TURKISH JOURNAL OF EMERGENCY MEDICINE

Original Article

Access this article online

Quick Response Code:

Website: https://journals.lww.com/TJEM

DOI: 10.4103/tjem.tjem_262_25

Evaluation of ChatGPT's performance on emergency medicine board examination questions

Mustafa Can Guzelce[1]*, Sefer Ozgur[1], Ilker Salli[2]

[1]Department of Emergency Medicine, Izmir University of Economics Medicalpoint Hospital, Izmir, Türkiye
[2]Department of Emergency Medicine, Izmir Tepecik Training and Research Hospital, Izmir, Türkiye
**Corresponding author*

Abstract:

OBJECTIVES: We aimed to evaluate the performance of a large language model (ChatGPT) in answering official sample questions from the Turkish Board of Emergency Medicine (TBEM). Two versions of the model, GPT-4 and GPT-4o, were assessed to explore consistency and accuracy across iterations.

METHODS: A cross-sectional observational study was conducted using 25 standardized multiple-choice questions publicly released by TBEM. Each question was manually entered into GPT-4 and GPT-4o through the OpenAI interface. Both models were prompted to select the best single answer from the provided options without additional clarification or training context. Model responses were evaluated for accuracy, consistency upon repetition, and domain-specific error types. This study is compliant with the STROBE statement and the MedinAI reporting guidelines.

RESULTS: GPT-4 correctly answered 20 out of 25 questions (80%) on the first attempt. On repetition, its score improved to 84%. GPT-4o also achieved a score of 88% (22/25) on its first attempt and showed consistent results upon a second evaluation, providing identical answers in both trials. Errors occurred in the domains of trauma during pregnancy, pediatric resuscitation, and adult resuscitation protocols. Both models demonstrated strong performance in fact-based domains and in questions involving image descriptions.

CONCLUSION: GPT-4 and GPT-4o performed above the TBEM passing threshold, showing solid accuracy across a range of emergency medicine topics. Both excelled in fact-based and image-related questions. However, they showed limitations in clinical reasoning, particularly in scenarios requiring nuanced judgment. These tools may support examination preparation but should not replace the expertise of trained emergency physicians.

Keywords:

Artificial intelligence, board examination, ChatGPT, emergency medicine, GPT-4, large language models, Turkish Board of Emergency Medicine

Submitted: 19-07-2025
Revised: 31-10-2025
Accepted: 04-11-2025
Published: 03-04-2026

ORCID:
MCG: 0000-0003-4141-2661
SO: 0000-0001-7908-6422
IS: 0000-0002-8954-8421

Address for correspondence:
Dr. Mustafa Can Guzelce, Department of Emergency Medicine, Izmir University of Economics Medicalpoint Hospital, Izmir, Türkiye.
E-mail: mcg@windowslive.com

Introduction

Artificial intelligence (AI) tools are playing a growing role in medical education, especially in preparation for standardized examinations. ChatGPT, developed by OpenAI, is a prominent example of a generative large language model that can interpret questions, synthesize knowledge, and generate complex written responses.[1]

ChatGPT has previously demonstrated high accuracy on the United States Medical Licensing Examination (USMLE)[2] and other board examinations, such as the American Board of Family Medicine (ABFM) and Emergency Medicine examinations. However, it has notable weaknesses,

How to cite this article: Guzelce MC, Ozgur S, Salli I. Evaluation of ChatGPT's performance on emergency medicine board examination questions. Turk J Emerg Med 2026;26:110-5.

Box-ED Section

What is already known on the topic?

- Large language models such as GPT-4 have shown high performance on board examinations like the United States Medical Licensing Examination and American Board of Emergency Medicine. Their performance on country-specific emergency medicine boards, such as the Turkish Board of Emergency Medicine (TBEM), has not been thoroughly assessed.

What this study adds?

- We report the first evaluation of ChatGPT's performance on TBEM examination using GPT-4 and GPT-4o models. GPT-4o demonstrated marginally improved accuracy and enhanced consistency across two trials. However, both models share similar weaknesses in context-sensitive clinical reasoning.

particularly in contextual clinical judgment and procedural decision-making.

