












# Evaluation of the accuracy of ChatGPT in answering asthma-related questions

Bruno Pellozo Cerqueira<sup>1</sup>, Vinicius Cappellette da Silva Leite<sup>1</sup>,  
Carla Gonzaga França<sup>1</sup>, Fernando Sergio Leitão Filho<sup>2</sup>, Sonia Maria Faresin<sup>2</sup>,  
Ricardo Gassmann Figueiredo<sup>3</sup>, Andrea Antunes Cetlin<sup>4</sup>,  
Lilian Serrasqueiro Ballini Caetano<sup>2</sup>, José Baddini-Martinez<sup>2</sup>

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.

**Submitted:** 25 November 2024.

**Accepted:** 19 May 2025.

Study carried out at the Escola Paulista de Medicina, Universidade Federal de São Paulo, São Paulo (SP) Brasil.

## ABSTRACT

**Objective:** To evaluate the quality of ChatGPT answers to asthma-related questions, as assessed from the perspectives of asthma specialists and laypersons. **Methods:** Seven asthma-related questions were asked to ChatGPT (version 4) between May 3, 2024 and May 4, 2024. The questions were standardized with no memory of previous conversations to avoid bias. Six pulmonologists with extensive expertise in asthma acted as judges, independently assessing the quality and reproducibility of the answers from the perspectives of asthma specialists and laypersons. A Likert scale ranging from 1 to 4 was used, and the content validity coefficient was calculated to assess the level of agreement among the judges. **Results:** The evaluations showed variability in the quality of the answers provided by ChatGPT. From the perspective of asthma specialists, the scores ranged from 2 to 3, with greater divergence in questions 2, 3, and 5. From the perspective of laypersons, the content validity coefficient exceeded 0.80 for four of the seven questions, with most answers being correct despite a lack of significant depth.

**Conclusions:** Although ChatGPT performed well in providing answers to laypersons, the answers that it provided to specialists were less accurate and superficial. Although AI has the potential to provide useful information to the public, it should not replace medical guidance. Critical analysis of AI-generated information remains essential for health care professionals and laypersons alike, especially for complex conditions such as asthma.

**Keywords:** Asthma; Artificial intelligence; Pulmonologists.

## INTRODUCTION

Artificial intelligence (AI) is a broad term referring to the ability of a computer system to simulate human intelligent behavior with a minimum of human intervention.<sup>(1)</sup> Although the use of the term AI is currently on the rise, the term has been used since the middle of the last century.<sup>(2)</sup>

ChatGPT, a generative pre-trained transformer developed by OpenAI, is currently one of the most widely used AI tools. ChatGPT is a natural language processing model trained on a variety of text data, being capable of generating human-like responses within seconds.<sup>(3)</sup> Its accessibility and ease of use have made it a subject of study in various fields of medicine.<sup>(4)</sup>

Asthma is one of the most common noncommunicable diseases, affecting over 300 million people worldwide. Because asthma is such a common disease, it is not unusual to hear a patient say that they have asthma on the basis of what they read on the internet, which is often superficial and inaccurate.<sup>(5)</sup> Despite its widespread occurrence, asthma is a disease whose management is complex and involves critical steps, beginning with proper diagnosis and disease staging.<sup>(6)</sup> This complexity often raises questions even among health care professionals,

who frequently rely on internet sources for quick access to relevant information.

Given the timeliness and relevance of this topic, the objective of the present study was to formulate questions addressing various aspects of asthma and pose them to ChatGPT, assessing the quality of the responses from two perspectives: those intended for laypersons and those intended for asthma specialists.

## METHODS

Two of the authors of the present study developed twenty-one questions addressing various aspects of asthma and then selected seven that they considered to be the most important and most commonly asked when consulting ChatGPT (Table 1). The two aforementioned authors have extensive experience in asthma management. They formulated the questions using the GINA as a reference. To obtain the most accurate answers, the paid version of ChatGPT (version 4) was used. The questions were asked between May 3, 2024 and May 4, 2024. To ensure the uniformity of the answers provided by ChatGPT, the questions were asked in the same format, with a request for answers to be approximately two pages long, thus guaranteeing

### Correspondence to:

Bruno Pellozo Cerqueira. Escola Paulista de Medicina, Universidade Federal de São Paulo, Rua Botucatu, 740, Vila Clementino, CEP 04023-062, São Paulo, SP, Brasil.

Tel.: 55 11 3385-4343. E-mail: bruno.pellozo@unifesp.br

Financial support: None.

consistent content. To reflect how the general population uses large language models, specific prompts were avoided. Since ChatGPT can retain information from previous interactions, the option not to save any of the conversations was selected, and the history was deleted after each response in order to minimize potential bias. Each question was asked twice at consecutive times after the previous chat had been cleared in order to assess the reproducibility of the answers.

After data collection, the answers were sent to specialists in pulmonology and asthma, along with a form, for independently assessing the reproducibility and quality of the answers. The experts were instructed to assess the answers on the basis of current guidelines and updates, particularly GINA guidelines.

A total of six experts evaluated the seven answers provided by ChatGPT from two perspectives: ChatGPT answers intended for a layperson; and ChatGPT answers intended for an asthma specialist. Each answer (item) was scored from 1 to 4 on a Likert scale, as follows: 1-totally correct; 2-correct but insufficient; 3-contains information that is correct and information that is incorrect; and 4-totally incorrect. The level of agreement among the six experts was analyzed by calculating the content validity coefficient (CVC)<sup>(7)</sup> for each item, as follows:

$$CVC = \frac{\left[ \left( \frac{\sum x}{J} \right) \right]}{\sqrt{mx}} - p$$

where  $x$  represents the mean scores;  $J$  represents the total number of judges (or experts, i.e., 6);  $\sqrt{mx}$  represents the highest possible score; and  $p$  represents the bias, which was calculated as follows:

$$p = \left( \frac{1}{J} \right)^J$$

As a rule, the cutoff point for acceptable item validity is 0.80 in content validity studies of scales measuring psychological phenomena. However, this is not applicable to all contexts, and we did not want to impose a fixed cutoff point in the present context. The program GraphPad Prism, version 8.0 (GraphPad Software, Inc., San Diego, CA, USA) was used in order to create Figures 1 and 2.

## RESULTS

Regarding the reproducibility of the answers provided by ChatGPT when questions were asked repeatedly, the expert panel in the present study considered that more than 85% of the answers were consistent. Those who considered that some of the questions lacked reproducibility attributed it to variability or misclassification of certain topics.

The results of the assessments are shown in Figure 1 (from the perspective of asthma specialists) and Figure 2 (from the perspective of laypersons). None of the experts assigned a score of 4 to any of the answers that they analyzed. From the perspective of

**Table 1.** Asthma-related questions asked to ChatGPT.

Number	Question
1	What is asthma?
2	How to diagnose asthma?
3	How is asthma severity classified?
4	How is asthma control classified?
5	What is the pharmacological treatment for asthma?
6	Is there a cure for asthma?
7	What are the risk factors for asthmatic patients experiencing poor outcomes in the future?

asthma specialists, the most prevalent scores were 2 (correct but insufficient) and 3 (contains information that is correct and information that is incorrect) across all questions, with question 3 receiving the highest proportion of low scores (100% scored 3). From the perspective of laypersons, a score of 1 (totally correct) was the most common, accounting for 55% of all possible answers. Question 1 had the highest number of responses that received a score of 1 (five of six).

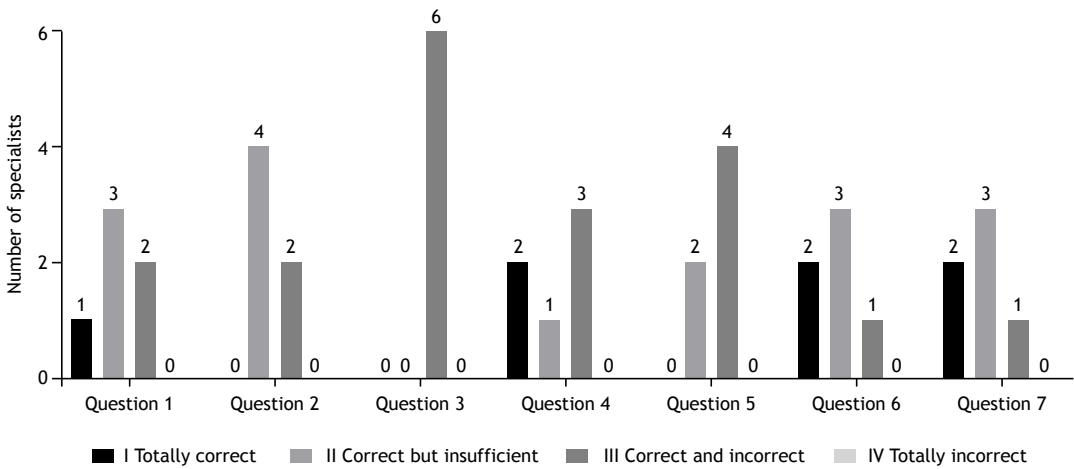
Table 2 shows the level of agreement among the six experts regarding the suitability of the answers provided by ChatGPT from the perspectives of asthma specialists and laypersons. From the perspective of asthma specialists, CVC values were satisfactory for questions 6 and 7, and reasonable for questions 1 and 4. For questions 2, 3, and 5, however, the experts disagreed with regard to the suitability of the answers provided by ChatGPT. For answers analyzed from the perspective of laypersons, there was greater agreement among the experts in comparison with those assessed from the perspective of asthma specialists, with CVC values exceeding 0.8 for four items: 1, 2, 6, and 7. Lower CVC values were observed for items 3, 4, and 5.

