Letter to the Editor

Authors' Reply: Critical Limitations in Comparing ChatGPT and DeepSeek for Orthopedic Assessment

Chirathit Anusitviwat[1], MD; Sitthiphong Suwannaphisit[2], MD; Jongdee Bvonpanttarananon[1]; Boonsin Tangtrakulwanich[1], MD

[1]Prince of Songkla University, Hat Yai, Songkhla, Thailand
[2]Navamindradhiraj University, Bangkok, Bangkok, Thailand

Corresponding Author:

Chirathit Anusitviwat, MD
Prince of Songkla University
15 Karnjanavanich Road
Hat Yai, Songkhla 90110
Thailand
Phone: 66 74451601
Email: pchirathit@gmail.com

Related Articles:
Comment on: https://formative.jmir.org/2025/1/e75607
Comment in: https://formative.jmir.org/2026/1/e90242/

JMIR Form Res 2026;10:e91470; doi: 10.2196/91470

Keywords: ChatGPT; large language model; LLM; orthopedic; multiple-choice question; MCQ

We thank you for the useful and constructive comments [1] on our article "Comparing ChatGPT and DeepSeek for Assessment of Multiple-Choice Questions in Orthopedic Medical Education: Cross-Sectional Study" [2]. This reply aims to address the concerning points that were brought up in the letter to the editor.

Misinterpretation of Reliability Statistics

According to our study, we administered the multiple-choice questions (MCQs) for ChatGPT and DeepSeek on a separate day. All data from the two large language models (LLMs) were measured by two assessors. Although two assessors were used for each LLM, the reported Cohen κ coefficient values represent within-model interrater reliability, not interrater reliability between the two LLMs [3]. Therefore, describing these results as agreement between the two models is inaccurate.

Linguistic Ambiguity and Generalizability

All MCQs used in our study were administered in English. No Thai language inputs or translations were used. Therefore, the performance differences between the two models reflect the model performance on English language medical questions rather than variability due to language translation or non-English linguistic processing.

Reproducibility and Interface Transparency

All models in our study were accessed via web-based user interfaces (UIs), not application programming interfaces. We acknowledge that web-based UIs may be subject to updates and lack version control. However, the web-based version of ChatGPT is easy to access and requires no software installation. It also allows quick testing and exploration without technical or cost barriers, making it well-suited for nontechnical users and educational studies [4]. Therefore, we used the web-based UI in our study.

Risk of Data Contamination

Even though these MCQs have been used for more than 5 years, the MCQs used in our study are from private orthopedic examinations. Thus, we believe that these items would not appear in public sources. Future research using newly created MCQs may be better for assessing the capability or efficacy of LLMs.

Data Reporting Discrepancy

Upon re-examination, we confirm that the correct accuracy for the pelvic and spine injury category (n=19) using the Reason function is indeed 16 of 19, corresponding to approximately 84.2%. The value of 68.8% reported in Table 2 was a typographical error. This error has been corrected through a published corrigendum [5].

Conflicts of Interest

None declared.

References

1. Ayas O, Acar A. Critical limitations in comparing ChatGPT and DeepSeek for orthopedic assessment. JMIR Form Res. 2026;10:e90242. [doi: 10.2196/90242]
2. Anusitviwat C, Suwannaphisit S, Bvonpanttarananon J, Tangtrakulwanich B. Comparing ChatGPT and DeepSeek for assessment of multiple-choice questions in orthopedic medical education: cross-sectional study. JMIR Form Res. Dec 19, 2025;9:e75607. [doi: 10.2196/75607] [Medline: 41418321]
3. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-282. [Medline: 23092060]
4. Park CR, Heo H, Suh CH, Shim WH. Uncover this tech term: application programming interface for large language models. Korean J Radiol. Aug 2025;26(8):793-796. [doi: 10.3348/kjr.2025.0360] [Medline: 40736411]
5. Anusitviwat C, Suwannaphisit S, Bvonpanttarananon J, Tangtrakulwanich B. Correction: comparing ChatGPT and DeepSeek for assessment of multiple-choice questions in orthopedic medical education: cross-sectional study. JMIR Form Res. Feb 26, 2026;10:e92549. [doi: 10.2196/92549] [Medline: 41747218]

Abbreviations

LLM: large language model
MCQ: multiple-choice question
UI: user interface

Edited by Amanda Iannaccio; This is a non–peer-reviewed article; submitted 15.Jan.2026; final revised version received 14.Feb.2026; accepted 26.Feb.2026; published 17.Mar.2026

Please cite as:
Anusitviwat C, Suwannaphisit S, Bvonpanttarananon J, Tangtrakulwanich B
Authors' Reply: Critical Limitations in Comparing ChatGPT and DeepSeek for Orthopedic Assessment
JMIR Form Res 2026;10:e91470
URL: https://formative.jmir.org/2026/1/e91470
doi: 10.2196/91470