ORIGINAL ARTICLE

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 ChatGPT-spawned answers revealed a precision tally by a median of 4 (mean 4.7, SD 2.6) and a completeness score by a median 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):13-20

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 specifications. [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 induction based

ORIGINAL ARTICLE

DOI: 10.63137/jsteam.744858

on human critique, 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 research received an exemption from IRB review since it did not involve any patient-specific data. 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. 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 prejudice.

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”. 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 verify the consistency of the results and to understand how responses might vary over time, an internal validation was performed. This involved re-submitting the same questions to ChatGPT, particularly those questions that initially received low accuracy scores (below 3). These re-submitted questions were processed by ChatGPT 8 to 17 days after the first set of responses, and the physicians re-evaluated the new answers.

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

ORIGINAL ARTICLE

DOI: 10.63137/jsteam.744858

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).

Results

Among 80 answers generated by Chat-GPT, the median precision 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 maximum precision level (correctness score of 6) was 30% (n= 24), and a score of almost perfect (precision score of 5) was 38.7% (n= 31). Conversely, completely incorrect (accuracy score of 1) was 8% (n=10). False responses, with a precision 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. Completeness and accuracy and were humbly correlated (Spearman's r = 0.3, 95% CI 0.4 to 0.6, $p < 0.01$, alpha = 0.04) through all questions.

The median precision score for all AI-created responses (n=240) across all of the three datasets (excluding answers that were re-graded) was 5.6 (SD 2.1, IQR 0.2, median 4.9), whereas the median completeness score was 3.2 (average 2.3, SD 1.4, IQR 0.8). All descriptive questions had a median precision of 4.2 (average 3.9, SD 2.2, IQR 1.5), while binary questions had a median precision of 5.8 (average 5.1, SD 1.8, IQR 1.5) (Mann Whitney U p=0.06). The descriptive questions had median precision 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, 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 precision 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).

Discussion

This investigation suggests that after a quarter year of existence, 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 10 specialties contributing 240 questions, highlights that (n=240) 52.1% of the AI responses were deemed "almost all accurate" or "accurate" (accuracy score by mean 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 hold out any statistical importance. 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.

An inner process of validation revealed ChatGPT's ability to significantly upgrade over a short span. This refinement is accredited to uninterrupted updates and advancement of the model's parameters and algorithms, as well as the

Table 1: Examples of answers generated by Chat-GPT to common medical questions

ORIGINAL ARTICLE

DOI: 10.63137/jsteam.744858

Difficulty level	Common medical condition	Question asked	Chat GPT generated answer	Precision rating	Totality rating
Easy-Descriptive	Anemia	What is the most likely cause resulting in anemia?	The most common reason for anemia worldwide is iron deficiency. Iron is a crucial component of hemoglobin, the protein in red blood cells responsible for transporting oxygen from the lungs to the rest of the body. When the body lacks sufficient iron, it cannot produce an adequate amount of hemoglobin, leading to a condition known as iron-deficiency anemia. Iron deficiency can result from various factors, including: 1. Inadequate Dietary Intake: Not consuming enough iron-rich foods, such as red meat, poultry, fish, beans, lentils, and fortified cereals. 2. Poor Iron Absorption: Certain conditions, like celiac disease or inflammatory bowel diseases, can interfere with the absorption of iron from the gastrointestinal tract. 3. Blood Loss: Chronic or acute blood loss can deplete the body's iron stores. Common sources of blood loss include gastrointestinal bleeding (from conditions like ulcers or colorectal cancer), heavy menstrual periods, and frequent blood donations. 4. Increased Iron Requirements: This can occur during periods of rapid growth, such as infancy, adolescence, and pregnancy.	6	3
Medium-Descriptive	Diabetes	What are the complications of untreated hyperglycemia?	Untreated hyperglycemia, or persistently high blood glucose levels, can lead to various complications, particularly in individuals with diabetes. Some of the potential complications include: 1. Cardiovascular Issues: Hyperglycemia is associated with an increased risk of cardiovascular diseases, such as coronary artery disease,	5	2

DOI: 10.63137/jsteam.744858

			heart attack, and stroke. 2. Neuropathy: Prolonged high blood sugar levels can damage nerves, leading to diabetic neuropathy. This can cause pain, numbness, tingling, and impaired sensation, particularly in the extremities. 3. Nephropathy: Hyperglycemia can contribute to kidney damage, leading to diabetic nephropathy. This condition may progress to chronic kidney disease and, in severe cases, kidney failure. 4. Retinopathy: Elevated blood sugar levels can cause damage to the blood vessels in the eyes, leading to diabetic retinopathy. This can result in vision impairment and, if left untreated, blindness. 5. Infections: Hyperglycemia can weaken the immune system, making individuals more susceptible to infections. Infections, especially in the feet, can be a significant concern for individuals with diabetes. 6. Delayed Wound Healing: Elevated blood glucose levels can impair the body's ability to heal wounds. This can lead to slow healing and an increased risk of infections. 7. Macrovascular Complications: Hyperglycemia contributes to the development of atherosclerosis, which involves the hardening and narrowing of large blood vessels. This can result in complications such as peripheral artery disease.		
Hard-Descriptive	Asthma	An 18-year-old patient with moderate persistent asthma should be treated with what regimen based on the latest asthma guidelines from the National	In general, the treatment of asthma is often based on the classification of asthma severity, which includes intermittent, mild persistent, moderate persistent, and severe persistent. Treatment plans may involve a combination of inhaled corticosteroids, long-acting beta-agonists, leukotriene modifiers, and other medications,	4	1

DOI: 10.63137/jsteam.744858

		Asthma Education and Prevention Program?	depending on the severity of the asthma. Consulting with a healthcare professional, such as an asthma specialist or pulmonologist, is essential to tailor the treatment plan to the individual patient's needs, taking into consideration factors like symptom control, lung function, and any potential side effects of medications.		