The Turkish Board of Emergency Medicine (TBEM) offers a standardized National Board Certification for emergency physicians in Türkiye. Their official sample questions represent core emergency medicine knowledge required in the local setting. To date, no study has evaluated ChatGPT's performance on the TBEM examination. The objective of this study was to provide a clear and systematic evaluation of ChatGPT (GPT-4 and GPT-4o) on TBEM sample questions, focusing on both accuracy and consistency. By doing so, this study contributes novel insights into the potential role of large language models in Turkish emergency medicine education, highlighting opportunities as well as limitations for their integration into examination preparation.

Methods

Study design and question set

This was a cross-sectional performance evaluation study conducted in June 2025 to benchmark the diagnostic accuracy of large language models using standardized board-style questions. All 25 official multiple-choice sample questions published by the TBEM were utilized. This study is compliant with the STROBE statement and the MedinAI[3] reporting guidelines [Supplementary File 1].

The primary outcome measure was defined as the overall accuracy, calculated as the proportion of correctly answered questions out of the total set. Secondary outcomes included consistency on repetition and domain-specific error categorization. The subsequent categorization of model responses by accuracy, consistency, and domain-specific error type served as the study's approach to external explainability and model interpretability, allowing us to infer the models' strengths and limitations in clinical reasoning.

Model evaluation procedure

Each question was manually entered into the OpenAI ChatGPT interface for both GPT-4 (2025 version) and GPT-4o, without any modification to wording or format. No custom prompts or additional instructions were used.

For GPT-4, all questions, including those with visual content, were uploaded with their corresponding images, and no supplementary text descriptions were provided beyond the question itself. Responses were generated based solely on the text descriptions provided in the question stems, as this version does not support image interpretation. In contrast, GPT-4o was tested using its multimodal capabilities, and the original image files (e.g., electrocardiograms [ECGs], radiologic images) were uploaded into the interface, allowing the model to directly interpret visual content.

We prompted each model to choose the single best answer from the five given options. No additional context, guidance, or external materials were provided. This study was conducted independently of OpenAI, which had no role in the study design, analysis, or reporting.

All evaluations were performed using the official ChatGPT platform developed by OpenAI. No plugins, browsing tools, or external application programming interfaces were enabled. All settings were kept at their default configurations. Each model was tested twice, and identical prompts were used for both rounds to ensure consistency.

Question categorization

The questions were categorized as either factual recall, testing isolated medical knowledge, or scenario-based, requiring clinical reasoning, prioritization, or interpretation.

Statistical analysis

Descriptive statistics were used to summarize the number and percentage of correct responses for each model. Accuracy was calculated as the proportion of correct answers out of 25 questions. Comparative analysis between GPT-4 and GPT-4o was performed using Fisher's exact test. $P < 0.05$ was considered statistically significant. All statistical analyses were conducted using IBM SPSS Statistics for Windows, Version 28.0 (IBM Corp., Armonk, NY, USA).

Ethical approval

This study was based solely on publicly available

official sample questions provided by the TBEM. It did not involve human participants, patient data, or any clinical interventions. As such, approval from an ethics committee or institutional review board was not required.

Results

GPT-4 correctly answered 20 of 25 questions (80%) on the first attempt and improved to 21 (84%) on repetition. GPT-4o achieved 88% on its first attempt and repeated the same score on the second evaluation, demonstrating answer consistency.

Of the 25 items, GPT-4 answered 14 out of 16 factual questions correctly (87.5%), while GPT-4o answered 15/16 factual questions (93.8%) and 6/9 scenario questions (66.7%), suggesting slightly stronger factual recall by GPT-4o and identical reasoning performance across both models.

Both models answered all four image-based questions correctly.

GPT-4 made errors in the following domains: pediatric resuscitation, adult resuscitation protocols, nephrology, acid–base disturbances, and trauma in pregnancy, each represented by one incorrect response. GPT-4o made only three errors in total, which occurred in the same categories: pediatric resuscitation, trauma in pregnancy, and adult resuscitation protocols.

GPT-4 revised one of its earlier mistakes on the second run, indicating some answer variability. GPT-4o demonstrated consistency by providing identical answers across both trials.

GPT-4o demonstrated a lower mean error rate (12%) compared to GPT-4 (20%), although this difference was not statistically significant (P = 0.4508), likely due to the small sample size. Most errors occurred in domains requiring nuanced clinical reasoning, which were pediatric resuscitation, trauma in pregnancy, and resuscitation protocol questions. GPT-4o also maintained consistent accuracy across repeated trials, whereas GPT-4 showed some variability in its responses [Figure 1].

