https://doi.org/10.1590/1806-9282.20231362

Performance of ChatGPT-4 in answering questions from the Brazilian National Examination for Medical Degree Revalidation

Somsri Wiwanitkit[1], Viroj Wiwanitkit[2*]

Dear Editor,

We follow the topic entitled "Performance of ChatGPT-4 in answering questions from the Brazilian National Examination for Medical Degree Revalidation[1]." The purpose of this study was to evaluate ChatGPT-4's performance in answering the 2022 Brazilian National Examination for Medical Degree Revalidation (Revalida) and its potential as a tool for providing feedback on the examination's quality. Two independent physicians entered all examination questions into ChatGPT-4 and compared their responses to the test solutions, determining whether they were adequate, inadequate, or indeterminate. Consensus was used to resolve disagreements. The study also used statistical analysis to evaluate performance across medical themes and to eliminate queries.

In the Revalida examination, ChatGPT-4 correctly answered 71 (87.7%) of the questions and mistakenly answered 10 (12.3%). The proportion of correct responses did not change statistically significantly across medical themes. However, in nullified questions, the model had a lower accuracy of 71.4%, and there was no statistical difference between the non-nullified and nullified groups. The reliance on the judgments of only two independent physicians to evaluate the accuracy of ChatGPT-4 is a potential weakness of this study. This raises the likelihood of subjective bias in their evaluations. Furthermore, the study does not provide extensive information on the criteria used to categorize the model's replies as adequate, inadequate, or uncertain, which may impair the evaluation's credibility.

Furthermore, the study does not provide extensive information on the criteria used to categorize the model's replies as adequate, inadequate, or uncertain, which may impair the evaluation's credibility. Furthermore, the study does not investigate the reasons for ChatGPT-4's wrong answers, which could have provided useful insights for enhancing the model's performance. Furthermore, the study does not address the potential constraints or obstacles of evaluating a medical examination utilizing a broad language model like ChatGPT-4. Overall, while the study gives some insights into ChatGPT-4's competence in answering the Revalida examination, the study's small number of evaluators and insufficient information on evaluation criteria are significant weaknesses. More studies with a larger and more diverse sample are needed. Modern approaches and a large training set are needed to remove bias and errors from chatbots[2,3]. This is due to the possibility of issues arising when relying solely on a huge data source. Employing chatbots poses ethical questions since it could lead to unforeseen or undesirable effects. To prevent the dissemination of harmful ideas and incorrect information, ethical controls and restrictions must be put in place as artificial intelligence language models advance[4].

AUTHORS' CONTRIBUTIONS

SW: Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **VW**: Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

[1]Private Academic and Editorial Consultant – Bangkok, Thailand.
[2]Saveetha University, Saveetha Institute of Medical and Technical Sciences, Saveetha Medical College – Chennai, India.
*Corresponding author: wviroj@yahoo.com

Conflicts of interest: the authors declare there is no conflicts of interest. Funding: none.
Received on October 05, 2023. Accepted on October 22, 2023.

REFERENCES

1. Gobira M, Nakayama LF, Moreira R, Andrade E, Regatieri CVS, Belfort R Jr. Performance of ChatGPT-4 in answering questions from the Brazilian National Examination for Medical Degree Revalidation. Rev Assoc Med Bras (1992). 2023;69(10):e20230848. https://doi.org/10.1590/1806-9282.20230848
2. Choi HS, Song JY, Shin KH, Chang JH, Jang BS. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. Radiat Oncol J. 2023;41(3):209-16. https://doi.org/10.3857/roj.2023.00633
3. Kleebayoon A, Wiwanitkit V. Comment on "how may ChatGPT impact medical teaching?" Rev Assoc Med Bras (1992). 2023;69(8):e20230593. https://doi.org/10.1590/1806-9282.20230593
4. Kleebayoon A, Wiwanitkit V. Artificial intelligence, chatbots, plagiarism and basic honesty: comment. Cell Mol Bioeng. 2023;16(2):173-4. https://doi.org/10.1007/s12195-023-00759-