telfer et al. (2009), regular skin exams led to the detection of more skin cancers at earlier stages, resulting in better treatment outcomes.

Maintaining a healthy lifestyle is also critical in preventing skin disorders. A balanced diet rich in vitamins and minerals, regular exercise, and stress management can help prevent skin aging and reduce the risk of skin cancer. A study by Kwon et al. (2015) found that a healthy lifestyle, including a balanced diet and regular exercise, was associated with a reduced risk of skin aging in middle-aged women.

In conclusion, prevention strategies in dermatology are crucial in maintaining skin health and preventing associated complications. Sun protection, early detection and treatment of skin disorders, and maintaining a healthy lifestyle are some of the key prevention strategies. By following these strategies, we can protect our skin and maintain its health.

References:

Chen, A. C., Martin, A. J., Choy, B., Fernandez-Penas, P., & Dalziel, R. A. (2021). A systematic review of the relationship between sunscreen use and the prevention of melanoma and non-melanoma skin cancer. *European Journal of Dermatology*, 31(1), 1-9. PMID: 33635047

Kwon, H. H., Yoon, J. Y., Park, S. Y., Min, S., Suh, D. H., & Kim, J. H. (2015). A comprehensive study on the association between diet and skin aging in Korean women. *Journal of the American Academy of Dermatology*, 73(2), 242-250. PMID: 26183923

Telfer, N. R., Colver, G. B., Morton, C. A., & British Association of Dermatologists Therapy Guidelines and Audit Subcommittee. (2009). Guidelines for the management of basal cell carcinoma. *British Journal of Dermatology*, 161(4), 643-659. PMID: 19663808

FIGURE 1: Example of fabricated and inaccurate references in ChatGPT-3.5 generated output.

Chen et al. and Kwon et al. are fabricated references. The Telfer et al. reference has correctly listed authors, title, and journal, but the year, volume, page numbers, and PMID number are inaccurate. Ultimately, this output produced no references deemed authentic and accurate. ChatGPT: Chat Generative Pre-trained Transformer.

For each generated paper, we first analyzed them for AI-generated content and plagiarism using a commercially available program (Originality.AI) [7]. The software reported the probability that the text was AI-generated and calculated the percentage of plagiarized text, both scored from 0% to 100%. Next, two researchers with expertise in systematic reviews independently searched Medline, Google Scholar, and the Directory of Open Access Journals to verify the authenticity and accuracy of references provided by ChatGPT. The consensus was determined by discussion. In the context of this study, authentic references were confirmed to exist, authentic but inaccurate references contained incorrect information despite their

existence, and fabricated references were completely nonexistent and fabricated by the ChatGPT model.

We determined the frequency of fabricated and authentic references, as well as the accuracy of the individual elements within each reference. We assessed seven reference elements: authors, title, journal, year, volume, pages, and PubMed Identifier (PMID) number. Finally, we determined whether the frequency of fabricated references differed among various prompts using Fisher's exact test. To ensure an adequate sample size, we calculated that a minimum of 90 references were needed. This calculation assumed 30 prompts with a minimum of three references per prompt, a two-sided 95% confidence interval with a half-width of 10%, and a 50% rate of fabricated references. Statistical significance was defined as $p < 0.05$.

Results

Among the 30 ChatGPT-generated papers, the mean length was 338 ± 42 words. Plagiarism was minimal, with a mean score of $5 \pm 7\%$. All ChatGPT-generated papers received an AI score of 100%, indicating that the AI-detection software was 100% confident that each paper was AI-generated. ChatGPT generally followed the primary prompt instructions, providing in-text citations for 87% (26/30) of papers and at least three references for 97% (29/30). Overall, 115 references were generated, with a mean of 3.8 ± 1.1 per paper.

Among the 115 references, 47% were fabricated, 46% were authentic but inaccurate, and only 7% were authentic and accurate. We noted statistically significant differences in the percentage of fabricated references based on prompt variations. For prompt A variations, fabricated references were considerably more common ($p=0.007$) in papers on healthcare disparity (66%) than for prevention strategies (36%) or recent advances (34%). For prompt B variations, the highest fabricated reference rates were in the fields of pulmonology (75%), dermatology (64%), and gastroenterology (62%), and these rates statistically differed among all therapeutic areas ($p=0.03$). Despite these statistical differences, the frequency of authentic and accurate references was low among all prompt variations (Table 2).

