

ChatGPT's Medical Exam Performance: Version and Language Analysis in General Surgery Fellowship Exam

 Suleyman Orman¹

¹ Department of General Surgery, Bursa Yuksek Ihtisas Training and Research Hospital, Bursa, Türkiye

What's known on the subject? and What does the study add?

ChatGPT performs well on English medical exams but struggles with non-English languages due to limited training data. This study evaluates ChatGPT on the Turkish GSFE, showing that ChatGPT-4 outperforms 3.5 and that language differences have minimal impact.

Abstract

Objective: The integration of Artificial Intelligence (AI) in medical education has the potential to revolutionize learning and assessment. This study evaluates the performance of ChatGPT-3.5 and ChatGPT-4 on the General Surgery Fellowship Examination (GSFE) in Turkey, comparing their accuracy in answering multiple-choice questions (MCQs) in Turkish and English.

Methods: 255 retired and publicly available GSFE questions (2011-2022) were analyzed. Questions were first presented in Turkish and subsequently translated into English for re-evaluation. ChatGPT-3.5 and ChatGPT-4 were prompted as if they were general surgeons answering the MCQs. The accuracy of responses was assessed, and statistical analyses were performed to identify significant differences between the bots and languages.

Results: In Turkish, ChatGPT-3.5 achieved 66.66% accuracy (170/255 correct answers), while ChatGPT-4 scored 69.41% (177/255). In English, ChatGPT-3.5 achieved 67.05% accuracy (171/255), and ChatGPT-4 scored 70.19% (179/255). Statistically significant differences were observed between ChatGPT-3.5 and ChatGPT-4 for both Turkish ($p < 0.05$) and English ($p < 0.05$) questions. However, language differences within the same versions were not statistically significant ($p > 0.05$). It was found that Cohen's Kappa number was quite high between pairwise comparisons between applications, and the consistency among the four applications showed similarly high agreement (Fleiss' Kappa = 0.95, $p < 0.001$).

Conclusion: ChatGPT-3.5 and ChatGPT-4 demonstrated satisfactory performance on GSFE questions, surpassing the minimum threshold for success in the examination. ChatGPT-4 outperformed ChatGPT-3.5 in both Turkish and English, highlighting the advancements in AI model development. This study underscores the promise of AI in medical education while emphasizing the need for further refinement to address linguistic diversity and domain-specific challenges.

Keywords: Artificial intelligence, ChatGPT, General surgery fellowship examination, Multiple-choice questions, Medical education.

Correspondence: Suleyman ORMAN MD, Department of General Surgery, Bursa Yuksek Ihtisas Training and Research Hospital, Bursa, Türkiye

E-mail: suleymanorm@hotmail.com ORCID-ID: orcid.org/0000-0003-4840-1279

Received: 10.02.2025 Accepted: 18.02.2025

Cite this article as: Orman S. ChatGPT's Medical Exam Performance: Version and Language Analysis in General Surgery Fellowship Exam Eur J Hum Health. 2025;1(1):1-9.

©Copyright 2025 by the European Journal of Human Health.

Licensed by Creative Commons Attribution-NonCommercial-NoDerivatives (CC BY-NC-ND) 4.0 International License.



Introduction

The role of artificial intelligence (AI) in medical education and examination assessment is rapidly expanding. The evolution of large language models (LLMs) has enabled the development of advanced AI tools such as ChatGPT (1,2). ChatGPT has demonstrated significant potential in supporting clinical decision-making, improving patient outcomes, and creating educational materials for medical students and professionals (3-6).

The General Surgery Fellowship Examination (GSFE) in Turkey has been conducted since 2011 to assess the knowledge of surgical specialists before they are admitted to advanced training programs. This multiple-choice exam comprehensively evaluates candidates' expertise in general surgery, including diagnosis, surgical treatment, and patient management. It is a critical gateway for surgeons aspiring to specialize in gastrointestinal surgery or surgical oncology (7). The GSFE, with its diverse and challenging content, provides a unique opportunity to analyze AI performance in real-world, high-stakes medical examinations.

Although ChatGPT's capabilities have been evaluated in some medical examinations, performance differences related to language and examination content have been highlighted as significant factors. Studies have shown that ChatGPT often performs well in English-language exams, reflecting the dominance of English in its training data (8-11). However, performance discrepancies are observed in non-English languages, such as Turkish, which are underrepresented in training datasets (12). This raises important questions about the effectiveness of AI in multilingual and culturally specific contexts.

