Turk J Public Health
Available online at: http://dergipark.org.tr/tjph/
Doi: 10.20518/tjph.1498611

TURKISH JOURNAL OF PUBLIC HEALTH

ORIGINAL ARTICLE

# Assessing ChatGPT's accuracy and reliability in medical education: a cross-sectional study

A Ra Vishal[1], A S Harshitha [2], Venkatesh Sindhu [3], R Abhivanth [1], MB Pavithra [1], Suwarna Madhukumar[1]

[1] Student, M.V.J Medical College and Research Hospital, Bangalore, India
[2] MD, M.V.J Medical College and Research Hospital, Department of Community Medicine, Bangalore, India

**Received:** 02.09.2024, **Accepted:** 29.03.2025

## Abstract

**Objective:** Artificial intelligence (AI), specifically ChatGPT, developed by Open AI provides human-like understanding and answers to a variety of domain questions and has the potential to transform medical education. However, its reliability in providing accurate clinical information is highly uncertain. This study is aimed at evaluating the accuracy and reliability of ChatGPT in answering multiple-choice questions (MCQs) and protocol-based questions in the field of medicine.

**Methods:** This cross-sectional study was conducted using mixed methods at MVJ Medical College and Research Hospital (April 2024), Hoskote, India, i.e. MCQs (n=228) and protocol-based questions (n=10) from all 19 MBBS Subjects from standard medical literature were used to test ChatGPT. Subject experts checked the responses for accuracy. Statistical analysis, by chi-square test, was performed using IBM SPSS Version 20.0 for Windows.

**Results:** The study findings stated that ChatGPT in easy and simple MCQs, had good accuracy, but its performance lowered with more complex questions, and overall answered about 57.02% of MCQs correctly. Protocol-based questions were given average scores, i.e. 6.35/10 for textbook accurate knowledge and 5.75/10 for real-life application.

**Conclusion:** ChatGPT shows potential as a tool for medical education, especially in recalling basic facts but, it should not be relied upon as a sole source of information, instead used in conjunction with traditional methods to ensure a comprehensive understanding of medical concepts.

**Keywords:** Application, ChatGPT, Knowledge, MCQS, Reliability

**Correspondence:** MD, Pavithra M B, M.V.J Medical College and Research Hospital, Department of Community Medicine, Bangalore, India. **E-mail:** pavi_mb@yahoo.co.in

**Cite This Article:** Vishal AR, Harshitha AS, Sindhu V, Abhivanth R, Pavithra MB, Madhukumar S . Assessing ChatGPT's Accuracy and Reliability in Medical Education: A Cross-Sectional Study. Turk J Public Health 2024;22(3): 11-17.

## INTRODUCTION

Artificial Intelligence (AI) emerged in 1955, with John McCarthy coining the term[1]. Since then, great progress has been made in AI, because of advances in machine learning (ML) and its subdisciplines. AI in healthcare is transforming diagnostic approaches and clinical decision-making, expanding its reach across all medical specialties [2,3].

Despite the growing interest in the application of AI in medical education, there is a notable gap in understanding the reliability of AI models, especially ChatGPT, i.e. Chat Generative Pre-trained Transformer[4] in this context although there is a considerable discourse about the potential of AI's ability to enhance the learning experience and limited empirical evidence of its effectiveness and accuracy for delivering educational content[5,6]. Given the growing use of AI models such as ChatGPT in the medical field, it is essential to investigate and systematically review their reliability to address the existing knowledge gap.

As medical research and study increase, the knowledge and database also increase over a wide area[7]. ChatGPT offers an opportunity to access all this data and give guidance immediately[8]. But, before such technology is embraced, it is important to ensure its accuracy and suitability for medical education[9].

AI-generated educational content risks misinformation due to probabilistic text generation, leading to occasional inaccuracies[13]. Over-reliance of AI can affect clinical reasoning and impair critical thinking. Complex medical issues are difficult for AI to handle, which emphasises the necessity of thorough testing before broad use in medical education.

