Letter to the Editor

Author's Reply: "Data Contamination in AI Evaluation"

ChulHyoung Park[1,2,3*], MD; Min Ho An[1,2,3*], MD; Gyubeom Hwang[1,2,3*], MD; Rae Woong Park[1,2,3,4], MD, PhD; Juho An[5], MD

[1]Department of Biomedical Informatics, School of Medicine, Ajou University, Suwon, Republic of Korea
[2]Center for Biomedical Informatics Research, Ajou University Medical Center, Suwon, Republic of Korea
[3]Department of Medical Sciences, Graduate School of Ajou University, Suwon, Republic of Korea
[4]BK21 R&E Initiative for Advanced Precision Medicine, Suwon, Republic of Korea
[5]Department of Emergency Medicine, School of Medicine, Ajou University, Suwon, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Juho An, MD
Department of Emergency Medicine
School of Medicine
Ajou University
164 Worldcup-ro
Yeongtong-gu
Suwon, 16499
Republic of Korea
Phone: 82 0312195016
Email: ermd.jh@gmail.com

Related Articles:
Comment on: https://medinform.jmir.org/2025/1/e68409/
Comment on: http://medinform.jmir.org/2025/1/e80987/

(JMIR Med Inform 2025;13:e82057) doi: 10.2196/82057

KEYWORDS

artificial intelligence; large language model; ChatGPT; emergency medicine; clinical performance examination; history taking; clinical reasoning; empathy; patient experience

We sincerely thank the author for the constructive commentary on our recent publication. Our study evaluated ChatGPT's performance across multiple dimensions—including history taking, diagnostic accuracy, communication skills, and empathic expression—through a clinical performance examination using simulated patients combined with written examinations [1].

In our study, the written examination was not intended to solely serve as a direct comparison of performance between ChatGPT and human physicians. Rather, it was included to support the interpretation of ChatGPT's communication skills and empathic responses observed during simulated patient interactions by providing additional context regarding the model's underlying clinical knowledge. A previous study has shown that patients may perceive ChatGPT's responses as empathic or trustworthy, even when those responses are clinically inappropriate [2]. However, effective clinical communication is not merely about verbal fluency or emotional tone; it must be grounded in adequate medical knowledge. For this reason, earlier studies evaluating artificial intelligence empathy have also assessed the clinical appropriateness of responses and compared them to those of human physicians [2,3].

Consistent with prior work, we also assessed the simulated patient conversations in terms of both clinical accuracy and empathic engagement, as evaluated by an emergency medicine professor. However, we recognize that physicians vary in their diagnostic styles and communication approaches. Subjective judgment from the evaluator may have influenced the ratings, especially given that the evaluated outputs were full conversations rather than single responses. To provide a complementary and more structured assessment, we incorporated a written test focused on 3 key domains: diagnosis, investigation, and treatment planning. Performance on this test may serve as a supporting element to help ensure that ChatGPT's interpersonal strengths were not misaligned with clinical reasoning.

As the author correctly pointed out, the questions in the written examination were adapted from a publicly available textbook published in 2018 [4]. We cannot rule out the possibility that ChatGPT was exposed to this material or similar content during pretraining, due to the limited transparency regarding its training data. Therefore, part of the model's performance on the written test may have been influenced by data contamination. We fully

acknowledge this methodological limitation and agree that the results from the written examination should be interpreted with caution.

We are truly grateful for the author's thoughtful engagement, which raises important considerations for future studies regarding the assessment of AI in clinical settings.

Acknowledgments

During the preparation of this manuscript, the authors used ChatGPT to assist with improving readability and correcting grammatical errors. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Conflicts of Interest

None declared.

References

1. Park C, An MH, Hwang G, Park RW, An J. Clinical performance and communication skills of ChatGPT versus physicians in emergency medicine: simulated patient study. JMIR Med Inform. Jul 17, 2025;13:e68409. [FREE Full text] [doi: 10.2196/68409] [Medline: 40674718]
2. Armbruster J, Bussmann F, Rothhaas C, Titze N, Grützner PA, Freischmidt H. "Doctor ChatGPT, can you help me?" The patient's perspective: cross-sectional study. J Med Internet Res. Oct 01, 2024;26:e58831. [FREE Full text] [doi: 10.2196/58831] [Medline: 39352738]
3. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. Jun 01, 2023;183(6):589-596. [FREE Full text] [doi: 10.1001/jamainternmed.2023.1838] [Medline: 37115527]
4. Shamil E, Ravi P, Mistry D. 100 Cases in Emergency Medicine and Critical Care. Boca Raton, FL. CRC Press; 2018.

Edited by A Iannaccio; this is a non–peer-reviewed article. Submitted 15.Aug.2025; accepted 20.Aug.2025; published 29.Sep.2025.

Please cite as:
Park C, An MH, Hwang G, Park RW, An J
Author's Reply: "Data Contamination in AI Evaluation"
JMIR Med Inform 2025;13:e82057
URL: https://medinform.jmir.org/2025/1/e82057
doi: 10.2196/82057
PMID: