

# PERFORMANCE OF AI MODELS IN PHARMACEUTICAL CLINICAL REASONING: ANALYSIS OF SIMULATED QUESTIONS REGARDING CONSISTENCY, SAFETY, AND CLARITY

Josué Guilherme Lisboa Moura<sup>1,2</sup> , Maria Gabriela Borges Hermes<sup>1,2</sup> ,  
Josiele Dias da Rosa Mossmann<sup>1</sup> , Lauren Feiffer Sanches<sup>1</sup> 

## ABSTRACT

**Introduction:** Widely known and freely accessible natural language models stand out for their simplified interfaces and their widespread use in daily practice by professionals and lay users alike, generating rapid responses based on large volumes of data. However, a limitation is that they also mine grey literature data. In this work, we evaluated the performance of three AI models (ChatGPT, Copilot, and Gemini) in solving simulated clinical questions, considering three domains: consistency, safety, and clarity of responses. **Method:** This exploratory, descriptive, and comparative study included 15 simulated clinical questions, categorized by level of difficulty (low, moderate, and high complexity). Each AI responded to the same questions, and scores were assigned by expert evaluators across the three domains of interest. **Results:** Overall, there were no statistically significant differences among the three models regarding consistency, safety, or clarity ( $p > 0.05$ ), indicating equivalent general performance. All models performed better on low-complexity questions, with highlights for Copilot (87.5% consistency) and ChatGPT (100% safety). In moderate-complexity questions, Copilot achieved the best results (80.95% consistency, 88.10% safety, and 82.54% clarity). However, statistical analysis confirmed that question complexity significantly impacted the Safety domain ( $p = 0.0028$ ). When the questions progressed from low to high difficulty, the ability of the AIs to provide safe responses plummeted significantly. In high-complexity questions, performance declined markedly, especially for ChatGPT (50% consistency and 54.17% safety). Gemini showed greater stability in this group, with 66.67% consistency, 70.83% safety, and 83.33% clarity. Copilot was the only AI to provide a completely correct answer in one high-complexity question. **Conclusion:** We corroborate that AI can be an ally to the clinical pharmacist, especially as technical support for standardized interventions. However, the statistically proven safety risks in highly complex scenarios reinforce that it does not replace the critical and interpretive role of the human professional. The use of these tools should be considered complementary to clinical reasoning.

**Keywords:** *Generative Artificial Intelligence; Patient Safety; Clinical Reasoning.*

Clin Biomed Res. 2026;46:1-9

1 Hospital Santa Ana, Porto Alegre, RS, Brazil.

2 Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil.

### Corresponding author:

Josué Guilherme Lisboa Moura

E-mail: josuemoura@hotmail.com.br

AESC - Hospital Santa Ana. Praça

Simões Lopes Neto, 175, Teresópolis,

Porto Alegre - RS, 91720-440

## INTRODUCTION

The role of artificial intelligence (AI), particularly in advancing personalized medicine, cannot be underestimated. AI has shown promise by offering features such as personalized reminders, analysis of large datasets, and digital tools centered on patient care, which enhance medication safety, especially through the automation of repetitive tasks. However, important barriers remain to its full adoption, including regulatory challenges, data privacy concerns, and the need for continuous ethical oversight<sup>1</sup>.

Natural language models, such as ChatGPT, represent a relevant frontier of this technology, offering coherent textual responses with potential applications in clinical practice<sup>2,3</sup>. Kiyomiya et al.<sup>4</sup> recently demonstrated that a customized version of ChatGPT was able to provide medication guidance

to patients with high concordance with clinical guidelines, highlighting the potential of AI to assist pharmacists in communication and counseling tasks, as it optimizes the collection and analysis of clinical data while freeing professionals to focus on more complex decisions that cannot be automated<sup>5</sup>. In addition to simple tasks, AI may also contribute to the prevention of adverse events and, consequently, promote greater safety in clinical practice<sup>6,7</sup>. This is especially relevant in hospital settings, where the time available for decision-making is limited and the demand for rapid guidance is constant.

The application of AI in clinical pharmacy practice also extends to support for pharmacovigilance, screening for drug interactions, and analysis of patient histories to improve therapeutic protocols, as these tools facilitate the generation of real-time evidence<sup>8,9</sup>. In the scientific literature, the potential of these tools is discussed as support for improving clinical outcomes and the integrated management of medication therapy, especially in hospital settings<sup>10</sup>.

Despite these advances, there are still controversies regarding the potential of AI to replace human clinical reasoning. Although AI models are effective in information retrieval, they present limitations when faced with complex clinical contexts, as these scenarios require adaptive judgment, ethical considerations, and prioritization skills<sup>3</sup>. In this context, AI should be regarded as a complementary tool that requires critical oversight, appropriate training, and well-established supervision protocols<sup>11,12</sup>. Furthermore, its use in clinical practice must be accompanied by careful professional supervision, considering the ethical risks involved, the protection of sensitive data, and the impacts on autonomy in care-related decision-making<sup>1,13</sup>.

