DOI: 10.5152/turkjnephrol.2026.251041

Original Article

Turkish Journal of Nephrology

Comparative Assessment of Chatbot Responses on Childhood Hypertension

Esra Ensari[1], Esra Nagehan Akyol Önder[2], Pelin Ertan[3]

[1]Department of Pediatric Nephrology, Ministry of Health Antalya City Hospital, Antalya, Türkiye
[2]Division of Pediatric Nephrology, Aksaray University Hospital, Aksaray, Türkiye
[3]Division of Pediatric Nephrology, Manisa Celal Bayar University School of Medicine, Manisa, Türkiye

ABSTRACT

Background: Hypertension (HT) is a serious public health concern. Although common in adults, the incidence of HT in children and adolescents is increasing. Natural language processing chatbots (NLPCs) are becoming more and more popular as the use of artificial intelligence in the health care sector continues to grow. The aim of this study is to evaluate the accuracy and reproducibility of answers to frequently asked questions (FAQs) about pediatric HT provided by NLPCs such as ChatGPT-4o, Gemini, and Copilot.

Methods: Fifty-five FAQs were collected from reputable health sources and grouped into 4 areas: basic information, diagnosis/tests, treatment, and prognosis. Each question was asked twice, 1 week apart, using ChatGPT-4o, Gemini, and Copilot. Based on established clinical guidelines, the answers were independently assessed by 2 pediatric nephrologists. Accuracy was rated on a 4-point scale, and reproducibility was determined by comparing the consistency of responses over time.

Results: Across all categories, ChatGPT-4o showed the highest accuracy (91.0%) and reproducibility (91.0%). Gemini and Copilot achieved an accuracy of 65.4% and 54.5%, respectively, with a lower level of consistency (Gemini: 52.7%; Copilot: 41.9%). A small proportion of totally inaccurate or misleading responses were also generated by Copilot.

Conclusion: Of the models evaluated, ChatGPT-4o provided the most accurate and reliable information about childhood HT. NLPCs should not replace professional medical advice, although they can support the education of patients and caregivers. The quality and safety of digital health interactions can be improved by incorporating clinical practice guidelines.

Keywords: Artificial intelligence, ChatGPT, children, Copilot, Gemini, hypertension

Corresponding author: Esra Ensari ✉ dresraensari@gmail.com

Received: April 17, 2025 **Revision Requested:** May 28, 2025
Last Revision Received: October 28, 2025 **Accepted:** November 9, 2025 **Publication Date:** April 7, 2026

Cite this article as: Ensari E, Onder ENA, Ertan P. Comparative assessment of chatbot responses on childhood hypertension. *Turk J Nephrol.* 2026;35(2):146-153.

INTRODUCTION

Childhood hypertension (HT) is defined as systolic or diastolic blood pressure (BP) readings above the threshold for age, height, and sex, taken on 3 separate occasions.[1] HT in children is a significant and increasing health problem. With the increasing incidence of obesity in children, the incidence of HT, especially the primary form, has increased in childhood. The incidence of HT in children has been described as 3.5%, but rates are likely to be higher as it may be overlooked during routine pediatric examination.[2] It is very important to define and treat HT in childhood because it can contribute to early atherosclerosis and the early development of cardiovascular disease.[3] Given the increasing incidence of HT in children, numerous guidelines have been created to promote a standardized approach to diagnosis and treatment.[1,4,5] Because lifestyle changes are the first line of treatment for HT, successful management is highly dependent on effective patient and caregiver education.[6] A study in adult hypertensive patients in the UK found that those with a good lifestyle had remarkably lower systolic BP than those with a poor lifestyle.[7] Many studies have confirmed that lifestyle changes such as regular exercise, weight loss, low-sodium diet,

This work is licensed under a Creative Commons Attribution 4.0 International License.