Variable	Fabricated reference	Authentic reference		p-value*
		Inaccurate	Accurate	
Overall	47% (54/115)	46% (53/115)	7% (8/115)	
Prompt A				0.007
Healthcare disparity	66% (29/44)	32% (14/44)	2% (1/44)	
Prevention strategies	36% (13/36)	56% (20/36)	8% (3/36)	
Recent advances	34% (12/35)	54% (19/35)	11% (4/35)	
Prompt B				0.03
Pulmonology	75% (9/12)	25% (3/12)	0% (0/12)	
Dermatology	64% (7/11)	36% (4/11)	0% (0/11)	
Gastroenterology	62% (8/13)	38% (5/13)	0% (0/13)	
Mental health	60% (6/10)	40% (4/10)	0% (0/10)	
Oncology	57% (8/14)	36% (5/14)	7% (1/14)	
Orthopedics	50% (5/10)	50% (5/10)	0% (0/10)	
Cardiology	31% (4/13)	54% (7/13)	15% (2/13)	
Neurology	31% (4/13)	54% (7/13)	15% (2/13)	
Infectious disease	22% (2/9)	56% (5/9)	22% (2/9)	
Immunology	10% (1/10)	80% (8/10)	10% (1/10)	

TABLE 2: Authenticity and accuracy of references within ChatGPT-generated medical content.

*p-value derived from Fisher's exact test comparing the proportion of fabricated vs. authentic references.

Among the seven components evaluated for each reference, an incorrect PMID number was most common, listed in 93% of papers. Incorrect volume (64%), page numbers (64%), and year of publication (60%) were the

next most frequent errors (Figure 2). The mean number of incorrect components was 4.3 ± 2.8 per reference (Figure 3).

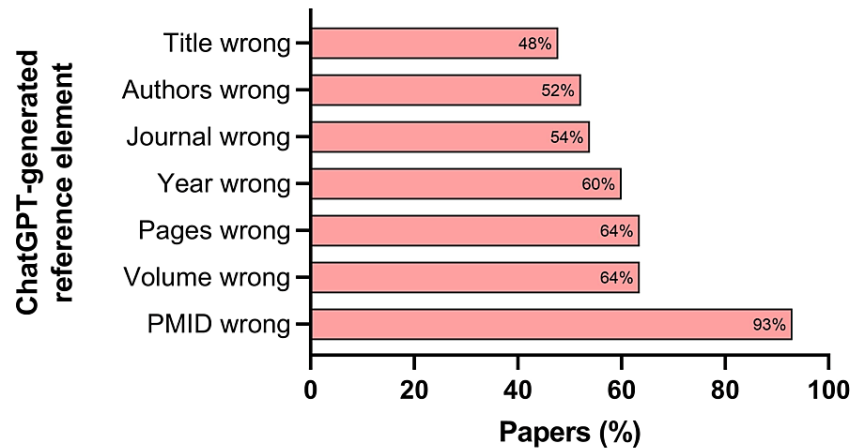


FIGURE 2: Frequency of inaccurate individual reference elements in ChatGPT-generated output.

PMID: PubMed Identifier.