Note: 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.

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 patient's queries and concerns swiftly and timely. [12]

Results after analyses of four recognizable query sets (asthma, diabetes, anemia, and ordinary conditions) demonstrated impartially consistent and top grades. 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 ongoing improvement of the model over weeks and even days.

Despite such encouraging findings, the generalizability of our interpretations is restricted by the restrained sample magnitude, consisting of thirty-three doctors belonging to a singular medical center for academia, and the quiz comprising 284 questions only, which possibly do not adequately represent the diversity of medical specialties and the myriads of potential 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 prejudice, 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). [13] 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. [14] The study was confined to one particular AI variant, ChatGPT, and possibly is not 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, gadgets like ChatGPT could aid as precious resources for swift medical information reclamation in fast-paced clinical settings, enhancing the efficiency of healthcare and aiding complicated decision-making. [15] 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. [16] 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. [17] However, depending only on the present-day publicly accessible edition

DOI: 10.63137/jsteam.744858

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., towering ChatGPT and similar language models hold the capability to quickly advance and revolutionize the circulation of healthcare information. Recent advancements, such as a GPT-style model of language exclusively trained about biomedical literature, achieving 51% accuracy on diverse biomedical question-answering assignments, underscore the aptitude of context-sensitive language generation models in practical healthcare applications. [18]

Proof of authenticity of healthcare knowledge given by AI requires more research involving bigger communities of medical specialists with varying question categories. [19] Furthermore, studies should also assess how AI-created healthcare data has advanced over time. Additional principal rumination covers privacy and ethical issues. As the data used to train these AI tools predetermine 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 briefed and dispense the latest statistics. Furthermore, text-based AI models might miss subtleties presented only in tabular or figure form instead of being explored expressly in theory. Lastly, upcoming endeavors ought to converge on building agile yet sturdy regulations 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

We highly acknowledge all the participating institutions and participants for their time.

Author Contribution

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

Funding

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

Conflict of Interest

None

Reference

1. Kumar A, Gond A. Natural Language Processing: Healthcare achieving benefits Via NLP. Int J Novel Res Devel. 2023; 8(12): 244-252.
2. Kotei E, Thirunavukarasu R. A Systematic Review of Transformer-Based Pre-Trained Language Models through Self-Supervised Learning. Information. 2023;14(3):187.
3. Lund BD, Wang T. Chatting about ChatGPT: how may AI and GPT impact academia and libraries? Lib Hi Tech News. 2023;40(3):26-9.
4. Vaishya R, Misra A, Vaish A. ChatGPT: Is this version good for healthcare and research? Diabetes & Metabolic Syndrome: Clin Res Rev. 2023;17(4):102744.
5. Yelne S, Chaudhary M, Dod K, Sayyad A, Sharma R. Harnessing the Power of AI: A Comprehensive Review of Its Impact and Challenges in Nursing Science and Healthcare. Cureus. 2023;15(11):e49252..
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.
7. King MR. The future of AI in medicine: a perspective from a Chatbot. Ann Biomed Eng. 2023;51(2):291-5.
8. Bansal G, Chamola V, Hussain A, Guizani M, Niyato D. Transforming Conversations with AI—A Comprehensive Study of ChatGPT. Cognit Comput [Internet]. 2024 January 24 [cited 2024 May 31:1-24. Available from: https://doi.org/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(11):101861.
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;7(1):e48023.
11. 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

DOI: 10.63137/jsteam.744858

Assessment. JMIR Med Ed. 2023;9(1):e45312.
12. Jahanshahi H, Kazmi S, Cevik M. Auto response generation in online medical chat services. J Healthc Inform Res. 2022;6(3):344-74.
13. Gore J, Shortland N, Power N, Brown O. Editorial: Naturalistic decision making (NDM): epistemic expertise in action. Front Psychol. 2023;14:1303098. doi: 10.3389/fpsyg.2023.1303098.
14. Wang YA, Eastwick PW. Solutions to the problems of incremental validity testing in relationship science. Pers Relationsh. 2020;27(1):156-75.
15. Kumar V. Digital enablers. The Economic Value of Digital Disruption: A Holistic Assessment for CXOs: Springer Nature; 2023. p. 1-110.
16. Javaid M, Haleem A, Singh RP. ChatGPT for healthcare services: An emerging stage for an innovative perspective. BenchCouncil Transactions on Benchmarks, Standards and Evaluations. 2023;3(1):100105.
17. Sapci AH, Sapci HA. Artificial intelligence education and tools for medical and health informatics students: systematic review. JMIR Med Ed. 2020;6(1): e19285.
18. 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. bioRxiv. 2024:2024.04. 19.588095.
19. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022;28(1):31-8.