



Access this article online

Quick Response Code:



Website:
<https://www.ovid.com/jnls/tjem>

DOI:
10.4103/tjem.tjem_295_26

Response to: "Evaluation of ChatGPT's performance on emergency medicine board examination questions"

Chee Fah Chong*

Emergency Department, Shin Kong Wu Ho-Su Memorial Hospital, Taipei City, Taiwan

*Corresponding author

Submitted: 13-06-2026

Accepted in Revised

Form: 17-07-2026

Published: 10-10-2026

ORCID:

CFC: 0000-0002-3665-8686

Address for correspondence:

Dr. Chee Fah Chong,
Emergency Department,
Shin Kong Wu Ho-Su
Memorial Hospital, No.
95 Wen-Chang Road,
Shih-Lin District, Taipei
City 111, Taiwan.
E-mail: m002202@
ms.skh.org.tw

Dear Editor-in-Chief,

We read with interest the recent article by Guzelce *et al.* evaluating GPT-4 and GPT-4o on Turkish Board of Emergency Medicine (TBEM) sample questions.^[1] While the study contributes a timely perspective to the growing literature on large language model (LLM) performance in medical education, several methodological concerns warrant consideration before broader interpretations are drawn.

Most critically, the entire evaluation rests on only 25 questions, a sample so small that a single correct or incorrect answer shifts performance by four percentage points. The observed accuracy difference between GPT-4 (80%) and GPT-4o (88%) was unsurprisingly nonsignificant ($P = 0.45$), yet the authors' conclusion that both models "performed above the TBEM passing threshold" risks conveying an unwarranted confidence. With several domains represented by only one question each, the domain-level error analysis presented in Figure 1 is statistically uninformative.

A less-discussed but equally important concern involves the asymmetric testing conditions. GPT-4 processed image-based questions using text descriptions embedded in question stems, while GPT-4o directly

interpreted uploaded images via its multimodal architecture. The fact that both models achieved identical scores on all four image-based questions does not validate GPT-4's image comprehension; it may simply reflect that the text stems already contained sufficient diagnostic information, inadvertently removing the visual challenge entirely. This confound undermines any meaningful comparison between the two architectures on visual reasoning.

Furthermore, only two testing iterations per model were performed. This is insufficient to characterize response stability, especially for GPT-4, which demonstrated answer variability between runs. More rigorous consistency testing, across randomized prompt sequences and temperature settings, would be necessary before drawing conclusions about real-world reliability.

Finally, the study's exclusive use of publicly available official sample questions raises a training data contamination concern that the authors do not address. These questions may have been included in GPT's pretraining corpus, artificially inflating observed accuracy beyond what the models would achieve on unseen examination content.

We commend the authors for addressing a regional gap in the artificial intelligence in medical education literature. However,

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License (CC BY-NC-ND), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

For reprints contact: WKHLRPMedknow_reprints@wolterskluwer.com

How to cite this article: Chong CF. Response to: "Evaluation of ChatGPT's performance on emergency medicine board examination questions". Turk J Emerg Med 2026;26:359-60.

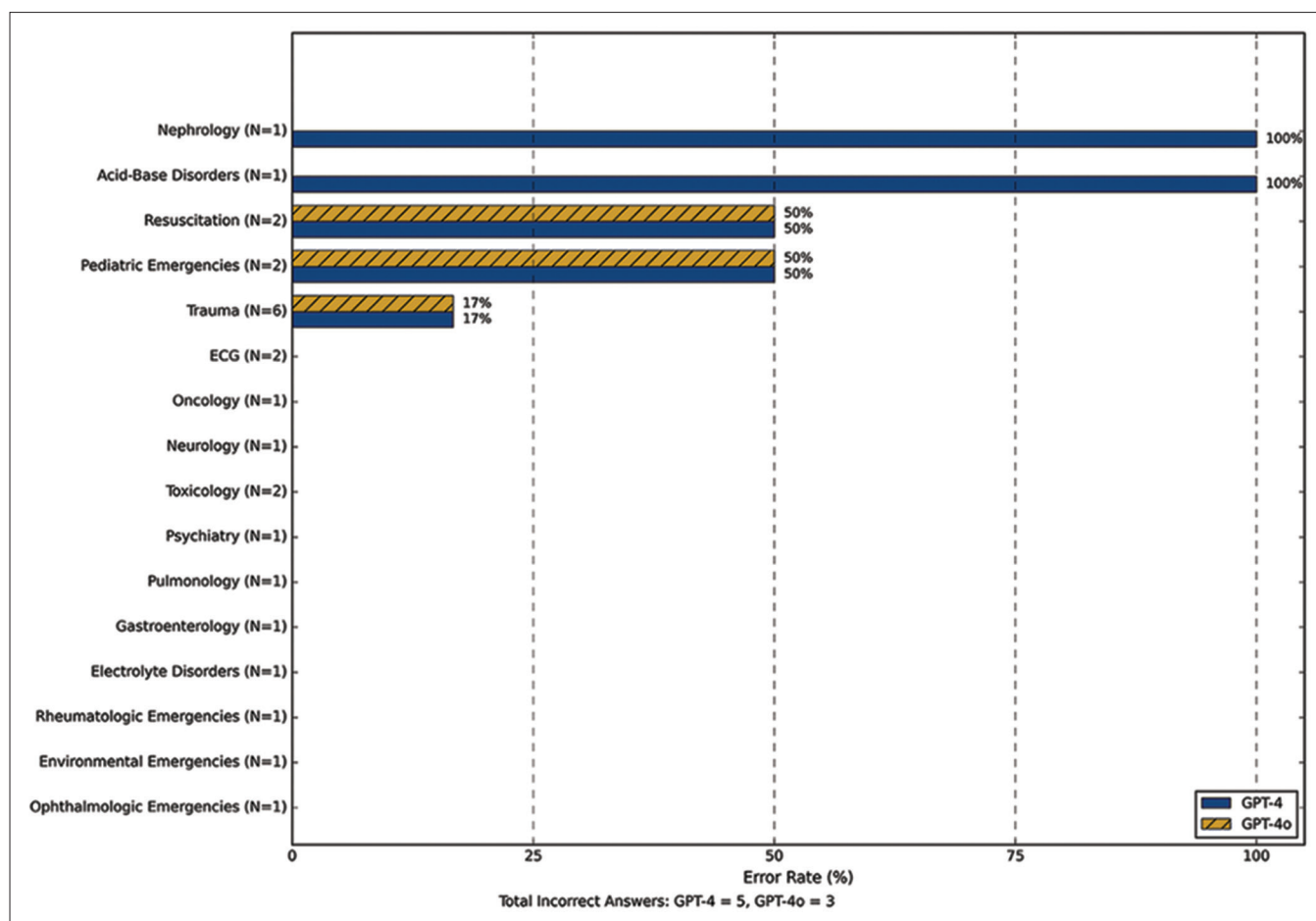


Figure 1: Comparison of GPT-4 and GPT-4o error rates across emergency medicine topics. (Source: *Turkish Journal of Emergency Medicine*, Vol. 26(2), pp. 110–115, 2026)

we urge readers and educators to interpret these findings as preliminary signals rather than performance benchmarks. Future studies should employ larger, stratified, and potentially confidential question sets, standardized prompt protocols, and blinded scoring methodologies to establish robust conclusions about LLM utility in specialty board preparation.

Author contribution

C-FC: Writing the original draft.

Conflict of interest

None Declared.

Funding

None.

Reference

1. 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.