

HIPERTENSIÓN & RIESGO VASCULAR

**Calidad de las respuestas
generadas por inteligencia
artificial sobre las crisis
hipertensivas en cuatro agentes
conversacionales: un estudio
transversal**

**Quality of AI-generated
responses in hypertensive crises
throughout four chat agents – A
cross-sectional study**

10.20960/HipertensRiesgoVasc.00003

10/28/2026

00003 OR

Quality of AI-generated responses in hypertensive crises throughout four chat agents – A cross-sectional study

Calidad de las respuestas generadas por inteligencia artificial sobre las crisis hipertensivas en cuatro agentes conversacionales: un estudio transversal

Lucas G. Pequeno, Érika A. Silva, Antonio L. C. Neto, Marcelo D. T. Melo, Guilherme A. T. Athayde

Federal University of Paraíba – UFPB. Brazil

Received: 30/07/2026

Accepted: 15/09/2026

Correspondence: Lucas G. Pequeno. Federal University of Paraíba – UFPB. Brazil

e-mail: lucasgpequeno7@gmail.com

Conflict of interest: The authors declare no conflict of interest.

Artificial Intelligence: Generative AI (ChatGPT and Gemini) was utilized in order to optimize the text legibility and to assist the making of Figure 1.

ABSTRACT

Background: the widespread accessibility of generative artificial intelligence (AI), particularly large language models (LLMs), is reshaping how the general public seeks cardiovascular education. LLMs have the potential to influence the patient journey, especially during acute events such as hypertensive crises.

Objectives: to assess the adequacy of four major LLMs, DeepSeek (V3.2), ChatGPT (5.3), Gemini (3), and Claude (Sonnet 4.6) in addressing inquiries regarding hypertensive crises

from the patient's viewpoint. The primary endpoints were four domains: accuracy, completeness, clarity and safety.

Methods: five standardized prompts in Portuguese, derived from clinical cases were inputted. To prevent contextual bias, new accounts were created and each prompt was entered into a fresh chat session. Two blinded cardiologists independently evaluated the responses using a 5-point Likert scale. Interrater reliability was descriptively analyzed. Non-parametric data were assessed using the Friedman test, while readability using Flesch reading ease.

Results: high performance was achieved across all domains. Interrater agreement showed a 48.7 % absolute concordance. The Friedman test revealed statistically significant raw differences among the four LLMs only in accuracy ($p = 0.01$) and safety ($p = 0.02$). *Post hoc* Dunn-Bonferroni correction showed no difference. All models provided responses aligned with current clinical guidelines, identifying red flags and suggesting appropriate initial management. Median Flesch reading ease was 48.30 (IQR, 40.85–52.77), suggesting a borderline difficult text.

Conclusions: AI is a powerful tool that can improve patient journey by providing accessible, high-quality medical information. Further research is needed to ensure reliability in health literacy.

Keywords: Artificial Intelligence. Hypertensive crisis. Patient Journey. Health literacy.

RESUMEN

Introducción: la amplia accesibilidad de la inteligencia artificial (IA) generativa, en particular de los modelos de lenguaje de gran escala (LLM), está transformando la forma en que la población busca información sobre salud cardiovascular. Los LLM pueden influir en el recorrido asistencial del paciente, especialmente durante eventos agudos como las crisis hipertensivas.

Objetivos: evaluar la adecuación de cuatro LLM principales, DeepSeek (V3.2), ChatGPT (5.3), Gemini (3) y Claude (Sonnet 4.6), para responder consultas sobre crisis hipertensivas desde la perspectiva del paciente. Los desenlaces primarios fueron exactitud, exhaustividad, claridad y seguridad.

Materiales y métodos: se introdujeron cinco *prompts* estandarizados en portugués, derivados de casos clínicos. Para evitar sesgo contextual, se crearon cuentas nuevas y cada *prompt* se ingresó en una sesión de chat independiente. Dos cardiólogos cegados evaluaron independientemente las respuestas mediante una escala de Likert de 5 puntos. Los datos no paramétricos se evaluaron con la prueba de Friedman.

Resultados: la concordancia absoluta entre evaluadores fue del 48,7 %. La prueba de Friedman mostró diferencias brutas estadísticamente significativas entre los cuatro LLM solo en exactitud ($p = 0,01$) y seguridad ($p = 0,02$). La corrección *post hoc* de Dunn-Bonferroni no mostró diferencias. Todos los modelos generaron respuestas acordes con las guías clínicas vigentes, identificando signos de alarma y sugiriendo un manejo inicial adecuado. La mediana del índice de facilidad de lectura de Flesch fue de 48,30 (RIC: 40,85-52,77), lo que indica un texto de dificultad moderada.

