Letter to the Editor

# Critical Limitations in Comparing ChatGPT and DeepSeek for Orthopedic Assessment

Orhan Ayas[1], MD; Alaeddin Acar[2], MD

[1]Department of Orthopedics and Traumatology, Fethi Sekin City Hospital, Elazığ, Turkey
[2]Department of Neurosurgery, Kulu State Hospital, Konya, Turkey

**Corresponding Author:**

Alaeddin Acar, MD
Department of Neurosurgery
Kulu State Hospital
No 4, 139518 Street, Dinek, Kulu
Konya 42770
Turkey
Phone: 90 542 472 37 23
Email: alaeacar@gmail.com

**Related Articles:**
Comment on: https://formative.jmir.org/2025/1/e75607
Comment in: https://formative.jmir.org/2026/1/e91470

***JMIR Form Res 2026;10:e90242;*** doi: 10.2196/90242

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

We read with great interest the study by Anusitviwat et al [1], which compared the performance of ChatGPT and DeepSeek in orthopedic examinations. While the study provides timely insights into the utility of large language models (LLMs) in medical education, we identified specific methodological and terminological limitations that warrant clarification to ensure the validity and reproducibility of the findings.

## Misinterpretation of Reliability Statistics

The authors state that the "interrater reliability between the two LLMs" was evaluated using the Cohen ϰ coefficient [1]. Mathematically, measuring the agreement between two independent raters (interrater) yields a single coefficient. However, the results report two separate values: ϰ of 0.81 for ChatGPT and ϰ of 0.78 for DeepSeek [1]. This finding, combined with the methodology stating that questions were input on "separate days" [1], indicates that the study actually measured intramodel consistency (test-retest reliability) rather than the agreement between the models. Labeling internal consistency as "interrater reliability" is terminologically inaccurate and misrepresents the statistical relationship between the two models.

## Linguistic Ambiguity and Generalizability

The manuscript does not specify the language of the input multiple-choice questions (Thai or English) used in the assessments. This omission is critical, as the impact of input language on LLM performance is well-documented. For instance, Noda et al [2] demonstrated that GPT-4V's accuracy on the Japanese Otolaryngology Board Examination significantly improved from 24.7% (Japanese input) to 47.3% when translated into English. This finding underscores that models optimized for English exhibit distinct performance disparities in non-English languages. Without clarifying whether the assessments were administered in the local language or English, it is impossible to determine if the reported accuracy gap between ChatGPT (80.4%) and DeepSeek (74.2%) stems from medical reasoning capabilities or linguistic processing proficiency.

## Reproducibility and Interface Transparency

The methodology reports the use of "Reason" and "DeepThink" functions but does not explicitly state whether the models were accessed via web-based user interfaces or application programming interfaces [1]. This distinction is vital for reproducibility. Web-based user interfaces are subject to opaque updates and lack the stability of controlled application programming interface environments. Without defining the access method and the specific prompt structures used, the experimental conditions cannot be replicated.

## Risk of Data Contamination

The authors note that the multiple-choice questions "have been used in orthopedic examinations for more than 5 years." This longevity significantly increases the risk of data contamination, as older items likely exist in public repositories within LLM training corpora, potentially conflating memorization with reasoning. To ensure validity, recent benchmarks use private datasets (Busch et al [3]) or questions postdating the model's training cutoff (Noda et al [2]). The absence of such controls in this study undermines the internal validity of the comparison.

## Data Reporting Discrepancy

Finally, we noted a minor discrepancy in Table 2. In the "Pelvic and spine injury" category (n=19), the accuracy for the Reason function is listed as 16 (68.8%) [1]. Mathematically, 16 of 19 corresponds to approximately 84.2%, not 68.8%. We respectfully invite the authors to clarify this value to ensure the precision of the tabulated data.

### Acknowledgments

Google Gemini was used for language editing.

### Conflicts of Interest

None declared.

### References

1. 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]
2. Noda M, Ueno T, Koshu R, et al. Performance of GPT-4V in answering the Japanese Otolaryngology Board Certification Examination questions: evaluation study. JMIR Med Educ. Mar 28, 2024;10:e57054. [doi: 10.2196/57054] [Medline: 38546736]
3. Busch F, Han T, Makowski MR, Truhn D, Bressem KK, Adams L. Integrating text and image analysis: exploring GPT-4V's capabilities in advanced radiological applications across subspecialties. J Med Internet Res. May 1, 2024;26:e54948. [doi: 10.2196/54948] [Medline: 38691404]

### Abbreviations

**LLM:** large language model

*Edited by Amanda Iannaccio; This is a non–peer-reviewed article; submitted 23.Dec.2025; accepted 26.Feb.2026; published 17.Mar.2026*

*Please cite as:*
*Ayas O, Acar A*
*Critical Limitations in Comparing ChatGPT and DeepSeek for Orthopedic Assessment*
*JMIR Form Res 2026;10:e90242*
*URL: https://formative.jmir.org/2026/1/e90242*
*doi: 10.2196/90242*