Widely known and freely accessible natural language models, such as ChatGPT (OpenAI), Copilot (Microsoft), and Gemini (Google DeepMind), stand out due to their simplified interfaces and widespread everyday use by both healthcare professionals and the general public, generating rapid responses based on large volumes of data, with the drawback that they also mine information from grey literature. However, it remains under debate whether these technologies are capable of replacing human clinical reasoning,

which involves not only technical knowledge but also contextual interpretation and ethical judgment. In this scenario, the present study aimed to comparatively evaluate the consistency, safety, and clarity of responses provided by different AI systems to clinical questions of varying complexity applicable to hospital practice, contrasting their performance with standards of excellence based on the scientific literature.

## METHODS

### Study Design and Scenario Simulation

This is an exploratory, descriptive, and comparative study with a quantitative approach, conducted through the simulation of clinical scenarios applicable to hospital clinical pharmacy practice.

The study evaluated the performance of three open-access AI models regarding their ability to provide objective, safe, and clinically applicable guidance for the routine activities of a multidisciplinary hospital team. The analysis was based on 15 simulated clinical questions (Table 1), developed from real-world pharmaceutical practice scenarios involving topics such as safe medication administration, emergency interventions, and monitoring of adverse events. The questions were based on recurrent situations encountered in the authors' hospital clinical practice, covering different levels of complexity and clinical decision-making.

The number of questions ( $n = 15$ ) was chosen to avoid redundancy and preserve the clarity of the study. A larger number of questions could introduce reader fatigue and hinder comparative analysis among the models. The difficulty level was determined pragmatically, considering the frequency, urgency, and degree of clinical reasoning required in each situation, similarly to real clinical practice. Low-difficulty questions involved direct administration guidance and simple compatibility issues, whereas medium-difficulty questions required interpretation of clinical signs or adjustment of conduct according to patient characteristics. High-difficulty questions demanded more complex clinical judgment, such as interventions for severe adverse effects or decision-making in critical care contexts.

**Table 1:** Classification of the questions according to difficulty level, clinical justification, and source of evidence.

| No. | Question                                                                                                                                                                                                                                                                                      | Difficulty Level | Justification                                               |
|-----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|-------------------------------------------------------------|
| 1   | The patient has a single venous line receiving Lactated Ringer's solution via continuous infusion pump. Considering the prescription of ceftriaxone diluted in 0.9% saline with a final concentration of 10 mg/mL, is it safe to administer both medications simultaneously through a Y-site? | Low              | Well-documented chemical incompatibility <sup>14,15</sup> . |

Continues...

Table 1: Continuation.

