

Assessment of the diagnostic accuracy of various artificial intelligence platforms: new technologies, better diagnostic accuracy?

Avaliação da precisão diagnóstica de diversas plataformas de inteligência artificial: novas tecnologias, melhor precisão diagnóstica?

Cindy Gomes^{1*}, Pedro Azevedo², and Ana Isabel Duarte¹

¹Pediatrics Department, Unidade Local de Saúde do Médio Tejo, Torres Novas; ²USF Pulsar, Unidade Local de Saúde de Coimbra, Coimbra, Portugal

Dear Editor,

Artificial intelligence (AI) stands as a transformative force across various domains in our daily lives, profoundly influencing numerous professional fields^{1,2}. From virtual assistants that streamline everyday tasks to sophisticated data analysis systems, artificial intelligence is becoming increasingly part of our daily activities. Medicine is no exception, and the gradual incorporation of AI into our practice seeks to mitigate its limited experience and potential errors within this realm².

These platforms hold the promise of revolutionizing clinical practice by enhancing diagnosis, treatment, and patient management. They could be considered as diagnostic aids, serving as a kind of “second opinion” in clinical settings. However, uncertainties and reservations persist regarding their efficacy and how they might assist medical professionals in their work.

In order to better understand whether AI platforms could potentially surpass human performance in diagnostic tasks, the authors based their opinion on a study that aimed to assess the viability of these technologies as reliable diagnostic tools. For this purpose, a selection of 100 clinical cases (including medical history, signs and symptoms, and the results of complementary diagnostic tests) published in the *Portuguese Journal*

of *Pediatrics* between 2019 and 2023 was analyzed and entered into three AI platforms to determine the correct diagnosis for each scenario. The results (Fig. 1) indicated that ChatGPT® achieved the highest diagnostic accuracy, with a success rate of 64.77%, followed by Bing®. In contrast, Gemini® exhibited the lowest level of diagnostic precision. Based on these findings, ChatGPT® outperformed the other platforms in diagnostic accuracy. However, when considering error rates, the performance across all three platforms displayed relatively comparable results.

There is no doubt that AI has the potential to become a valuable asset in the field of medicine. Nevertheless, it is essential to acknowledge that it still has certain limitations. The effectiveness and accuracy of its interventions are intrinsically tied to the quality and breadth of the information provided by the user^{1,3}.

In other words, the more objective and detailed the collection of medical history, the higher the likelihood of success for the platform.

However, there remains a crucial “component” that cannot be encapsulated in words: clinical intuition. This limitation represents a fundamental drawback of this technology, not only in healthcare but across various domains.

*Correspondence:

Cindy Gomes

E-mail: cindycarvalhogomes@hotmail.com

Received: 29-10-2024

Accepted: 30-12-2024

<https://pjp.spp.pt>

Available online: 20-03-2025

Port J Pediatr. 2025;56(4):247-248

DOI: 10.24875/PJP.24000086

2184-3333 / © 2024 Portuguese Society of Pediatrics. Published by Permayer. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