Discussion

GPT-4o's 88% accuracy was above the TBEM passing threshold within this limited sample and should be interpreted as a preliminary, hypothesis-generating finding. This aligns with performance levels reported in prior studies, including Pastrak *et al*., who demonstrated that GPT-4 achieved a high level of accuracy on emergency medicine board-style questions.[4] Our study contributes to the growing body of research on large language model utility in medical education by comparing GPT-4 and GPT-4o using localized, domain-specific examination questions. Given the restricted question pool, conclusions are not definitive, but the results provide exploratory insights into the potential of these models.

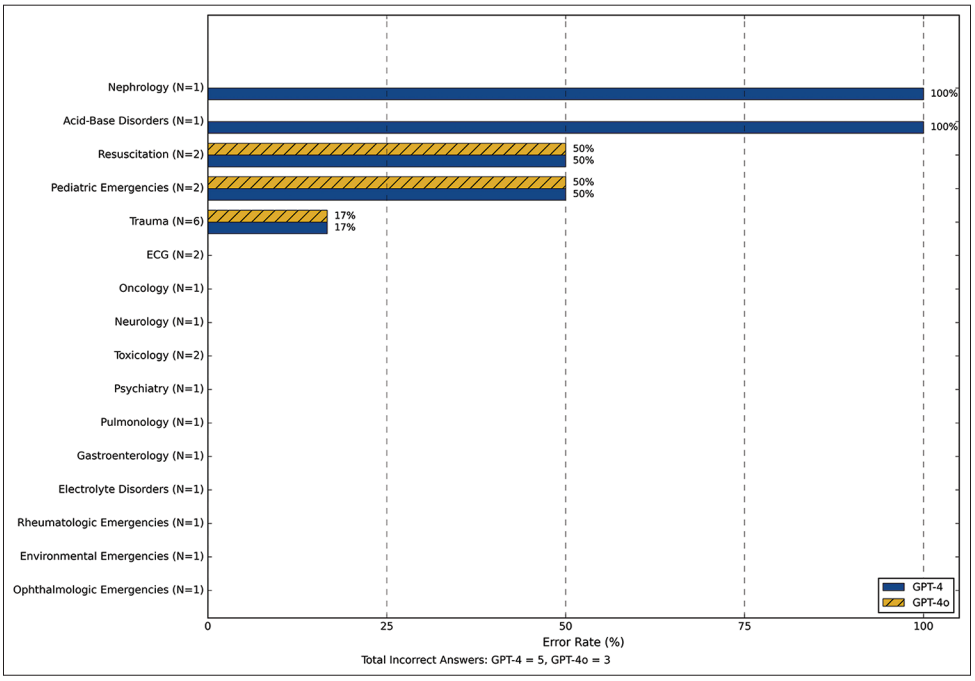

Figure 1: Comparison of GPT-4 and GPT-4o error rates across emergency medicine topics. Topics are labeled with the number of questions (N). Error rates represent the proportion of incorrect responses per topic. Errors in nephrology and acid-base disorders were knowledge-based, while errors in resuscitation, pediatric emergencies, and trauma were scenario-based

While overall accuracy was similar, GPT-4o showed a subtle edge in factual recall (93.8% vs. 87.5%) and zero response variability. Its limitations emerged in clinical reasoning domains, particularly trauma management during pregnancy and pediatric dosing. This pattern aligns with broader findings from AI research in medical education, where contextually nuanced decision-making continues to challenge current model architectures.

Our results also align with the findings of Gilson *et al*.,[2] who demonstrated that GPT-4 performed at or above the passing threshold on all three steps of the USMLE, often outperforming GPT-3.5 by a substantial margin. That study also emphasized GPT-4's improved reasoning and consistency, reinforcing its suitability as an adjunctive tool in medical education.

Our findings are consistent with those of Lee *et al*.,[5] who assessed GPT models on a Korean emergency medicine board question bank. In that study, GPT-4 achieved an accuracy rate of 75.6%, significantly higher than GPT-3.5 (56.9%), and demonstrated strong performance even on higher-order clinical reasoning questions. This parallels our observation of GPT-4 and GPT-4o attaining passing-level accuracy on TBEM questions, particularly in factual and structured scenarios. However, both studies noted reduced performance in nuanced, protocol-driven cases.