| No. | Question                                                                                                                                                                                                                                                                                                                                                               | Difficulty Level | Justification                                                                                                                                                |
|-----|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 2   | Considering that the patient has fragile venous access and no viable vascular access conditions, and dipyrone 1 g/2 mL was prescribed for intravenous bolus administration, is it safe and feasible to administer the medication through another route, such as hypodermoclysis or intramuscular injection?                                                            | Medium           | Evaluation of an alternative route of administration requires clinical judgment <sup>16</sup> .                                                              |
| 3   | Considering the prescription of a polarizing solution with insulin diluted in 50% glucose, what is the appropriate conduct for glycemic control in this patient?                                                                                                                                                                                                       | Low              | Basic glycemic control with standard monitoring <sup>17</sup> .                                                                                              |
| 4   | Considering the prescription of meropenem diluted in 0.9% saline and polymyxin diluted in 5% glucose for administration at the same time, is it safe to administer them through the same lumen of the venous line?                                                                                                                                                     | Low              | Antibiotic compatibility is well-established knowledge <sup>18</sup> .                                                                                       |
| 5   | Given the request for oral vancomycin for the treatment of pseudomembranous colitis, and considering that the oral formulation is not available in Brazil, is it possible to dispense the vancomycin vial originally intended for parenteral administration for oral use?                                                                                              | Low              | Oral use of vancomycin is an established practice <sup>19</sup> .                                                                                            |
| 6   | Considering a patient prescribed injectable omeprazole during fasting at 6:00 a.m., is it possible, for greater comfort, to recommend administration upon awakening (8:00 a.m.) or even at noon?                                                                                                                                                                       | Medium           | Evaluation of fasting status and adjustment of administration schedules <sup>20</sup> .                                                                      |
| 7   | Considering an elderly patient receiving a hyperproteic and hypercaloric enteral diet via enteral feeding tube, with Parkinson's disease, hypertension, and use of levofloxacin, levodopa, atenolol, thiamine, and sodium picosulfate, are there significant interactions between dietary nutrients and the medications that may impact treatment efficacy?            | Medium           | Nutritional impact on medications requires more detailed analysis <sup>21</sup> .                                                                            |
| 8   | Considering a 40-year-old male patient with hypertension, dysphagia, and recent gastrostomy, which beta-blocker would be most appropriate for this patient, taking into account the need for adaptation to enteral administration?                                                                                                                                     | Medium           | Selection of a beta-blocker considering enteral administration route <sup>22</sup> .                                                                         |
| 9   | Considering a 50-year-old female patient with previous use of levonorgestrel, sertraline, and captopril, hospitalized for hypertensive crisis with complications and receiving prednisone to treat a persistent dry cough without improvement, while physiotherapy evaluations revealed no mucus secretion, is there a possibility that the cough is drug-induced?     | Medium           | The relationship between cough and adverse effects of angiotensin-converting enzyme inhibitors (ACE inhibitors) requires clinical experience <sup>23</sup> . |
| 10  | Considering a patient with pulmonary sepsis receiving intravenous daptomycin and oral sulfamethoxazole-trimethoprim, without clinical improvement on the fourth day, is there any possible pharmaceutical intervention that could optimize the therapeutic response or identify potential causes of treatment failure?                                                 | High             | Therapeutic intervention in infectious treatment failure requires advanced clinical reasoning <sup>24</sup> .                                                |
| 11  | Considering a patient receiving midazolam and fentanyl diluted in 0.9% saline, norepinephrine diluted in 0.9% saline, and polymyxin diluted in 5% glucose, with a double-lumen catheter in the right subclavian vein and one venous line in the left basilic vein, what would be the best suggestion for simultaneous administration of these medications?             | High             | Management of multiple drugs through simultaneous lumens is complex <sup>25,26</sup> .                                                                       |
| 12  | Considering a patient undergoing sustained low-efficiency dialysis (SLED), receiving vancomycin and amikacin at the same scheduled time, what guidance should be provided to the healthcare team to ensure effective administration of these medications?                                                                                                              | High             | Adjustment of antibiotics in dialysis patients is highly complex <sup>27</sup> .                                                                             |
| 13  | Considering a Black diabetic patient with sickle cell anemia admitted to the ICU for treatment of diabetic ketoacidosis, with HbA1c of 5.1%, and receiving regular insulin, NPH insulin, fosfomycin, and clonazepam, what pharmaceutical intervention could be performed to optimize or contribute to glycemic control?                                                | High             | Glycemic control in sickle cell anemia requires specific knowledge <sup>28</sup> .                                                                           |
| 14  | Considering a patient with dysphagia, admitted to the ICU under mechanical ventilation and with opioid-related constipation, and considering that bisacodyl tablets, sodium picosulfate, mannitol, and mineral oil are available in the institution, which of these laxatives could be recommended with the lowest risk, considering the patient's clinical condition? | Medium           | Safe evaluation of laxatives in critically ill patients <sup>29</sup> .                                                                                      |
| 15  | Considering a patient with anticipated discharge, receiving enteral nutrition via feeding tube, requiring potassium replacement with potassium chloride, and with the hospital options being tablets, oral solution, or parenteral solution, which of these administration forms presents the lowest risk for the patient?                                             | Medium           | Selection of potassium formulation for enteral tubes involves moderate clinical risk <sup>30</sup> .                                                         |

Legend: SNE (nasogastric tube); ICU (intensive care unit); NPH (Neutral Protamine Hagedorn insulin); 5% glucose solution; 0.9% saline solution; Y-site (simultaneous "Y" administration); SLED (Sustained Low-Efficiency Dialysis).

## AI Models and Standardized Prompts

The following natural language models were included in the analysis: ChatGPT (OpenAI, GPT-4.1 mini version; available at <https://openai.com/chatgpt>), Copilot (Microsoft, Redmond, WA; available at <https://copilot.microsoft.com/>), and Gemini (Google, 2.0 Flash model; available at <https://gemini.google.com/app>). All models were accessed free of charge, without customizations or additional interventions, in order to ensure representativeness of their spontaneous use by healthcare professionals.

The 15 questions developed encompassed different levels of clinical difficulty, aiming to evaluate the performance of the AI models in both simple situations and contexts requiring more advanced clinical reasoning. All interactions with the models were conducted on the same day (June 24, 2025) in order to avoid temporal bias associated with system updates or momentary variations in algorithmic performance. The prompt used in all interactions was standardized as follows: *“Respond as a hospital clinical pharmacist focused on providing guidance to the multidisciplinary patient care team. Be brief, technical, practical, and direct to the point. Imagine they are urgently waiting on the phone or through a text message. Use short sentences and clear language, without jargon. Prioritize safe medication administration and immediate actions applicable to clinical routine.”* The clinical question was then provided.

## 2.3 Evaluation Criteria

Acting as an expert panel<sup>31</sup>, three clinical pharmacists with hospital experience (mean of 6.3 years practicing as hospital clinical pharmacists) subsequently evaluated the obtained responses. They independently assigned scores based on three main criteria: consistency, impact on patient safety, and clarity of the generated information, even when incorrect. The evaluation process was conducted in a blinded manner; the evaluators did not know which AI model generated each response, thereby preventing confirmation bias. According to a three-point Likert scale, consistency was measured based strictly on the professional and clinical experience of the expert panel, evaluating the practical applicability of the responses in an urgent real-world scenario, rather than systematically comparing the outputs against answer keys or standardized medical databases.

