THE ATLANTIC Journal of Medical Science and Research

Atlantic J Med Sci Res 2025;5(1):9-12
DOI: 10.5455/atjmed.2024.12.023

Original Article

Mevlut Hakan Goktepe[1], Fatma Erdeo[2]

[1]Necmettin Erbakan University, Faculty of Medicine, Department of Internal Medicine, Konya, Türkiye
[2]Necmettin Erbakan University, Nezahat Keleşoğlu Faculty of Health Sciences, Department of Physiotherapy and Rehabilitation, Türkiye

Received: 16 December, 2024
Accepted: 04 March, 2025
Published: 22 March, 2024

Corresponding Author: Fatma Erdeo, Necmettin Erbakan University, Nezahat Keleşoğlu Faculty of Health Sciences, Department of Physiotherapy and Rehabilitation, Türkiye
Email: fatmacobanerdeo@hotmail.com

CITATION
Goktepe MH, Erdeo F. Can ChatGPT pass internal diseases exams?. Atlantic J Med Sci Res. 2025;5(1):9-12. DOI: 10.5455/atjmed.2024.12.023

Can ChatGPT pass internal diseases exams?

Abstract

Aim: Chat Generative Pre-trained Transformer (ChatGPT) is a large multimodal language model created by OpenAI. Although ChatGPT is successful in the medical field, especially in answering medical questions in English, the number of studies evaluating the performance of ChatGPT in Turkish is limited. The aim of this study was to investigate the success of ChatGPT in a theoretical exam measuring professional knowledge in medical education and to explore the existing research evidence on the performance of ChatGPT in exams.
Materials and Methods: The free version 1.2023.256 of ChatGPT was used in this descriptive study. 150 questions from the 2022 2nd semester final exam were used. The 6 questions that the AI could not fully recognise and were abbreviations were removed. Each question was added to ChatGPT and the answer was recorded by comparing it with the correct answer. Enterprise Learning Management Planning System (KEYPS) and Excel were used to calculate accuracy rates for each question type. The answer was received in writing.
Results: ChatGPT answered 104 questions correctly and 46 questions incorrectly out of 150 questions. The item difficulty index of the incorrect questions was 0.68. Normally, this index should be around 0.50. Unanswered questions were categorised as difficult questions.
Conclusion: As a result, ChatGPT's ability to know the questions was not good enough. ChatGPT performed well on negatively worded questions and case questions. ChatGPT can be a useful tool for learning.

Keywords: Artificial intelligence, educational measurement, medicine, exam

INTRODUCTION

Chat Generative Pre-trained Transformer (ChatGPT) is a 175 billion parameter natural language processing model trained on large amounts of data that uses deep learning algorithms to produce human-like responses (1). ChatGPT is designed as a general-purpose conversational tool covering a wide range of topics in everyday life. This makes it potentially useful for customer service, chatbots and many other applications. Since its launch, it has attracted attention by providing automatic responses in the style of Shakespeare's sonnets (2,3).

There are different studies on artificial intelligence in the field of medicine. Although there are artificial intelligence algorithms that read computer tomographs, identify and diagnose skin lesions, its place in the marketing and education system is being investigated. (4-7). ChatGPT gained popularity among students, faculty members and researchers at the beginning of 2023. This popularity also brings some concerns. ChatGPT may increase cheating in online exams and minimise critical thinking skills (8). ChatGPT can help to prepare various articles and abstracts, summarise relevant information and provide suggestions, and create drafts for any assignment (9). Through ChatGPT, texts can be written on a range of topics in various disciplines (10).

Well-accepted and widely used multiple-choice question (MCQ) patterned exams are used in medical education, which can promote learning strategies (11). In medical schools, medical

students face very challenging exams at various examination stages in their professional careers. They are looking for various ways to consolidate their basic knowledge and prioritize study resources and approaches directly related to their exams (12). A study reported that well-structured MCQs were good at testing higher skills (11,13). MCQs in the form of knowledge, comprehension, application, analysis, synthesis and evaluation were determined according to Bloom's taxonomy (14). Well-written MCQs can support student engagement in higher-level cognitive reasoning. The purpose of this study is to evaluate the success of artificial intelligence AI in medical education and to make a prediction about the future.