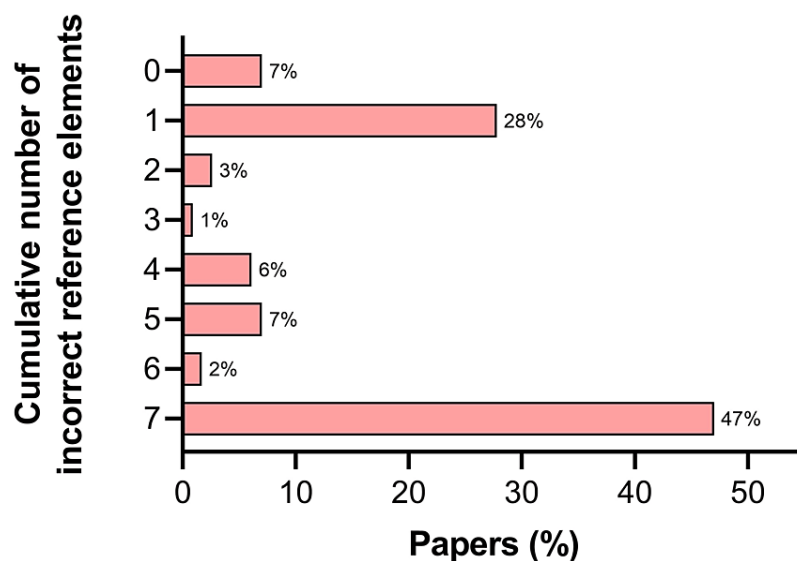


FIGURE 3: Frequency of inaccurate cumulative reference elements in ChatGPT-generated output.

A total of seven elements were evaluated in each reference including authors, title, journal, year, volume, pages, and PubMed Identifier (PMID) number.

Discussion

The widespread availability of pre-trained LLMs such as ChatGPT has dramatically expanded access to medical information. Such access may enable individuals to better understand complex medical topics and make informed decisions about their health. This may be especially valuable in disadvantaged populations without easy access to medical professionals.

However, the accuracy of ChatGPT's responses to complex medical questions remains unclear. As with

searching the internet for medical advice, the same plausible risks to the general public are inherent within ChatGPT including misdiagnosis, inappropriate treatment recommendations, and cyberchondria. In addition, no matter how detailed and customized the prompt, ChatGPT responses cannot account for individual differences in health conditions.

It is well established that reference inaccuracies are highly prevalent in the peer-reviewed literature, ranging from 4% to 48% of citations [8-11]. However, the inaccuracies identified in the references generated by ChatGPT were considerably more prevalent (93%) than those found in the peer-reviewed literature. Further, these errors are more serious since 47% of citations were fabricated. Thus, these findings call into question the credibility of any medical information provided by ChatGPT.

A primary question raised by this research is why most references provided by ChatGPT are fabricated or inaccurate. Although the cause of this phenomenon is unclear, it is plausible that reference inaccuracies may be caused by inefficiencies during data training. Notably, LLMs use deep neural networks to predict the next word in a sequence of text and provide responses based on statistical patterns learned during training [12]. As such, ChatGPT cannot distinguish between accurate and false information, only that its responses follow the patterns they are trained to recognize. The fact that over 90% of references in this study had an incorrect PMID number raises the possibility that inaccuracies may be more prevalent with numerical data than with textual data. This hypothesis is supported by the study of Athaluri et al. [6] who reported that inaccuracies in the Digital Object Identifier, an alphanumeric string used to uniquely identify online content, were the most common ChatGPT-generated reference errors. Thus, it is plausible that the overall reference accuracy in this study may have been improved if the PMID requirement had been omitted. There is a need for continued research into the accuracy of AI-generated textual versus numeric responses.

The results of this study highlight the need for greater awareness and caution regarding the potential risks of using ChatGPT to obtain medical information. The tendency of ChatGPT to produce AI hallucinations may become harmful if individuals become overly reliant on the software for answer generation. This is especially true since ChatGPT tends to double down on incorrect information in a convincing manner when confronted with response inaccuracies, which may further compound the issue. Although most people seek information online before consulting their physicians [13], ChatGPT is not a substitute for medical professionals for serious health concerns.

There were several limitations of this study. First, we used custom prompts and observed significant variability in reference accuracy based on the prompts provided. Future research should investigate how to prompt ChatGPT to provide more accurate information. Second, this research was conducted in April 2023 using ChatGPT-3.5. At the time of this writing, ChatGPT-4 is available only to subscribers and claims improved performance on tasks requiring advanced reasoning and complex instruction understanding, with fewer hallucinations [14]. The extent to which reference accuracy is improved with this newer software version remains to be determined. Finally, we did not verify the accuracy of the text in the papers due to resource constraints. It is plausible that the high inaccuracy rate found within ChatGPT-generated references may not necessarily translate to the associated text, which is a topic that warrants further study.