This study is among the first to compare the performance of two versions of ChatGPT (3.5 and 4.0) on the GSFE in

both Turkish and English. It aims to analyze the impact of language and model versions on exam accuracy, offering insights into the potential and limitations of ChatGPT in medical education. The findings of this study are expected to contribute not only to understanding the performance of language models in medical examinations but also to improving AI applications for low-resource languages like Turkish. These results may pave the way for enhancing the role of AI in global medical education and professional assessments.

Methods

Study Design

This observational study evaluated the performance of ChatGPT-3.5 and ChatGPT-4 on multiple-choice questions (MCQs) from the GSFE in Turkey. 255 retired and publicly available questions from the 2011-2022 examinations were analyzed.

Bot Evaluation:

ChatGPT-3.5 and ChatGPT-4 were prompted to answer the questions as if they were general surgeons. A standardized instruction was used for consistency across all inputs: "As healthcare professionals, we plan to use you for research purposes. Assuming that you are a general surgeon, answer the following five-choice multiple-choice question by selecting the correct option." (Figure 1) ChatGPT applications were obtained from online stores providing services in Turkey. ChatGPT paid applications were used in the research. A new ChatGPT window was used for each new question asked. In this way, the effects of contextual memory on the answers to the questions were tried to be minimized. Different versions of ChatGPT were used on two different computers using WINDOWS 11 as the operating system.

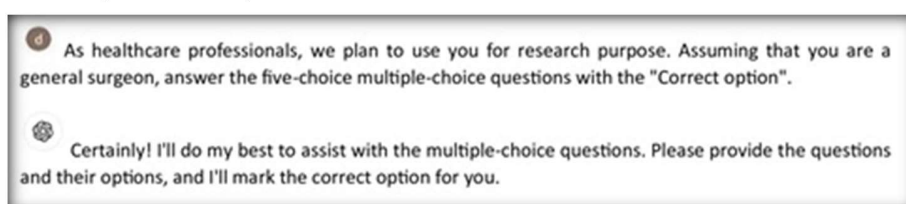
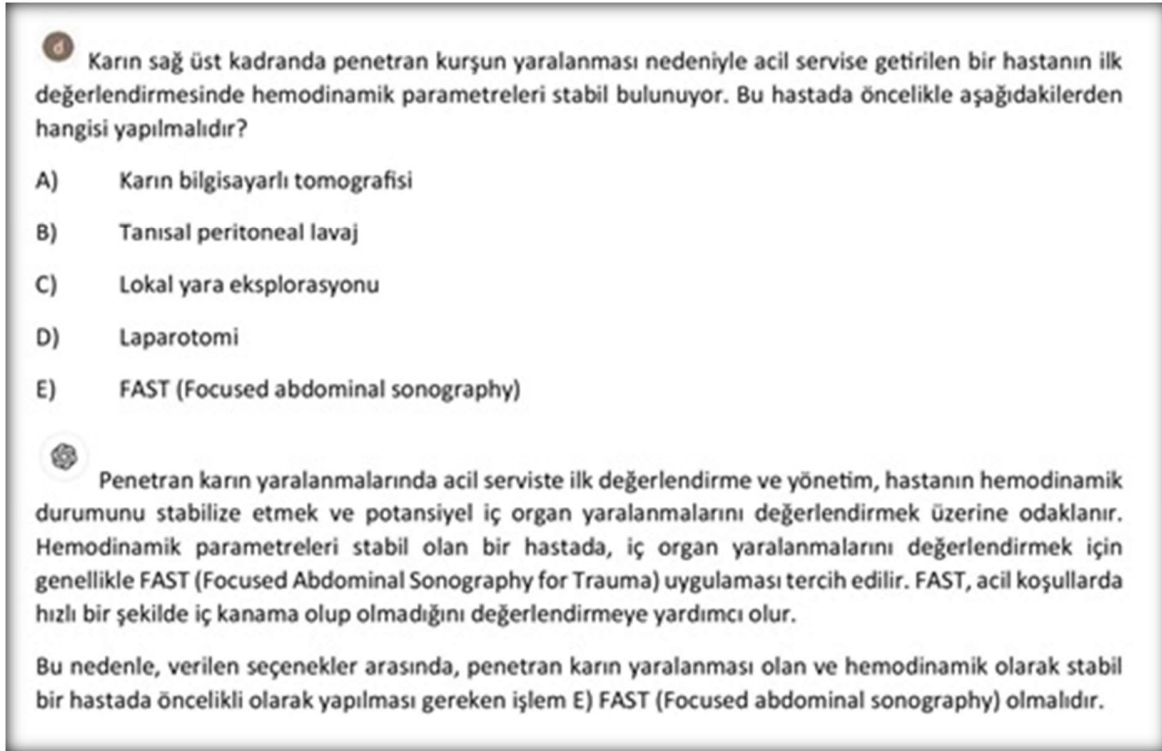