Research Questions

1. What is ChatGPT's success in multiple choice questions?
2. What is the success of ChatGPT in single replies?
3. Is there any evidence of ChatGPT's strengths or weaknesses in the medical field?

MATERIALS AND METHODS

Since this study was not a study on humans or animals, there was no need to obtain ethical approval and consent forms (15). Since an AI platform was included, sample size estimation was not necessary. On September 15, 2023, we entered multiple choice professional practice questions into ChatGPT4. To examine ChatGPT, we used 150 students' pre-administered questions across three datasets. The questions were asked in Turkish. Additionally, Corporate Education Management Planning System (KEYPS) was used to determine the difficulty level of the questions. KEYBS is a system that allows institutions to organize joint educational activities, compare their curricula and conduct joint assessment and evaluation. In this system, the difficulty levels of the questions were determined by the item difficulty index. The Item Difficulty index is the average of scores for an item scored dichotomously as 1 or 0. It is the ratio of the number of individuals who answered an item correctly to the total number of individuals in the group. It takes values between 0 and 1 (Table 1). A total of 144 questions were processed (Figure 1). A question with a picture was removed from Quiz 1. In Quiz 2 and 3, the questions could not be answered even though they were asked in different ways. Additional information requested. An example question is as follows (Figure 2).

Table 1. Item difficulty index

Item difficulty index	Description
0.00-0.20	Very difficult
0.21-0.40	Difficult
0.41-0.60	Medium difficulty
0.61-0.80	Easy
0.81-1.00	Very easy

Question: Serum creatinine of a 29-year-old male patient with 1-2 mm in diameter, red, raised, widespread lesions in the back and waist area, as well as occasional complaints of severe pain in the fingers and toes, and hypohidrosis, serum creatinine: 1.8 mg/dL, GFR: Proteinuria of 54ml/min/1.73 m2 and 1 g/day is observed in the 24-hour urine analysis. Which of the following is the most likely diagnosis in this patient?

A) Von Hipple Lindau
B) Tuberous Sclerosis
C) Fabry Disease
D) Bartter Syndrome
E) Psödohipoaldosteronizm

ChatGPT4: This patient's symptoms resemble typical symptoms of Fabry Disease. Fabry Disease is an X-chromosome-linked disease caused by the deficiency or absence of the alpha-galactosidase A enzyme. Symptoms of the disease include pain, hypohidrosis (decreased sweating), small, red-pink lesions on the skin (angiokeratomas), kidney failure, and proteinuria. These symptoms coincide with those reported by the patient. However, genetic tests are required to make a definitive diagnosis. Therefore, it is important for the patient to be evaluated by a healthcare professional. This information should not be used as a substitute for medical advice provided by a healthcare professional.

Figure 1. Sample question

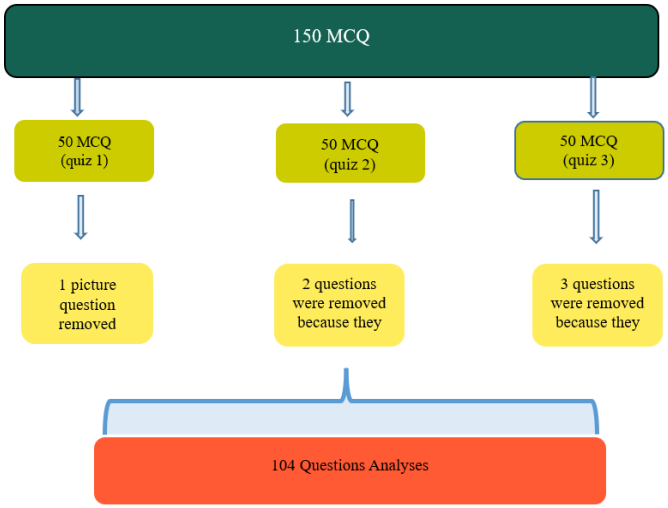

Figure 2. Selection of questions

Data Analysis

Data analysis was performed using SPSS22. Data were expressed as mean±standard deviation, and comparison was made using the Chi-Square test. 0.05 was taken as the significance level.

