International Journal of Advances in Medicine
Prakash R et al. Int J Adv Med. 2025 Sep;12(5):478-481
http://www.ijmedicine.com
pISSN 2349-3925 | eISSN 2349-3933

DOI: https://dx.doi.org/10.18203/2349-3933.ijam20252532

Original Research Article

Comparing the performance of ChatGPT and Chatsonic on PLAB-style questions: a cross-sectional study

Rashmi Prakash[1], Kritika Pathak[2]*, Pooja Manjula[3], Samia Abdul Moiz[4], Khawar Tariq Mehmood[5]

[1]Department of Medicine, Adichunchanagiri Institute of Medical Sciences, B.G. Nagara, Karnataka, India
[2]Department of Medicine, HNB Medical Education University, Dehradun, Uttarakhand, India
[3]Department of Medicine, Government Vellore Medical College, Vellore, Tamil Nadu, India
[4]Department of Medicine, Mahadevappa Rampure Medical College, Kalaburagi, Karnataka, India
[5]Department of Medicine, Aster Hospital Br of Aster Dm Healthcare FZC, Al Raffa, Dubai, United Arab Emirates

Received: 06 June 2025
Revised: 23 July 2025
Accepted: 28 July 2025

***Correspondence:**
Dr. Kritika Pathak,
E-mail: 99kritikapathak@gmail.com

ABSTRACT

Background: Artificial Intelligence (AI), particularly large language models like ChatGPT and Chatsonic, has garnered significant attention. These models, trained on massive datasets, generate human-like responses. Studies have assessed their performance on professional and licensing examinations, as well as medical examinations, with varying levels of competency. Assessment of ChatGPT and ChatSonic's competence in addressing PLAB-oriented queries.

Method: We conducted an independent cross-sectional study in May 2023 to evaluate the performance of ChatGPT and Chatsonic on the PLAB-1 Exam. The study used 180 multiple-choice questions from a mock test on the 'Pastest' platform and excluded questions with images, tables, or unanswered by AI. The responses of the two AI models, correct answers, and question difficulty statistics were recorded and compared. The performance of the two AI software packages was assessed based on the recorded metrics.

Results: Out of 180 questions, 141 were included and 39 excluded. ChatGPT outperformed Chatsonic, answering 78% of questions correctly compared to the latter's 66%. ChatGPT achieved 85% accuracy in answering easy questions, while Chatsonic performed poorly across all levels, answering 75% of easy questions, 64% of average questions, and only 38% of difficult questions.

Conclusions: ChatGPT outperformed Chatsonic in all dataset categories and showed non-statistically significant superior performance across difficulty levels. Both AI models' accuracy decreased with increasing question difficulty.

Keywords: Medicine, Data science, Artificial intelligence, Generative AI, ChatGPT, Chatsonic, Multiple choice questions

INTRODUCTION

The domain of generative artificial intelligence (AI) has experienced significant progress in recent times with the advent of large-language models (LLMs). These are computer programs capable of executing a range of operations involving natural language processing, such as language translation and content generation.[1,2] They are accomplished through the use of deep learning algorithms that have been trained on vast amounts of data, allowing them to produce responses that closely resemble human language.[1] Examples of such models include Chat Generative Pre-trained transformer (ChatGPT) and Chatsonic.[3,4] The professional and linguistic assessments board test, commonly referred to as the PLAB test, is designed for doctors who have received their medical training abroad and seek to practice under limited registration in the UK. The exam evaluates a candidate's readiness to serve in a senior house officer capacity at a UK NHS hospital.[5] A study conducted by Kung et al assessed ChatGPT's performance characteristics on the United States Medical Licensing Examination (USMLE) to gauge its capabilities against biological and clinical questions of standardized complexity and difficulty. The results showed that ChatGPT's increased precision helped it approach or exceed the USMLE passing mark.[6]

Aims and objectives

The aim was to compare the performances of ChatGPT and Chatsonic in generating correct responses to PLAB-style questions.

METHODS

We conducted an independent cross-sectional study in May 2023 to compare the performance of ChatGPT3 and Chatsonic4 on the questions similar to the PLAB-1 Exam. Ethics Committee Approval was not obtained for this study, as it did not involve any human interaction or participation.

The study was conducted using questions obtained from a mock test available on the 'Pastest' website, which is a learning platform for PLAB and other medical licensing exams and is recommended by the General Medical Council (GMC) in the UK.7 This website provides question banks that are often used by medical students and offers statistics on question difficulty and exam performance in relation to the user base. Prior to commencing the study, permission was obtained via email for the use of a mock question set from 'Pastest'.7 Questions of all difficulty levels were included. Questions that had images or tables or both (not readable by the AI software) and unanswered by either ChatGPT or Chatsonic were excluded from the study. A total of 180 questions were used.

The 180 questions were divided based on the level of difficulty into “easy”, “average” and “difficult”–as graded by the ‘Pastest’, based on the correct responses by students to these questions. These questions were then typed one by one on ChatGPT and Chatsonic, and responses generated were entered in an Excel sheet. The number of correctly answered questions by both AI software based on difficulty level was totaled and percentages were calculated. For statistical analysis, two proportion z-test was used. Level of significance was kept as <0.05.