limited alcohol consumption, and adequate potassium intake effectively improve BP control.[8,9] Improving lifestyle is efficient in controlling BP and preventing complications. However, the content of HT self-management is quite rich and the details are complex; a systematic intervention framework has not yet been established. In out-patient clinics with high patient density, this information may not be explained to patients in detail. Self-management and behavioral changes are difficult processes. There are a number of digital health tools like mobile health applications, digital pill boxes, and recently, chatbots to support HT self-management. Chatbots are dialogue systems that mimic human conversational features and have been increasingly adopted in healthcare areas. Since patients cannot always reach their doctors, they can ask natural language processing chatbots (NLPCs) questions about their illnesses, how they should use their medications, or in which cases they should go to the hospital. While ready access to information can enhance health literacy, it has also exposed the risk of information being misused and the spread of misinformation. Hence, many studies have aimed to evaluate the quality of these NLPCs.[10-12] In this study, the aim was to compare the quality of various NLPCs, including ChatGPT-4o, Gemini, and Copilot, in answering frequently asked questions (FAQs) by patients and their parents regarding HT.

MATERIALS AND METHODS

Subjects and Study Design

In this study, 55 questions related to childhood HT were identified from FAQs by patients on the Cleveland Clinic, Mayo Clinic, and National Health Service UK websites (Table 1). This study was based exclusively on publicly available information from institutional websites and did not involve human participants or identifiable personal data. Therefore, informed consent was not required. The identified questions were asked to 3 selected NLPCs: ChatGPT-4o (GPT-4, OpenAI, San Francisco, California, United States), Copilot (Bing Chat, Microsoft, Redmond, Washington, United States), and Gemini (PaLM 2, Google, Mountain View, California, United States). Questions were divided into 4 subcategories: basic information, diagnosis/tests, treatment, and prognosis. Artificial intelligence supported technologies were not used in the writing phase of this study.

MAIN POINTS

- ChatGPT-4o showed superior performance in both accuracy and repeatability when answering frequently asked questions on childhood hypertension (HT).
- Natural language chatbots should not replace clinician guidance or evidence-based resources, although they may improve access to HT-related information.
- This study supports the potential of natural language processing chatbots, and ChatGPT-4o in particular, as a complementary tool for the education of caregivers and patients in the field of pediatric kidney and cardiovascular health.

Table 1. Categorization of Frequently Asked Questions About Childhood Hypertension

Basic Information
What is BP?
What does high BP mean?
How common is childhood HT?
What are the types of childhood HT?
What causes high BP in children and teens?
What are the risk factors for childhood HT?
How does high BP affect children?
What are the symptoms of high BP in children and teens?
What is normal BP for children?
What is high BP for children?
I have high BP, will my child have HT too?
Diagnosis and Tests
What type of doctor should I consult for my child's high BP?
Is my child's HT a symptom of an underlying medical condition?
What are the diagnostic tests for childhood HT? How is childhood HT diagnosed?
When should my child see their healthcare provider about high BP?
What is ABPM? How does it work?
What are the harms and benefits of having the ABPM test?
Treatment
What are the treatment options for childhood HT? How is high BP in children treated?
What are lifestyle changes for chilhood HT?
What foods should my child with HT eat?
What foods should my child with HT avoid eating?
What's an appropriate level of physical activity?/ What exercises can I/my child with high BP do?
Which foods contain high salt?
When we can stop using antihypertensive medication?
What are the medication treatments for childhood HT?
What are ACEI drugs?
Are ACEI drugs suitable for all hypertensive children?
How does ACEI drug work for HT?
What are the side effects of ACEI drugs?
What will my doctor do if my BP does not decrease despite taking ACEI drug?
What medication should be used if ACEI drug has side effects on my hypertensive child?
What are ARB drugs? What are their mechanisms of action?