Figure 1: Invitation to bots.

Dataset Preparation and Translation:

The GSFE dataset included diagnosis, surgical treatment, and patient management questions. Questions were first

inputted into the bots in Turkish. All questions were then translated into English by a native speaker. (Figure 2-5)



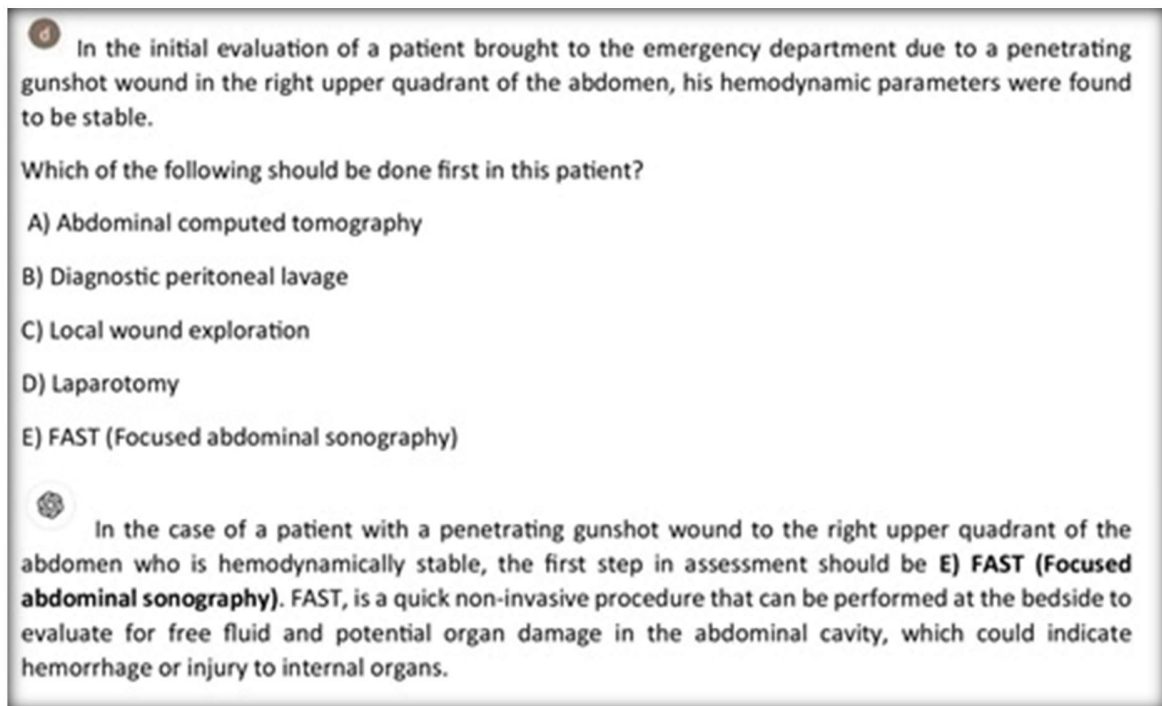
d Karın sağ üst kadranda penetran kurşun yaralanması nedeniyle acil servise getirilen bir hastanın ilk değerlendirmesinde hemodinamik parametreleri stabil bulunuyor. Bu hastada öncelikle aşağıdakilerden hangisi yapılmalıdır?