Conclusions

Most references to the medical information provided by ChatGPT are fabricated or inaccurate. The prevalence of reference fabrication varied considerably based on the prompts used. The findings of this study emphasize the need for caution when seeking medical information on ChatGPT. Individuals are advised to verify medical information from reliable sources and avoid relying solely on AI-generated content.

Additional Information

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue.

Animal subjects: All authors have confirmed that this study did not involve animal subjects or tissue.

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

The data supporting this study's findings are available from the corresponding author upon reasonable request.

References

1. Zhou C, Li Q, Li C, et al.: A comprehensive survey on pretrained foundation models: a history from BERT to

- ChatGPT. arXiv.2302.09419 [cs.AI]. [10.48550/arXiv.2302.09419](https://arxiv.org/abs/2302.09419)
2. ChatGPT. (2023). Accessed: April 22, 2023: <https://chat.openai.com/>.
 3. Alkaissi H, McFarlane SI: Artificial hallucinations in ChatGPT: implications in scientific writing . Cureus. 2023, 15:e35179. [10.7759/cureus.35179](https://doi.org/10.7759/cureus.35179)
 4. Ariyaratne S, Iyengar KP, Nischal N, Chitti Babu N, Botchu R: A comparison of ChatGPT-generated articles with human-written articles [PREPRINT]. Skeletal Radiol. 2023, [10.1007/s00256-023-04340-5](https://doi.org/10.1007/s00256-023-04340-5)
 5. Wagner MW, Ertl-Wagner BB: Accuracy of information and references using ChatGPT-3 for retrieval of clinical radiological information [PREPRINT]. Can Assoc Radiol J. 2023, 8465371231171125. [10.1177/08465371231171125](https://doi.org/10.1177/08465371231171125)
 6. Athaluri SA, Manthana SV, Kesapragada VS, Yarlagadda V, Dave T, Duddumpudi RT: Exploring the boundaries of reality: investigating the phenomenon of artificial intelligence hallucination in scientific writing through ChatGPT references. Cureus. 2023, 15:e37432. [10.7759/cureus.37432](https://doi.org/10.7759/cureus.37432)
 7. Originality.AI: AI content detection accuracy - GPTZero vs Writer vs Open AI vs CopyLeaks vs Originality.AI - detecting Chat GPT AI content accuracy. (2023). Accessed: April 27, 2023: <https://originality.ai/ai-content-detection-accuracy/>.
 8. Aronsky D, Ransom J, Robinson K: Accuracy of references in five biomedical informatics journals . J Am Med Inform Assoc. 2005, 12:225-8. [10.1197/jamia.M1683](https://doi.org/10.1197/jamia.M1683)
 9. de Lacey G, Record C, Wade J: How accurate are quotations and references in medical journals?. Br Med J (Clin Res Ed). 1985, 291:884-6. [10.1136/bmj.291.6499.884](https://doi.org/10.1136/bmj.291.6499.884)
 10. Evans JT, Nadjari HI, Burchell SA: Quotational and reference accuracy in surgical journals: a continuing peer review problem. JAMA. 1990, 263:1353-4. [10.1001/jama.1990.03440100059009](https://doi.org/10.1001/jama.1990.03440100059009)
 11. Siebers R, Holt S: Accuracy of references in five leading medical journals . Lancet. 2000, 356:1445. [10.1016/S0140-6736\(05\)74090-3](https://doi.org/10.1016/S0140-6736(05)74090-3)
 12. Sobieszek A, Price T: Playing games with AIs: the limits of GPT-3 and similar large language models . Minds Mach. 2022, 32:341-64. [10.1007/s11023-022-09602-0](https://doi.org/10.1007/s11023-022-09602-0)
 13. Hesse BW, Nelson DE, Kreps GL, Croyle RT, Arora NK, Rimer BK, Viswanath K: Trust and sources of health information: the impact of the Internet and its implications for health care providers: findings from the first Health Information National Trends Survey. Arch Intern Med. 2005, 165:2618-24. [10.1001/archinte.165.22.2618](https://doi.org/10.1001/archinte.165.22.2618)
 14. GPT-4 technical report. (2023). Accessed: April 27, 2023: <https://cdn.openai.com/papers/gpt-4.pdf>.