EURASIAN JOURNAL OF EMERGENCY MEDICINE
Original Article
DOI: 10.4274/eajem.galenos.2025.32644
Eurasian J Emerg Med.

Evaluation of ChatGPTs Performance in Türkiye's First Emergency Medicine Sub-Specialization Exam

Hüseyin Mutlu, Kamil Kokulu, Ekrem Taha Sert, Muhammed Ali Topuz

Aksaray University Faculty of Medicine, Department of Emergency Medicine, Aksaray, Türkiye

Abstract

Aim: This study aims to evaluate ChatGPT's performance in Türkiye's Emergency Medicine Sub-Specialization Exam by assessing its success in answering both standalone and scenario-based questions through repeated testing.

Materials and Methods: This study utilized 60 multiple-choice questions from the Emergency Medicine Sub-Specialization Exam, comprising 30 standalone questions (50%) and 30 scenario-based questions (50%). Each question was presented to ChatGPT five times on different days, with all tests being conducted by researchers using the same computer. The latest version of ChatGPT, based on the GPT-4 architecture and extensively trained on medical texts and journals as of October 2023, was employed to ensure the highest available level of medical knowledge.

Results: ChatGPT achieved an overall accuracy rate of 85%, correctly answering 255 out of 300 questions across five trials. The accuracy rates for the five trials were 85% (51/60), 86.7% (52/60), 86.7% (52/60), 85% (51/60), and 81.7% (49/60), respectively, with no statistically significant difference between trials (p=0.94). ChatGPT demonstrated significantly higher accuracy in standalone questions compared to scenario-based questions [91.3% (137/150) vs. 78.7% (118/150), p=0.002]. Notably, ChatGPT exhibited consistent accuracy in interpreting visual data and correctly answering the two radiology-related questions across all five trials.

Conclusion: ChatGPT demonstrated high performance and consistency in Türkiye's first Emergency Medicine Sub-Specialization Exam, particularly excelling in standalone questions and radiological image interpretation. While the system is generally promising, its lower performance on scenario-based questions highlights the need for further development of clinical reasoning skills. These findings suggest potential applications of artificial intelligence systems in medical education and assessment, while emphasizing the necessity for improvements in clinical decision-making abilities.

Keywords: ChatGPT, artificial intelligence, emergency medicine, sub-specialty examination, medical education

Introduction

Recent advancements in artificial intelligence (AI) technology have introduced significant potential applications in the medical field. Specifically, Large Language models (LLMs), a subset of AI systems developed for natural language processing, have shown promise in medical knowledge evaluation and clinical decision-making (1). Among these, ChatGPT, developed by OpenAI and launched in November 2022, has attracted considerable attention for its performance in medical education (2).

Early studies evaluating ChatGPT's performance in medicine focused on the United States Medical Licensing Examination, where the system achieved success rates exceeding 60% (3). Similarly, studies conducted in Europe reported its successful performance in various specialty board exams (4). In the field of emergency medicine, ChatGPT has demonstrated that it can be used in successful triage of mass casualty events, and has shown promising results in Taiwan's Emergency Medicine Sub-Specialization Exam (5,6). While these findings highlight AI's potential in medical knowledge evaluation, they also underscore the need for further research on its role in clinical reasoning and decision-making processes.

Sub-Specialization Exams comprehensively assess both theoretical knowledge and clinical reasoning skills of the

Corresponding Author: Muhammed Ali Topuz MD, Aksaray University Faculty of Medicine, Department of Emergency Medicine, Aksaray, Türkiye
E-mail: alitopuzmd@gmail.com **ORCID ID:** orcid.org/0000-0003-2677-1260

Cite this article as: Mutlu H, Kokulu K, Sert ET, Topuz MA. Evaluation of ChatGPTs performance in Türkiye's first emergency medicine sub-specialization exam. Eurasian J Emerg Med. [Epub Ahead of Print].

Received: 04.02.2025
Accepted: 14.03.2025
Epub: 05.05.2025

candidates (7,8). Previous studies have reported ChatGPT's success in Sub-Specialization Exams for Medical Specializations (YDUS) and evaluations of sub-specialty trainees (7-9). In this study, we aim to evaluate ChatGPT's performance in Türkiye's first Emergency Medicine in YDUS (ED-YDUS). Specifically, we analyzed the system's performance on standalone and scenario-based questions, assessed its consistency across repeated tests, and examined its accuracy in radiological image interpretation.