A) Karın bilgisayarlı tomografisi
B) Tanısal peritoneal lavaj
C) Lokal yara eksplorasyonu
D) Laparotomi
E) FAST (Focused abdominal sonography)

e Penetran karın yaralanmalarında acil serviste ilk değerlendirme ve yönetim, hastanın hemodinamik durumunu stabilize etmek ve potansiyel iç organ yaralanmalarını değerlendirmek üzerine odaklanır. Hemodinamik parametreleri stabil olan bir hastada, iç organ yaralanmalarını değerlendirmek için genellikle FAST (Focused Abdominal Sonography for Trauma) uygulaması tercih edilir. FAST, acil koşullarda hızlı bir şekilde iç kanama olup olmadığını değerlendirmeye yardımcı olur.

Bu nedenle, verilen seçenekler arasında, penetran karın yaralanması olan ve hemodinamik olarak stabil bir hastada öncelikli olarak yapılması gereken işlem E) FAST (Focused abdominal sonography) olmalıdır.

Figure 2: ChatGPT-3.5 with Turkish question inputted.



d In the initial evaluation of a patient brought to the emergency department due to a penetrating gunshot wound in the right upper quadrant of the abdomen, his hemodynamic parameters were found to be stable.

Which of the following should be done first in this patient?

A) Abdominal computed tomography
B) Diagnostic peritoneal lavage
C) Local wound exploration
D) Laparotomy
E) FAST (Focused abdominal sonography)

e In the case of a patient with a penetrating gunshot wound to the right upper quadrant of the abdomen who is hemodynamically stable, the first step in assessment should be **E) FAST (Focused abdominal sonography)**. FAST, is a quick non-invasive procedure that can be performed at the bedside to evaluate for free fluid and potential organ damage in the abdominal cavity, which could indicate hemorrhage or injury to internal organs.

Figure 3: ChatGPT-4 with English question inputted.

d In the initial evaluation of a patient brought to the emergency department due to a penetrating gunshot wound in the right upper quadrant of the abdomen, his hemodynamic parameters were found to be stable.

Which of the following should be done first in this patient?

- A) Abdominal computed tomography
- B) Diagnostic peritoneal lavage
- C) Local wound exploration
- D) Laparotomy
- E) FAST (Focused abdominal sonography)

e In the initial evaluation of a patient with a penetrating gunshot wound to the abdomen who is hemodynamically stable, the appropriate first step is to assess for internal injuries or bleeding using a non-invasive method. Given the stability of the patient's hemodynamic parameters, the use of Focused Assessment with Sonography for Trauma (FAST) is the recommended initial diagnostic tool.

FAST (Focused Abdominal Sonography for Trauma) is a rapid, non-invasive bedside ultrasound examination used to detect free fluid (indicative of bleeding) in the abdominal cavity. This test helps to quickly determine whether there is a need for further diagnostic procedures or immediate surgical intervention.

Here's why the other options are less suitable as the initial step:

Abdominal computed tomography (CT): While CT is a highly detailed imaging method and can provide comprehensive information about internal injuries, it is typically used after initial assessment with FAST if there is a need for more detailed evaluation. It is not usually the first step in hemodynamically stable trauma patients due to its time requirement and potential delays.

Diagnostic peritoneal lavage (DPL): This is an invasive procedure that has become less common with the advent of FAST. It was historically used to detect intra-abdominal bleeding but is now generally reserved for cases where FAST is inconclusive.

Local wound exploration: This procedure is typically reserved for cases where there is clear indication of injury that requires direct assessment, which would be determined after initial imaging and assessment.

Laparotomy: This is an invasive surgical procedure that is usually indicated if the patient shows signs of significant internal bleeding or instability, or if diagnostic imaging suggests major organ injuries. It is not the first step in stable patients.

Therefore, for a stable patient with a penetrating abdominal wound, the first step should be:

E) FAST (Focused abdominal sonography)

Figure 4: ChatGPT-3.5 with English question inputted.

d Karın sağ üst kadranda penetran kurşun yaralanması nedeniyle acil servise getirilen bir hastanın ilk değerlendirmesinde hemodinamik parametreleri stabil bulunuyor. Bu hastada öncelikle aşağıdakilerden hangisi yapılmalıdır?