Although all image-based questions were answered correctly by both models, the mechanisms behind their success differed. GPT-4 relied solely on textual descriptions embedded in question stems to simulate visual interpretation. In contrast, GPT-4o directly processed the original images using its multimodal architecture. This distinction is notable considering the findings by Brin *et al*.,[6] who reported that GPT-4V achieved over 80% accuracy in radiological image analysis, correctly identifying anatomical structures and detecting abnormalities in most cases. These findings support the growing relevance of multimodal AI in domains that combine text and visuals, such as emergency medicine board examinations, where ECGs, X-rays, and magnetic resonance imagings frequently appear.

A recent study from Qatar adds further perspective to our findings. Iftikhar *et al*.[7] examined ChatGPT's performance on emergency medicine residency examinations and found that while the model performed reasonably well overall, it tended to falter in questions that required clinical judgment, prioritization, or ethical reasoning. Our results are consistent with that study, particularly in areas such as trauma and pediatric care, where nuanced decision-making played a larger role. The authors concluded that ChatGPT could be a helpful tool for examination preparation, especially for reinforcing core knowledge, but emphasized that its use should remain firmly within the bounds of guided learning. Like in our study, the key takeaway is not that these tools can replace clinical expertise but that they may have a growing place in how we approach medical education.

Our findings also resonate with recent work by Wei,[8] who prospectively compared several large language models on radiology board examinations and found that GPT-4 achieved the highest accuracy (83.3%), significantly outperforming alternative models such as Claude, Bard, and Gemini Pro. Importantly, their study demonstrated that GPT-4 maintained strong performance across both lower-order and higher-order cognitive tasks, while other models showed marked declines in complex diagnostic reasoning. This finding is consistent with our observation that, despite strong factual recall, performance in nuanced clinical reasoning particularly in domains such as pediatric resuscitation and trauma protocols remains a limitation of current models. Together, these results reinforce the notion that GPT-4's relative advantage is consistent across specialties, but they also underscore the ongoing need for domain-specific training datasets and careful integration of AI tools into specialty-specific education.

Our observations are also in line with the findings by Chen *et al*.,[9] who assessed ChatGPT's performance on neurology practice questions and found accuracy rates between ~65.8% (first attempt) and ~75.3% (over multiple attempts). Their study illustrated that ChatGPT's performance improved with repeated querying and that performance varied considerably across subspecialties, especially in areas requiring interpretive reasoning or specialized knowledge. These patterns complement our results, which showed strong factual recall but limitations in nuanced clinical reasoning tasks. Together, such evidence underscores that although large language models show promise in standardized test settings, their reliability in more complex, domain-specific clinical reasoning remains to be refined.

Similarly, Goodings *et al*.[10] examined ChatGPT-4's performance on the ABFM examination and reported accuracy rates of 87%–89% across different configurations, with no significant differences between the standard and customized versions. Their findings reinforce the notion that ChatGPT-4 can reach or surpass the passing threshold in board-style examinations across specialties, even without domain-specific fine-tuning. This consistency across contexts supports our observation that large language models demonstrate reliable factual recall,

though clinical reasoning and nuanced decision-making remain areas of challenge.

In cardiology, Huwiler *et al*.[11] conducted an experimental assessment of AI-powered chatbots on a multiple-choice cardiology board examination and found that while all cardiology fellows achieved near-perfect scores, most chatbots performed well below the passing threshold. Only Jasper Quality and ChatGPT-4 reached passing-level accuracy, with median top-1 accuracies for other chatbots around 47%. Their study highlights the considerable gap that remains between human expertise and general-purpose chatbots in specialized board examinations, underscoring that even advanced models may require fine-tuning and domain-specific adaptation to achieve reliable performance in high-stakes medical assessments.

For learners, these models can serve as helpful tools for self-testing and review ahead of high-stakes examinations. Their strength in factual domains and image interpretation supports their educational utility. However, their limited contextual reasoning warrants caution when used without expert oversight.

The question set in this study included a representative sampling of emergency medicine domains, but performance may vary with broader content exposure. Nevertheless, the restricted sample size means findings should be viewed as exploratory, underscoring the need for further research with larger datasets.

Limitations

This study was limited by its use of a small pool of only 25 official TBEM sample questions. While reflective of core TBEM content, it does not cover the full spectrum of the curriculum, and some domains were represented by a single question, limiting detailed subanalysis. As such, the study should be regarded as an exploratory evaluation rather than a definitive assessment. The models could not interact with actual images, and the prompts did not simulate real-time, multistep clinical reasoning. Finally, the study did not evaluate how these models perform under varying prompt designs or with retrieval-augmented support, which are increasingly relevant in AI-assisted medical learning environments.