RESULTS

A total of 178 people took the exam (Table 1). The minimum and maximum scores they received and their average scores are shown in the table. The scores ChatGPT received and the difficulty levels of the questions are shown in Table 2.

Table 2. Descriptive analyses of student's quizes

	N	Min	Max	Mean	SD
Quiz 1	75	28	90	70.60	11.75
Quiz 2	46	26	84	63.46	11.49
Quiz 3	57	26	90	66.95	13.12

Student and ChatGPT responses were compared. The chi-square test gave results with χ^2=3.41, degrees of freedom (df)=2, and a significance level of 0.05. This result shows that the relationship between the 2 variables is not significant (p=0.300) (Table 3).

Table 3. ChatGPT analyses

	Quiz 1		Quiz 2		Quiz 3	
	Right	**Wrong**	**Right**	**Wrong**	**Right**	**Wrong**
ChatGPT 4	35	14	42	4	27	20
ChatGPT 4	70 score		84 score		54 score	
	X	**SD**	**X**	**SD**	**X**	**SD**
Item difficulty index	69.49	23.93	70.60	11.92	24.24	11.13

DISCUSSION

ChatGPT is becoming more popular with its rapid development. Trained using deep learning techniques and online resources using all data until September 2021. In Türkiye, intern students at medical faculties are given a standard test that takes an average of 1.5 hours, although it varies from university to university. Students answer 50 multiple choice questions. The test consists of questions that are both knowledge-based and cover all areas of medicine, including questions regarding the clinical management of patients, professional conduct and medical ethics (16-18). In this study, we evaluated the success of ChatGPT in the professional practice exam in the field of internal medicine. Our aim was to determine ChatGPT's success rate in this exam and evaluate possible risks and advantages. According to the results of the study, ChatGPT was successful in internal medicine questions by receiving 70 and 84 points in Quiz 1 and Quiz 2. The item difficulty index of these questions is in the range of 0.61-0.80. It is defined as an easy question.

He failed Quiz3 with 54 points. When the item difficulty index is examined, it is seen that it consists of difficult questions in the range of 0.21-0.40.

ChatGPT aims to generate human-like language through deep learning. In addition to its use for educational purposes, its ability to answer questions related to medical exams is mind-boggling (19,20). In Jin et al.'s study, it achieved only 37% accuracy when evaluating a dataset of 12,723 questions from Chinese medical exams (16). Oh et al. reported a lower accuracy rate of 29% in the medical domain in 2019 by analyzing 454 questions (18). Gilson et al. showed that ChatGPT performed comparable to previous models and even surpassed them when compared to questions of similar difficulty and content (19). These results support our findings.

The greatest strength of our study is that it has a comprehensive data set consisting of 150 exam questions in the department of internal medicine. In addition, the difficulty levels of the questions in this data set were determined by the item difficulty index. While he answered incorrectly to some of the questions with a high item difficulty index, he also answered correctly to some questions with a low item difficulty index. This finding of a significant difference in performance between question formats is consistent with the literature that ChatGPT is less successful in MCQs (21). In our study, the success rate was 25% when only 2 out of 8 choice questions were answered correctly. These differences observed in the accuracy of ChatGPT's answers for single-choice and multiple-choice questions can be thought of as ChatGPT's inability to evaluate each answer independently. Instead, it is designed to evaluate the available options and prioritize the logically correct answer.

There was no significant difference in the correct answer rate according to the knowledge level of the items. However, this may differ with other tests. As a result, ChatGPT has succeeded in exams in the field of internal medicine. If the literature is examined from the beginning, ChatGPT has developed very rapidly, especially in the last two years. Today's rapid progress in the field of medicine poses a risk in terms of exam security. Instead, the advantages of ChatGPT can be benefited from with education-oriented studies.

Limitation

The limiting feature of our study is that it was studied with a single artificial intelligence language model. To obtain a general result, it is important to repeat the study with different chat models such as GPT-3.5, GPT-4, Bard, Claude, and Bing and draw a common conclusion.