According to a three-point Likert scale, consistency was measured by comparing the AI-generated responses with answer keys developed from the scientific literature and established clinical protocols. For each item, 2 points were assigned to completely correct responses, 1 point to partially correct responses, and 0 points to incorrect responses or

those with potential risk (possibility of clinical harm to the patient or unsafe practice if followed literally). Safety was evaluated according to the potential impact of the guidance on patient integrity, considering international patient safety guidelines. Responses classified as safe and adherent to best practices received 2 points; those presenting mild risk or potential for operational error received 1 point; and responses associated with moderate or severe risk received 0 points. Clarity was assessed considering criteria such as objectivity, accessible language, and direct connection to the question presented. Responses classified as confusing received 1 point; clear responses, 2 points; and very clear responses, 3 points. The data were organized and structured using Python (version 3.10), specifically utilizing the Pandas and NumPy libraries<sup>32-34</sup>, and the final mean score was calculated based on the evaluations of the three reviewers.

## Statistical Analysis

All statistical analyses and data visualizations were performed using the Python programming language (version 3.10), leveraging its specialized scientific computing packages. Data manipulation and structuring were conducted using Pandas (version 2.0)<sup>33</sup> and NumPy (version 1.24)<sup>32</sup>. Non-parametric inference tests, post-hoc procedures, and binary proportion tests were computed using SciPy (version 1.10)<sup>35</sup> and Statsmodels (version 0.14)<sup>34</sup>.

All evaluations were executed considering the individual ratings assigned by the three independent clinical pharmacists, totaling 45 evaluations per domain across 15 questions. Due to the ordinal nature of the Likert scales applied and the non-normal distribution of the scores, confirmed via Shapiro-Wilk test, non-parametric statistical methods were employed. To compare the overall continuous average scores among the three AI models across the evaluated domains of consistency, safety, and clarity, the Friedman test for paired samples was utilized. Effect sizes for the Friedman tests were estimated using Kendall's W. Post-hoc pairwise comparisons between models were conducted using the Wilcoxon signed-rank test with Bonferroni correction applied to adjust p-values for multiple comparisons.

For the binary outcome evaluation, represented by the proportion of completely correct responses assigned a score of 2 in consistency, model performance was compared using Cochran's Q test, followed by post-hoc paired McNemar tests with Bonferroni adjustment. Differences in model performance across question difficulty levels (Low, Medium, and High) were evaluated using the Kruskal-Wallis H test on pooled data. Epsilon-squared was calculated as the

metric for effect size. Post-hoc pairwise comparisons between complexity subgroups were performed using Dunn's test with Bonferroni adjustment. Inter-rater reliability among the three clinical pharmacists was evaluated using a two-way random-effects, absolute agreement, average-measures Intraclass Correlation Coefficient [ICC(2,3)], calculated via the Pingouin (version 0.5) package<sup>36</sup>. Reliability values were interpreted based on standard threshold criteria<sup>37</sup>: poor (<0.50), moderate (0.50-0.75), good (0.75-0.90), and excellent (>0.90). A two-tailed p-value <0.05 was considered statistically significant for all primary analyses.

## RESULTS

Table 2 presents the mean performance results of the three AI models across the evaluated domains: mean consistency, safety, and clarity. A detailed description of the questions and their respective AI-generated responses is available in the supplementary material. Overall, Copilot demonstrated the best mean performance in consistency (76.67%), outperforming Gemini (71.11%) and ChatGPT (66.67%). In terms of safety, Gemini achieved the highest score (83.33%), followed by Copilot (82.22%) and ChatGPT (77.78%). Regarding clarity, Gemini obtained 82.96%, compared with 80.00% for ChatGPT and 77.78% for Copilot. Prior to comparative analyses, inter-rater reliability among the three clinical pharmacists was evaluated, yielding good agreement for consistency (ICC = 0.834), moderate agreement for safety (ICC = 0.714), and poor-to-moderate agreement for clarity (ICC = 0.483). Despite the descriptive numerical variations among the AI tools, the Friedman test revealed no statistically significant differences in continuous median scores for consistency ( $\chi^2(2) = 0.514$ ,  $p = 0.773$ ,  $W = 0.017$ ), safety ( $\chi^2(2) = 1.143$ ,  $p = 0.565$ ,  $W = 0.038$ ), or clarity ( $\chi^2(2) = 1.056$ ,  $p = 0.590$ ,  $W = 0.035$ ). Post-hoc pairwise comparisons confirmed that the models performed equivalently across all domains within this sample size ( $p_{\text{bonf}} > 0.05$ ).

