World Journal of Urology (2024) 42:104
https://doi.org/10.1007/s00345-024-04844-2

LETTER TO THE EDITOR

Letter to the editor, "How does artificial intelligence master urological board examinations?"

Junjun Wang[1] · **Xing Yun**[2]

Received: 24 January 2024 / Accepted: 24 January 2024

Dear Editor,

I am writing to provide a response to the article by Kollitsch [1] et al. titled "How does artificial intelligence master urological board examinations?". The authors performed a comparative analysis of various large language models (LLMs) to assess their accuracy and reliability in answering urological knowledge-based questions and revealed a correlation between question complexity and the performance of LLMs, underscoring the importance of conducting further research in specific subdomains of urology.

In addition to the evaluation presented in this paper, we would like to provide further insights into the implications and future directions. First, the findings of this study indicated that ChatGPT-4 and Bing AI consistently outperformed ChatGPT-3.5 in terms of RoCA scores. However, the reliability of responses across multiple rounds varied. This suggests the need for careful assessment of the reliability of LLM-generated responses in the context of medical education and knowledge acquisition. Second, the study uncovered a consistent trend across all three LLMs, indicating a decrease in test accuracy as question complexity increased. This underscores the importance of training LLMs on a wide range of medical literature and resources to enhance their performance in tackling complex questions. Third, the study's findings suggest that the adaptive learning capacity of LLMs may have limitations. Therefore, there is a significant need for continuous updates, ongoing training, and active maintenance of LLMs to ensure their reliability and effectiveness in acquiring medical knowledge. Furthermore, the study highlights significant concerns regarding the quality and consistency of responses generated by LLMs. It underscores the necessity for additional research to comprehensively evaluate the reliability and the reasoning quality of responses generated by LLMs, especially in the context of medical education and knowledge assessment.

In conclusion, the authors' research enhances our understanding of the potential of LLMs in medical education and clinical practice, highlighting the need for further research to assess the performance of LLMs in specific subdomains within urology and other medical disciplines.

Author contributions JJW and XY: contributed the writing of this letter.

Funding This study received no funding.

Availability of data and materials Not applicable.

Declarations

Conflict of interest The authors declare that they have no competing interests.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Reference

1. Kollitsch L, Eredics K, Marszalek M et al (2024) How does artificial intelligence master urological board examinations? A comparative analysis of different Large Language Models' accuracy and reliability in the 2022 In-Service Assessment of the European Board of Urology. World J Urol 42(1):20

This comment refers to the article available online at https://doi.org/10.1007/s00345-023-04749-6.

✉ Xing Yun
13306716716@163.com

[1] Department of Urology, The First People's Hospital of Xiaoshan District, Xiaoshan Affiliated Hospital of Wenzhou Medical University, Hangzhou, China

[2] Department of Medical Insurance, The First People's Hospital of Xiaoshan District, Xiaoshan Affiliated Hospital of Wenzhou Medical University, Hangzhou, China

Published online: 26 February 2024