## DISCUSSION

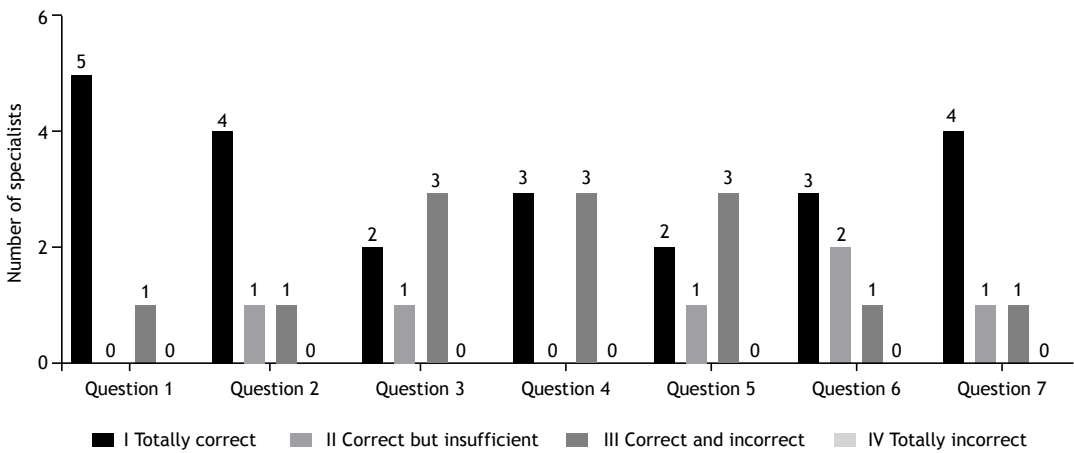
The widespread availability of information on the internet has resulted in a growing reliance on knowledge sources without adequate critical analysis, even among health care professionals. This trend presents significant challenges, particularly when clinical management is complex, as is the case with asthma.

Despite being a common condition, asthma requires a multidimensional approach for effective management. The latest GINA report, published in 2024, outlines two potential therapeutic pathways organized into steps, emphasizing the complexity of treatment.<sup>(8)</sup> Despite the fact that most pulmonologists are well-equipped to manage disease effectively, many patients are treated by primary care physicians. This is due to the high prevalence of asthma. Primary care physicians often lack specialized and comprehensive training in asthma care, and this can impact the quality of care provided to patients.

Reddel et al.<sup>(9)</sup> evaluated the degree of heterogeneity in the diagnosis and management of asthma and COPD in a cohort of more than 11,000 patients. The



**Figure 1.** Expert opinion on the quality of ChatGPT answers from the perspective of asthma specialists.



**Figure 2.** Expert opinion on the quality of ChatGPT answers from the perspective of laypersons.

**Table 2.** Agreement among judges regarding the suitability of ChatGPT answers from the perspectives of asthma specialists and laypersons.

Question	CVC	
	Specialist	Layperson
1	0.71	0.92
2	0.67	0.88
3	0.50	0.71
4	0.71	0.75
5	0.58	0.71
6	0.79	0.83
7	0.79	0.88

CVC: content validity coefficient.

findings revealed significant variability in diagnosis and treatment, with evidence of under or overtreatment in relation to disease severity. Notably, approximately 50% of those patients were managed in primary care, a proportion that is likely similar to or even lower than that observed in Brazil. Although there is a lack of specific data on this topic, it is plausible that primary care physicians are increasingly relying on AI tools to manage complex conditions such as asthma.

Our analysis revealed that although ChatGPT frequently provided correct information, it also made errors or offered insufficient responses for the required level of expertise. Although ChatGPT did not achieve a CVC with consistently positive concordance values, it can still be useful if the information that it provides is critically evaluated.