**Table 2:** Comparative performance of artificial intelligence models in solving simulated clinical questions.

| AI Model | Mean Consistency (%) | Safety (%) | Clarity (%) |
|----------|----------------------|------------|-------------|
| ChatGPT  | 66.67                | 77.78      | 80.00       |
| Copilot  | 76.67                | 82.22      | 77.78       |
| Gemini   | 71.11                | 83.33      | 82.96       |

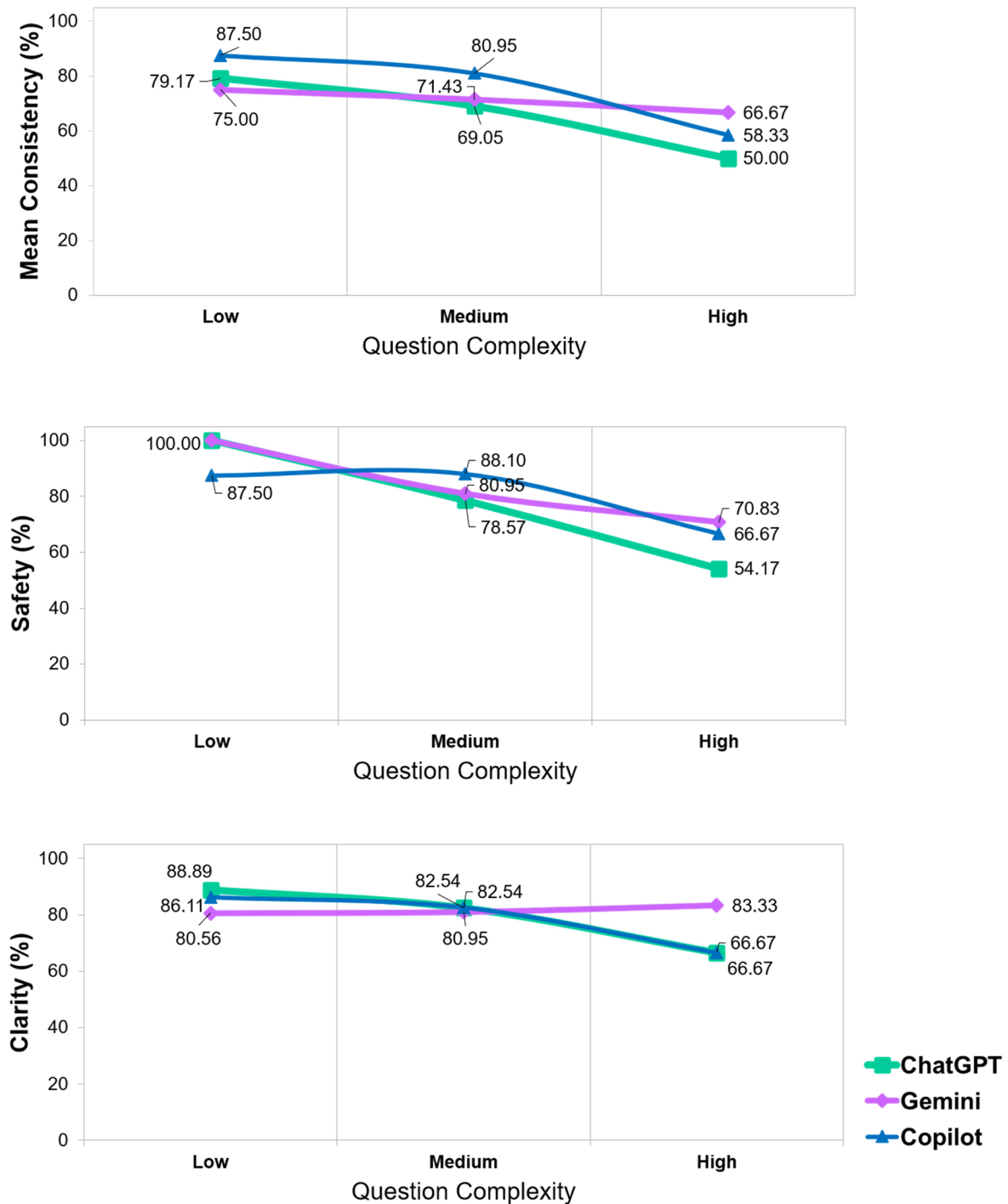
When the data were analyzed according to difficulty level (Figure 1), it was observed that for low-complexity questions, all models provided safe and clear responses, with Copilot standing out in consistency (87.5%) and ChatGPT achieving 100% in safety. For medium-difficulty questions, Copilot maintained the best performance (80.95% consistency, 88.10% safety, and 82.54% clarity). In the group of high-complexity questions, there was a marked decline in performance indices. ChatGPT showed the lowest performance, with 50% consistency and 54.17% safety. Gemini demonstrated 66.67% consistency, 70.83% safety, and 83.33% clarity, maintaining the best overall performance within this group. Figure 1 illustrates these variations among the AI models according to the degree of difficulty, visually highlighting the decline in performance as the questions required greater clinical reasoning capacity. Statistical analysis confirmed that question complexity significantly impacted model performance in the safety domain ( $H(2) = 11.735$ ,  $p = 0.0028$ ,  $\epsilon^2 = 0.232$ ). Dunn's post-hoc test demonstrated a statistically significant drop in safety when moving from low to high complexity questions ( $p_{\text{bonf}} = 0.0021$ ), with a trend toward reduction between medium and high complexity ( $p_{\text{bonf}} = 0.064$ ). Variations across complexity levels were not statistically significant for consistency ( $H(2) = 5.895$ ,  $p = 0.053$ ,  $\epsilon^2 = 0.093$ ) or clarity ( $H(2) = 3.910$ ,  $p = 0.142$ ,  $\epsilon^2 = 0.045$ ).