Materials and Methods

In this observational study, we evaluated ChatGPT's performance on the ED-YDUS. In Türkiye, YDUS has been administered by the Student Selection and Placement Center (ÖSYM) across various sub-specialties since 2010. The first ED-YDUS was conducted on December 15, 2024 (10). The exam consists of 60 multiple-choice questions, each with five answer options, and is prepared in Turkish by ÖSYM. While designing the exam, ÖSYM refers to standard reference textbooks in emergency medicine, including Tintinalli's Emergency Medicine: Comprehensive Study Guide, 9th Edition, and Rosen's Emergency Medicine: Concepts and Clinical Application, 10th Edition. In this study, we used ChatGPT-4 Omni (ChatGPT-4o), considered to have the highest level of medical knowledge among its peers (11). ChatGPT-4o is an advanced LLM developed by OpenAI using the GPT-4 architecture (12), extensively trained on medical datasets, including texts and journals, up to September 2024 (13).

YDUS exam questions were obtained from the official website (https://ais.osym.gov.tr/bireyselgiriş/yandalsoruları) between December 15 and 25, 2024. The 60 multiple-choice questions were independently evaluated by two authors (Kamil Kokulu and Hüseyin Mutlu) and categorized into two groups: standalone questions and scenario-based questions. In cases where the two authors had differing decisions in categorization, a third author (Ekrem Taha Sert) reviewed the question, and a final decision was reached. Of the total questions, 30 (50%) were classified as standalone, while the remaining 30 (50%) were categorized as scenario-based.

Each question was presented to ChatGPT-4 one time on five separate days between December 25 and 31, 2024 by one of the authors (Muhammed Ali Topuz), using the same computer. For each question, five responses were generated. This methodology aligns with previous research where each question is presented three times to assess consistency and stability in responses generated by LLMs (10,14). All data, including the official answer key provided by ÖSYM, along with ChatGPT-4's responses, were systematically recorded in a Microsoft Excel 2023 document (Version 16.73, Microsoft Corporation, Redmond, WA).

Since this study did not involve human or animal subjects, ethical committee approval was not required.

Statistical Analysis

Statistical analysis was conducted using SPSS software (Version 26.0, SPSS Inc., Chicago, IL, USA). A $p<0.05$ was considered statistically significant. Categorical variables were presented as absolute numbers and percentages. Comparisons between categorical variables were performed using the chi-square test or Fisher's exact test as appropriate. Agreement between ChatGPT-4o's responses was evaluated using Cohen's kappa and Fleiss's kappa coefficients.

Results

ChatGPT correctly answered 255 out of 300 YDUS questions, achieving an overall accuracy rate of 85%. When presented with YDUS questions for the first time, ChatGPT demonstrated an accuracy rate of 85% (51/60). The accuracy rates for subsequent trials were 86.7% (52/60), 86.7% (52/60), 85%

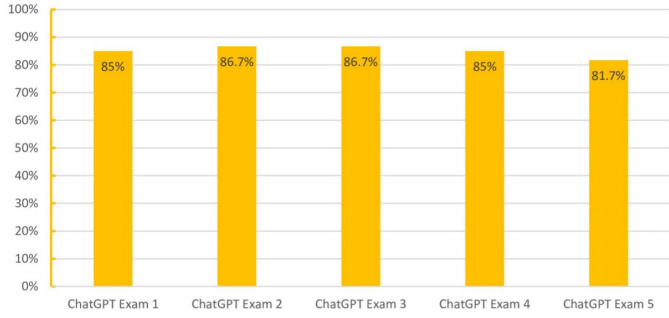

Figure 1. ChatGPTs Performances in Emergency Situation Medical YDUS Exam

ED-YDUS: Emergency Medicine in Sub-Specialization Exams for Medical specializations