Conclusiones: la IA es una herramienta poderosa que puede mejorar el recorrido asistencial del paciente al proporcionar información médica accesible y de alta calidad.

Palabras clave: Inteligencia artificial. Crisis hipertensiva. Recorrido asistencial del paciente. Alfabetización en salud.

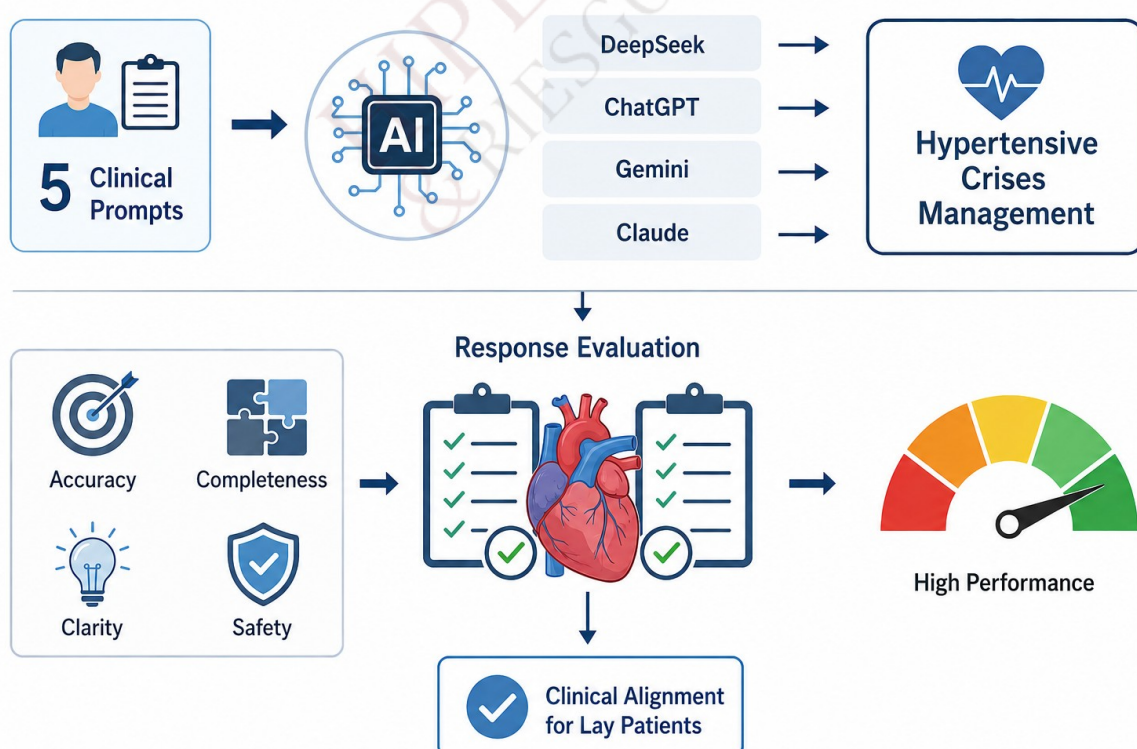


Figure 1. Central illustration: AI in hypertensive crises (Source: Elaborated by the author; 2026).

INTRODUCTION

In the era of big biomedical data and Artificial Intelligence (AI), the recent emergence of flexible natural language processing models, including ChatGPT, has been considered an inflection point in AI-driven healthcare (1). Unlike traditional AI models, large language models (LLMs) are pre-trained on massive datasets, enabling them to learn complex linguistic patterns and adapt to diverse tasks through fine-tuning or prompting (2). LLMs can rapidly synthesize data from multiple sources and suggest differential diagnoses and management plans. Some LLMs have already been integrated into electronic medical record software as assistive tools for summarization and communication (3). However, integrating these technologies into medical education poses challenges and raises concerns about accuracy, ethical implications and detriments to critical thinking. Despite the surge in interest, there remains a notable gap in the comprehensive evaluation of LLMs within medical educational settings (4).

Hypertensive crisis is characterized by an acute and severe elevation of blood pressure (BP), accompanied or not by acute hypertension-mediated organ damage (5). The primary management goal in these cases is twofold: to alleviate symptoms and to reduce blood pressure in a controlled manner (6). Hypertensive crises account for 0.45 % to 0.59 % of all emergency department visits (5). Thus, the unique demands of healthcare require LLMs to interpret and generate specialized medical content with high reliability and contextual relevance (2). In this matter, the aim of this study is to address the usability of generative AI in cardiovascular education, especially in hypertensive crises, focusing on the patient journey and health literacy.