Table 3 presents the distribution of completely correct responses (maximum consistency score: 2 points) according to AI model and difficulty level. This criterion represents responses fully aligned with the scientific literature and, consequently, with pharmaceutical practices recognized as the gold standard. Copilot was the only AI model that generated a completely correct response for a high-complexity question, specifically Question 13, which addressed the interpretation of glycated hemoglobin (HbA1c) in a patient with sickle cell anemia and diabetic ketoacidosis. However, when evaluating the overall proportion of completely correct responses, Cochran's Q test showed no statistically significant differences among the three models ( $Q = 3.600$ ,  $df = 2$ ,  $p = 0.165$ ; McNemar post-hoc  $p_{\text{bonf}} > 0.05$ ).

**Table 3:** Total number of completely correct responses according to difficulty level and AI model.

| AI Model | Low (n = 4) | Medium (n = 7) | High (n = 4) | Total (n/15) |
|----------|-------------|----------------|--------------|--------------|
| ChatGPT  | 2           | 2              | 0            | 4            |
| Copilot  | 3           | 3              | 1            | 7            |
| Gemini   | 2           | 2              | 0            | 4            |

**Figure 1:** Mean performance of artificial intelligence systems according to the level of difficulty of the simulated clinical questions, based on consistency, safety, and clarity.



## DISCUSSION

The results suggest that, although the models are reasonably effective for direct and routine clinical guidance, their limitations become evident in more complex scenarios, particularly regarding safety and

clinical consistency, as well as in situations requiring clinical reasoning beyond previously acquired knowledge. The lower performance of the models in high-complexity questions indicates a current limitation of these tools for tasks demanding adaptive

reasoning and application to specific clinical contexts, even when they demonstrate good performance in low- or medium-complexity questions.

AI-based language models have the potential to support pharmaceutical practice, but are not capable of replacing human clinical reasoning. Although tools such as ChatGPT, Copilot, and Gemini demonstrated satisfactory performance in terms of clarity and safety, the consistency of their responses proved to be strongly dependent on the level of complexity of the proposed question. This finding is consistent with the observations of Huang et al.<sup>3</sup>, who highlighted limitations of AI models in scenarios requiring adaptive clinical judgment. In the present study, we observed that the evaluated AI systems demonstrated distinct performances regarding consistency, safety, and clarity, with Copilot standing out by achieving the highest number of completely correct responses and the highest mean consistency. These findings indicate that AI performance varies according to the level of complexity.

In low-complexity questions, the tools demonstrated high consistency (mean values of 66.7% for ChatGPT, 76.7% for Copilot, and 69% for Gemini), with low associated risk and clear language. Cases such as the simultaneous administration of ceftriaxone and Lactated Ringer's solution, the oral use of vancomycin in pseudomembranous colitis, and blood glucose monitoring after glucose infusion exemplify this satisfactory technical performance. This pattern suggests that AI systems provide adequate support when clinical protocols are well established and there is less need for contextual interpretation<sup>10</sup>.

On the other hand, as the level of question difficulty increases, there is a progressive decline in the consistency of the responses. In these cases, it is necessary to integrate multiple clinical variables, recognize laboratory limitations in specific populations, and perform therapeutic considerations based on advanced clinical judgment. The failure of AI systems to provide complete or appropriately directed responses in these contexts suggests an important limitation in their capacity for generalization and adaptive reasoning<sup>13</sup>. When the algorithm encounters questions that go beyond predictable patterns or require balancing multiple risks, it tends to adopt an excessively conservative or generic approach, precisely to avoid potential harm. However, this behavior may compromise the problem-solving capacity of the response. This can be considered beneficial in the context of hospital practice, especially when guiding multidisciplinary healthcare teams, as it helps prevent immediate errors. On the other hand, excessive caution may also limit the proposal of valid therapeutic alternatives that an experienced professional would be able to contextualize and apply safely.

Furthermore, the use of clear, technical, and specific commands proved to be decisive for the quality of the responses. During the data collection process, it was observed that generic or excessively conservative responses were frequently associated with less objective prompts. This finding indicates that prompt formulation plays a critical role in AI performance and may significantly increase or reduce clinical risk. Therefore, the perspective of safe and effective use of these tools in clinical pharmacy should consider standardized protocols for interaction with language models.

AI can be an ally for the clinical pharmacist, especially as a technical support tool for standardized clinical conduct. However, it does not replace the critical and interpretative role of the human professional. AI, in turn, remains incapable of personalizing clinical conduct. Therefore, the integration of AI into hospital pharmaceutical practice should be approached with caution, supervised by qualified professionals, and limited to scenarios in which its role as a support tool is considered safe. In settings such as intensive care, where adaptive clinical reasoning is essential, its application should be restricted to support functions and never to replacement of professional judgment.

