ORIGINAL ARTICLE
Journal of Science Technology Education Art and Medicine
DOI: 10.63137/jsteam.744858

Evaluating the Precision and Dependability of Medical Answers Generated by ChatGPT

Zain Abidin[a]*

[a]Cooper Medical School, Rowan University, USA

*Corresponding address: *Cooper Medical School, Rowan University, USA*
Email: zainwildelake@gmail.com

Received: 26 January 2024 / **Revised:** 03 May 2024 / **Accepted:** 15 May 2024 / **Available Online:** 10 June 2024

ABSTRACT

Objective: This study focuses on the assessment of the precision and depth of ChatGPT's feedback to medical questions posed by physicians, providing preliminary evidence of its reliability in offering precise and comprehensive information.
Methods: This research involved 10 physicians formulating questions for ChatGPT without patient-specific data. Approximately 29% of the 35 invited doctors participated, creating eight questions each. The questions covered easy, medium, and hard levels, with yes/no or descriptive responses. ChatGPT's responses were evaluated by physicians for accuracy and completeness using established Likert scales. An internal validation re-submitted questions with low accuracy scores, and statistical measures analyzed the outcomes, revealing insights into response consistency and variation over time.
Results: The analysis of 80 first-round ChatGPT responses revealed a median accuracy score of 4 (mean 4.7, SD 2.6) and a median completeness score of 2 (mean 1.8, SD 1.5). Notably, 30% of responses achieved the highest accuracy score (6), and 38.7% were rated nearly all correct (5), while 8% were deemed completely incorrect (1). Inaccurate answers were more common for physician-rated hard questions. Completeness varied, with 45% considered comprehensive, 37.5% adequate, and 17.5% incomplete. A modest correlation (Spearman's r = 0.3) existed between accuracy and completeness across all questions.
Conclusion: Integrating models of language like ChatGPT in the practice of medicine shows promise, but cautious considerations are crucial for safe use. While AI-generated responses display commendable accuracy and completeness, ongoing refinement is needed for reliability.
Keywords: Artificial Intelligence; Assessment; Decision Making; Healthcare

Citation: Abidin Z. Evaluating the Precision and Dependability of Medical Answers Generated by ChatGPT. J Sci Technol Educ Art Med. 2024;1(1):16-22

Introduction

The employment of natural language processing (NLP) techniques in the medical field offers a significant opportunity to improve access to healthcare information for both professionals and patients. [1] Among these techniques, Large Language Models (LLMs) stand out for their ability to mimic human text. These models are distinct from traditional supervised deep learning models in that they utilize a more efficient learning approach. This approach involves initial training through self-supervision on extensive unlabeled data, followed by specific fine-tuning on smaller, labeled datasets for specialized tasks. [2]

One AI-based tool that has recently gained attention is Chat-Generative Pre-Trained Transformer (ChatGPT). This tool is developed on the foundation of the Generative Pre-Trained Transformer-3.5 (GPT-3.5) framework, an LLM boasting more than 175 billion parameters. [3] ChatGPT's training includes a wide array of internet materials like articles, books, and websites. Its capability for communicative tasks is honed

through principles of reinforcement learning from human feedback (RLHF), allowing it to understand the subtleties of user intent and to effectively tackle diverse tasks, possibly including those related to healthcare. [4]

As the volume of medical-related data grows and clinical-based decision-making becomes more complex, NLP technologies could help doctors in making rapid and better choices, thereby enhancing healthcare quality and efficiency. [5] Remarkably, ChatGPT has demonstrated capabilities close to the qualifying standards for the United States Medical Licensing Exam (USMLE) regardless of no training in any specialty. [6] This underlines its possible utility in healthcare education and practical hospital support. Moreover, technological progress has shifted the dynamics of knowledge acquisition. Patients increasingly use AI chatbots and search engines as favorable, readily usable sources of health information, moving away from relying solely on medical professionals. [7]

ChatGPT and similar AI chatbots are now capable of engaging in detailed conversations and providing seemingly authoritative answers to complex medical inquiries. [8] However, despite its promising capabilities, ChatGPT sometimes generates plausible yet inaccurate responses. [9] This calls for caution in its application in clinical settings and research. The dependability and precision of these tools, particularly in responding to open-ended medical questions commonly posed by doctors and patients, have not been thoroughly evaluated.

This study aimed to assess the accuracy and depth of ChatGPT's replies to medical questions posed by physicians, providing preliminary evidence of its reliability in offering precise and comprehensive information. Furthermore, the study will shed light on the limitations inherent in AI-generated medical advice.