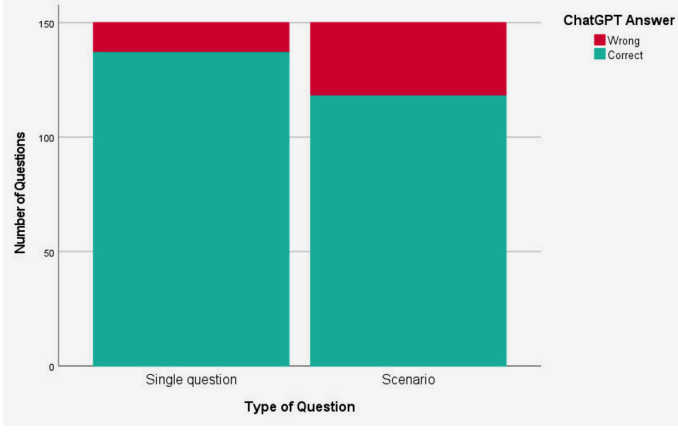

Figure 2. Accuracy of ChatGPT in ED-YDUS. ChatGPT 5 answers created for each question independently and scenario based

ED-YDUS: Emergency Medicine in Sub-Specialization Exams for Medical specializations

(51/60), and 81.7% (49/60), with no statistically significant difference in accuracy across trials (p=0.94) (Figure 1). ChatGPT demonstrated significantly better performance on standalone questions compared to scenario-based questions, with accuracy rates of 91.3% (137/150) and 78.7% (118/150), respectively (p=0.002, Pearson's chi-square test) (Figure 2). Additionally, ChatGPT consistently provided correct answers to both questions involving X-ray or computed tomography images and requiring radiological image analysis across all five trials (Figure 3).

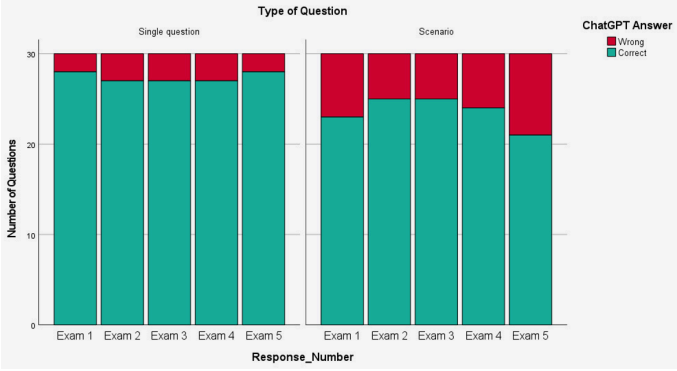

Figure 3. Five different exam ChatGPTs performances independent and scenario-based ED-YDUS questions

ED-YDUS: Emergency Medicine in Sub-Specialization Exams for Medical specializations

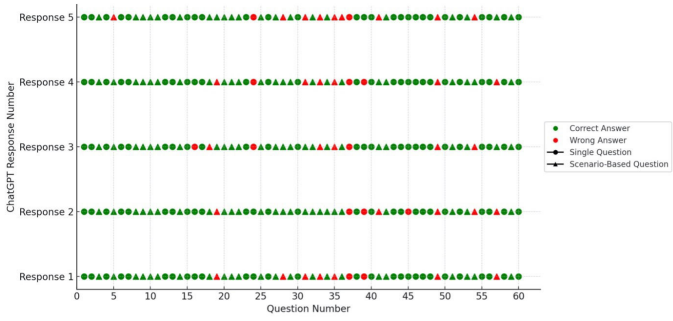

Figure 4. ChatGPT answer correctness by response number and question type

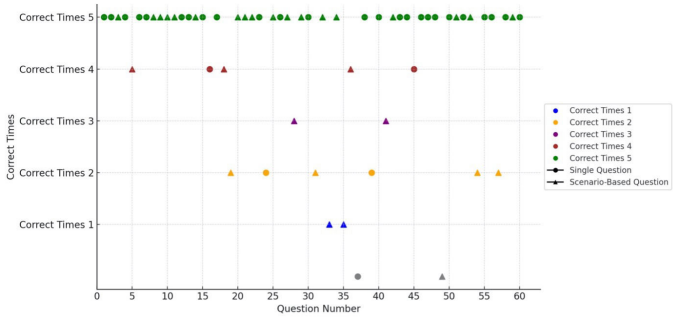

Figure 5. ChatGPT answer correctness by correct times and question type

Almost perfect agreement was found between the five answers (answers 1,2,3,4, and 5) given by ChatGPT for each question, when the answers correctly marked (or selected) by ChatGPT for each question (choice A,B,C,D, or E) were analyzed [Fleiss' kappa =0.83, 95% confidence interval (CI): 0.79-0.87, p<0.001] (Figure 4).

