Commentary

The Evaluation of Generative AI Should Include Repetition to Assess Stability

Lingxuan Zhu[1*], MD; Weiming Mou[2*], MD; Chenglin Hong[1], MD; Tao Yang[3], MD; Yancheng Lai[1], MD; Chang Qi[4], MEng; Anqi Lin[1], MD; Jian Zhang[1], MD; Peng Luo[1], MD

[1]Department of Oncology, Zhujiang Hospital, Southern Medical University, Guangzhou, China

[2]Department of Urology, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[3]Department of Medical Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

[4]Institute of Logic and Computation, TU Wien, Austria

[*]these authors contributed equally

Corresponding Author:
Peng Luo, MD
Department of Oncology
Zhujiang Hospital
Southern Medical University
253 Industrial Avenue
Guangzhou
China
Phone: 86 020 61643888
Email: luopeng@smu.edu.cn

Related Article:
Comment on: https://mhealth.jmir.org/2024/1/e51526/

Abstract

The increasing interest in the potential applications of generative artificial intelligence (AI) models like ChatGPT in health care has prompted numerous studies to explore its performance in various medical contexts. However, evaluating ChatGPT poses unique challenges due to the inherent randomness in its responses. Unlike traditional AI models, ChatGPT generates different responses for the same input, making it imperative to assess its stability through repetition. This commentary highlights the importance of including repetition in the evaluation of ChatGPT to ensure the reliability of conclusions drawn from its performance. Similar to biological experiments, which often require multiple repetitions for validity, we argue that assessing generative AI models like ChatGPT demands a similar approach. Failure to acknowledge the impact of repetition can lead to biased conclusions and undermine the credibility of research findings. We urge researchers to incorporate appropriate repetition in their studies from the outset and transparently report their methods to enhance the robustness and reproducibility of findings in this rapidly evolving field.

(JMIR Mhealth Uhealth 2024;12:e57978) doi: 10.2196/57978

KEYWORDS

large language model; generative AI; ChatGPT; artificial intelligence; health care

Since OpenAI released ChatGPT-3.5, there has been a growing interest within the medical community regarding the prospective applications of this general pretrained model in health care [1-7]. Using ChatGPT as a search keyword in the PubMed database, the results show that 2075 papers discussing ChatGPT were published in 2023. As the leading journal in the field of digital medicine, JMIR Publications Inc published a total of 115 papers related to ChatGPT in the year 2023. It should be noted that this is a quick and simple search that may not comprehensively capture all relevant articles, but it provides a general reflection of the growing interest and research on ChatGPT in the medical field. For example, Gilson et al [8] explored the performance of ChatGPT on the United States Medical Licensing Examination (USMLE) step 1 and step 2 exams, discovering that ChatGPT's performance exceeded the passing score for third-year medical students in step 1. More studies are exploring ChatGPT's performance on other medical exams, such as the Japanese and German Medical Licensing Examinations [9,10], the Otolaryngology-Head and Neck Surgery Certification Examinations [11], and the UK Standardized Admission Tests

[12]. Beyond examinations, many articles have discussed the potential applications of ChatGPT in medicine from various perspectives. Shao et al [13] examined the suitability of using ChatGPT for perioperative patient education in thoracic surgery within English and Chinese contexts. Cheng et al [14] investigated whether ChatGPT could be used to generate summaries for medical research, and Hsu et al [15] evaluated whether ChatGPT could correctly answer basic medication consultation questions. However, we would like to point out that as a relatively new technology, there are some differences in evaluating the potential application of generative artificial intelligence (AI) like ChatGPT in health care that require additional attention from researchers.