Materials and Methods

This study received an exemption from Institutional Review Board review at Cooper Medical School, Rowan University, as it did not involve human participants, patient data, or identifiable health information. Data were collected between July 2023 and November 2023 using the publicly available version of ChatGPT based on GPT-3.5 (OpenAI, San Francisco, CA, USA), accessed via chat.openai.com. The investigation employed a set of questions formulated by 10 physicians from various specialties. Of the 35 doctors initially invited, around 29% participated. These participants were either faculty members or recent graduates from diverse international locations. The directive for these doctors was to create queries based on clear, non-controversial information derived from current medical guidelines, reflecting the knowledge base as of early 2021, which aligns with the training data cut-off for ChatGPT. Each doctor was tasked with devising eight questions, yielding a total of 80 unique questions. Half of these questions were designed to have straightforward yes/no or correct/incorrect responses, categorized into easy, medium, and hard levels based on the doctor's subjective assessment. The remaining questions required either descriptive responses or a compilation of multiple correct answers, also classified into the same three levels of difficulty. The doctors were instructed against pre-testing the quizzes on ChatGPT themselves to decrease the chances of bias.

For consistency, a single researcher (ZA) inputs all queries into the ChatGPT system, instructing the AI to provide detailed responses and to integrate relevant medical guidelines where applicable, using the prompt “Be distinct and embrace any relevant healthcare guidelines”. Each question was entered as a new conversation to avoid contextual carryover between responses. The responses generated via AI were then evaluated by the doctors who formulated the queries. These physicians, considering their expertise in their respective fields, assessed the responses using two established scales: one for accuracy and another for completeness.

The accuracy assessment was conducted using a six-point Likert scale, ranging from 1 (completely incorrect) to 6 (entirely correct). The completeness of the answers was evaluated on a three-point Likert scale, where 1 indicated an incomplete answer that missed significant elements, 2 denoted adequacies in covering all aspects of the question, and 3 signified a comprehensive response that provided additional context or information beyond the basic query. Responses deemed completely incorrect in terms of accuracy (receiving a score of 1) were not evaluated for completeness.

To assess response consistency over time, the complete set of 80 questions was submitted to ChatGPT on three separate occasions, spaced 8–17 days apart, yielding a total of 240 evaluated responses (80 questions × 3 rounds). The same physicians evaluated all three rounds of responses to their own questions. This design allowed comparison of ChatGPT's performance across time

points to determine whether the model's outputs varied with ongoing updates.

The outcomes were presented descriptively, using statistical measures such as mean, median, standard deviation, and intra-quartile range. Differences between groups were analyzed using either the Kruskal-Wallis test or the Mann-Whitney U test. The re-evaluated answers were juxtaposed using the signed rank test by Wilcoxon. In specific data subsets, such as those concerning immunotherapy/melanoma and common conditions, intra-rater reliability was assessed with kappa statistics for all marks (1-3 for completeness and 1-6 for accuracy). Additionally, an exploratory and condensed analysis was conducted to gauge common agreement (accuracy scores for accuracy into categories of 1-2, 3-4, and 5-6). A two-sided significance level of $\alpha = 0.05$ was applied throughout. All analyses were performed in IBM SPSS Statistics version 29.0 (IBM Corp., Armonk, NY, USA).

Results

Of the 35 physicians invited across multiple specialties, 10 (28.6%) agreed to participate. The participating physicians represented general internal medicine (n = 4), oncology (n = 2), endocrinology (n = 2), pulmonology (n = 1), and hematology (n = 1). Among the 80 first-round responses generated by Chat-GPT, the median accuracy score was 4 (mean 4.7, SD 2.6), and the median completeness score was 2 (mean 1.8, SD 1.5) (Table 1). Amongst them, the highest accuracy rating (score of 6) was 30% (n= 24), and a score of near-perfect rating (accuracy score of 5) was 38.7% (n= 31). Conversely, completely incorrect (accuracy score of 1) was 8% (n=10). False responses, with an accuracy score of 2 or lower (n=14), were mostly in answer to physician-rated difficult questions with either two answers (n=9, 11.25%) or explanatory answers (n=5, 6%) but were divided across all groups. Furthermore, the thoroughness of the responses was assessed, with 45% (n=36) rated as thorough, 37.5% (n = 30) as sufficient, and 17.5% (n=14) as unfinished. Accuracy and completeness were modestly correlated (Spearman's r = 0.30; 95% CI 0.09–0.49; $p < 0.01$) across all questions. When all three evaluation rounds were pooled (n = 240 responses to 80 unique questions), the mean accuracy score was 5.6 (SD 2.1) with a median of 4.9 (IQR 1.5), whereas the mean completeness score was 2.3 (SD 0.7) with a median of 2.0 (IQR 1.0).