METHODS

An observational, comparative cross-sectional study was conducted in April 2026, using AI chats via a web-based browser. Given the exploratory nature of this pilot study, a convenience sample size was utilized as proof-of-concept. Five standardized prompts in Portuguese were made, adapted from validated and published clinical cases that depicted

hypertensive crises (7-11). To prevent contextual bias, new accounts were created in each agent, and every prompt was inputted into an exclusive fresh chat session. The four models evaluated were DeepSeek (V3.2), Chat-GPT (5.3), Gemini (3) and Claude (Sonnet 4.6). All twenty responses were catalogued and randomized for the subsequent analysis of two independent blinded cardiologists. The analysis took place with a 5-point Likert scale that assessed 4 endpoints: a) accuracy; b) completeness; c) clarity; and d) safety.

Statistical analysis

Categorical variables were presented as frequencies and percentages, while continuous variables were presented as median and interquartile ranges (IQR) for non-normally distributed data. The weighted Cohen's Kappa was performed to evaluate interrater reliability. Non-parametric data were analyzed using the Friedman test to compare scores between the four models. *Post hoc* analysis was calculated via Dunn-Bonferroni correction. Readability was assessed using Flesch reading ease via ALT platform (12). All analyses were performed using specialized packages in R Foundation for Statistical Computing software (v4.5.3), with a significance level set at 5 %.

RESULTS

Due to ceiling effect from highly skewed score distributions, interrater reliability via weighted Cohen's Kappa yielded an artificially low index ($\kappa = -0.04$, $p = .69$). However, in descriptive analysis the raters showed strong consensus with 97.50 % of the matched scores ranging from 4 to 5, and with an absolute concordance in 48.75 % and 51.25 % respectively of ratings differing by only 1 point. The applied Friedman test primarily revealed significant statistical differences in accuracy ($\chi^2 = 11.73$, $p = .01$) and safety ($\chi^2 = 10.36$, $p = .02$) domains, while no difference was shown in completeness ($\chi^2 = 4.78$, $p = .19$) and clarity ($\chi^2 = 5.77$, $p = .12$). *Post hoc* using Dunn-Bonferroni for multiple comparisons unveiled no significant differences across the four domains. The Flesch reading ease showed a median of 48.30 (IQR, 40.85–52.77) for all four chat agents combined, which indicates that the responses are delivered in a “difficult” complexity (Table I).

Table I. Frequencies of 5-point Likert scores across models in four domains

	DeepSeek (n = 10)	ChatGPT (n = 10)	Gemini (n = 10)	Claude (n = 10)	Total (n = 40)
Accuracy					
Totally disagree	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)
Partially disagree	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)
Neutral	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)
Partially agree	1 (10.0 %)	1 (10.0 %)	6 (60.0 %)	5 (50.0 %)	13 (32.5 %)
Totally agree	9 (90.0 %)	9 (90.0 %)	4 (40.0 %)	5 (50.0 %)	27 (67.5 %)
Completeness					
Totally disagree	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)
Partially disagree	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)
Neutral	0 (0 %)	0 (0 %)	1 (10.0 %)	1 (10.0 %)	2 (5.0 %)
Partially agree	2 (20.0 %)	4 (40.0 %)	5 (50.0 %)	4 (40.0 %)	15 (37.5 %)
Totally agree	8 (80.0 %)	6 (60.0 %)	4 (40.0 %)	5 (50.0 %)	23 (57.5 %)
Clarity					
Totally disagree	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)
Partially disagree	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)
Neutral	0 (0 %)	0 (0 %)	1 (10.0 %)	0 (0 %)	1 (2.5 %)
Partially agree	0 (0 %)	2 (20.0 %)	4 (40.0 %)	3 (30.0 %)	9 (22.5 %)
Totally agree	10 (100.0 %)	8 (80.0 %)	5 (50.0 %)	7 (70.0 %)	30 (75.0 %)
Safety					
Totally disagree	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)
Partially disagree	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)
Neutral	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)
Partially agree	0 (0 %)	0 (0 %)	4 (40.0 %)	3 (30.0 %)	7 (17.5 %)
Totally agree	10 (100.0 %)	10 (100.0 %)	6 (60.0 %)	7 (70.0 %)	33 (82.5 %)

Source: Elaborated by the author (2026).