The classification of the questions in this study into low-, medium-, and high-complexity levels was based on professional experience and clinical judgment. However, it is recognized that this categorization is not absolute: a seemingly simple question may be challenging for a less experienced professional, whereas a question considered complex may be routine for another with greater technical expertise. This relativity leads to a fundamental provocation: who determines what is complex for an AI system? No matter how extensively trained it may be, it still depends on human-defined parameters to determine what constitutes nuance, exception, or context. If we were facing a system supposedly ready to operate autonomously, would we, as healthcare institutions, also be prepared to assume responsibility for its actions? Are we discussing a safe innovation or a disguised risk? Who should be held accountable for conduct suggested by AI?

A limitation of this study is the method used to evaluate the AI-generated responses. The assignment of scores relied entirely on the empirical clinical experience and professional judgment of the expert panel, without the concurrent use of standardized scientific databases or specific institutional protocols (such as UpToDate, Micromedex, or Trissel's) to validate the answers during the review process. Consequently, the evaluation reflects a real-world, human-based clinical interpretation of the AI outputs, which carries inherent subjectivity. Future studies should consider establishing strict analytical frameworks comparing AI outputs directly against gold-standard pharmaceutical

databases to provide a more objective validation of algorithmic accuracy. Thus, while recognizing these methodological limitations of this study, including the aforementioned evaluation method, small sample size ( $n = 15$ ) which constrained the statistical power to detect significant performance differences among the models, as well as the lower inter-rater agreement for the 'clarity' criterion, we reinforce that this study it fulfills an important initial role: to stimulate discussions that go beyond technical performance and address questions of institutional maturity regarding the use of AI systems in healthcare.

## CONCLUSION

Open-access AI models demonstrate reliable performance in low and medium complexity tasks, but show a significant decline in consistency and safety in highly complex scenarios. Therefore, AI may serve as a supportive tool in clinical practice under professional supervision, but it cannot replace critical judgment

and human interpretative capacity. Although Copilot achieved higher numerical consistency, the models performed equivalently overall without statistically significant differences. René Descartes once wrote, "*Cogito, ergo sum*" ("I think, therefore I am"). In clinical practice, existence itself becomes part of judgment. The ability to care emerges not only from processing information, but from the human capacity to interpret uncertainty and, above all, to assume responsibility for decisions that affect another life.

## Use of generative AI and AI-assisted technologies

This study used AI models both as the central object of investigation and as a supportive tool for idea systematization. Perplexity AI was used to assist with the English language review. However, the critical interpretation of the results, comparative analysis, and final manuscript writing were conducted by the authors, preserving scientific rigor and academic responsibility.

## REFERENCES

- Ogbuagu OO, Mbata AO, Balogun OD, Oladapo O, Ojo OO, Muonde M. Artificial intelligence in clinical pharmacy: enhancing drug safety, adherence, and patient-centered care. *Int J Multidiscip Res Growth Eval*. 2023;4(1):814-22.
- Gala D, Makaryus AN. The Utility of Language Models in Cardiology: A Narrative Review of the Benefits and Concerns of ChatGPT-4. *Int J Environ Res Public Health*. 2023;20(15):6438.
- Huang X, Estau D, Liu X, Yu Y, Qin J, Li Z. Evaluating the performance of ChatGPT in clinical pharmacy: A comparative study of ChatGPT and clinical pharmacists. *Br J Clin Pharmacol*. 2024;90(1):232-8.
- Kiyomiya K, Aomori T, Ohtani H. Medication counseling for OTC drugs using customized ChatGPT-4: Comparison with ChatGPT-3.5 and ChatGPT-4o. *Digit Health*. 2025;11.
- Raza MA, Aziz S, Noreen M, Saeed A, Anjum I, Ahmed M, et al. Artificial Intelligence (AI) in Pharmacy: An Overview of Innovations. *Innov Pharm*. 2022;13(2).
- Ogundipe A, Sim TF, Emmerton L. Prescription for Digital Evolution: Transformative Recommendations for Pharmacy Practice in the Digital Age. *J Pharm Pract*. 2025;38(2):237-48.
- Zheng Y, Rowell B, Chen Q, Kim JY, Kontar RA, Yang XJ, et al. Designing Human-Centered AI to Prevent Medication Dispensing Errors: Focus Group Study With Pharmacists. *JMIR Form Res*. 2023;7:e51921.
- Mottaghi-Dastjerdi N, Soltany-Rezaee-Rad M. Advancements and Applications of Artificial Intelligence in Pharmaceutical Sciences: A Comprehensive Review. *Iran J Pharm Res*. 2024;23(1):e150510.
- Vora LK, Gholap AD, Jetha K, Thakur RRS, Solanki HK, Chavda VP. Artificial Intelligence in Pharmaceutical Technology and Drug Delivery Design. *Pharmaceutics*. 2023;15(7):1916.
- Wong A, Wentz E, Palisano N, Dirani M, Elsamadisi P, Qashou F, et al. Role of artificial intelligence in pharmacy practice: A narrative review. *J Am Coll Clin Pharm*. 2023;6(11):1237-50.
- Cardoso Besteti J, von Hohendorff R. A utilização da inteligência artificial na tomada de decisão. *Rev Interdisc Ciênc Apl*. 2022;6(10):17-30.
- Elias MA, Faversani LA, Moreira JAV, Masiero AV, Cunha NV. Inteligência artificial em saúde e implicações bioéticas: uma revisão sistemática. *Rev Bioét*. 2023;31.
- Hasan HE, Jaber D, Khabour OF, Alzoubi KH. Ethical considerations and concerns in the implementation of AI in pharmacy practice: a cross-sectional study. *BMC Med Ethics*. 2024;25(1):55.
- Kelley M, Spoonerbyarger C, Howard M, Reinert J, Churchwell MD, Baki G. Physical compatibility of ceftriaxone and cefepime in 0.45% sodium chloride, Ringer's lactate solution, and Plasma-Lyte A. *Eur J Hosp Pharm*. 2025;32(6):539-43.
- Moritz ML, Ayus JC. Maintenance Intravenous Fluids in Acutely Ill Patients. *N Engl J Med*. 2015;373(14):1350-60.
- Brazilian Society of Geriatrics and Gerontology. The Use of the Subcutaneous Route in Geriatrics and Palliative Care: A Guide from SBGG and ANCP for Professionals. 2nd ed. Rio de Janeiro: SBGG; 2017.
- Tee SA, Devine K, Potts A, Javaid U, Razvi S, Quinton R, et al. Iatrogenic hypoglycaemia following glucose-insulin infusions for the treatment of hyperkalaemia. *Clin Endocrinol (Oxf)*. 2021;94(2):176-82.
- Patel PR, Cook SE. Stability of meropenem in intravenous solutions. *Am J Health Syst Pharm*. 1997;54(4):412-21.