All descriptive questions had a median accuracy of 4.2 (average 3.9, SD 2.2, IQR 1.5), while binary questions had a median accuracy of 5.8 (average 5.1, SD 1.8, IQR 1.5) (Mann Whitney U p=0.06). The descriptive questions had median accuracy scores of 5.5 (mean 5.1, SD 2.5, IQR 2), 4.7 (mean 4.4, SD 1.8, IQR 1.6), and 3.8 (mean 4.8, SD 1.7, IQR 1.9) for the easy, medium, and hard questions, respectively (Kruskal-Wallis p = 0.3). There was a significant difference between the groups (Kruskal-Wallis p = 0.02) in the binary questions. The median accuracy scores for the easy, medium, and hard questions were 6 (mean 5.5, SD 1.8, IQR 2), 5 (mean 4.4, SD 2.2, IQR 1.8), and 3.5 (mean 2.7, SD 1.4, IQR 1.9). For the repeated-question subsets, intra-rater agreement across rounds was fair to moderate. In the immunotherapy/melanoma subset, agreement was $\kappa = 0.34$ for accuracy and $\kappa = 0.29$ for completeness, while in the common-conditions subset, agreement was $\kappa = 0.47$ for accuracy and $\kappa = 0.41$ for completeness. In the exploratory condensed analysis of accuracy (1–2, 3–4, 5–6), agreement increased to $\kappa = 0.53$ for immunotherapy/melanoma and $\kappa = 0.61$ for common conditions.

Table 1: Examples of answers generated by ChatGPT to common medical questions

Difficulty Level	Common Medical Condition	Question Asked	ChatGPT generated answer	Accuracy Score	Completeness Score
Easy (Descriptive)	**Anemia**	What is the most likely cause of anemia?	Iron deficiency is the most likely cause of anemia worldwide. It can arise from poor intake, reduced absorption, chronic blood loss, or increased demand during growth and	6	3

			pregnancy.		
Medium (Descriptive)	**Diabetes**	What are the complications of untreated hyperglycemia?	Untreated hyperglycemia can injure small and large blood vessels. Key complications include retinopathy, nephropathy, neuropathy, cardiovascular disease, recurrent infections, delayed wound healing, and peripheral vascular disease.	5	2
Hard (Descriptive)	**Asthma**	For an 18-year-old with moderate persistent asthma, what regimen is recommended by the current National Asthma Education and Prevention Program? guidance?	For moderate persistent asthma in an 18-year-old, a preferred Step 3 regimen is a low-dose inhaled corticosteroid with formoterol used for both maintenance and relief, known as single maintenance and reliever therapy. Alternatives include a low-dose inhaled corticosteroid with long-acting beta-agonist, along with a short-acting beta-agonist as needed, or a medium-dose inhaled corticosteroid.	4	1

† The accuracy scale employed a six-point Likert scale ranging from 1 (completely incorrect) to 6 (correct). The completeness scale, comparatively, utilized a Likert scale with three points: 1 (insufficient), 2 (adequate, addressing every side of the quiz with the least required information), and 3 (comprehensive, addressing all sides of the question with additional information or context beyond expectations). Responses rated as completely incorrect (score of 1) on the accuracy scale were exempt from assessment for comprehensiveness.

Discussion

This study suggests that approximately three months after its public release, ChatGPT shows the potential to provide error-free and extensive medical-related information. Nevertheless, it does not prove to be entirely reliable. The analysis by 10 physicians across five specialties, evaluating 240 responses generated from 80 unique questions across three rounds, highlights that (n=240) 52.1% of the AI responses were deemed "almost all accurate" or "accurate" (mean accuracy score 4.4, median 5). A majority of the answers (53.5%) were considered thorough (median 3, mean completeness score 2.4), indicating thorough responses with added context or information. The examination of this data reveals that the median of accuracy scores tended to be higher than the mean of said scores, underscoring instances where the chatbot provided notably incorrect information. Consequently, utilizing the latest model of ChatGPT for the dissemination of healthcare expertise should acknowledge its potential to arrive at entirely mistaken conclusions, presented convincingly.