When the responses given by ChatGPT were categorised as "correct" and "incorrect", there was moderate agreement between the five responses given by ChatGPT for each question (Fleiss' kappa =0.50, 95% CI: 0.42-0.58, p<0.001) (Figure 5).

Of the exam questions: 12 Option A, 13 Option B, 11 Option C, 12 Option D, 12 Option E were correct answers. When the options marked by ChatGPT were analyzed, ChatGPT selected Option A 61 times, Option B 61 times, Option C 57 times, Option D 60 times, and Option E 61 times.

Discussion

The evaluation of medical knowledge and clinical reasoning skills of AI systems has become an increasingly important area of research in recent years (7,15). Our study is the first to evaluate the performance of ChatGPT in the first ED-YDUS in Türkiye. ChatGPT's overall success rate of 85% in five different exams and its consistent performance in repeated exams show the potential of AI systems in medical knowledge assessment. It is particularly noteworthy that all questions requiring radiological image analysis were answered correctly. This success parallels the development of the ability of AI systems to interpret visual medical data (7).

Ghanem et al. (9) reported that ChatGPT was less successful in scenario-based questions (55.56%) than in standalone questions (65.83%). Similarly, the study by Takagi et al. (16) found that ChatGPT performed less well in complex scenarios requiring clinical judgement. In our study, we also found that the performance of ChatGPT on standalone questions (91.3%) was statistically significantly higher than that on scenario-based questions (78.7%) (p=0.002). This finding suggests that AI systems are successful in assessing isolated medical knowledge, but have room for improvement in analysing complex clinical scenarios. This suggests that clinical decision-making processes should be further optimised in future versions of AI systems (1).

The fact that the distribution of ChatGPT's answers is balanced (A: 61, B: 61, C: 57, D: 60, E: 61) and that this distribution is similar to the distribution in the real answer key of the exam (A: 12, B: 13, C: 11, D: 12, E: 12) shows that the system does not respond randomly and makes a consistent evaluation. This finding supports the idea that AI systems have an objective and

systematic approach to the evaluation of medical information. A similar consistency was observed by Skalidis et al. (4) in their evaluation of the European Cardiology Board Examination, and it was emphasised that AI showed a systematic approach, choosing answers (2).

When analysing the performance of ChatGPT in repeated exams, success rates of 85%, 86.7%, 86.7%, 85%, and 81.7% were obtained from the first to the fifth exam, respectively. The study by Lee et al. (17) on Basic Life Support and Advanced Cardiovascular Life Support while Kokulu et al. (12) examined Paediatric Advanced Life Support. Both studies showed that ChatGPT-4 performed consistently in repeated assessments. This consistency suggests that AI systems can be used as a reliable tool in the assessment of medical information (18).

Ghanem et al. (9) reported that ChatGPT had little success with questions containing images, and Panthier and Gatinel (19) stated that questions containing images should be removed from scoring in their study conducted in the Ophthalmology Board Exam. In the study conducted by Toyama et al. (20) in the Radiology Board Exam, it was emphasised that AI systems still have areas requiring further development in image interpretation. In contrast to these studies, one of the interesting findings of our study was that ChatGPT correctly answered both questions requiring radiological image analysis, in all five repetitions. The success of ChatGPT in answering questions involving images shows the development of AI systems' skills in interpreting visual medical data. This is a remarkable development, especially given the importance of rapid and accurate interpretation of radiological images in emergency medicine.

Study Limitations

Our study shows that the latest GPT-4 model makes significant progress in addressing professional exam questions. However, there are some limitations to our study. First, as previous research has shown, the language in which questions are asked can have a significant impact on results. Since there may be a difference in meaning between the languages, when questions are asked in Turkish or translated into English, the accuracy of the answers may be affected. Secondly, if the sources of the questions (Tintinalli's Emergency Medicine: A Comprehensive Study Guide, 9th Edition, and Rosen's Emergency Medicine: Concepts and Clinical Practice, 10th Edition) were not taught to GPT-4, this may have influenced the accuracy of the answers. Third, since there were no Emergency Medicine Sub-Specialization study questions, they could not be uploaded to GPT-4. Finally, as analyses of the ED-YDUS results have not been published, the performance of GPT-4o could not be compared with actual exam results.