Conclusion

GPT-4 and GPT-4o performed above the TBEM passing threshold within the limited pool of available questions and demonstrated strong accuracy across a representative sample of emergency medicine board items. Both models excelled in fact-based content and image interpretation tasks, with GPT-4o showing slightly greater consistency. They exhibited limitations in clinical reasoning, particularly in scenarios requiring nuanced judgment, such as pediatric resuscitation and trauma protocols. These results support the use of large language models as supplementary tools in examination preparation but reinforce that these cannot substitute for the clinical expertise and decision-making skills of trained emergency physicians.

Acknowledgments

We thank the TBEM for providing access and permission to use their official sample questions.

Author contributions statement

- MCG: Conceptualization, methodology, formal analysis, investigation, writing - original draft, review and editing, supervision, and project administration
- SÖ and İŞ: Conceptualization, methodology, investigation, software, resources, data curation, visualization, and writing – review and editing.

Conflicts of interest

None Declared.

Ethical approval

This study used publicly available official sample questions and did not involve human participants, patient data, or interventions. Therefore, ethical committee approval was not required.

Funding

None.

References

1. OpenAI. ChatGPT. San Francisco (CA): OpenAI; 2025. Available from: https://chat.openai.com/. [Last accessed on 2025 Mar 22].
2. 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.
3. Bottacin WE, de Souza TT, Reis WC, Melchiors AC. Guidelines for reporting artificial intelligence studies in medicines, pharmacotherapy, and pharmaceutical services: MedinAI development, validation and statement. Int J Clin Pharm 2025;47:945-56.
4. Pastrak M, Kajitani S, Goodings AJ, Drewek A, LaFree A, Murphy A. Evaluation of ChatGPT performance on emergency medicine board examination questions: Observational Study. JMIR AI 2025;4:e67696.
5. 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.
6. Brin D, Sorin V, Barash Y, Konen E, Glicksberg BS, Nadkarni GN, *et al*. Assessing GPT-4 multimodal performance in radiological image analysis. Eur Radiol 2025;35:1959-65.
7. Iftikhar H, Anjum S, Bhutta ZA, Najam M, Bashir K. Performance of ChatGPT in emergency medicine residency exams in Qatar: A comparative analysis with resident physicians. Qatar Med J 2024;2024:61.
8. Wei B. Performance evaluation and implications of large language models in radiology board exams: Prospective comparative analysis. JMIR Med Educ 2025;11:e64284.

9. Chen TC, Multala E, Kearns P, Delashaw J, Dumont A, Maraganore D, *et al*. Assessment of ChatGPT's performance on neurology written board examination questions. BMJ Neurol Open 2023;5:e000530.
10. Goodings AJ, Kajitani S, Chhor A, Albakri A, Pastrak M, Kodancha M, *et al*. Assessment of ChatGPT-4 in family medicine board examinations using advanced AI learning and analytical methods: Observational study. JMIR Med Educ 2024;10:e56128.
11. Huwiler J, Oechslin L, Biaggi P, Tanner FC, Wyss CA. Experimental assessment of the performance of artificial intelligence in solving multiple-choice board exams in cardiology. Swiss Med Wkly 2024;154:3547.