The present study also focused on evaluating content generated by ChatGPT from the perspective of a layperson, given the growing trend of self-diagnosis and the increasing search for health information online. Before the widespread use of AI, Google was the primary source for such inquiries, being informally referred to as "Dr. Google." Today, given the easy access to AI-powered tools, it is inevitable that patients will turn to such platforms for information. In a study involving 607 participants, approximately 80% expressed willingness to use ChatGPT for self-diagnosis.<sup>(10)</sup> This finding underscores the importance of assessing the accuracy of the content that ChatGPT provides to laypersons. The present study shows that, from the perspective of a layperson, the responses were mostly correct but often insufficient, suggesting

the potential of ChatGPT to inform the public. However, it is crucial to emphasize that ChatGPT should not be used as a substitute for a medical diagnosis.

Several studies have evaluated the role of ChatGPT in medical specialties such as urology and oncology, yielding results similar to ours.<sup>(11,12)</sup> Yeo et al.<sup>(11)</sup> assessed the quality of ChatGPT responses using brief prompts, with evaluations carried out by two specialists. They found that although most responses were partially correct, they often lacked completeness and included accurate and inaccurate information. These findings are consistent with those of our study. Notably, the accuracy of the responses varied by topic, such as general concepts, diagnosis, and treatment, with higher accuracy observed in specific areas. This was also evident in our analysis. Ayers et al.<sup>(13)</sup> investigated whether ChatGPT could provide responses comparable in quality and empathy to those of physicians. Interestingly, 80% of patients preferred the responses generated by ChatGPT, suggesting good accuracy and a patient-friendly approach to addressing their questions.

In a recent study conducted in Denmark, the authors analyzed 26 asthma-related questions.<sup>(14)</sup> The results showed that ChatGPT could provide adequate responses, albeit with some inconsistencies. However, because the aforementioned study was a brief report, it lacked methodological details, such as the version of the software used and the dates when the queries were made. Additionally, there was no evaluation of the responses from the perspective of the lay public.

In our study, we observed a significant inconsistency in the responses provided by ChatGPT, particularly regarding the classification of asthma severity (question 3). According to the American Thoracic Society/European Respiratory Society task force<sup>(15)</sup> and several international guidelines, asthma severity should currently be classified retrospectively, on basis of the level of treatment required to achieve disease control and prevent exacerbations. In other words, severity is determined by the dose of inhaled corticosteroids needed to manage symptoms rather than by the mere presence or intensity of those symptoms. However, ChatGPT incorrectly described the classification as follows: "The classification of asthma severity is generally divided into four categories: mild intermittent, mild persistent, moderate persistent, and severe persistent. This categorization helps determine the appropriate treatment regimen and is based on the frequency and intensity of symptoms, nighttime awakenings, the use of short-acting beta-agonists for quick relief, and the impact on normal activities." For example, GINA defines severe asthma as asthma that remains uncontrolled despite high doses of inhaled corticosteroids combined with long-acting bronchodilators, or that requires chronic use of oral corticosteroids. In contrast, mild asthma is characterized by symptom control achieved with low doses of inhaled corticosteroids, such as budesonide at a maximum daily dose of 400 µg. Those definitions are different from those provided by ChatGPT.

There was a high level of consistency in the evaluations made by specialists regarding certain responses. For example, question 6 ("Is there a cure for asthma?") received widespread agreement among reviewers. The response provided by ChatGPT was as follows: "There is currently no cure for asthma. However, the condition can be effectively managed through a combination of treatments and strategies, allowing many individuals with asthma to lead normal and active lives." This is consistent with current medical understanding and was deemed accurate and appropriate by most of the asthma specialists in the present study.

The present study has some limitations. One major limitation is the use of short prompts. Although more specific prompts generally produce higher-quality responses, we chose to use concise commands to replicate everyday usage scenarios. However, this decision may have compromised the quality of the generated content to some extent. Additionally, the evaluation of the content was subjective, given that it involved different specialists conducting the assessments. To reduce this bias, we calculated the CVC and included a substantial number of reviewers, thus strengthening our analysis. Another significant limitation is related to the evaluation of layperson-oriented content by specialists. To assess public understanding more accurately, a different methodology would be necessary, although it would not align with the current study design. The objective of our evaluation was to have experts review the content based on what they consider essential for patients to understand about the disease. Finally, since ChatGPT is a constantly evolving model, the responses provided at a later time may vary.

In conclusion, ChatGPT has the potential to generate informative responses for general audiences, with satisfactory agreement among reviewers in certain areas. However, when evaluated by specialists—particularly regarding more complex clinical concepts—the responses were often interpreted with greater variability and deemed less accurate. Given the known risk of AI-generated "hallucinations," which are plausible but incorrect or misleading pieces of information, it is crucial to emphasize that language models such as ChatGPT should not be used as the sole source of health information. This caution is especially important for managing diseases such as asthma, which require individualized and nuanced care. To promote the safe and effective use of AI tools in clinical practice and health education, we recommend the following: use AI-generated information as a complementary tool rather than a replacement for professional medical advice; health care professionals should guide and supervise the use of these tools, especially when patients or caregivers are involved; developing frameworks for validating and curating AI-generated content may help ensure alignment with current clinical guidelines and reduce the spread of misinformation; and educational strategies should include digital health literacy to empower users to critically evaluate the reliability of AI responses.

Ultimately, although ChatGPT can serve as a starting point for health-related inquiries, human oversight remains essential to maintain the quality and safety of medical information.

## AUTHOR CONTRIBUTIONS

BPC, LSBC, and JBM: designed the study. BPC, VCSL, and CGF: asked ChatGPT the study questions

and wrote the manuscript. FSLF, SMF, RGF, ACVAC, LSBC, and JBM: assessed the answers provided by ChatGPT. BPC: performed the statistical analysis. All authors reviewed and approved the final version of the manuscript.

## CONFLICTS OF INTEREST

None declared.

## REFERENCES

1. Sheikh H, Prins C, Scrijvers E. Artificial Intelligence: Definition and Background. In: Mission AI. Research for Policy. [monograph on the Internet]. Springer; 2023. [cited 2024 Nov 18]. Available from: [https://link.springer.com/chapter/10.1007/978-3-031-21448-6\\_2](https://link.springer.com/chapter/10.1007/978-3-031-21448-6_2)
2. Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. *Gastrointest Endosc.* 2020;92(4):807-812. <https://doi.org/10.1016/j.gie.2020.06.040>
3. OpenAI [homepage on the Internet]. OpenAI; c2015-2025 [updated 2022 Nov 30; cited 2024 Nov 18]. Introducing ChatGPT. Available from: <https://openai.com/index/chatgpt/>
4. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell.* 2023;6:1169595. <https://doi.org/10.3389/frai.2023.1169595>
5. Papi A, Brightling C, Pedersen SE, Reddel HK. Asthma. *Lancet.* 2018;391(10122):783-800. [https://doi.org/10.1016/S0140-6736\(17\)33311-1](https://doi.org/10.1016/S0140-6736(17)33311-1)
6. Kwah JH, Peters AT. Asthma in adults: Principles of treatment. *Allergy Asthma Proc.* 2019;40(6):396-402. <https://doi.org/10.2500/aap.2019.40.4256>
7. Hernández-Nieto R. Contributions to Statistical Analysis: The Coefficients of Proportional Variance, Content Validity and Kappa. Charleston, SC: BookSearch Publishing; 2002. 228 p.
8. Global Initiative for Asthma [homepage on the Internet]. Bethesda: Global Initiative for Asthma; c2024 [cited 2024 Nov 1]. Global Strategy for Asthma Management and Prevention (2019 update). Available from: <https://ginasthma.org/>
9. Reddel HK, Vestbo J, Agustí A, Anderson GP, Bansal AT, Beasley R, et al. Heterogeneity within and between physician-diagnosed asthma and/or COPD: NOVELTY cohort. *Eur Respir J.* 2021;58(3):2003927. <https://doi.org/10.1183/13993003.03927-2020>
10. Shahsavari Y, Choudhury A. User Intentions to Use ChatGPT for Self-Diagnosis and Health-Related Purposes: Cross-sectional Survey Study. *JMIR Hum Factors.* 2023;10:e47564. <https://doi.org/10.2196/47564>
11. Yeo YH, Saman JS, Ng WH, Ting PS, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol.* 2023;29(3):721-732. <https://doi.org/10.3350/cmh.2023.0089>
12. Warren CJ, Payne NG, Edmonds VS, Voleti SS, Choudry MM, Punjani N, et al. Quality of Chatbot Information Related to Benign Prostatic Hyperplasia. *Prostate.* 2025;85(2):175-180 <https://doi.org/10.1002/pros.24814>
13. 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.* 2023;183(6):589-596. <https://doi.org/10.1001/jamainternmed.2023.1838>
14. Høj S, Thomsen SF, Ulrik CS, Meteran H, Sigsgaard T, Meteran H. Evaluating the scientific reliability of ChatGPT as a source of information on asthma. *J Allergy Clin Immunol Glob.* 2024;3(4):100330. <https://doi.org/10.1016/j.jacig.2024.100330>
15. Chung KF, Wenzel SE, Brozek JL, Bush A, Castro M, Sterk PJ, et al. International ERS/ATS guidelines on definition, evaluation and treatment of severe asthma. *Eur Respir J.* 2014;43(2):343-73. <https://doi.org/10.1183/09031936.00202013>