





Accuracy of ChatGPT in answering asthma-related questions

Hinpetch Daungsupawong¹, Viroj Wiwanitkit²

The publication on "Evaluation of the accuracy of ChatGPT in answering asthma-related questions"⁽¹⁾ is interesting. That study, which evaluated the quality of ChatGPT's responses to asthma-related questions, is noteworthy because it demonstrates the potential and limitations of large language model (LLM) in interacting with both medical professionals and the general public. However, thorough study exposes statistical limitations, confounding factors, and points that must be reinterpreted in order to broaden the conversation.

Although a Likert scale and content validity coefficient (CVC) were employed to test interrater consistency, the limited number of questions and six raters may not accurately reflect the diversity of real-world situations. Using more powerful statistical methods, such as the intraclass correlation coefficient (ICC), may improve interrater reliability. Furthermore, rather than merely reporting averages or CVC numbers, tests capable of meaningfully comparing subgroups should be used to analyze disparities between physician and patient opinions.

The perceptions of medical professionals and the general population differ significantly. Physicians may anticipate detailed responses that address guidelines and empirical data, whereas the general public may prefer simple, accessible explanations. This means that the "quality" score is based on the rater's expectations rather than the actual correctness of the data. Furthermore, raters' familiarity with the LLM could be a confounder in the perceived quality of the response.

Given that ChatGPT obtained a score of 2-3 from professionals but a high CVC from laypeople, it appears that the model has the capacity to communicate basic asthma knowledge to patients but lacks the depth required for academic or complex patient treatment. According to a new interpretation, ChatGPT's strength rests in its role as a supplementary health communication tool, rather than in clinical decision-making itself.

This study raises further questions, such as whether employing personalized prompts will enable ChatGPT to give better guideline-based responses. How would real patients with chronic asthma rank the quality of their responses compared to laypeople? Could merging expert and patient assessments result in a new criterion for determining an AI's "medical quality"? Would comparing ChatGPT to other LLMs, such as Claude or Gemini, affect the results for the same question?

FINANCIAL SUPPORT

None.

AUTHOR CONTRIBUTIONS

HP: conception, drafting, reviewing, and approval of the final version of the manuscript. VW: conception, supervision, and approval of the final version of the manuscript.

CONFLICTS OF INTEREST

None declared.

REFERENCES

1. Cerqueira BP, Leite VCDS, França CG, Leitão Filho FS, Faresin SM, Figueiredo RG, et al. Evaluation of the accuracy of ChatGPT in answering asthma-related questions. J Bras Pneumol. 2025;51(3):e20240388. <https://doi.org/10.36416/1806-3756/e20240388>
2. Department of Research Analytics, Saveetha Dental College and Hospitals, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, India.

1. Private Academic Consultant, Phonhong, Lao People's Democratic Republic.

2. Department of Research Analytics, Saveetha Dental College and Hospitals, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, India.

Authors' Reply

Bruno Pellozo Cerqueira¹, Vinicius Cappellette da Silva Leite¹, Carla Gonzaga França¹, Fernando Sergio Leitão Filho², Sônia Maria Faresin², Ricardo Gassmann Figueiredo³, Andrea Antunes Cetlin⁴, Lilian Serrasqueiro Ballini Caetano², José Baddini-Martinez²

We read with great interest the correspondence regarding our recently published article and would sincerely like to thank the authors for their thoughtful and insightful comments. We greatly appreciate the opportunity to discuss the scope and implications of our work further.

We recognize that, like any exploratory study, our study has inherent limitations, some of which we addressed in the original manuscript. We intentionally limited the number and the wording of questions to keep the article concise and focused on the main asthma-related issues.

Regarding the statistical approach, our team, together with the statistical advisors, selected the content validity coefficient (CVC) to assess agreement among evaluators. We determined that this method was appropriate for the objectives of our study and that it provided a reliable measure of inter-rater agreement.

We also agree that the perceptions of medical professionals and laypeople may differ considerably, and we recognize this as the primary limitation of our study. To incorporate evaluations from the general population, it would have been necessary to conduct pre- and post-tests, requiring a methodology distinct

from what was proposed. The objective of our study, however, was to assess the perspectives of physicians experienced in managing asthma patients in both private and public outpatient clinics, focusing on what they consider essential for patients to understand about the disease. While we acknowledge that this approach is subjective and limited, we still regard it as valuable data that adds to the discussion.

The additional questions raised represent promising directions for future research. The decision to avoid highly specific prompts was made to simulate real-life interactions better; however, studies utilizing personalized prompts may produce different outcomes. Involving real patients with chronic asthma and integrating both expert and patient assessments could establish new and meaningful criteria for evaluating AI-generated medical information. Furthermore, comparative analyses of various large language models, including those developed for scientific or medical applications such as OpenEvidence, constitute an important next step in this field.

The authors' valuable contribution is appreciated. Their observations highlight significant avenues for further research, and such scientific dialogue enhances the understanding and development of large language model applications in the medical field.

1. Escola Paulista de Medicina, Universidade Federal de São Paulo, São Paulo (SP) Brasil.

2. Divisão de Pneumologia, Escola Paulista de Medicina, Universidade Federal de São Paulo, São Paulo (SP) Brasil.

3. Divisão de Pneumologia, Universidade Estadual de Feira de Santana, Feira de Santana (BA) Brasil.

4. Divisão de Pneumologia, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto (SP) Brasil.