DISCUSSION

In our study, four major generative AI large language models (LLMs) demonstrated consistently high performance and revealed similar patterns of data processing and information retrieval, providing reliable guidance for the initial management of hypertensive crises in lay patients. However, it is known that thorough medical history and physical examination are essential for the correct diagnosis of a hypertensive crisis. If a life-threatening condition is identified, appropriate treatment should be initiated concurrently with the investigation (5). Hence, the inherent barrier in relying purely on generative AI in these cases. In a systematic review, these models have demonstrated competence in multiple standardized clinical exams, potentially making them a valuable reference for medical students and educators (4). Our study also showed the potential of these LLMs, since adequate responses were observed in the majority of the input prompts. Another recent study revealed the feasibility of using LLMs to assess patients with rare and life-threatening cardiac conditions, showing the strength of AI even in complex and extreme situations (3).

The most concerning finding is the complexity of delivered responses. These models are likely trained on high-level academic literature and may struggle to translate acute cardiovascular concepts into accessible language, potentially compromising health literacy for the lay patient's journey. In this context, the observed high performance might be attenuated by limited effectiveness in conveying information to the end user.

This study has several limitations. First, the small sample size and limited number of expert evaluators diminished the statistical power of our quantitative analysis. Second, this is a cross-sectional study with a temporal snapshot of programs that are updated regularly, limiting the generalizability of the findings. In addition, although a 5-point Likert scale was employed, the resultant ceiling effect restricted reliability testing. Nonetheless, these results still provide relevant and up-to-date findings.

CONCLUSION

All the AI agents exhibited a pattern of a secure response, regarding hypertensive crisis management, which may support future integration into cardiology education. These early findings highlight the fast dissemination of novel tools that can improve health literacy

among lay public. However, larger scale studies are needed to ensure safe implementation and to confirm their positive impact on health outcomes.

REFERENCES

1. Park SH, Pinto-Powell R, Thesen T, Lindqwister A, Levy J, Chacko R, et al. Preparing healthcare leaders of the digital age with an integrative artificial intelligence curriculum: a pilot study. *Med Educ Online* 2024;29(1):2315684. DOI: 10.1080/10872981.2024.2315684
2. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC Med Inform Decis Mak* 2025;25(1):117. DOI: 10.1186/s12911-025-02954-4
3. O'Sullivan JW, Palepu A, Saab K, Weng W-H, Amponsah DK, Cheng E, et al. A large language model for complex cardiology care. *Nat Med* 2026;32(2):616-23. DOI: 10.1038/s41591-025-04190-9
4. Lucas HC, Upperman JS, Robinson JR. A systematic review of large language models and their implications in medical education. *Med Educ* 2024;58(11):1276-85. DOI: 10.1111/medu.15402
5. Brandão AA, Rodrigues CIS, Bortolotto LA, Armstrong AC, Mulinari RA, Feitosa ADM, et al. Brazilian Guidelines of Hypertension – 2025. *Arq Bras Cardiol* 2025;122(9):e20250624. DOI: 10.36660/abc.20250624
6. Kırık S, Efgan MG, Kanter E, Ermete Güler E, Duman Şahan T, Gezer HK. Electrophysiological Response in Hypertensive Crisis: An Evaluation of the ICEB Score. *Medicina* 2026; 62(3):501. DOI: 10.3390/medicina62030501
7. Shukla D, Virani AY. Reversible Loss of Vision in Malignant Hypertension. *N Engl J Med* 2016;374(14):1368. DOI: 10.1056/nejmicm1508601
8. Singhal AB, Gonzalez RG, Chwalisz BK, Mukerji SS. Case 26-2020: A 60-Year-Old Woman with Altered Mental Status and Weakness on the Left Side. *N Engl J Med* 2020;383(8):764-73. DOI: 10.1056/NEJMcpc2004976
9. Senecal EL, Rosenfield K, Caldera AE, Passeri JJ. Case 36-2011. *N Engl J Med* 2011;365(21):2021-8. DOI: 10.1056/nejmcpc1103565
10. Iliopoulos O, Chan-Smutko G, Gonzalez RG, Louis DN, Stone JR. Case records of the Massachusetts General Hospital. Case 23-2006. A 36-year-old man with numbness in

the left [corrected] hand and hypertension. N Engl J Med 2006;355(4):394-402. DOI: 10.1056/NEJMcpc069014

11. Case records of the Massachusetts General Hospital. Weekly clinicopathological exercises. Case 26-2001. Hypertensive encephalopathy with impaired renal function in a 67-year-old woman with polymyositis. N Engl J Med 2001;345(8):596-605. DOI: 10.1056/NEJMcpc010026
12. Moreno GC, Souza M, Hein N, Hein AK. ALT: A software for readability analysis of Portuguese-language texts. ArXiv (Cornell University). October 2022. DOI: 10.48550/arxiv.2210.00553

HIPERTENSIÓN
& RIESGO VASCULAR