Reply to Kleebayoon and Wiwanitkit

In Reply:

We thank Drs Kleebayoon and Wiwanitkit for their thoughtful letter regarding our recently published article, "Evaluation of a Large Language Model on the American Academy of Pediatrics' PREP Emergency Medicine Question Bank."[1] Our study evaluated the performance of a publicly available large language model (LLM) on pediatric emergency medicine (PEM) board examination questions, contributing to the growing dialogue on the role of ChatGPT and other LLMs in medical education. The authors raised important points about the implications of our findings and opportunities for future work, which we address below.

We concur with the authors' observation that assessing accuracy through majority voting may mask nuances in reasoning errors. Previous studies have examined these limitations. Gilson et al[2] analyzed LLM performance on United States Medical Licensing Examination (USMLE) questions, identifying logical errors, defined as responses that correctly identify pertinent information but fail to apply it appropriately, as the primary reason for incorrect answers (42%). Similarly, Kung et al[3] assessed ChatGPT's consistency across different question formats, highlighting its ability to maintain concordance and provide insightful reasoning when justifications were required.

Regarding performance across topics, we acknowledge that sample sizes for individual categories were limited, precluding rigorous comparisons. While prior work has demonstrated that ChatGPT has passing performance on the general pediatrics board examination questions,[4,5] approaches which attempt to examine responses in greater detail for pediatric or pediatric subspecialty question banks are lacking in the literature. Nuanced analyses of reasoning patterns for these kinds of questions warrant further investigation.

Recent comparative analyses, including by Yanagita et al,[6] demonstrate ChatGPT's accuracy on medical licensing examinations in different contexts, offering insights into its strengths and gaps relative to human performance. Such studies could help contextualize ChatGPT's utility in real-world medical education.

Future directions proposed by the authors, such as integrating adaptive questioning and personalized prompts, merit further exploration. Adaptive learning techniques, as discussed by Singhal et al,[7] can enhance model performance by enabling dynamic problem-solving tailored to learner needs. In addition, combining LLMs with interactive simulation platforms could simulate case-based learning, fostering deeper understanding and application of clinical concepts.[8]

The authors' suggestion of leveraging LLMs as collaborative tools in medical education aligns with emerging models of AI-human interaction. Karimov et al[9] highlighted ChatGPT's potential to assist clinicians as a just-in-time reference while underscoring the importance of human oversight to ensure accuracy. Such hybrid models could guide learners through iterative problem-solving processes, enhancing educational outcomes while maintaining safety and reliability.

The authors raise excellent points with respect to future opportunities comparing the rationale of ChatGPT with expert-based respondents, and the use of LLM in a collaborative system. As LLMs evolve, their performance in standardized testing and clinical reasoning tasks will improve. Addressing current limitations, such as topic-specific knowledge gaps and reasoning inconsistencies, will enhance their utility in medical education. We appreciate the opportunity to address these thoughtful points and welcome further discussion to advance this field.

Sriram Ramgopal, MD*†
Srinivasan Suresh, MD, MBA†
*Department of Pediatrics, Division of Emergency Medicine, Ann & Robert H. Lurie Children's Hospital of Chicago Northwestern University Feinberg School of Medicine, Chicago, IL
†Department of Pediatrics, Divisions of Health Informatics and Emergency Medicine, University of Pittsburgh School of Medicine & UPMC Children's Hospital of Pittsburgh, Pittsburgh, PA

S.R. is supported by the National Institutes of Health/National Heart, Lung, and Blood Institute (K01HL169921).
Disclosure: The authors declare no conflict of interest.
DOI: 10.1097/PEC.0000000000003342

REFERENCES

1. Ramgopal S, Varma S, Gorski JK, et al. Evaluation of a large language model on the American Academy of Pediatrics' PREP Emergency Medicine Question Bank. *Pediatr Emerg Care*. 2024;40:871–875.
2. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9:e45312.
3. Kung TH, Cheatham M, Medenilla A. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digital Health*. 2023;2:e0000198.
4. Le M, Davis M. ChatGPT yields a passing score on a pediatric board preparatory exam but raises red flags. *Glob Pediatr Health*. 2024;11:2333794X241240327.
5. Suresh S, Misra SM. Large language models in pediatric education: current uses and future potential. *Pediatrics*. 2024;154:e2023064683.
6. Yanagita Y, Yokokawa D, Uchida S, et al. Accuracy of ChatGPT on medical questions in the national medical licensing examination in Japan: evaluation study. *JMIR Form Res*. 2023;7:e48023.
7. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620:172–180.
8. Misra SM, Suresh S. Artificial intelligence and objective structured clinical examinations: using ChatGPT to revolutionize clinical skills assessment in medical education. *J Med Educ Curric Dev*. 2024;11:23821205241263475.
9. Karimov Z, Allahverdiyev I, Agayarov OY, et al. ChatGPT vs UpToDate: comparative study of usefulness and reliability of Chatbot in common clinical presentations of otorhinolaryngology–head and neck surgery. *Eur Arch Otorhinolaryngol*. 2024;281:2145–2151.