The main objective of this study is to evaluate the accuracy and reliability of ChatGPT (GPT-3.5, free version, accessed via OpenAI's web interface, April 2024) in handling both multiple-choice questions (MCQs) and Protocol-based questions from standard medical literature on a variety of topics, and also to make recommendations on integrating ChatGPT and similar AI technologies in medical education. Prior studies, such as Jin et al[10]. and Han et al[11]., examined AI performance on USMLE-style MCQs. This study expands on their findings by including protocol-based questions, assessing ChatGPT's accuracy, reliability, and error patterns across medical domains. It also addresses a key gap by recommending optimizations for medical education, an underexplored area.

## METHOD

The cross-sectional study was conducted at MVJ Medical College and Research Hospital, Hoskote, Bangalore, focusing on evaluating the reliability of the ChatGPT to answer medically relevant questions. Before doing the study, ethical clearance was obtained from the institutional ethical committee.

The study used a mixed-methods approach, including multiple-choice questions (MCQs) and protocol-based questions to assess the general reliability of the ChatGPT.

ChatGPT (GPT-3.5, free version, accessed via OpenAI's web interface, April 2024) was used in this study. As the free version does not allow user-controlled parameter adjustments (such as temperature or penalty settings), responses were generated under default system settings. All responses were recorded in a controlled academic environment to

maintain consistency across multiple runs.

A total of 228 MCQs were made, with 12 questions for each of the 19 subjects of the Bachelor of Medicine, Bachelor of Surgery syllabus, and were classified equally according to difficulty, i.e. 4 easy, 4 medium, and 4 hard questions in each subject, from recognized medical literature, such as standard medical textbooks (e.g., Harrison's Internal Medicine[14], Bailey & Love's Surgery[15]), reputed journals ensuring that it had a diverse and representative set of questions. The MCQs were and reviewed by subject matter experts of their respective disciplines for validity. This sample size was chosen to balance statistical robustness and feasibility while covering a wide range of medical topics for a comprehensive analysis of the performance of ChatGPT in various contexts, increasing the generality of the findings.

In addition, 10 protocol-based questions were developed to assess understanding of ChatGPT theoretically and as applied to practical matters and define grading criteria for subject matter experts to evaluate responses provided by ChatGPT.

These questions were designed to evaluate real-world application, and while a larger sample could improve reliability, this number was determined based on expert feasibility and the complexity of evaluating long-form responses. Future studies may expand this dataset for greater generalizability.

A standardized interaction protocol was followed to maintain consistency. ChatGPT was given each question without additional context beyond what a student would receive. Each question was presented as a standalone prompt to ChatGPT with no additional contextual information. For MCQs, ChatGPT was instructed: 'Select only one correct answer and provide no explanation', then a MCQ question with four options A, B, C, D. The response was noted, compared with the premade key and graded subject-wise and difficulty-wise. For protocol-based questions, prompts were structured as: 'Provide a detailed response based on standard clinical guidelines', then the problem statement was given and the response was noted. Each protocol-based response was graded by two independent subject matter experts (Professors/Associate Professors in relevant medical disciplines) using a structured rubric. The rubric assessed two aspects: (i) 'Textbook Knowledge Accuracy' (factual correctness, alignment with standard medical texts). (ii) 'Real-Life Applicability' (practicality of response, alignment with clinical guidelines). Scores were assigned on a 10-point scale, and inter-expert discrepancies were resolved through discussion.

Ensuring reproducibility in studies involving large language models is crucial, as AI-generated responses can vary due to updates and underlying model parameters. Prior research has proposed structured benchmarking frameworks to improve the reliability of AI assessments in public health and medical applications[16].

Statistical analysis was performed using IBM SPSS Version 20.0 for Windows. A chi-square test was used to determine statistical significance in MCQ accuracy across difficulty levels and subjects. For protocol-based questions, mean scores and standard deviations were calculated to assess variability in responses. Confidence intervals (CI) for ChatGPT's accuracy were not explicitly

calculated in this study but should be explored in future research.