The most significant difference affecting the evaluation of ChatGPT compared to traditional AI models known to people is the randomness inherent in the responses generated by ChatGPT. Common perception holds that for a given input, an AI model should produce the same output consistently each time. However, for natural language models like ChatGPT, this is not the case. ChatGPT generates a response by predicting the next most likely word, followed by each subsequent word. The process of generating responses involves a certain degree of randomness. If you access ChatGPT using the application programming interface, you can also control the degree of randomness in the generated responses with the temperature parameter. Even with the same input, the responses provided by ChatGPT will not be the same, and sometimes may even be completely contradictory. Therefore, when evaluating ChatGPT's performance, it is necessary to generate multiple responses to the same input and assess these responses collectively to explore ChatGPT's performance accurately; otherwise, there is a high likelihood of drawing biased conclusions. For example, as one of the earliest studies published, Sarraju et al [4] asked the same question three times and assessed whether the three responses given by ChatGPT to the same question were consistent. As OpenAI made the ChatGPT application programming interface accessible, it became feasible to ask the same question many more times. In a recent study investigating whether ChatGPT's peer-review conclusions are influenced by the reputation of the author's institution, von Wedel et al [16] conducted 250 repeated experiments for each question to mitigate the effects of ChatGPT's randomness. However, not all researchers have recognized this aspect. For instance, in a study where ChatGPT was asked to answer the American Heart Association Basic Life Support and Advanced Cardiovascular Life Support exams, they found that ChatGPT could not pass either examination [17]. However, that study only asked the question once without repeating, which means that the randomness of ChatGPT could have had an impact on the experiment, affecting the reliability of the conclusions. In another improved study, researchers acknowledged the impact of ChatGPT's randomness, asking each question three times. Compared to earlier results, ChatGPT's performance in this study significantly improved, and it could pass the Basic Life Support exam [18], further underscoring the importance of repetitions. Therefore, it is inappropriate to evaluate ChatGPT's performance based on a single response if one aims to draw rigorous, scientifically meaningful conclusions. Just as biological experiments typically require three repetitions for validity, without repetition, it becomes challenging to determine whether the observed phenomenon is an inherent characteristic of the model or merely a random occurrence. Additionally, for models intended for clinical practice applications, whether for patient education, diagnosis, or support in clinical documentation writing, we hope that ChatGPT can always provide correct and harmless responses. Repetition also allows us to evaluate the model's stability and further assess its application value. However, we noticed that many recent manuscripts we reviewed were not aware of this, thus affecting the reliability of the conclusions.

Therefore, in research on the application of generative AI like ChatGPT in health care, appropriate repetition should be included to comprehensively evaluate the model's performance by assessing the stability of the model in the task set by the author. This should be considered from the beginning of the research. Since models like ChatGPT will continue to be upgraded, if the authors only realize the need for repetition when revising the manuscript, there will be a considerable time gap between the authors' supplementary analysis and the original analysis. The model has likely been upgraded during this period, introducing new uncertainties into the research. Alternatively, the authors need to completely redo the analysis from scratch during the manuscript revision process, wasting time and effort. Therefore, we hope that future researchers will recognize the necessity of repeated experiments from the start and report in the manuscript how the repetition was carried out in the study [19].

Conflicts of Interest

None declared.

References

1. Grünebaum A, Chervenak J, Pollet SL, Katz A, Chervenak FA. The exciting potential for ChatGPT in obstetrics and gynecology. Am J Obstet Gynecol. Jun 2023;228(6):696-705. [doi: 10.1016/j.ajog.2023.03.009] [Medline: 36924907]
2. Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: the end of the consulting infection doctor? Lancet Infect Dis. Apr 2023;23(4):405-406. [doi: 10.1016/S1473-3099(23)00113-5] [Medline: 36822213]
3. Zhu L, Mou W, Luo P. Potential of large language models as tools against medical disinformation. JAMA Intern Med. Apr 01, 2024;184(4):450. [doi: 10.1001/jamainternmed.2024.0020] [Medline: 38407861]
4. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA. Mar 14, 2023;329(10):842-844. [FREE Full text] [doi: 10.1001/jama.2023.1044] [Medline: 36735264]