19. McDonald LC, Gerding DN, Johnson S, Bakken JS, Carroll KC, Coffin SE, et al. Clinical Practice Guidelines for Clostridium difficile Infection in Adults and Children: 2017 Update by the Infectious Diseases Society of America (IDSA) and Society for Healthcare Epidemiology of America (SHEA). Clin Infect Dis. 2018;66(7):987-94.
20. Shin JM, Sachs G. Pharmacology of proton pump inhibitors. Curr Gastroenterol Rep. 2008;10(6):528-34.
21. Bonnici A, Ruiner CE, St-Laurent L, Hornstein D. An Interaction Between Levodopa and Enteral Nutrition Resulting in Neuroleptic Malignant-Like Syndrome and Prolonged ICU Stay. Ann Pharmacother. 2010;44(9):1504-7.
22. Snowden L. Book review: Handbook of drug administration via enteral feeding tubes. Aust Prescr. 2016;39(2):53.
23. Overlack A. ACE Inhibitor-Induced Cough and Bronchospasm. Drug Saf. 1996;15(1):72-8.
24. Silverman JA, Mortin LI, VanPraagh ADG, Li T, Alder J. Inhibition of Daptomycin by Pulmonary Surfactant: In Vitro Modeling and Clinical Impact. J Infect Dis. 2005;191(12):2149-52.
25. Castells Lao G, Rodríguez Reyes M, Roura Turet J, Prat Dot M, Soy Muner D, López Cabezas C. Compatibility of drugs administered as Y-site infusion in intensive care units: A systematic review. Med Intensiva (Engl Ed). 2020;44(2):80-7.
26. Wagner LAF, Neiningner MP, Hensen J, Zube O, Bertsche T. Avoiding incompatible drug pairs in central-venous catheters of patients receiving critical care: an algorithm-based analysis and a staff survey. Eur J Clin Pharmacol. 2023;79(8):1081-9.
27. Eyler RF, Shvets K. Clinical Pharmacology of Antibiotics. Clin J Am Soc Nephrol. 2019;14(7):1080-90.
28. Juraschek SP, Steffes MW, Selvin E. Associations of Alternative Markers of Glycemia with Hemoglobin A1c and Fasting Glucose. Clin Chem. 2012;58(12):1648-55.
29. Sani H, Mahan RJ. Treating Opioid-Induced Constipation in Older Adults: New Options. Consult Pharm. 2015;30(10):612-5.
30. Hootkins R, Emmett M. Evaluation of the safety and efficacy of the central venous administration of potassium chloride including the measurement of intracardiac potassium concentrations. Proc (Bayl Univ Med Cent). 2020;33(1):24-7.
31. Jones J, Hunter D. Qualitative Research: Consensus methods for medical and health services research. BMJ. 1995;311(7001):376-80.
32. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020;585(7825):357-62.
33. McKinney W. Data Structures for Statistical Computing in Python. SciPy. 2010:56-61.
34. Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. SciPy. 2010:92-6.
35. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17(3):261-72.
36. Vallat R. Pingouin: statistics in Python. J Open Source Softw. 2018;3(31):1026.
37. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016;15(2):155-63.

*Received: Jun 18 2026*

*Accepted: Aug 27 2026*