## RESULTS

The study findings indicate ChatGPT's performance levels in various aspects of medical education. In Table I, when tested with multiple-choice questions, ChatGPT proved to be relatively more accurate on easy questions (n=54/76; 71.05%), and less on hard questions (n=29/76; 38.16%). Overall ChatGPT answered about 57.02% of all MCQs correctly (p: 0.0004), this shows that there is a statistically significant relationship between difficulty level and the agreement between standard reference books and ChatGPT responses.

**Table 1.** Performance of ChatGPT on Multiple-Choice Questions (MCQs) of Varying Difficulty Levels

| Difficulty | No of Questions | No. of Correct Responses |
|---|---|---|
| Easy | 76 | 54 (71.1%) |
| Medium | 76 | 47 (61.8%) |
| Hard | 76 | 29 (38.2%) |
| Total | 228 | 130 (57.0%) |

Chi-Square Value: 17.86, DF: 2, P Value: 0.0004

In Table II, MCQ responses vary across medical disciplines, where high accuracy is observed in Radiology, Surgery, Anatomy, Pathology. Subjects like Dermatology (41.67%) and Community Medicine (41.67%) had lower accuracy, possibly due to the complexity of diagnostic reasoning required and the variability in treatment guidelines across different geographic regions, which may not be well-represented in ChatGPT's training data.

In Table III ChatGPT's responsiveness to protocol-based questionnaires was examined and graded in categories of textbook accurate knowledge and real-life applicability of that knowledge out of a score of 10 each by experts in the concerned topic. The average score for all questions was 6.35/10 for knowledge and 5.75/10 for application, i.e. there wasn't much difference in the scores of knowledge and application (P Value: 0.7837). While ChatGPT performed well on structured management protocols (e.g., CPR steps: 80% accuracy), its accuracy declined in decision-heavy scenarios (e.g., triage: 60%, hospital waste management: 35%). This suggests that AI performs better in well-defined protocols but struggles with contextual decision-making, possibly due to a lack of real-world clinical experience.

**Table 2.** Distribution of Multiple Choice Question (MCQ) Response Accuracy Across Medical Disciplines

| Subject | No of Questions | No. of Correct Responses |
|---|---|---|
| **Anaesthesia** | 12 | 8 (66.7%) |
| **Dermatology** | 12 | 5 (41.7%) |
| **ENT** | 12 | 6 (50.0%) |
| **Medicine** | 12 | 5 (41. 7%) |
| **Ob&G** | 12 | 8 (66.7%) |
| **Ophthalmology** | 12 | 6 (50.0%) |
| **Orthopaedics** | 12 | 6 (50.0%) |
| **Paediatrics** | 12 | 6 (50.0%) |
| **Psychiatry** | 12 | 7 (58.3%) |
| **Radiology** | 12 | 10 (83.3%) |
| **Surgery** | 12 | 9 (75.0%) |
| **Anatomy** | 12 | 9 (75.0%) |
| **Biochemistry** | 12 | 8 (66.7%) |
| **Physiology** | 12 | 6 (50.0%) |
| **Microbiology** | 12 | 4 (33.3%) |
| **Pathology** | 12 | 9 (75.0%) |
| **Pharmacology** | 12 | 6 (50.0%) |
| **Community Medicine** | 12 | 5 (41.7%) |
| **Forensic Medicine** | 12 | 7 (58.3%) |
| **Total** | 228 | 130 (57.0%) |

