THE MEDICAL BULLETIN OF
SISLI ETFAL HOSPITAL

DOI: 10.14744/SEMB.2025.05046
Med Bull Sisli Etfal Hosp 2025;59(3):440-441

Letter to the Editor

Comment on "Performance of AI Models vs. Orthopedic Residents in Turkish Specialty Training Development Exams in Orthopedics"

Hinpetch Daungsupawong,[1] Viroj Wiwanitkit[2]

[1]Private Academic Consultant, Phonhong, Lao People's Democratic Republic
[2]Department of Community Medicine, Dr. D Y Patil Vidyapeeth (Deemed to be University), D Y Patil Medical College, Hospital and Research Centre, Pune, India

Please cite this article as "Daungsupawong H, Wiwanitkit V. Comment on "Performance of AI Models vs. Orthopedic Residents in Turkish Specialty Training Development Exams in Orthopedics". Med Bull Sisli Etfal Hosp 2025;59(3):440-441".

Dear Editor,

The publication on "Performance of AI Models vs. Orthopedic Residents in Turkish Specialty Training Development Exams in Orthopedics[1]" is hereby discussed. This study aims to be both modern and crucial in the era of AI's increasing role in medical education and clinical decision-making. It compares the performance of large language models (LLMs)—ChatGPT-4o, Gemini, Bing AI, and DeepSeek—with orthopaedic residents on the Specialty Training Development Exams (UEGS) from 2010 to 2021. While the results indicate that AI outperforms humans in terms of "accuracy," there are still some statistical and research design criticisms that need further consideration.

The first restriction is that the comparison is retroactive, with AI answering prior questions without taking into account environmental factors that may affect resident learning, such as curriculum changes, data access, or test stress levels. In contrast, AI responds to inquiries without regard for the environment. Furthermore, while one-way ANOVA is adequate for group comparisons, it does not offer critical markers such as effect size, standard deviation, or confidence interval, making it impossible to judge clinical, rather than statistical, significance.

A closer look suggests that "accuracy" may not be enough to evaluate AI's capabilities in a healthcare setting. Qualitative tools, such as specific rubrics or expert assessments, should be used to examine factors such as "depth of explanation" and "clinical consistency". Assuming that an AI is equivalent to a physician with 5 years of expertise is a subjective evaluation with no concrete benchmarks. Furthermore, ChatGPT-4o's lower-than-expected performance could be attributed to the model's emphasis on providing answers in general situations rather than highly specialized exams.

The appropriateness of utilizing AI in clinical skills testing or assessments is a hotly debated topic in academia. A correct answer does not imply that it can "think critically" or evaluate the circumstances of a genuine patient. Furthermore, it is worth considering if artificial intelligence should be created to "replace" or "support" physician decision-making. Clear ethical criteria should be provided for the use of these language models in teaching and evaluation situations, especially in disciplines that need extensive knowledge, experience, and judgment, such as orthopedic surgery.

Address for correspondence: Hinpetch Daungsupawong, MD. Private Academic Consultant, Phonhong, Lao People's Democratic Republic
E-mail: hinpetchdaung@gmail.com

Submitted Date: August 21, 2025 **Accepted Date:** September 25, 2025 **Available Online Date:** October 13, 2025

Disclosures

Conflict of Interest: The authors declare no conflict of interest.

Funding Statement: The authors declared that this work received no special funding.

Author contributions: Concept – H.D., V.W.; Design – H.D., V.W.; Supervision – H.D., V.W.; Materials – H.D., V.W.; Data Collection and/or Processing – H.D., V.W; Analysis and/or Interpretation – H.D., V.W.; Literature Review – H.D., V.W.; Writing – H.D., V.W.; Critical Review – H.D., V.W.

Use of AI for Writing Assistance: The authors declared that generative artificial intelligence tools such as ChatGPT (OpenAI) and AI-assisted translation software such as DeepL were used during the manuscript preparation.

References

1. Ipek E, Sulek Y, Balkanli B. Performance of AI Models vs. Orthopedic Residents in Turkish Specialty Training Development Exams in Orthopedics. Sisli Etfal Hastan Tip Bul. 2025;59:151–5. [CrossRef]

Author's Reply

Dear Editor,

We sincerely thank the readers for their thoughtful comments and constructive criticism on our recently published article. We are pleased to clarify several points raised.

First, by design, our comparison was retrospective. Residents' scores were drawn from UEGS examinations administered over multiple years, whereas the AI models' responses were generated on the same items within a short time window using a standardized, text-only protocol without personalization or web browsing. Consequently, environmental factors that may influence resident learning—such as curricular changes, access to archives/study materials, or exam-related stress—were not directly measured. In contrast, the AI systems were run under identical conditions (same model/version, prompt, and settings) and were therefore not exposed to such human environmental variability. Our findings should thus be interpreted as a comparison against a reference that is deliberately insulated from human contextual fluctuations.

Second, for multiple-group comparisons we used one-way ANOVA. We agree that reporting effect sizes and confidence intervals would strengthen clinical interpretation. We welcome this suggestion and, at the Editor's request, can provide effect sizes (e.g., η^2/Hedges' g) and 95% confidence intervals as a supplementary file. Moreover, we plan to report these metrics systematically in future work to enhance interpretability.

Third, we share the view that, in healthcare settings, simple "right/wrong" scoring does not fully capture the clinical value of AI. The scope of our study was a performance comparison on identical question sets; therefore, qualitative dimensions such as "explanatory depth" and "clinical coherence" were not systematically scored. Especially for open-ended or interpretive items, we believe future studies should incorporate blinded expert ratings based on a pre-specified rubric to evaluate the depth and coherence of AI-generated answers, accompanied by reliability reporting (e.g., ICC).

Fourth, our study does not claim general equivalence. Any phrasing suggesting "equivalence to a five-year physician" refers solely to relative performance on specific exam items. Clinical competence is multidimensional (e.g., EPAs, patient safety, team communication, decision-making under time pressure, and legal/ethical considerations) and cannot be inferred from item-level accuracy alone.

As acknowledged in our Limitations, although AI models may demonstrate high accuracy in certain contexts, they may not yet reflect the contextual understanding and nuanced clinical reasoning required in real-world practice. In future work, we intend to evaluate these qualitative dimensions using a standardized rubric, blinded expert grading, and reliability statistics (ICC).

We also agree that a correct answer alone does not demonstrate a model's capacity for critical thinking or for assessing the circumstances of a real patient. Our study presents a performance comparison on the same item sets; it does not assert high-stakes claims about clinical reasoning or bedside proficiency. Our stance is that AI should support—not replace—physician decision-making within appropriate boundaries. We likewise concur on the need for clear ethical principles governing the use of AI in teaching and assessment, particularly in specialties that demand extensive knowledge, experience, and judgment such as orthopedic surgery.

In summary, our findings illustrate AI's potential for item-level performance; however, in clinical practice AI should function as a decision-support tool under explicit ethical guidelines and institutional oversight. We are grateful for the opportunity to clarify these points and strongly support ongoing efforts to integrate AI into medical education responsibly.

Sincerely,

The Authors

Enver Ipek, Yusuf Sulek, Bahadir Balkanli

Department of Orthopedics, University of Health Sciences Türkiye, Sisli Hamidiye Etfal Training and Research Hospital, Istanbul, Türkiye

E-mail: enveripek88@gmail.com

Doi: 10.14744/SEMB.2025.48107