In general, accuracy remained relatively high across different question types and difficulty levels. More challenging questions exhibited slightly lower accuracy scores compared to easier ones, hinting at a possible restraint in handling complicated medical questions. ChatGPT was able to pass the Japanese medical board exam with a 72% result which reinforces its potential for solving easy to medium questions. [10] Similarly, this AI chatbot was able to solve more than 60% of the questions included in four different types of USA medical board exams including NBME-Free-Step1, and NBME-Free-Step2, AMBOSS-Step1, and AMBOSS-Step2. [11] We did not find studies that investigated ChatGPT's response to easy, medium

and hard questions separately. However, in this study, the difference did not reach any statistical significance. Overall, the performance across question types (descriptive or binary) and difficulty levels was comparable, suggesting ChatGPT's potential applicability to a variety of open-ended questions with varying difficulty levels.

The internal validation revealed ChatGPT's ability to significantly upgrade over a short span. This improvement may be attributable to uninterrupted updates and advancement of the model's parameters and algorithms, as well as the influence of recurrent user feedback resulting in reinforcement learning. This feature is being used by researchers and entrepreneurs in devising algorithms for AI medical software, where AI chats with patients and answers their most common medical-related questions. The AI collects recorded conversations and presents them to the consultant in an efficient manner. This helps the consultant respond to patients' queries and concerns swiftly and in a timely manner. [10]

Results after analyses of four recognizable query sets (asthma, diabetes, anemia, and ordinary conditions) demonstrated relatively consistent and high scores. The dataset of common conditions exhibited relatively greater accuracy scores, implying that the availability of more training data for common conditions may contribute to improved scores. However, it's worth noting that scoring on this dataset was repeated later, potentially reviewing the ongoing improvement of the model over weeks and even days.

Despite such encouraging findings, the generalizability of our interpretations is restricted by the limited sample size, consisting of 10 physicians from a single academic medical centre evaluating 80 unique questions (240 total responses across three rounds), which possibly do not adequately represent the diversity of medical specialties and the wide range of possible questions within them. Moreover, there are notable biases, including selection bias stemming from the cohort being confined to academic practitioners, in addition to a potential bias of respondents. Additional restraints include the personalized selection of presented queries and the lack of a method of validation to authenticate the precision of the given facts and figures. The research depended on doctors' own reported, subjective rankings, introducing potential bias, especially considering variations in judgments among physicians (e.g., the distinction between nearly all correct vs more correct than incorrect (5 vs. 4) might be subtle). [11] Questions selected by physicians tended to have unchallenged and clear answers, aligned with current guidelines, potentially deviating from the inquiries that patients and the general public might pose, lacking explicit knowledge regarding factors such as prior therapies, cancer staging, and sites of metastases, which can significantly influence responses. [12] The study was confined to one particular AI variant, ChatGPT, and may not be universally applicable to additional variants of AI, especially with specialized healthcare training.

The research presents preliminary confirmation supporting the potential of systems using AI to address non-multiple choice clinical questions of the real world. With further validation, tools like ChatGPT could serve as valuable resources for rapid medical information retrieval in fast-paced clinical settings, enhancing the efficiency of healthcare and aiding complicated decision-making. [13] Healthcare providers should think about how patients might put these tools into use and how ChatGPT is set up to make the right suggestions and send patients to licensed healthcare professionals. [14] To help patients and healthcare providers make educated decisions about when and how to employ AI-powered tools, healthcare education should include relevant instruction on the potential advantages, drawbacks, and hazards associated with these tools. [15] However, depending only on the present-day publicly accessible edition of ChatGPT for medical information is cautioned against. If instructed by credible specialists on a wide-ranging dataset of scrutinized medical information, such as pharmacology databases, medical literature, electronic healthcare accounts, etc., training ChatGPT and similar language models on curated medical datasets could substantially improve the accuracy and reliability of AI-generated healthcare information. Recent advancements, such as a GPT-style model of language exclusively trained on biomedical literature, achieving 51% accuracy on diverse biomedical question-answering assignments, underscore the aptitude of context-sensitive language generation models in practical healthcare applications. [16]

Validation of AI-generated healthcare information requires more research involving bigger communities of medical specialists with varying question categories. [17] Furthermore, studies should also assess how AI-created healthcare data has advanced over time. Additional considerations include privacy and ethical issues. As the data used to train these AI tools predetermines their reliability, efforts are made regarding the incorporation of reliable sources of medical

information, such as pharmacology databases, medical literature, and evidence from existing sources, to guarantee that AI variants are properly trained and provide current, evidence-based information. Furthermore, text-based AI models might miss subtleties presented only in tabular or figure form instead of being explored expressly in theory. Lastly, future efforts should focus on developing a robust regulatory framework and standards regarding the effective and secure application of AI in medicine.