CONCLUSION

As a result, GPT-4 passed exam I and exam II with good results, achieving the passing grade of a qualified Turkish medical candidate. GPT-4 received a lower grade in exam III, as it mainly included practical questions. Our study shows that ChatGPT can help understand medical information in Turkish and has the potential to be an electronic health infrastructure supporting medical development in Turkish.

Conflict of Interests
The authors declare that there is no conflict of interest in the study.

Financial Disclosure
The authors declare that they have received no financial support for the study.

Ethical Approval
Since this study was not a study on humans or animals, there was no need to obtain ethical approval and consent forms.

REFERENCES

1. Scott K. Microsoft teams up with OpenAI to exclusively license GPT-3 language model, Sep 22, 2020. https://blogs.microsoft.com/blog/2020/09/22/microsoft-teams-up-with-openai-to-exclusively-license-gpt-3-language-model/ access date 21.03.2025.
2. Bowman E. A new AI chatbot might do your homework for you. But it's still not an A+ student, December 19, 2022. https://www.npr.org/2022/12/19/1143912956/chatgpt-ai-chatbot-homework-academia access date 21.03.2025.
3. Gilson A, Safranek C, Huang T, et al. How does ChatGPT perform on the medical licensing exams? The implications of large language models for medical education and knowledge assessment. MedRxiv. December 26, 2022. doi: 10.1101/2022.12.23.22283901. [Epub ahead of print].
4. Xiao D, Meyers P, Upperman JS, Robinson JR. Revolutionizing healthcare with ChatGPT: an early exploration of an AI language model's impact on medicine at large and its role in pediatric surgery. J Pediatr Surg. 2023;58:2410-5.
5. Fink MA, Bischoff A, Fink CA, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. Radiology. 2023;308:e231362.
6. Rogasch JMM, Metzger G, Preisler M, et al. ChatGPT: can you prepare my patients for [18^{F}]FDG PET/CT and explain my reports?. J Nucl Med. 2023;64:1876-9.
7. Arviani H, Tutiasri RP, Fauzan LA, Kusuma A. ChatGPT for marketing communications: friend or foe?. Kanal: Jurnal Ilmu Komunikasi. 2023;12:1-7.
8. Rahman M, Watanobe Y. ChatGPT for education and research: opportunities, threats, and strategies. Appl Sci. 2023;13:5783.
9. Khlaif ZN, Mousa A, Hattab MK, et al. The potential and concerns of using AI in scientific research: ChatGPT performance evaluation. JMIR Med Educ. 2023;9:e47049.
10. Huang J, Tan M. The role of ChatGPT in scientific communication: writing better scientific review articles. Am J Cancer Res. 2023;13:1148-54.
11. Ali R, Sultan AS, Zahid N. Evaluating the effectiveness of 'MCQ development workshop using cognitive model framework: A pre-post study. J Pak Med Assoc. 2021;71:119-21.
12. Kenwright D, Dai W, Osborne E, et al. "Just tell me what I need to know to pass the exam!" Can active flipped learning overcome passivity?. Asia Pac Sch. 2017;2:1-6.
13. Khan MU, Aljarallah BM. Evaluation of modified essay questions (MEQ) and multiple choice questions (MCQ) as a tool for assessing the cognitive skills of undergraduate medical students. Int J Health Sci (Qassim). 2011;5:39-43.
14. Zaidi NLB, Grob KL, Monrad SM, et al. Pushing critical thinking skills with multiple-choice questions: does bloom's taxonomy work?. Acad Med. 2018;93:856-9.
15. Wang X, Gong Z, Wang G, et al. ChatGPT Performs on the Chinese National Medical Licensing Examination. J Med Syst. 2023;47:86.
16. Jin D, Pan E, Oufattole N, et al. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl Sci. 2021;11:6421.
17. 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), Varna, Bulgaria. INCOMA Ltd. 2019;418-22.
18. Biswas S. ChatGPT and the future of medical writing. Radiology. 2023;307:e223312.
19. Gilson A, Safranek CW, Huang T, 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 assessment. JMIR Med Educ. 2023;9:e45312.
20. Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. arXiv. 30 Dec 2022. doi: 10.48550/arXiv.2212.14882. [Epub ahead of print].
21. Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. J Educ Eval Health Prof. 2023;20:1.