A) Karın bilgisayarlı tomografisi
B) Tanısal peritoneal lavaj
C) Lokal yara eksplorasyonu
D) Laparotomi
E) FAST (Focused abdominal sonography)

e Hastanın ilk değerlendirilmesinde hemodinamik olarak stabil olduğu ve karın bölgesinde penetran bir mevcut yaralanma olduğu bildirilmiştir. Bu tür bir durumda öncelikle yaralanmanın yaygınlığını ve iç organlara etkisini anlamak için non-invaziv ve hızlı bir yöntem olan FAST (Focused Abdominal Sonography for Trauma) uygulanmalıdır. FAST, iç organ yaralanmaları, serbest sıvı ve kanama varlığını hızla tespit etmek için kullanılır ve acil müdahale gerekliliğini belirlemede yardımcı olur.

Bu nedenle doğru seçenek;

E) FAST (Focused Abdominal Sonography for Trauma)

Figure 5: ChatGPT-4 with Turkish question inputted.

Outcome Measures:

The primary outcome was accuracy, the percentage of correct answers compared to the answer key. Responses

were also reviewed for linguistic errors and other inaccuracies. (Figure 6,7)

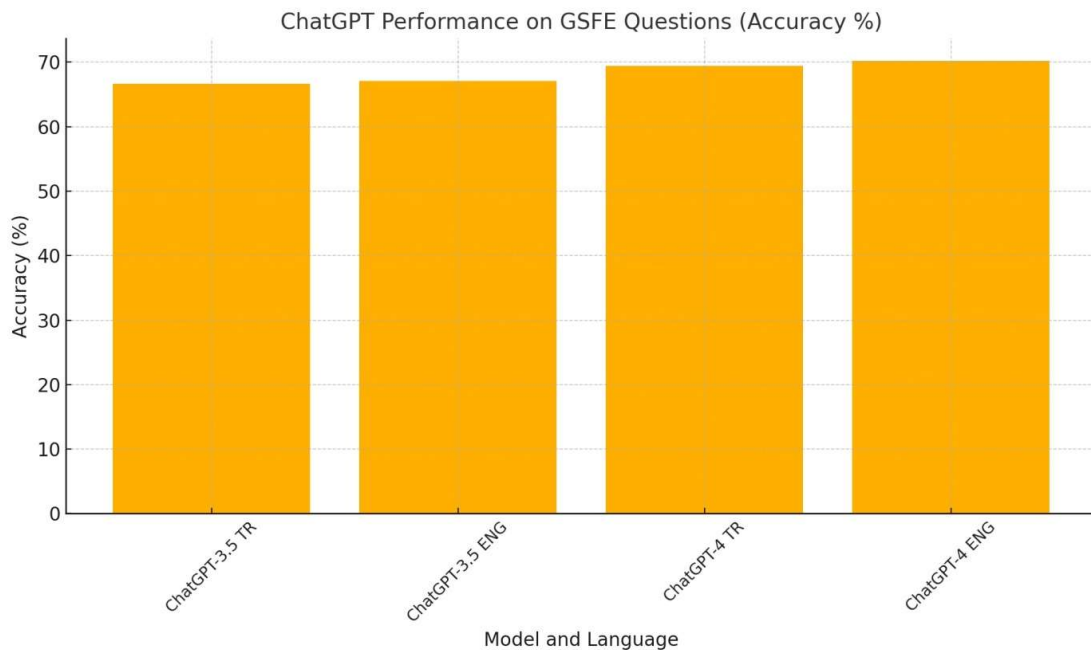


Figure 6: Comparison of bots accuracy rates (%)