RESULTS

A dataset consisting of 180 multiple-choice questions was chosen to evaluate the performance of the two AI systems. These questions closely resembled the format of the PLAB 1 exam and were sourced from a widely-used question bank called Pastest. Around 32 (17.78%) questions with tables or images; 5 (2.78%) questions unanswered by ChatGPT and 2 (1.11%) questions unanswered by Chatsonic were excluded from the study (Table 1). A total of 141 questions were included for comparison of responses generated by ChatGPT and Chatsonic. Out of 141 questions, 110 (78.01%) questions were correctly answered by ChatGPT and 93 (65.95%) by Chatsonic. Using Two proportion z-test, this difference is statistically significant, p = 0.0339 (p <0.05).

However, on the basis of level of difficulty, ChatGPT answered more questions correctly compared to Chatsonic at all levels–easy, average, and difficult. However, this difference was not statistically significant (Table 3).

Table 1: The number of questions included in the study and reason for exclusion.

Category	Count	%
Images/tables	32	17.78
Unanswered by ChatGPT	5	2.78
Unanswered by Chatsonic	2	1.11
Unanswered by both	0	0
Included	141	78.33
Total	180	100

Table 2: Number of correct answers given by ChatGPT and Chatsonic according to difficulty level.

Difficulty level	No. of questions	Correct answers	
		ChatGPT	Chatsonic
Easy	67	57	50
Average	58	44	37
Difficult	16	9	6
Total	141	110	93

Table 3: Comparison of ChatGPT and ChatSonic performance.

Difficulty level	Proportion of correct answers by		P value	Significance
	ChatGPT	Chatsonic		
Easy	0.85	0.75	0.1963	Not Significant
Average	0.76	0.64	0.2249	Not Significant
Difficult	0.56	0.38	0.4786	Not Significant

DISCUSSION

The current study compared the performance of ChatGPT versus Chatsonic to assess performance in PLAB-styled questions and revealed that the correctly answered questions by ChatGPT was significantly greater than that by Chatsonic. However, the difference was not significant for easy, average, and difficult questions. AI aims to develop machines capable of performing tasks requiring human intelligence, such as learning and understanding abstract concepts. The development of large language models has advanced AI but concerns about bias and potential security risks exist. Human intelligence is versatile due to its evolutionary history, adaptability, creativity, emotional intelligence, and ability to comprehend complex abstract concepts. AI has the potential to benefit healthcare, but caution is advised due to associated risks and failures.[8]

Language models have shown remarkable success in tasks such as summarization, translation, and question answering. This success can be attributed to two factors: the transformer architecture, which has a large number of parameters and self-attention mechanisms, and a two-stage training process. In the first stage, LLMs use self-supervised learning to learn from large amounts of unannotated data, eliminating the need for manual annotation. In the second stage, LLMs are fine-tuned on small, task-specific datasets to perform specific tasks with high accuracy.[8]

Recently, several studies have assessed the performance of artificial intelligence models in various examinations across different educational levels, including medical entrance and specialty examinations. A study by Gilson et al, found that ChatGPT achieved a competency rate of 44-64.4% on four sets of USMLE format questions.[9] Its performance decreased as the difficulty of the questions increased in one set. However, ChatGPT outperformed InstructGPT by an average of 8.15% in all sets. In one set of standard USMLE questions, ChatGPT achieved an accuracy rate of over 60%, equivalent to a third-year medical student's passing score. ChatGPT also provided logical justifications and informational context for its answers.[9] Our study, which used a different question format (PLAB), had similar results, with ChatGPT outperforming the second AI model.

Kung et al conducted a study to evaluate ChatGPT's ability to process complex medical info.[6] Using USMLE Step 1, Step 2 CK, Step 3 format questions, which had standardized complexity. Excluding indeterminate responses, ChatGPT achieved 75.0% accuracy for step 1, 61.5% for step 2 CK, and 68.8% for step 3 when questions were encoded as open-ended. When questions were encoded as multiple-choice single-answers with justification, ChatGPT achieved 64.5% competence in step 1, 52.4% in step 2, and 65.2% in step 3. The study concluded that ChatGPT approached or exceeded passing scores of all three examinations. ChatGPT also demonstrated the ability to provide explanations for selected answers with high concordance and additional insights that may aid in learning.

AI can provide assistance to both instructors and students in various ways, such as generating course materials, offering suggestions, and performing language translation for instructors, and serving as a virtual tutor for students, assisting with activities like question answering, summarizing information, facilitating collaboration, and providing feedback.[10]

The efficiency of large language models (LLMs) in medical education is determined by the training data, which can be diverse but may not always include the most recent or specialized medical knowledge. This limitation affects the competency and expertise of the information generated by LLMs, as medical knowledge is constantly evolving. While LLMs can provide general information, they are unable to adapt in real-time to new medical findings and may lead to a superficial understanding of complex medical concepts.[11,12]

ChatGPT's performance in AHA BLS and ACLS exams was studied by Fijacko et al.[13] It achieved 68% and 64% competency in 25-question BLS exams and 68.4% and 76.3% accuracy in 38-question ACLS exams, excluding image-based questions.[13] Although the results were qualitatively similar to our study, the sample questions were not equivalent or similar and were not at the passing level of 84% for these examinations.

