Thorac Res Pract. 2026;27(4):266-267 **DOI:** 10.4274/ThoracResPract.2026.2026-6-15

Letter to the Editor

Reply Letter to the Editor: Free Large Language Model-based Chatbots Can Help to Align Statistical Tests with the Study Design and Avoid Pseudo-replication

Ömer F. Karakoyun[1], Halil E. Koyuncuoğlu[1], Ömer H. Sağnıç[2], Mehmed E. Özdemir[3], Yalçın Gölcük[4], Birdal Yıldırım[4]

[1]Clinic of Emergency Medicine, Muğla Training and Research Hospital, Muğla, Türkiye
[2]Clinic of Emergency Medicine, Isparta City Hospital, Isparta, Türkiye
[3]Department of Artificial Intelligence, Muğla Sıtkı Koçman University, Graduate School of Natural and Applied Sciences, Muğla, Türkiye
[4]Department of Emergency Medicine, Muğla Sıtkı Koçman University Faculty of Medicine, Muğla, Türkiye

Cite this article as: Karakoyun ÖF, Koyuncuoğlu HE, Sağnıç ÖH, Özdemir ME, Gölcük Y, Yıldırım B. Reply Letter to the Editor: Free large language model-based chatbots can help to align statistical tests with the study design and avoid pseudo-replication. *Thorac Res Pract.* 2026;27(4):266-267

KEYWORDS

Friedman test, Kruskal-Wallis test, large language models, pulmonary embolism, clinical decision support

Received: 17.06.2026 ***Revision Requested:*** 19.06.2026 ***Last Revision Received:*** 23.06.2026

Accepted: 23.06.2026 ***Publication Date:*** 24.07.2026

DEAR EDITOR,

We thank the author(s) of the Letter to the Editor entitled "Letter to the Editor Re: Free Large Language Model-based Chatbots Can Help to Align Statistical Tests with the Study Design and Avoid Pseudo-replication" for their careful reading of our article, "AI in Patient Care: Evaluating Large Language Model Performance Against Evidence-Based Guidelines for Pulmonary Embolism," and for highlighting an important statistical issue.[1,2]

The main concern relates to the use of the Kruskal-Wallis test for comparing scores from four large language model (LLM)-based chatbots. We acknowledge that, because the same ten clinical questions were submitted to each model, the observations are better considered as blocked or repeated measures rather than independent groups. Therefore, a repeated-measures non-parametric approach is more appropriate for the overall comparison.[3,4]

To address this point, we reanalyzed the model scores using the Friedman test, treating each question as a block and the four LLMs as related conditions. The Friedman test showed no statistically significant difference among the four models [χ^2=4.50, P = 0.212]. Since the overall test was not significant, no post-hoc pairwise comparisons were performed.[5] This result is consistent with the main conclusion of our original analysis: although ChatGPT-4o achieved the highest total score and DeepSeek-V2 the lowest, no model demonstrated statistically significant overall superiority.

We agree that the statistical analysis section could have more clearly reflected the repeated-measures structure of the data. Nevertheless, the reanalysis did not change the principal interpretation of the study. The evaluated LLMs demonstrated model-specific strengths and limitations across different components of pulmonary embolism management, but none consistently outperformed the others across all domains.

Accordingly, the relevant Methods and Results statements in the original article should be corrected through an erratum, replacing the Kruskal-Wallis analysis with the Friedman test and reporting the updated result [$\chi^2(3)$=4.50, P = 0.212].

Corresponding author: Prof. Birdal Yıldırım, MD, e-mail: birdalgul@gmail.com

We appreciate the opportunity to clarify this issue and believe that this discussion contributes constructively to the methodological rigor of studies evaluating AI-based clinical decision-support tools.

Footnotes

Authorship Contributions

Concept: Ö.F.K., H.E.K., Ö.H.S., M.E.Ö., Y.G., B.Y., Design: Ö.F.K., H.E.K, Y.G., B.Y.,Data Collection or Processing: H.E.K., Ö.H.S., M.E.Ö., Y.G., Literature Search: Ö.F.K., H.E.K., Ö.H.S., Y.G. Analysis or

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

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

REFERENCES

1. Metze K, da Silva CS, Lorand-Metze I, de Mattos AC. Letter to the Editor: Free large language model-based chatbots can help to align statistical tests with the study design and avoid pseudo-replication. *Thorac Res Pract*.2026;27(4):264-265. **[Crossref]**
2. Karakoyun ÖF, Koyuncuoğlu HE, Sağnıç ÖH, Özdemir ME, Gölcük Y, Yıldırım B. AI in patient care: evaluating large language model performance against evidence-based guidelines for pulmonary embolism. *Thorac Res Pract*. 2026;27(1):38-46. **[Crossref]**
3. Kim HY. Statistical notes for clinical researchers: nonparametric methods for comparing three or more groups. *Restor Dent Endod*. 2014;39(4):329-332. **[Crossref]**
4. Sheldon MR, Fillyaw MJ, Thompson WD. The use and interpretation of the Friedman test in the analysis of ordinal-scale data in repeated measures designs. *Physiother Res Int*. 1996;1(4):221-228. **[Crossref]**
5. Eisinga R, Heskes T, Pelzer B, Te Grotenhuis M. Exact p-values for pairwise comparison of Friedman rank sums, with application to comparing classifiers. *BMC Bioinformatics*. 2017;18:68. **[Crossref]**