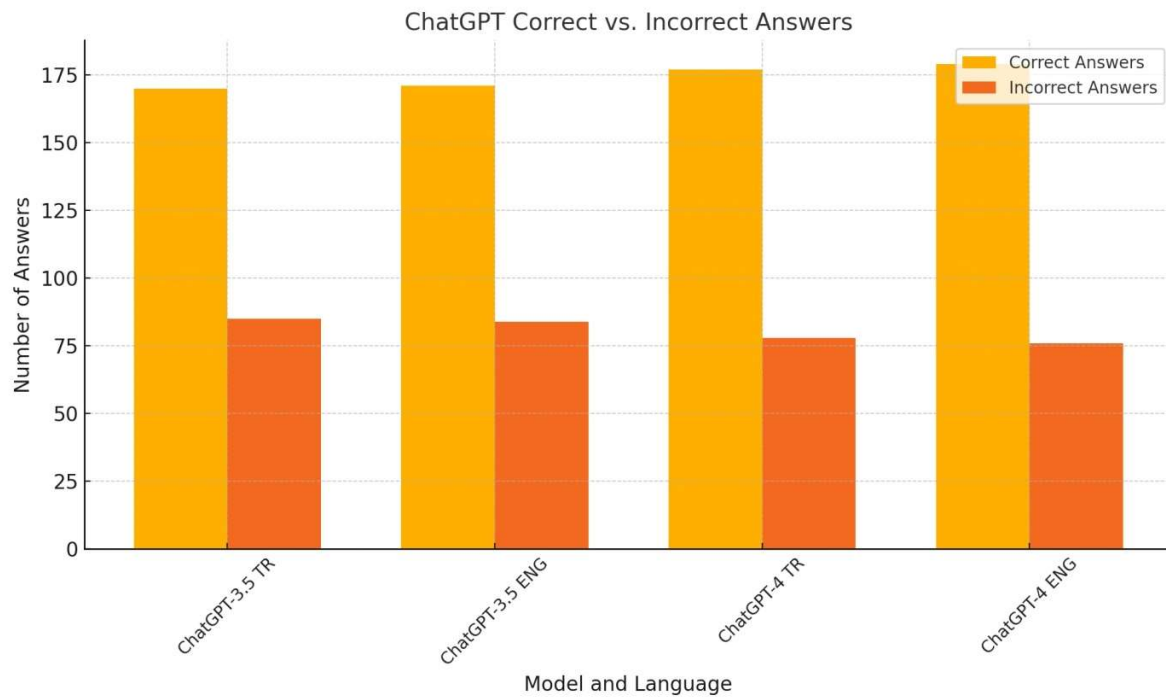


Figure 7: Number of correct and incorrect answers of bots

Statistical Analysis:

The accuracy of the responses was analyzed using Pearson χ^2 tests to compare performance between bot versions (3.5 vs. 4.0) and between languages (Turkish vs. English). Pairwise consistency between different versions of ChatGPT was calculated with Cohen's Kappa test. Consistency between four different versions was calculated with Fleiss' Kappa test. Statistical significance was set at $p < 0.05$.

Results

The performance of ChatGPT-3.5 and ChatGPT-4 was evaluated using 255 MCQs from the GSFE in Turkish and English.

For the Turkish questions, ChatGPT-3.5 correctly answered 170 out of 255 questions, achieving an accuracy of 66.66%. ChatGPT-4 answered 177 questions correctly, with an

accuracy of 69.41%. The difference in performance between ChatGPT-3.5 and ChatGPT-4 for Turkish questions was statistically significant ($p < 0.05$), indicating that ChatGPT-4 outperformed ChatGPT-3.5 in this language.

For the English questions, ChatGPT-3.5 achieved slightly higher accuracy than in Turkish, correctly answering 171 out of 255 questions (67.05%). ChatGPT-4 answered 179 questions correctly in English, with an accuracy of 70.19%. Similar to the Turkish questions, ChatGPT-4 significantly outperformed ChatGPT-3.5 for English questions ($p < 0.05$).

When comparing the same version of ChatGPT across languages (Turkish vs. English), both ChatGPT-3.5 and ChatGPT-4 showed slightly higher accuracy in English. However, these differences were not statistically significant ($p > 0.05$), suggesting that language did not have a substantial impact on the performance of either version (Table 1).

Table 1: ChatGPT performance results

Model	n	Correct Answers	Incorrect Answers	Accuracy (%)
ChatGPT-3.5 TR	255	170	85	66.66%
ChatGPT-3.5 ENG	255	171	84	67.05%
ChatGPT-4 TR	255	177	78	69.41%
ChatGPT-4 ENG	255	179	76	70.19%

ChatGPT-3.5TR: ChatGPT-3.5 with Turkish questions inputted, ChatGPT-4 TR: ChatGPT-4 with Turkish questions inputted, ChatGPT-3.5 ENG: ChatGPT-3.5 with English questions inputted, ChatGPT-4 ENG: ChatGPT-4 with English questions inputted

It was found that Cohen’s Kappa number was quite high similarly high agreement (Fleiss’ Kappa = 0.95, $p < 0.001$) between pairwise comparisons between applications, and (Table 2). the consistency among the four applications showed