According to Zhu et al, ChatGPT achieved an overall competency of 84% in BLS exams with three responses per question.[14] This competency improved to 96% and 92.1% for BLS and ACLS exams, respectively, when incorrect questions were re-entered as open-ended questions. This finding suggests limitations in the Fijacko et al. study and raises questions about AI model consistency. These considerations are important for future AI model performance research in standardized medical examinations.[14]

Al-Shakarchi et al and Haq et al evaluated ChatGPT's performance on the United Kingdom Medical Licensing Assessment-Applied Knowledge Test (UKMLA-AKT), which is set to replace the PLAB in 2024 as a standardized licensing exam.[15] The study found that ChatGPT achieved a 73.3% competency rate on 191 questions from the UK Medical School Council website sample practice material. Although the passing range for PLAB is not publicly known, it is estimated to be between 58.9% and 68.5% and ChatGPT met or exceeded the passing range in most areas of clinical practice, including achieving 100% in questions related to certain specialties.[16]

The study has some limitations. It did not assess the explanations provided by ChatGPT or ChatSonic in response to their answers. The sample size of 180 questions, of which only 141 were assessed, was relatively small compared to the vast amount of medical information available. Therefore, a larger sample size is necessary for a more accurate assessment. AI models may produce different responses to similar questions when phrased differently or when they are open ended. However, these scenarios were not assessed in this study. There is a lack of data on the performance of AI in PLAB and PLAB-format UK licensing medical examinations. Our study did not include questions with images and tables nor did it evaluate the content of the explanations provided by the AI models for each generated response. Furthermore, we did not assess the performance in specific areas of clinical practice or subjects. Therefore, additional large-scale and in-depth studies are necessary to evaluate the performance of AI models in PLAB and PLAB-format medical examinations more accurately.

CONCLUSION

ChatGPT outperformed Chatsonic in overall performance, but there were no significant differences in accuracy across difficulty levels. Our study offers insights for AI-assisted medical evaluations in healthcare education and professional assessments. Further research and model improvements are necessary to enhance AI effectiveness and reliability in practical settings.

Declaration

The authors declare the use of ChatGPT and Chatsonic as research tools in the methodology of this study. The questions were entered into these models to generate data for comparative analysis. The human authors were responsible for the study design, data collection, analysis, and validation of all results. The authors did not use any generative AI or AI-assisted technology for the writing or editing of this manuscript.

Funding: No funding sources
Conflict of interest: None declared
Ethical approval: Not required

REFERENCES

1. Iannantuono GM, Bracken-Clarke D, Floudas CS, Roselli M, Gulley JL, Karzai F. Applications of large language models in cancer care: current evidence and future perspectives. Front Oncol. 2023;13:1268915.
2. Deoghare S. An interesting conversation with ChatGPT about acne vulgaris. Indian Dermatol Online J. 2024;15(1):137-40.
3. Pedersen FH. Ownership at OpenAI: From the Perspective of Enterprise Foundation Governance. 2024.
4. Ahuja M. Analysis and comparison of generative AI chatbot applications. 2024.
5. Schunder E, Adam P, Higa F, Remer KA, Lorenz U, Bender J, et al. Phospholipase PlaB is a new virulence factor of Legionella pneumophila. International J Med Microbiol. 2010;300(5):313-23.
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 Digital Health. 2023;2(2):568.
7. Schunder E. Untersuchungen zur Funktion und Lokalisation der Phospholipase A/Lysophospholipase A (PlaB) von Legionella pneumophila. 2010.
8. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. healthcare (Basel). 2023;11(6):887.
9. Gilson A, Safranek CW, Huang T. How Does ChatGPT Perform on the United States Medical Licensing Examination. The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023;9:45312.
10. Currie GM. Academic integrity and artificial intelligence: is ChatGPT hype, hero or heresy. Semin Nucl Med. 2023;53(5):719-30.
11. Lo CK. What Is the Impact of ChatGPT on Education. A Rapid Review of the Literature. Education Sci. 2023;13(4):410.
12. 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;185:109732.
13. 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;185:109732.
14. Zhu L, Mou W, Yang T, Chen R. ChatGPT can pass the AHA exams: Open-ended questions outperform multiple-choice format. Resuscitation. 2023;188:109783.
15. Al-Shakarchi NJ, Haq IU. ChatGPT performance in the UK medical licensing assessment: how to train the next generation. mayo clinic proceedings: Digital Health. 2023;1(3):56.
16. McManus IC, Wakeford R. PLAB and UK graduates' performance on MRCP(UK) and MRCGP examinations: data linkage study. BMJ. 2014;348:2621.

Cite this article as: Prakash R, Pathak K, Manjula P, Moiz SA, Mehmood KT. Comparing the performance of ChatGPT and Chatsonic on PLAB-style questions: a cross-sectional study. Int J Adv Med 2025;12:478-81.