Thorac Res Pract. 2026;27(4):264-265 **DOI:** 10.4274/ThoracResPract.2026.2026-6-3

Letter to the Editor

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

Konradin Metze[1], Cauê Sousa da Silva[1], Irene Lorand-Metze[2], Amilcar Castro de Mattos[1,3]

[1]Department of Pathology, State University of Campinas Faculty of Medical Sciences, São Paulo, Brazil
[2]Department of Internal Medicine, State University of Campinas Faculty of Medical Sciences, São Paulo, Brazil
[3]Laboratory of Pathology, Pontifical Catholic University of Campinas, Campinas, São Paulo, Brazil

Cite this article as: 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

KEYWORDS

Clinical problems, eHealth, artificial intelligence, biostatistics, large language models

Received: 07.06.2026

Accepted: 19.06.2026 ***Publication Date:*** 24.07.2026

DEAR EDITOR,

In a recent contribution,[1] the authors compared the quality of replies provided by different large language-based chatbots (LLMs) to questions related to diagnosis, treatment, and prognosis of pulmonary thromboembolism. The study design was based on 10 questions, all submitted to four different chatbots and their replies were scored by specialists. This study is composed of blocks of 4 dependent answers always related to one question. Therefore, the Kruskal-Wallis test, which has been created for independent data, is not adequate. When incorrectly applied for nested data, this test will provoke pseudo-replication.[2] The test power will be distorted, thus provoking unreliable results. Adequate solutions would be tests for repeated measures, such as mixed models, Friedman tests with post-hoc tests and alpha-error corrections. Eventually, ANOVA tests for dependent data, when assumptions were met.

Here we tried to find out whether freely available LLM-based chatbots could help to select the adequate statistical tests. The published abstract was sent to 14 different chatbots with the following prompt:

“Please analyze this text and look for major problems, such as methodological errors, inconsistencies, or contradictions. Pay special attention to the statistical tests and examine whether they are well aligned with the study design. Please suggest corrections. Should we change a test by another one?“

The following 12 algorithms recognized the nested study design and suggested substitution by the above-mentioned adequate test. appointing it as the most important error: ChatGPT Free, Claude Sonnet 4.6, Consensus Corpus All, Copilot Smart, Deep Seek Fast DeepThink, Gemini 3.5 Flash, Grok Fast, Julius 1.2 Lite, Meta AI Instant, Notebook LM Free, Perplexity Learn Tutor. Furthermore, other questions regarding test power, data distribution, counting procedures, post-hoc tests, alpha-error correction, and inter-rater variability were also correctly discussed.

The remaining two bots (Bohrium Expert and Mistral Le Chat Balanced) did not recognize the nested study design and therefore did not indicate correction of the most important error.

Corresponding author: Konradin Metze, MD, e-mail: kmetze@fcm.unicamp.br

Since replies of different LLMs to the same prompt or repeated submissions may lead to diverging results,[3-5] we always recommend to consult several LLMs and eventually, to discuss conflicting answers with the chatbots until a final solution can be reached.[6]

Footnotes

Authorship Contributions

Concept: K.M., A.C.D.M., Design: K.M., C.S.D.S., Data Collection or Processing: K.M., C.S.D.S., A.C.D.M., Literature Search: K.M., C.S.D.S., I.L-M., A.C.D.M., Analysis or Interpretation: K.M., C.S.D.S., Writing: K.M., I.L-M., A.C.D.M.

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. 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]**
2. Lazic SE. The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? *BMC Neurosci*. 2010;11:5. **[Crossref]**
3. Metze K, Morandin-Reis RC, Lorand-Metze I, Florindo JB. Bibliographic research with ChatGPT may be misleading: the problem of hallucination. *J Pediatr Surg*. 2024;59(1):158. **[Crossref]**
4. Metze K, Morandin-Reis RC, Lorand-Metze I, Florindo JB. Bibliographic research with large language model ChatGPT-4: instability, hallucinations and sometimes alerts. Clinics (Sao Paulo). 2024;79:100409. **[Crossref]**
5. Metze K, Morandin-Reis RC, de Ávila Reis MF, da Silva Fago M, Florindo JB. Misinformation, false positives and delegation of tasks - large language models should not be used for the detection of retracted literature - A study of 21 Chatbots. *J Clin Anesth*. 2025;107:112032. **[Crossref]**
6. Metze K, Mattos AC, Lorand-Metze I. Generative artificial intelligence sycophancy and critical thinking - clues for introducing chatbots in the classroom. *Plast Reconstr Surg*. 2026 Mar 10. [Epub ahead of print]. **[Crossref]**