Conclusion

Incorporating models of language like ChatGPT into the application of medicine shows initial potential, but careful deliberations are essential for ensuring secure and effective utilization. Although the AI-generated responses demonstrated commendable completeness scores and accuracy across diverse question types, specialties, and levels of difficulty, ongoing refinement is necessary to enhance the dependability and resilience of such tools before their seamless incorporation into clinical settings. Patients and medical practitioners should remain cognizant of the shortcomings and actively verify medical information generated by AI through dependable sources. This research serves as a fundamental step in initiating an authentic base for the integration of healthcare-related AI language models, emphasizing the crucial need for continuous evaluation and regulatory measures.

Acknowledgments

The author thanks the participating physicians for their time in formulating questions and evaluating responses.

Author Contribution

ZA conceived the idea, collected data, and wrote the manuscript.

Data Availability Statement

All relevant data are within the manuscript. Additional data supporting this study are available from the corresponding author upon reasonable request.

Ethical Considerations

This study received an exemption from Institutional Review Board review at Cooper Medical School, Rowan University, as it did not involve human participants, patient data, or identifiable health information. All data were generated by a publicly available AI system (ChatGPT) in response to hypothetical medical questions formulated by volunteer physicians.

Funding

The research did not receive funding from any profit / non-profit organization.

Conflict of Interest

The author declares no conflicts of interest.

References

1. Zhou B, Yang G, Shi Z, Ma S. Natural language processing for smart healthcare. IEEE Rev Biomed Eng. 2024;17:4-18. doi:10.1109/RBME.2022.3210270.
2. Kotei E, Thirunavukarasu R. A systematic review of transformer-based pre-trained language models through self-supervised learning. Information. 2023;14(3):187. doi:10.3390/info14030187.
3. Lund BD, Wang T. Chatting about ChatGPT: how may AI and GPT impact academia and libraries? Libr Hi Tech News. 2023;40(3):26-9. doi:10.1108/LHTN-01-2023-0009.
4. Vaishya R, Misra A, Vaish A. ChatGPT: is this version good for healthcare and research? Diabetes Metab Syndr. 2023;17(4):102744. doi:10.1016/j.dsx.2023.102744.
5. Kasula BY. Advancements in AI-driven healthcare: a comprehensive review of diagnostics, treatment, and patient care integration. Int J Mach Learn Sustain Dev. 2024;1(1):1-5.
6. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198.
7. King MR. The future of AI in medicine: a perspective from a Chatbot. Ann Biomed Eng. 2023;51(2):291-5. doi:10.1007/s10439-022-03121-w.
8. Bansal G, Chamola V, Hussain A, Guizani M, Niyato D. Transforming conversations with AI—a comprehensive study of ChatGPT. Cogn Comput. 2024;16(5):2487-510. doi:10.1007/s12559-023-10236-2.
9. Kocoń J, Cichecki I, Kaszyca O, Kochanek M, Szydło D, Baran J, et al. ChatGPT: jack of all trades, master of none. Inf Fusion. 2023;99:101861. doi:10.1016/j.inffus.2023.101861.

10. Jahanshahi H, Kazmi S, Cevik M. Auto response generation in online medical chat services. J Healthc Inform Res. 2022;6(3):344-74. doi:10.1007/s41666-022-00118-x.
11. Wong E, Williams O, Williams ZM, Báez-Mendoza R. Naturalistic generative narratives reveal effects of social characteristics on decision-making. Front Psychol. 2024;15:1412131. doi:10.3389/fpsyg.2024.1412131.
12. Wang YA, Eastwick PW. Solutions to the problems of incremental validity testing in relationship science. Pers Relatsh. 2020;27(1):156-75. doi:10.1111/pere.12309.
13. Kumar V. Digital enablers. In: The economic value of digital disruption: a holistic assessment for CXOs. Singapore: Springer; 2023. p. 1-110. doi:10.1007/978-981-19-8148-7_1.
14. Javaid M, Haleem A, Singh RP. ChatGPT for healthcare services: an emerging stage for an innovative perspective. BenchCouncil Trans Benchmarks Stand Eval. 2023;3(1):100105. doi:10.1016/j.tbench.2023.100105.
15. Sapci AH, Sapci HA. Artificial intelligence education and tools for medical and health informatics students: systematic review. JMIR Med Educ. 2020;6(1):e19285. doi:10.2196/19285.
16. Doneva SE, Qin S, Sick B, Ellendorff T, Goldman J-P, Schneider G, et al. Large language models to process, analyze, and synthesize biomedical texts: a scoping review. Discov Artif Intell. 2024;4(1):107. doi:10.1007/s44163-024-00197-2.
17. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022;28(1):31-8. doi:10.1038/s41591-021-01614-0.