Chi-Square Value: 17.18, DF: 18, P Value: 0.5107

**Table 3.** Expert Evaluation of ChatGPT's Responsiveness to Protocol-Based Questionnaires

| Sl No | Question | Knowledge | Application | P Value |
|---|---|---|---|---|
| 1 | Management of Organophosphate Poisoning | 08.0 | 07.0 | 0.6056 |
| 2 | Management of Snake Bite | 06.0 | 06.0 | 1.0000 |
| 3 | Steps of CPR | 08.0 | 08.0 | 1.0000 |
| 4 | Triage in Disaster Management | 09.0 | 06.0 | 0.1213 |
| 5 | Treatment of Diarrhoea (Plan B) | 07.0 | 05.0 | 0.3613 |
| 6 | Hospital Waste Management | 03.5 | 03.5 | 1.0000 |
| 7 | Management of Heat Stroke | 05.0 | 07.0 | 0.6594 |
| 8 | Treatment of Haemorrhagic Shock | 06.5 | 06.5 | 1.0000 |
| 9 | Treatment of Anaphylactic Shock | 03.5 | 03.5 | 1.0000 |
| 10 | Management of Myocardial Infarction | 07.0 | 05.0 | 0.3613 |
| Average | | 06.35 | 05.75 | 0.7837 |

## DISCUSSION

In this study, the reliability of the ChatGPT was assessed by answering multiple-choice medical questions (MCQs) and protocol-based questions. For MCQs, ChatGPT gave high accuracy for easy questions, but lowered accuracy for medium and hard questions, just over half of the MCQs were answered correctly. The questions performed well in subjects such as radiology, surgery, pathology, anatomy, etc while their accuracy was lower in dermatology, medicine, and community medicine. Although its performance on the MCQs was moderate, its ability to answer protocol-based questions accurately and appropriately was not consistent where ChatGPT got only an average score. This suggests that although ChatGPT may be useful in some tasks, such as recalling basic facts, it may not be reliable in more complex clinical situations that require clinical consideration.

A similar study done by Jin et al [10] used 12723 MCQ questions and got 36.7% correct responses, while another by Han et al [11], got 29% on using 454 USMLE MCQ questions. Our study found a higher MCQ accuracy (57.02%), which may be due to differences in question complexity and dataset selection. Unlike these studies, our work also includes protocol-based questions, providing insights into ChatGPT's clinical reasoning abilities beyond factual recall.

The strength of this study is its methodical technique used in this study to evaluate the performance of the ChatGPT across a range of clinical topics and questions. Comprehensive assessment across all 19 MBBS subjects, unlike previous studies focusing only on USMLE MCQs. Mixed-methods approach, incorporating both MCQs and protocol-based questions. Use of standardized grading rubrics for protocol-based questions to enhance consistency.

Limitations of the study include the small sample of protocol-based questions of only 10 cases limit generalizability. Given the complexity of protocol-based questions, a larger sample size was not feasible for detailed expert grading. Future studies should expand this dataset for greater generalizability. Increasing the sample size would enhance statistical robustness. And reliance on subjective grading by subject matter experts may introduce a potential bias. Future studies should use inter-rater reliability scores. Moreover, the study of the performance of

ChatGPT was examined only in a controlled setting and did not assess its usefulness in a real-world clinical setting. Since ChatGPT is continuously updated, responses may vary over time. This study represents a snapshot of its performance and highlights the need for ongoing validation as AI models evolve.

## CONCLUSION

Based on the findings, it can be concluded that although ChatGPT shows potential as a complementary tool for medical education, it should not be relied upon as the sole source of information. AI tools such as ChatGPT should be approached with a lot of caution by medical students and professionals and used along with traditional teaching methods to ensure a proper understanding of correct medical concepts.

Further research is required to further investigate the validity and reliability of ChatGPT, and its limitations in medical education, and to explore the integration of other AI tools into existing medical curricula and clinical practice to determine its practical benefits and impact on patient care. AI tools should be used as supplementary aids, not as standalone sources of medical knowledge. Institutions could incorporate AI-generated MCQs for self-assessment, with faculty moderation to correct misinformation. Future studies should compare ChatGPT's performance with other LLMs (e.g., GPT-4, Claude, Med-PaLM) using a standardized evaluation framework.

## ACKNOWLEDGMENT

**Conflict of Interest:** The authors declare that they have no conflict of interest.

**Financial support:** The authors received no financial support for this article's research, authorship and /or publiscation.