Supplementary File 1: MedinAI Reporting Checklist

Section	Item	Location in manuscript (Page No.)
1. Study	1.1. Study design	P. 2 (Methods Section)
	1.2. Basal reporting guideline that was used	P. 3 (Cites STROBE)
	1.3. Study location	P. 1, P. 2 (Context of Turkish Board of Emergency Medicine - TBEM)
	1.4. Study timeline	P. 2
	1.5. Study setting	P. 3 (Independent academic evaluation via OpenAI interface)
2. Core of the AI Model	2.1. Stage of the AI development/life cycle	NA
	2.2. Collaborative development (multidisciplinary and interdisciplinary team)	P. 3 (Conducted independently of OpenAI)
	2.3. Classification of the type of high-level algorithm used	P. 1, P. 2 (Large language model)
	2.4. Specification of algorithm methodologies	P. 1, P. 3 (GPT-4 and GPT-4o versions)
	2.5. Performance metrics	P. 1, P. 3 (Accuracy, consistency upon repetition)
	2.6. Predictive analysis capabilities and analysis	P. 3 (Statistical tests used)
	2.7. Publication of source code and models	NA (Proprietary model)
	2.8. Explainability	P. 3 (Via domain-specific error categorization)
	2.9. Domain generalization techniques applied to avoid domain shift	NA
	2.10. Operation in real-time or through batch processing	P. 3 (Manually entered/batch processing)
	2.11. External validation	P. 1 (First evaluation on TBEM)
3. Legal and Regulatory Bias	3.1. Legislative jurisdiction considered during development and year	NA
4. Core of the AI-based Application	4.1. Interoperability and integration with healthcare systems	NA
	4.2. Presence of a user-friendly interface (UI)	P. 3 (OpenAI ChatGPT interface)
	4.3. User experience (UX) tests	NA
5. Data	5.1. Introductory description of data	P. 1, P. 2 (25 standardized questions from TBEM)
	5.2. Statement confirming the use of real-world or synthetic data	P. 2 (Official sample questions)
	5.3. Data acquisition	P. 2 (Utilized 25 official sample questions)
	5.4. Authorization for data use	P. 8 (Acknowledgments)
	5.5. Data privacy and security	P. 3 (No patient data)
	5.6. Data anonymization	NA
	5.7. Data governance practices	P. 2 (Publicly available questions)
	5.8. Data cleaning and preparation	P. 3 (Entered manually as-is)
	5.9. Handling of missing data	NA (Complete set of questions used)
	5.10. Data subsets	P. 3 (Factual vs. scenario-based)
	5.11. Adequate population and/or data representativeness or disclaimer	P. 7 (Disclaimer on small pool/exploratory)
	5.12. Detection of seasonal variations in data	NA
	5.13. Compliance with data protection regulations	P. 3 (No human/patient data involved)
	5.14. Data sharing to promote transparency and reproducibility	P. 2 (Publicly available data from TBEM)
6. Trustworthy Use of AI by Healthcare Professionals	6.1. Responsibilities assigned to the AI and those assigned to the healthcare professionals	P. 1, P. 8 (Supplementary tool, not replacement)
	6.2. Professionals' responsibilities in verifying AI recommendations using other techniques	P. 8 (Implied need for expert oversight)
	6.3. The extent to which professionals should trust AI at its current stage of development	P. 7 (Caution warranted without expert oversight)
	6.4. Considerations on education and training of healthcare professionals in the use of the AI	P. 2, P. 8 (Potential role in education)
7. Model Update, Quality, and Cost-effectiveness	7.1. Handling of model drift/degradation and adaptation to new scientific evidence	P. 1, P. 4 (Comparison of two versions/ consistency testing)
	7.2. Specification of roles and accountability for model maintenance and quality across stakeholders	NA
	7.3. Channel for receiving user and/or patient feedback	NA
	7.4. Cost-effectiveness considerations considering economic, clinical, and humanistic outcomes	NA
	7.5. Continuous monitoring post-implementation	NA
	7.6. Post-marketing surveillance actions	NA

Contd...

Supplementary File 1: Contd...

Section	Item	Location in manuscript (Page No.)
8. Ethics	8.1. Ethics committee/IRB approval or exemption	P. 3 (Not required)
	8.2. Detailed informed consent process	NA
	8.3. Procedure and methodology on promotion of respect for patient autonomy and patient privacy	NA
	8.4. Procedure and methodology for data override or deletion upon patient/organization request	NA
	8.5. Robust testing to prevent unfavorable outcomes influenced by socioeconomic factors, racial and ethnic identity, population groups, vulnerable groups, and genetic variability	P. 7 (Addressed in Limitations/Discussion)
	8.6. Procedure and methodology used to ensure that AI is accessible and beneficial to all groups of people, regardless of their background, to promote fairness and equity (justice allocation of AI utilization)	NA
	8.7. Comprehensive disclose of funding sources, including contributions from industry, government grants, and institutional support	P. 3 (Reported as independent of OpenAI)
9. Domains related to Medicines as Products	9.1 through 9.10	NA
10. Domains related to Services for Medicines and Pharmacotherapy	10.1 through 10.15	NA
11. Domain related to Ethical Considerations in Medication and Pharmacotherapy	11.1 through 11.23	NA
12. Domain related to Risk Prediction and Prognosis	12.1 through 12.10	NA
13. Domain related to the Diagnosis of Diseases	13.1 through 13.11	NA
14. Domain related to Image Analysis	14.1 through 14.15	NA

NA: Not applicable, IRB: Institutional review board, AI: Artificial intelligence