Table 2: Cohen’s Kappa and Fleiss’ Kappa values among the used bots

	Kappa	95% CI	p
ChatGPT-3.5 Turkish vs ChatGPT-4 Turkish	0.94	0.89-0.97	<0.001
ChatGPT-3.5 English vs ChatGPT-4 English	0.93	0.87-0.97	<0.001
ChatGPT-3.5 Turkish vs ChatGPT-3.5 English	0.99	0.97-1.0	<0.001
ChatGPT-4 Turkish vs ChatGPT-4 English	0.98	0.95-1.0	<0.001
Fleiss’ Kappa	0.95	0.93-0.96	<0.001

Overall, both ChatGPT versions exceeded the minimum success threshold for the GSFE exam, set at 65%, demonstrating their potential as tools for medical education. ChatGPT-4 consistently outperformed ChatGPT-3.5, reflecting advancements in the model’s accuracy and problem-solving abilities. Additional analysis revealed that the incorrect answers were more common in complex, case-based questions requiring multi-step reasoning and numerical calculations, such as dosage adjustments or laboratory value interpretations. These challenges were observed across

both languages and versions.

In conclusion, ChatGPT-4 exhibited better performance compared to ChatGPT-3.5, with notable improvements in both Turkish and English. Although the models performed slightly better in English, the differences were not statistically significant, indicating reliable performance in both languages. These findings underscore the potential utility of ChatGPT in medical education while highlighting areas where further improvements could enhance its applicability.

Discussion

This study demonstrates the ability of ChatGPT models to perform satisfactorily on the GSFE, with accuracy rates exceeding 65% and surpassing the minimum passing threshold. ChatGPT-4 consistently outperformed ChatGPT-3.5 in both Turkish and English questions, highlighting the advancements achieved with the newer version. These findings align with prior studies evaluating ChatGPT's performance on medical exams in English, such as the USMLE, where its accuracy rates varied but were generally satisfactory (8,9). However, the performance in non-English languages, such as Turkish, remains a notable area for improvement due to the limited representation of such languages in its training data (12). It has been observed that ChatGPT bots pass the exam in some native languages while performing poorly in others (13-15).

The observed discrepancies between Turkish and English accuracy rates, though statistically insignificant, suggest the model's robustness across languages. However, the slightly higher performance in English indicates that ChatGPT's capabilities are influenced by the prevalence of English training data. Expanding the dataset with Turkish medical literature and context-specific clinical information could improve performance in underrepresented languages. Such efforts are critical to ensure the equitable utility of AI tools in diverse linguistic and cultural settings.

ChatGPT's performance highlights its potential as a tool for medical education. With accuracy rates above 65%, it can assist students in preparing for examinations, offering immediate feedback and enhancing their understanding of complex medical concepts. Educators could leverage ChatGPT to create diverse question banks, evaluate curriculum effectiveness, and simulate realistic exam scenarios. Moreover, its interactive capabilities enable its use as a personalized tutor, explaining intricate medical topics in an accessible manner, which could significantly enhance self-directed learning.

To enhance the applicability of ChatGPT in medical

education and assessments, several future directions are recommended. First, incorporating more diverse and region-specific datasets, particularly in Turkish and other underrepresented languages, could address existing disparities. Second, exploring its performance across different medical specialties, such as cardiology or neurology, would broaden its scope of utility. Third, investigating its potential in open-ended question formats and interactive learning environments could provide insights into its adaptability in medical training programs. Lastly, fine-tuning the model to improve its reasoning and numerical calculation capabilities would enhance its effectiveness in clinical applications.

Limitations of the Study

This study has some limitations. Querying a separate computer for each separate bot version and language could have reduced contextual memory-related problems. The online purchase of the ChatGPT version from Turkey may have also led to a lack of access to resources for artificial intelligence. Additionally, we cannot comment on whether the AI could reach a Turkish answer to a question asked in English.

Despite these promising applications, ChatGPT exhibited limitations. Questions requiring multi-step reasoning or numerical calculations, such as dosage adjustments and laboratory interpretations, were challenging for both versions. Additionally, contextual understanding was occasionally insufficient, particularly for nuanced clinical scenarios. These findings are consistent with prior research that noted similar challenges in language model performance on complex tasks.

