

Interpreting LLM-Based Trauma Triage Simulations: Reference Standards and Clinical Validation

Büyük Dil Modeli Tabanlı Travma Triyaj Simülasyonlarının Yorumlanması: Referans Standartlar ve Klinik Validasyon

✉ Ahmet Aykut

University of Health Sciences Türkiye, Van Training and Research Hospital, Clinic of Emergency Medicine, Van, Türkiye

Keywords: Large language models, trauma triage, clinical validation

Anahtar Kelimeler: Büyük dil modelleri, travma triyajı, klinik validasyon

Dear Editor,

I read with interest the article by Çavuşoğlu Yalçın et al. (1), "Comparison of AI-Based Triage Systems in High-Energy Thoracic Trauma: A Pilot Study Using ChatGPT and Gemini." The authors address an important question: whether large language models can support early decision-making in high-acuity thoracic trauma. Their comparison of ChatGPT-4 and Gemini 1.5 across 30 synthetic thoracic trauma cases is a useful pilot contribution, particularly because thoracic trauma requires rapid recognition and prioritization of immediately treatable conditions such as tension pneumothorax, massive hemothorax, flail chest, and cardiac tamponade (1,2).

The study is strongest when interpreted as an assessment of guideline-concordant early management recommendations rather than as a direct test of clinical triage. This distinction is important. In the emergency department, triage involves prioritizing real patients under time pressure, incomplete information, changing physiology, crowding, and resource constraints. In contrast, the evaluated task involved structured synthetic vignettes with defined clinical findings and an immediate management question (1). Such simulations are useful for early model testing, but they do not reproduce the operational setting in which triage decisions are made. Describing the outcome as "guideline-concordant management recommendation"

may therefore be more precise than describing it as "triage performance."

A second point concerns the interpretation of the model-to-model receiver operating characteristic (ROC) analysis. The authors appropriately describe this analysis as exploratory rather than as a clinical diagnostic metric, with Gemini 1.5 evaluated against ChatGPT-4 outputs and an area under the curve of 0.62 (1). This framing is important. However, treating one large language model's (LLM) binary outputs as the reference standard for another does not establish clinical discrimination; it primarily characterizes inter-model divergence. In a study in which model responses had already been judged against guideline-based expert assessment, clinical accuracy should remain anchored to that expert-adjudicated standard rather than to another language model. The ROC result should therefore be interpreted as a descriptive comparison of model behavior, not as an independent measure of clinical performance.

Several methodological limitations relevant to this interpretation are already acknowledged by the authors, including the synthetic design, single-query evaluation, absence of formal quantification of inter-rater variability, and need for validation using real patient data (1). My additional concerns, therefore, relate not to whether these limitations were recognized but to how future evaluations can more clearly separate reproducibility from clinical validity.



Address for Correspondence: Ahmet Aykut, MD, University of Health Sciences Türkiye, Van Training and Research Hospital, Clinic of Emergency Medicine, Van, Türkiye

E-mail: ahmet.aykut@gmail.com **ORCID ID:** orcid.org/0009-0001-3173-8994

Received: 06.06.2026 **Accepted:** 07.09.2026 **Epub:** 15.10.2026

Cite this article as: Aykut A. Interpreting LLM-based trauma triage simulations: reference standards and clinical validation. Hamidiye Med J. [Epub Ahead of Print]



Copyright© 2026 The Author(s). Published by Galenos Publishing House on behalf of University of Health Sciences Türkiye, Hamidiye Faculty of Medicine. This is an open access article under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND) International License.

Reproducibility would be strengthened by disclosure of the exact prompts, precise model versions, access dates, and other model settings, all needed to reproduce the outputs. Clinical validity, in turn, may be better characterized by prespecified error categories that reflect the consequences of discordant recommendations. In thoracic trauma, failure to recommend immediate decompression, inappropriate imaging before stabilization, delayed tube thoracostomy, and unnecessary escalation of transfer are not clinically equivalent.

Recent reporting guidance for studies of medical large language models supports this direction. MI-CLEAR-LLM emphasizes transparent reporting of model specification, access, prompting, and evaluation procedures, while TRIPOD-LLM provides broader reporting guidance for studies using large language models in biomedical settings (3,4). In emergency medicine, these requirements are especially relevant because a superficially correct recommendation may still be unsafe if timing, prioritization, or escalation logic is wrong. A recent critical appraisal of large language model studies in emergency medicine similarly found that many studies remain limited by incomplete prompt reporting, insufficient use of real-world clinical data, limited expert benchmarking, and weak linkage to clinical usefulness (5).

The work by Çavuşoğlu Yalçın et al. (1) should therefore be viewed as a constructive starting point. It shows that large language models can generate structured, guideline-aligned responses in selected thoracic trauma scenarios while also demonstrating substantial inter-model divergence. Consistent with the authors' discussion, the next step

should shift from synthetic concordance testing to clinically grounded validation using real patient data. In parallel, a clearer distinction between simulated management concordance and operational triage, transparent reporting of prompts and model configuration, and a clinically meaningful interpretation of different error types would strengthen future evaluations. Until such validation is available, these systems are best framed as experimental adjuncts to structured reasoning and education, rather than as operational triage tools.

Footnotes

Conflict of Interest: No conflict of interest was declared by the author.

Financial Disclosure: The author declared that this study received no financial support.

References

1. Çavuşoğlu Yalçın N, Sarı Akyüz M, Karataş O, Keskin O, Özkaya M. Comparison of AI-based triage systems in high-energy thoracic trauma: a pilot study using ChatGPT and Gemini. *Hamidiye Med J*. 2026;7:50-59. [\[Crossref\]](#)
2. American College of Surgeons Committee on Trauma. *Advanced Trauma Life Support (ATLS) student course manual*. 10th ed. Chicago (IL): American College of Surgeons; 2018. [\[Crossref\]](#)
3. Park SH, Suh CH, Lee JH, Kahn CE, Moy L. Minimum reporting items for clear evaluation of accuracy reports of large language models in healthcare (MI-CLEAR-LLM). *Korean J Radiol*. 2024;25:865-868. [\[Crossref\]](#)
4. Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat Med*. 2025;31:60-69. [\[Crossref\]](#)
5. Aykut A, Yıldırım C, Günsoy E. Large language models in emergency medicine: a critical appraisal of validity, reproducibility, and clinical utility (2020-2025). *Eurasian J Emerg Med*. 2026;25:117-124. [\[Crossref\]](#)