ARCHIVES OF THE TURKISH SOCIETY OF CARDIOLOGY

Official journal of the TURKISH SOCIETY OF CARDIOLOGY

Reply to the Letter to the Editor: Can Large Language Models Guide Aortic Stenosis Management? A Comparative Analysis of ChatGPT and Gemini AI

Editöre Mektup Yanıtı: Büyük Dil Modelleri Aort Darlığı Yönetimine Rehberlik Edebilir mi? ChatGPT ve Gemini Yapay Zekanın Karşılaştırmalı Analizi

LETTER TO THE EDITOR REPLY
EDİTÖRE MEKTUP YANITI

To the Editor,

We sincerely thank the author for their thoughtful and constructive comments[1] regarding our article[2] titled "Can Large Language Models Guide Aortic Stenosis Management? A Comparative Analysis of ChatGPT and Gemini AI." Their insights contribute meaningfully to the ongoing discussion on the role of large language models (LLMs) in cardiovascular decision-making.

First, we appreciate the author's emphasis on the critical impact of prompt phrasing and contextual completeness on LLM performance. As noted, clinical scenario questions that lack explicit diagnostic or anatomical detail may yield variable model outputs. This observation is particularly relevant for Questions 33 and 35, which the author highlights. We agree that Question 33, in its isolated phrasing, did not explicitly reference aortic stenosis (AS), and that Question 35 intentionally provided limited procedural details to evaluate general decision-making tendencies rather than to simulate a comprehensive real-world case. These points offer valuable guidance for refining future question design, especially for studies aiming to better approximate clinical complexity.

Nonetheless, the primary objective of our study was not to reproduce full clinical scenarios, but rather to compare how two accessible LLMs interpret standardized, open-ended prompts under uniform methodological constraints. All questions were deliberately phrased in a zero-shot, decontextualized format to minimize prompt-engineering bias and to isolate each model's intrinsic behavior. Within this framework, the observed differences in performance—particularly ChatGPT's lower compliance in selected scenarios—reflect the models' reasoning capabilities at the time of evaluation rather than limitations of guideline knowledge per se.

Importantly, we wish to underscore a point also raised in the Letter:[1] LLM outputs are inherently sensitive to model versioning.[3] Our study used ChatGPT-4 (GPT-4-Turbo) in May 2025, whereas the Letter presents examples from ChatGPT-5.1, a substantially more advanced generation with improved clinical reasoning and contextual stability.[4] In the limitations section of our article, we explicitly acknowledged that LLM performance evolves rapidly and that findings represent a "snapshot" of a specific model version at a fixed time point. The differences illustrated by the author—namely, markedly improved responses with ChatGPT-5.1—further reinforce our argument that continuous benchmarking is essential and that LLMs cannot be evaluated as static tools.

Despite these considerations, we appreciate the positive remarks regarding the study's contribution and its value as a foundation for future research. We fully agree that upcoming work should incorporate richer clinical contexts, multimodal data, and mechanistic analyses of LLM decision pathways.

We thank the author once again for their constructive engagement and believe their observations will help inform more refined and clinically meaningful assessments of LLMs in AS management.

Ali Sezgin[1]

Veysel Ozan Tanık[1]

Bülent Özlek[2]

[1]*Department of Cardiology, Ankara Etlik City Hospital, Ankara, Türkiye*
[2]*Department of Cardiology, Muğla Sıtkı Koçman University, Faculty of Medicine, Muğla, Türkiye*

Corresponding author:
Bülent Özlek
bulent_ozlek@hotmail.com

Cite this article as: Sezgin A, Tanık VO, Özlek B. Reply to the Letter to the Editor: Can Large Language Models Guide Aortic Stenosis Management? A Comparative Analysis of ChatGPT and Gemini AI. *Turk Kardiyol Dern Ars.* 2026;54(1):85–86.

DOI: 10.5543/tkda.2025.16285

References

1. Çalışkan MU, Yaman Akkuş NM. Re: Can Large Language Models Guide Aortic Stenosis. *Turk Kardiyol Dern Ars*. 2026;54(1):83–84.
2. Sezgin A, Tanık VO, Akdoğan M, et al. Can Large Language Models Guide Aortic Stenosis Management? A Comparative Analysis of ChatGPT and Gemini AI. *Turk Kardiyol Dern Ars*. 2025;53(8):599-606. [CrossRef]
3. Funk PF, Hoch CC, Knoedler S, et al. ChatGPT's Response Consistency: A Study on Repeated Queries of Medical Examination Questions. *Eur J Investig Health Psychol Educ*. 2024;14(3):657-668. [CrossRef]
4. Sanli AN, Turan B, Tekcan Sanli DE. Advances in Large Language Model Performance: A Comparative Study of ChatGPT-4 and ChatGPT-5 on ABSITE Questions. Am Surg. October 18, 2025:31348251390958. doi: 10.1177/00031348251390958. [Epub ahead of print]. [CrossRef]