This study used retired and publicly available GSFE questions, which may not fully reflect the current level of difficulty or scope of the examination. Future evaluations should include newly designed or real-time exam questions to assess AI performance more accurately. Furthermore, the focus on multiple-choice questions limits the generalizability of the findings to other exam formats, such as open-ended or essay-style questions, which may demand

higher levels of reasoning and contextual comprehension.

Conclusion

This study highlights the potential of ChatGPT as a valuable tool in medical education and examination preparation. ChatGPT-3.5 and ChatGPT-4 demonstrated satisfactory performance on the GSFE, with accuracy rates exceeding 65% and surpassing the minimum passing threshold. ChatGPT-4's superior performance across Turkish and English questions reflects the advancements in AI model development, particularly in understanding and solving complex medical scenarios.

While the study underscores ChatGPT's robust multilingual capabilities, it also reveals areas for improvement, especially in handling multi-step reasoning, numerical calculations, and nuanced clinical scenarios. Addressing these limitations, particularly through the inclusion of more diverse and region-specific training data, could enhance ChatGPT's applicability in non-English languages and specialty-specific domains.

ChatGPT's interactive and dynamic capabilities make it an appealing resource for medical students and educators, offering opportunities to create realistic mock exams, explain complex concepts, and provide instant feedback. However, further research is required to evaluate its performance in open-ended question formats, explore its utility in various medical specialties, and optimize its reasoning capabilities.

In conclusion, ChatGPT represents a promising advancement in AI-driven medical education and assessment. With continued refinement and adaptation, it has the potential to transform traditional learning and evaluation methods, bridging gaps in accessibility and providing a globally impactful resource for medical professionals and students alike.

References

1. Park YJ, Jerng SE, Yoon S, Li J. 1.5 million materials narratives generated by chatbots. *Sci Data*. 2024;11:1060.
2. Datt M, Sharma H, Aggarwal N, Sharma S. Role of ChatGPT-4 for Medical Researchers. *Ann Biomed Eng*. 2024;52:1534-1536.

3. Touma NJ, Caterini J, Liblk K. Is ChatGPT ready for primetime? Performance of artificial intelligence on a simulated Canadian urology board exam. *Can Urol Assoc J*. 2024;18:329-332.
4. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer*. 2018;18:500-510.
5. Talyshinskii A, Juliebø-Jones P, Zeeshan Hameed BM, Naik N, Adhikari K, Zhanbyrbekuly U, et al. ChatGPT as a Clinical Decision Maker for Urolithiasis: Compliance with the Current European Association of Urology Guidelines. *Eur Urol Open Sci*. 2024;69:51-62.
6. Gertz RJ, Bunck AC, Lennartz S, Dratsch T, Iuga AI, Maintz D, et al. GPT-4 for Automated Determination of Radiological Study and Protocol based on Radiology Request Forms: A Feasibility Study. *Radiology*. 2023;307:e230877.
7. ÖSYM [Measurement, Selection, and Placement Center]. 2024-YDUS guide and application information. Available from: <https://www.osym.gov.tr/TR,29307/2024-ydus-1-donem-kilavuz-ve-basvuru-bilgileri.html>.
8. Sumbal A, Sumbal R, Amir A. Can ChatGPT-3.5 Pass a Medical Exam? A Systematic Review of ChatGPT's Performance in Academic Testing. *J Med Educ Curric Dev*. 2024;11:23821205241238641.
9. 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.
10. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations. *Radiology*. 2023;307:e230582.
11. Giannos P. Evaluating the limits of AI in medical specialization: ChatGPT's performance on the UK Neurology Specialty Certificate Examination. *BMJ Neurol Open*. 2023;5:e000451.
12. Seghier ML. ChatGPT: not all languages are equal. *Nature*. 2023;615:216.
13. Ngo A, Gupta S, Perrine O, Reddy R, Ershadi S, Remick D. ChatGPT 3.5 fails to write appropriate multiple choice practice exam questions. *Acad Pathol*. 2023;11:100099.
14. Wójcik S, Rulkiewicz A, Pruszczyk P, Lisik W, Poboży M, Domienik-Kartowicz J. Reshaping medical education: Performance of ChatGPT on a PES medical examination. *Cardiol J*. 2024;31:442-450.
15. 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.