Conclusion

ChatGPT exhibited remarkable success in the first ED-YDUS exam in Türkiye with an overall success rate of 85% and an especially high performance in standalone questions with a success rate of 91.3%. Interestingly, it showed consistent performance in all five exams and correctly answered questions requiring radiological image analysis. The high success rate of ChatGPT in ED-YDUS, shows that AI systems can be used as a potential tool in medical education and knowledge assessment. However, the system's relatively low performance in scenario-based questions and clinical decision making suggests that AI cannot yet replace human clinical reasoning, but it can only be used as a supporting tool. Future research should focus on refining the clinical reasoning capabilities of AI systems and optimising their role in medical education.

Ethics

Ethics Committee Approval: Since the study did not involve human or animal subjects, ethics committee approval was not required.

Informed Consent: Since the study did not involve human or animal subjects, ethics committee approval was not required.

Footnotes

Authorship Contributions

Surgical and Medical Practices: H.M., K.K., Concept: H.M., K.K., M.A.T., Design: H.M., K.K., M.A.T., Data Collection or Processing: M.A.T., K.K., Analysis or Interpretation: H.M., K.K., E.T.S., Literature Search: H.M., K.K., E.T.S., M.A.T., Writing: H.M., K.K.

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

Financial Disclosure: There are no financial conflicts of interest to disclose.

References

1. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2:230-43.
2. Sharma S, Pajai S, Prasad R, Wanjari MB, Munjewar PK, Sharma R, et al. A critical review of ChatGPT as a potential substitute for diabetes educators. Cureus. 2023;15:e38380.
3. 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 assessment. JMIR Med Educ. 2023;9:e45312.
4. Skalidis I, Cagnina A, Luangphiphat W, Mahendiran T, Muller O, Abbe E, et al. ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story? Eur Heart J Digit Health. 2023;4:279-81.
5. Gan RK, Ogbodo JC, Wee YZ, Gan AZ, González PA. Performance of Google bard and ChatGPT in mass casualty incidents triage. Am J Emerg Med. 2024;75:72-8.

6. Liu CL, Ho CT, Wu TC. Custom GPTs Enhancing performance and evidence compared with GPT-3.5, GPT-4, and GPT-4o? A study on the emergency medicine specialist examination. Healthcare (Basel). 2024;12:1726.
7. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28:924-33.
8. Noda R, Izaki Y, Kitano F, Komatsu J, Ichikawa D, Shibagaki Y. Performance of ChatGPT and Bard in self-assessment questions for nephrology board renewal. Clin Exp Nephrol. 2024;28:465-9.
9. Ghanem D, Nassar JE, El Bachour J, Hanna T. ChatGPT Earns American Board Certification in Hand Surgery. Hand Surg Rehabil. 2024;43:101688.
10. https://www.osym.gov.tr/TR,5149/2010-ydus-sonbahar-donemi-ek-yerlestirme-tercihlerin-alinmasi-27122010.html
11. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. Commun Med (Lond). 2023;3:141.
12. Kokulu K, Demirtaş MS, Sert ET, Mutlu H. ChatGPT and pediatric advanced life support: a performance evaluation. Resuscitation. 2024;205:110451.
13. He K, Mao R, Lin Q, Ruan Y, Lan X, Feng M, et al. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. Information Fusion. 2025;102963.
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. Hoch CC, Wollenberg B, Lüers JC, Knoedler S, Knoedler L, Frank K, et al. ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. Eur Arch Otorhinolaryngol. 4271-8.
16. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: Comparison Study. JMIR Med Educ. 2023;9:e48002.
17. Lee GU, Hong DY, Kim SY, Kim JW, Lee YH, Park SO, et al. Comparison of the problem-solving performance of ChatGPT-3.5, ChatGPT-4, Bing Chat, and Bard for the Korean emergency medicine board examination question bank. Medicine (Baltimore). 2024;103:e37325.
18. Delahanty RJ, Kaufman D, Jones SS. Development and evaluation of an automated machine learning algorithm for ın-hospital mortality risk adjustment among critical care patients. Crit Care Med. 2018;46:e481-e488.
19. Panthier C, Gatinel D. Success of ChatGPT, an AI language model, in taking the French language version of the European Board of Ophthalmology examination: A novel approach to medical knowledge assessment. J Fr Ophtalmol. 2023;46:706-11.
20. Toyama Y, Harigai A, Abe M, Nagano M, Kawabata M, Seki Y, et al. Performance evaluation of ChatGPT, GPT-4, and Bard on the official board examination of the Japan Radiology Society. Jpn J Radiol. 2024;42:201- 7.