**Ethical Declaration:** The research/study approved by the Institutional Review Board of the relevant college , number MVJMC&RH/IEC-125/2024, dated 14-02-24

**Author Contribution:** ARV, ASH, AVS, AR, PMB, SM: Conception and design of paper. ARV, ASH, AVS, AR, PMB, SM: Data Collection, analysis and interpretation. ARV: Draft Manuscript. ARV, ASH, AVS, AR, PMB, SM: Critical Revision of Article. ARV, ASH, AVS, AR, PMB, SM: Final Approval of the version intended for publication.

**Thanks:** Mr Suresha, Statistician for his valuable insight.

## REFERENCES

1. McCarthy, J., Minsky, M.L., Rochester, N. and Shannon, C.E. 2006. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine*. 27, 4 (Dec. 2006), 12. DOI:https://doi.org/10.1609/aimag.v27i4.1904.
2. Chen J. Playing to our human strengths to prepare medical students for the future. *Korean J Med Educ*. 2017;29(3):193-197. doi:10.3946/kjme.2017.65
3. Meskó B, Hetényi G, Győrffy Z. Will artificial intelligence solve the human resource crisis in healthcare? *BMC Health Serv Res*. 2018;18(1):545. Published 2018 Jul 13. doi:10.1186/s12913-018-3359-4
4. OpenAI. ChatGPT [Internet]. OpenAI API; 2022
5. Savery M, Abacha AB, Gayen S, Demner-Fushman D. Question-driven summarization of answers to consumer health questions. *Sci Data*. 2020;7(1):322. Published 2020 Oct 2. doi:10.1038/s41597-020-00667-z
6. Gutiérrez BJ, McNeal N, Washington C, Chen Y, Li L, Sun H, et al. Thinking about GPT-3 in-context learning for biomedical IE? Think again. arXiv.

Preprint posted online on November 5, 2022. [doi: 10.48550/arXiv.2203.08410]

7. Kolachalama, V. B., & Garg, P. S. (2018). Machine learning and medical education. *NPJ digital medicine*, *1*(1), 54.

8. Zarei M, Mamaghani HE, Abbasi A, Hosseini M. Application of artificial intelligence in medical education: A review of benefits, challenges, and solutions. Medicina Clínica Práctica. doi:10.1016/j.mcpsp.2023.100422

9. Sun L, Yin C, Xu Q, Zhao W. Artificial intelligence for healthcare and medical education: a systematic review. *Am J Transl Res*. 2023;15(7):4820-4828. Published 2023 Jul 15

10. Jin, Di & Pan, Eileen & Oufattole, Nassim & Weng, Wei-Hung & Fang, Hanyi & Szolovits, Peter. (2021). What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. Applied Sciences. 11. 6421. 10.3390/app11146421.

11. Ha LA, Yaneva V. Automatic question answering for medical MCQs: can it go further than information retrieval? In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019). 2019 Presented at RANLP 2019; September 2-4, 2019; Varna, Bulgaria p. 418-422. [doi: 10.26615/978-954-452-056-4_049

12. Xu X, Chen Y, Miao J. Opportunities, challenges, and future directions of large language models, including ChatGPT in medical education: a systematic scoping review. J Educ Eval Health Prof [Internet]. 2024;21:6. [doi: 10.3352/jeehp.2024.21.6]

13. Mackey BP, Garabet R, Maule L, Tadesse A, Cross J, Weingarten M. Evaluating ChatGPT-4 in medical education: an assessment of subject exam performance reveals limitations in clinical curriculum support for students. Discov Artif Intell [Internet]. 2024;4(1). [doi: 10.1007/s44163-024-00135-2]

14. Harrison TR, Braunwald E. Harrison's principles of internal medicine. 15th ed. New York, NY: McGraw-Hill; 2002

15. O'Connell PR, McCaskie AW, Sayers RD, editors. Bailey & love's short practice of surgery - 28th edition. 28th ed. London, England: CRC Press; 2023.

16. Espinosa L, Salathé M. Use of large language models as a scalable approach to understanding public health discourse. PLOS Digit Health [Internet]. 2024;3(10):e0000631 [doi: 10.1371/journal.pdig.0000631]