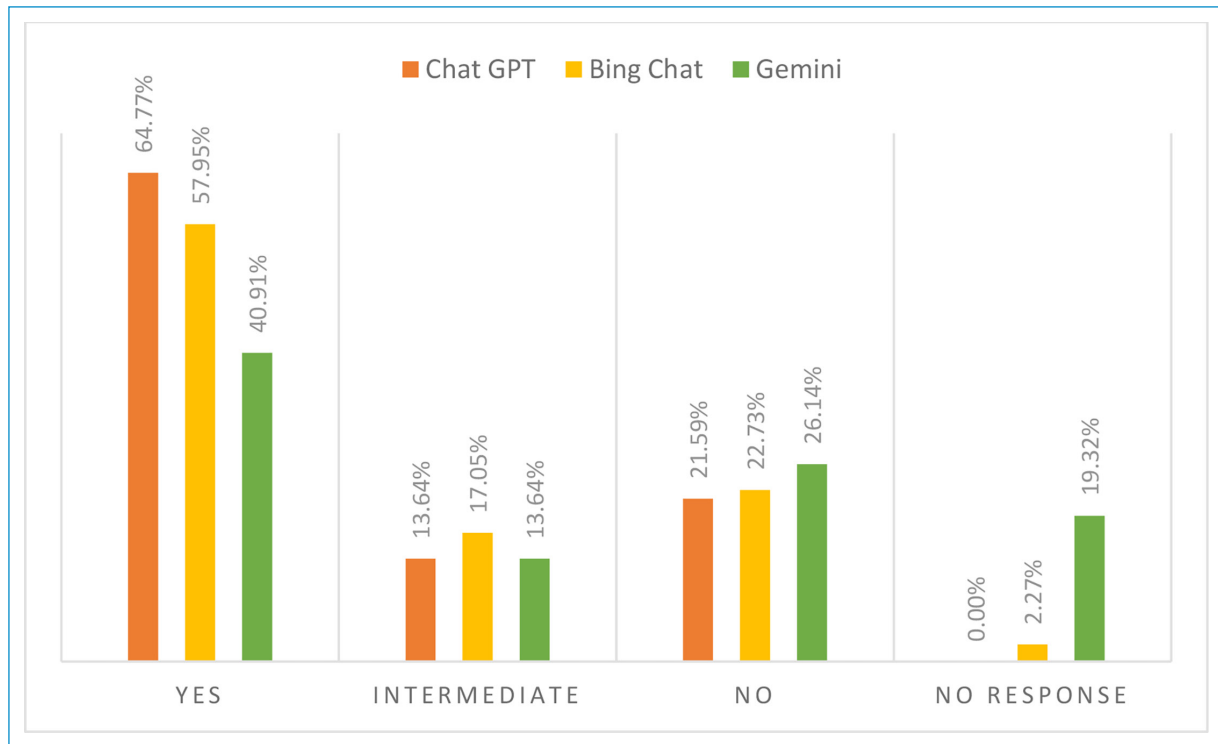


Figure 1. Distribution of responses across the three artificial intelligence platforms.

No matter how advanced AI becomes, can it ever fully replace physicians until it overcomes this limitation, or will it remain a complementary tool in clinical practice? To what extent is it justifiable to rely on a tool whose diagnostic capability depends solely on the most comprehensive objective data, while disregarding the sensory and intuitive inputs integral to a clinical consultation? Furthermore, from an ethical standpoint, is it acceptable or viable to use a platform with such diagnostic capabilities, knowing it could potentially fail in nearly a third of cases?

These questions remain unanswered, and only time will reveal whether these platforms will ever replace healthcare professionals or whether they will remain as auxiliary tools to enhance clinical outcomes.

Author contributions

C. Gomes: conceptualization; formal analysis; investigation; methodology; resources; writing. P. Azevedo: formal analysis; methodology; resources; writing. Ana I. Duarte: conceptualization; supervision; validation; writing.

Funding

None.

Conflicts of interest

None.

Ethical considerations

Protection of humans and animals. The authors declare that no experiments involving humans or animals were conducted for this research.

Confidentiality, informed consent, and ethical approval. The study does not involve patient personal data nor requires ethical approval. The SAGER guidelines do not apply.

Declaration on the use of artificial intelligence. The authors declare that no generative artificial intelligence was used in the writing of this manuscript.

References

1. Choi IH, Seo GH, Park JY, Kim YM, Cheon CK, Kim YM, et al. Evaluation of users' level of satisfaction for an artificial intelligence-based diagnostic program in pediatric rare genetic diseases. *Medicine* [Internet]. 2022 Jul 15 [cited 2024 Nov 2];101(28):E29424. Available from: <https://pubmed.ncbi.nlm.nih.gov/35838999/>
2. Uchikov P, Khalid U, Vankov N, Kraeva M, Kraev K, Hristov B, et al. The Role of Artificial Intelligence in the Diagnosis and Treatment of Ulcerative Colitis. *Diagnostics (Basel)* [Internet]. 2024 May 1 [cited 2024 Nov 2];14(10). Available from: <https://pubmed.ncbi.nlm.nih.gov/38786302/>
3. Soellner M, Koenigstorfer J. Compliance with medical recommendations depending on the use of artificial intelligence as a diagnostic method. *BMC Med Inform Decis Mak* [Internet]. 2021 Dec 1 [cited 2024 Nov 2];21(1):236. Available from: <https://pmc.ncbi.nlm.nih.gov/articles/PMC8344186/>