The Evaluation of Generative AI Should Include Repetition to Assess Stability.

Lingxuan Zhu, Weiming Mou, Chenglin Hong, Tao Yang, Yancheng Lai, Chang Qi, Anqi Lin, Jian Zhang, Peng Luo

Submitted to: JMIR mHealth and uHealth
on: March 01, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

The Evaluation of Generative AI Should Include Repetition to Assess Stability.

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

[1]Department of Oncology Zhujiang Hospital Southern Medical University Guangzhou CN
[2]Department of Urology Shanghai General Hospital Shanghai Jiao Tong University School of Medicine shanghai CN
[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 Bejing CN
[4]Institute of Logic and Computation TU Wien AT
[*]these authors contributed equally

Corresponding Author:
Peng Luo
Department of Oncology
Zhujiang Hospital
Southern Medical University
luopeng@smu.edu.cn
Guangzhou
CN

Abstract

The increasing interest in the potential applications of generative AI models like ChatGPT-3.5 in healthcare has prompted numerous studies exploring 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 Preprints 01/03/2024:57978)
DOI: https://doi.org/10.2196/preprints.57978

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
Please make my preprint PDF available to anyone at any time (recommended).
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
✓ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?
✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

The Evaluation of Generative AI Should Include Repetition to Assess Stability.

Article Type: Invited comment.

Lingxuan Zhu[1,2†], Weiming Mou [3†], Chenglin Hong [1], Tao Yang[4], Yancheng Lai[1], Chang Qi[5], Anqi Lin [1], Jian Zhang [1*], Peng Luo [1*]

[1] Department of Oncology, Zhujiang Hospital, Southern Medical University, Guangzhou, 510282, China.

[2] Department of Etiology and Carcinogenesis, National Cancer Center/ National Clinical Research Center for Cancer/Cancer Hospital, Changping laboratory, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China.

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

[4] 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.

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

*** Correspondences to:**

Peng Luo and Jian Zhang,

Department of Oncology, Zhujiang Hospital, Southern Medical University, 253 Industrial Avenue, Guangzhou, 510282, Guangdong

Tel: 0086-18588447321 (Peng Luo); 0086-13925091863 (Jian Zhang)

mail: luopeng@smu.edu.cn (Peng Luo); zhangjian@i.smu.edu.cn (Jian Zhang)

†: Lingxuan Zhu and Weiming Mou have contributed equally to this work and share first authorship.

Main Body

Since OpenAI released ChatGPT-3.5, there has been a growing interest within the medical community regarding the prospective applications of this general pre-trained model in healthcare[1–7]. Using ChatGPT as search keywords in the PubMed database, the results show that a total of 2,075 papers discussing ChatGPT were published in 2023 As the leading journal in the field of digital medicine, JMIR Publications group published a total of 115 papers related to ChatGPT in the year 2023. For example, Gilson et al. 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 the Step 1 [8]. There are more studies 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. examined the suitability of using ChatGPT for perioperative patient education in thoracic surgery within English and Chinese contexts[13]. Cheng et al. investigated whether ChatGPT could be utilized to generate summaries for medical research[14], and Hsu et al. evaluated whether ChatGPT could correctly answer basic medication consultation questions[15]. However, we would like to point out that as a relatively new technology, there are some differences in evaluating the potential application of generative AI like ChatGPT in healthcare that require additional attention from researchers.

The most significant difference affecting the evaluation of ChatGPT compared to traditional artificial intelligence 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 API, 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 exactly 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. asked the same question three times and assessed whether the three responses given by ChatGPT to the same question were consistent[4]. As OpenAI opened up access to the ChatGPT API, 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, Wedel et al. conducted 250 repeated experiments for each question to mitigate the effects of ChatGPT's randomness[16]. However, not all researchers have recognized this aspect. For instance, in a study where ChatGPT was asked to answer the American Heart Association (AHA) Basic Life Support (BLS) and Advanced Cardiovascular Life Support (ACLS) exams to test whether it could pass these 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 was significantly improved, and it could pass the BLS 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 healthcare, 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 very 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. It is likely that the model has 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].

Statements & Declarations

Acknowledgements

Not applicable.

Funding

Not applicable.

Competing Interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

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 2023 Jun;228(6):696–705. PMID:36924907

2. Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: the end of the consulting infection doctor? Lancet Infect Dis 2023 Apr;23(4):405–406. PMID:36822213

3. Zhu L, Mou W, Luo P. Potential of Large Language Models as Tools Against Medical Disinformation. JAMA Intern Med 2024 Feb 26; doi: 10.1001/jamainternmed.2024.0020

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 2023 Mar 14;329(10):842–844. doi: 10.1001/jama.2023.1044

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 2023 Apr 19;21(1):269. doi: 10.1186/s12967-023-04123-5

6. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. Lancet Digit Health 2023 Apr;5(4):e179–e181. PMID:36894409

7. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023 Mar;5(3):e107–e108. PMID:36754724

8. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. 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 2023 Feb 8;9:e45312. PMID: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 2024 Feb 8;10:e50965. PMID: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 2023 Oct 13;7:e48023. PMID:37831496

11. Long C, Lowe K, Zhang J, Santos AD, Alanazi A, O’Brien D, Wright ED, Cote D. A Novel Evaluation Model for Assessing ChatGPT on Otolaryngology-Head and Neck Surgery Certification Examinations: Performance Study. JMIR Med Educ 2024 Jan 16;10:e49970. PMID: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 2023 Apr 26;9:e47737. PMID:37099373

13. Shao C-Y, Li H, Liu X-L, Li C, Yang L-Q, Zhang Y-J, Luo J, Zhao J. Appropriateness and Comprehensiveness of Using ChatGPT for Perioperative Patient Education in Thoracic Surgery in Different Language Contexts: Survey Study. Interact J Med Res 2023 Aug 14;12:e46900. PMID:37578819

14. Cheng S-L, Tsai S-J, Bai Y-M, Ko C-H, Hsu C-W, Yang F-C, Tsai C-K, Tu Y-K, Yang S-N, Tseng P-T, Hsu T-W, Liang C-S, Su K-P. Comparisons of Quality, Correctness, and Similarity Between ChatGPT-Generated and Human-Written Abstracts for Basic Research: Cross-Sectional Study. J Med Internet Res 2023 Dec 25;25:e51229. PMID:38145486

15. Hsu H-Y, Hsu K-C, Hou S-Y, Wu C-L, Hsieh Y-W, Cheng Y-D. Examining Real-World Medication Consultations and Drug-Herb Interactions: ChatGPT Performance Evaluation. JMIR Med Educ 2023 Aug 21;9:e48433. PMID:37561097

16. von Wedel D, Schmitt RA, Thiele M, Leuner R, Shay D, Redaelli S, Schaefer MS. Affiliation Bias in Peer Review of Abstracts by a Large Language Model. JAMA 2024 Jan 16;331(3):252–253. doi: 10.1001/jama.2023.24641

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 2023 Apr 1;185:109732. doi: 10.1016/j.resuscitation.2023.109732

18. Zhu L, Mou W, Yang T, Chen R. ChatGPT can pass the AHA exams: Open-ended questions outperform multiple-choice format. Resuscitation 2023 Jul 1;188:109783. doi: 10.1016/j.resuscitation.2023.109783

19. Chen J, Zhu L, Mou W, Liu Z, Cheng Q, Lin A, Zhang J, Luo P. STAGER checklist: Standardized Testing and Assessment Guidelines for Evaluating Generative AI Reliability. arXiv; 2023. doi: 10.48550/arXiv.2312.10074