5. Zhu L, Mou W, Chen R. Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? J Transl Med. Apr 19, 2023;21(1):269. [FREE Full text] [doi: 10.1186/s12967-023-04123-5] [Medline: 37076876]
6. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. Lancet Digit Health. Apr 2023;5(4):e179-e181. [FREE Full text] [doi: 10.1016/S2589-7500(23)00048-1] [Medline: 36894409]
7. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health. Mar 2023;5(3):e107-e108. [FREE Full text] [doi: 10.1016/S2589-7500(23)00021-3] [Medline: 36754724]
8. 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 (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. Feb 08, 2023;9:e45312. [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
9. Meyer A, Riese J, Streichert T. Comparison of the performance of GPT-3.5 and GPT-4 with that of medical students on the written German Medical Licensing Examination: observational study. JMIR Med Educ. Feb 08, 2024;10:e50965. [FREE Full text] [doi: 10.2196/50965] [Medline: 38329802]
10. Yanagita Y, Yokokawa D, Uchida S, Tawara J, Ikusaka M. Accuracy of ChatGPT on medical questions in the National Medical Licensing Examination in Japan: evaluation study. JMIR Form Res. Oct 13, 2023;7:e48023. [FREE Full text] [doi: 10.2196/48023] [Medline: 37831496]
11. Long C, Lowe K, Zhang J, Santos AD, Alanazi A, O'Brien D, et al. A novel evaluation model for assessing ChatGPT on Otolaryngology-Head and Neck Surgery Certification Examinations: performance study. JMIR Med Educ. Jan 16, 2024;10:e49970. [FREE Full text] [doi: 10.2196/49970] [Medline: 38227351]
12. Giannos P, Delardas O. Performance of ChatGPT on UK standardized admission tests: insights from the BMAT, TMUA, LNAT, and TSA examinations. JMIR Med Educ. Apr 26, 2023;9:e47737. [FREE Full text] [doi: 10.2196/47737] [Medline: 37099373]
13. Shao C, Li H, Liu X, Li C, Yang L, Zhang Y, et al. Appropriateness and comprehensiveness of using ChatGPT for perioperative patient education in thoracic surgery in different language contexts: survey study. Interact J Med Res. Aug 14, 2023;12:e46900. [FREE Full text] [doi: 10.2196/46900] [Medline: 37578819]
14. Cheng S, Tsai S, Bai Y, Ko C, Hsu C, Yang F, et al. Comparisons of quality, correctness, and similarity between ChatGPT-generated and human-written abstracts for basic research: cross-sectional study. J Med Internet Res. Dec 25, 2023;25:e51229. [FREE Full text] [doi: 10.2196/51229] [Medline: 38145486]
15. Hsu H, Hsu K, Hou S, Wu C, Hsieh Y, Cheng Y. Examining real-world medication consultations and drug-herb interactions: ChatGPT performance evaluation. JMIR Med Educ. Aug 21, 2023;9:e48433. [FREE Full text] [doi: 10.2196/48433] [Medline: 37561097]
16. von Wedel D, Schmitt RA, Thiele M, Leuner R, Shay D, Redaelli S, et al. Affiliation bias in peer review of abstracts by a large language model. JAMA. Jan 16, 2024;331(3):252-253. [doi: 10.1001/jama.2023.24641] [Medline: 38150261]
17. Fijačko N, Gosak L, Štiglic G, Picard CT, John Douma M. Can ChatGPT pass the life support exams without entering the American Heart Association course? Resuscitation. Apr 2023;185:109732. [doi: 10.1016/j.resuscitation.2023.109732] [Medline: 36775020]
18. Zhu L, Mou W, Yang T, Chen R. ChatGPT can pass the AHA exams: open-ended questions outperform multiple-choice format. Resuscitation. Jul 2023;188:109783. [doi: 10.1016/j.resuscitation.2023.109783] [Medline: 37349064]
19. Chen J, Zhu L, Mou W, Liu Z, Cheng Q, Lin A, et al. STAGER checklist: Standardized Testing and Assessment Guidelines for Evaluating Generative AI Reliability. arXiv. Preprint posted online on Decemeber 8, 2023. 2024. [doi: 10.48550/arXiv.2312.10074]

Abbreviations

AI: artificial intelligence
USMLE: United States Medical Licensing Examination

Edited by L Buis; this is a non–peer-reviewed article. Submitted 01.03.24; accepted 30.04.24; published 06.05.24.

Please cite as:
Zhu L, Mou W, Hong C, Yang T, Lai Y, Qi C, Lin A, Zhang J, Luo P
The Evaluation of Generative AI Should Include Repetition to Assess Stability
JMIR Mhealth Uhealth 2024;12:e57978
URL: https://mhealth.jmir.org/2024/1/e57978
doi: 10.2196/57978
PMID: 38688841