

Response to: *Potential and limitations of large language models in acute chest pain triage*

Arian Zaboli,¹ Gianni Turcato²

¹Innovation, Research and Teaching Service (SABES-ASDAA), Teaching Hospital of the Paracelsus Medical Private University (PMU), Bolzano, Italy; ²Department of Internal Medicine, Intermediate Care Unit, Hospital Alto Vicentino (AULSS-7), Santorso, Italy

Key words: ChatGPT, emergency department, chest pain, MACE, artificial intelligence.

Correspondence: Arian Zaboli Innovation, Research and Teaching Service (SABES-ASDAA), Teaching Hospital of the Paracelsus Medical Private University (PMU), via T. Edison, 10^o; Bolzano, Italy. E-mail: zaboliarian@gmail.com

Dear Editor-in-Chief,

We would like to thank Shinde and colleagues for their thoughtful comments on our article, “Evaluating the predictive accuracy of ChatGPT in risk stratification for chest pain in the emergency department”.^{1,2} We are grateful for their appreciation of both the rationale of the study and the dynamic methodology adopted, in which ChatGPT was evaluated across sequential phases of the emergency department assessment.²

We agree with the authors that Large Language Models (LLMs) may represent an interesting and accessible form of clinical decision support. However, we believe that their use in emergency care requires particular caution. Emergency departments are complex, time-sensitive, high-risk environments, where diagnostic uncertainty, incomplete information, overcrowding, and rapid clinical deterioration are frequent. In this context, even apparently small errors in risk stratification may have relevant consequences for patient safety.¹

A key issue is that much of the current evidence on LLMs in emergency medicine is still based on simulated cases, retrospective scenarios, or highly controlled experimental settings.^{3,4} Although these studies are useful for exploring the potential of these tools, they do not fully reproduce the complexity of real-world emergency care. Clinical practice is not a static exercise in which all relevant information is immediately available; rather, it is a dynamic process in which data are progressively collected, interpreted, questioned, and integrated with clinical judgment.¹ For this reason, future research should move beyond testing LLMs in simulated or idealized contexts and should focus on pragmatic, prospective, and workflow-embedded evaluations.

Our study was designed precisely with this concern in mind. Rather than asking ChatGPT to evaluate isolated clinical vignettes, we assessed its performance using real patients and progressively available clinical information, including history, ECG interpretation, and serial troponin testing. The improvement in predictive accuracy across the three phases suggests that LLMs may benefit from dynamic reassessment.¹ Nevertheless, the persistence of clinically relevant misclassifications, including patients with MACE classified as non-urgent or not at risk, confirms that current LLMs should not be considered reliable stand-alone tools for chest pain risk stratification.¹

We also agree that, in this specific field, traditional machine learning models have shown stronger discriminatory performance than LLMs.⁵ The key challenge, therefore, may not be whether LLMs can replace existing risk prediction systems, but how different forms of Artificial Intelligence (AI) can be made clinically usable, transparent, safe, and integrated into emergency department workflows. From this perspective, it may be more valuable to invest resources in making high-performing machine learning models more interpretable, operationally simple, and integrated with electronic health records, rather than assuming that conversational LLMs are the most appropriate solution for every clinical decision-support problem.

At the same time, LLMs may still have a role, particularly as interfaces that help clinicians interact with complex predictive systems, summarize clinical information, or support structured reassessment. However, this potential role must be carefully distinguished from autonomous clinical reasoning. The intrinsic risks of LLMs, including hallucinations, inconsistency, overconfidence, and non-transparent reasoning processes, are especially relevant in emergency medicine, where patient safety depends on timely and accurate decisions.

Future studies should therefore evaluate not only accuracy metrics, but also clinical impact, safety, explainability, integration with human judgment, and effects on workflow. The most relevant research question is no longer whether LLMs can perform well in simulated scenarios, but how AI systems can be safely interpreted, governed, and used in real emergency care alongside physicians and nurses.

In conclusion, we thank Shinde and colleagues for their constructive contribution.² We share their view that AI may contribute to the future of emergency medicine, but we believe that its development must be guided by pragmatic evidence, clinical realism, and a clear focus on patient safety. LLMs should be studied not as a universal solution, but as one possible component of a broader, carefully validated ecosystem of human-AI collaboration in the emergency department.

References

1. Malalan F, Zaboli A, Fiore A, et al. Evaluating the predictive

- accuracy of ChatGPT in risk stratification for chest pain in the emergency department. *Emerg Care J* 2025;21:13829
2. Shinde V, Kanani P, Bhimani K. Potential and limitations of large language models in acute chest pain triage. Response to Evaluating the predictive accuracy of ChatGPT in risk stratification for chest pain in the emergency department. *Emerg Care J* 2026;22:15130.
 3. Günay S, Öztürk A, Özerol H, Yiğit Y. Comparison of emergency medicine specialist, cardiologist, and ChatGPT in electrocardiography assessment. *Am J Emerg Med* 2024;80:51-60.
 4. Günay S, Öztürk A, Yiğit Y. The accuracy of Gemini, GPT-4, and GPT-4o in ECG analysis: a comparison with cardiologists and emergency medicine specialists. *Am J Emerg Med*. 2024;84:68-73.
 5. Rjoob K, Bond R, Finlay D, et al. Machine learning and the electrocardiogram over two decades: Time series and meta-analysis of the algorithms, evaluation metrics and applications. *Artif Intell Med* 2022;132:102381.

Received: 19 May 2026; Accepted: 28 May 2026; Early view: 10 July 2026.

Contributions: Arian Zaboli, conceptualization; writing - review & editing; writing - original draft; Gianni Turcato, writing - review & editing; writing - original draft.

Data availability statement: not applicable. No datasets were generated or analysed.

Funding statement: not applicable.

Conflict of interest disclosure: no conflict of interest.

Publisher's note: all claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).