(*Continued*)

Table 1. Categorization of Frequently Asked Questions About Childhood Hypertension (*Continued*)

Are there any side effects associated with ARB drugs?
Should any special care be taken when using antihypertensive drug?
What are CCB drugs? What are their mechanisms of action?
When should my child take ACEI drug?
When should my child take ARB drug?
When should my child take CCB drug?
What are the side effects of CCB drugs?
When to stop using antihypertensive drug?
What happens if I/my child overdose on enalapril/ACEI?
What are diuretic drugs?
Should any special care be taken when using antihypertensive drug?
What to do if my child's high BP does not come down with medication?
Which device should I use to measure the child's BP? Do digital devices measure BP accurately?
How often should I measure the BP of my child who is taking antihypertensive medication?
Prognosis
What are the complications of HT in children?
What are the long-term outcomes of childhood HT?
If my child loses weight will her high BP go down?
How does high uric acid affect HT?
How does HT affect persons mental health?
How should I encourage my child to follow a diet?
How should I encourage my child to use BP medication?
What are the consequences if my child does not follow her diet and her BP does not decrease?
What should I do to prevent high BP in my child?/How can I prevent high BP in my child?

ABPM, ambulatory blood pressure monitoring; ACEI, angiotensin-converting enzyme inhibitor; ARB, angiotensin receptor blocker; BP, blood pressure; CCB, calcium channel blocker; FAQ, frequently asked question; HT, hypertension.

Each question was asked in a separate chat tab. On December 18, 2024, and again on December 25, 2024, the questions were presented in the same order to the ChatGPT-4o, Gemini, and Copilot chatbots. Responses were evaluated by 2 pediatric nephrologists using the 2017 American Academy of Pediatrics (AAP),[1] 2016 European Society of HT (ESH),[5] and 2022 American Heart Association[13] guidelines, and disagreements were resolved by a third pediatric nephrologist. Questions were asked in English. The reliability of the responses of the NLPCs was measured by the Reliability and Accuracy Rating Scale. The responses were listed as comprehensive/accurate (4) responses that were comprehensive, precise, and fully answered the question; incomplete/partially accurate (3) for responses that were somewhat accurate but lacking in detail or completeness; accurate and inaccurate/misleading (2) responses that were both accurate and included inaccurate or potentially misleading information; and completely inaccurate/irrelevant (1) for responses that were inaccurate, irrelevant, or did not address the question. The answers were divided into 2 groups according to their consistency and repeatability. If the answers fell into different categories, the question was considered unrepeatable, reflecting the variability in the answer accuracy. The reproducibility of chatbot responses was evaluated to assess the consistency of their outputs across repeated sessions. Reproducibility was defined as identical categorical scoring across 2 independent evaluation sessions using the same set of FAQs. All chatbot outputs were generated under standardized conditions using default parameters (e.g., temperature, randomness, and model version) and identical prompts to ensure consistency. Each chatbot session was conducted separately to avoid cross-model influence.

Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics for Windows, version 26.0 (IBM Corp., Armonk, NY, USA). Descriptive statistics were presented as counts and percentages, and 95% CIs for proportions were calculated using the Wilson method. Inter-rater agreement between the 2 evaluators was assessed using Cohen's κ statistic, with values interpreted as follows: <0.20 poor, 0.21-0.40 fair, 0.41-0.60 moderate, 0.61-0.80 substantial, and >0.80 almost perfect agreement.

RESULTS

The scoring results for ChatGPT-4o were as follows: 5 (9.0%) incomplete/partially correct and 50 (91.0%) comprehensive/correct. There were no inaccurate/misleading or completely inaccurate/irrelevant responses. ChatGPT-4o answered questions correctly at the highest rate (91.0%) among all NLPCs. Overall accuracy and comparative performance of the chatbots are shown in Figure 1. ChatGPT-4o provided comprehensive/correct responses to 10 (91.0%) questions in the basic information category, 5 (83.3%) related to diagnosis/tests, 27 (93.1%) in management/treatment, and 8 (88.8%) in prognosis. Table 2 shows the accuracy rates of ChatGPT-4o responses. Accuracy rates across subcategories, including basic information, diagnosis/tests, treatment, and prognosis, are illustrated in Figure 2.

Gemini provided 36 (65.4%) comprehensive/correct responses, 15 (27.3%) incomplete/partially correct responses, and 4 (7.3%) inaccurate/misleading responses. No completely inaccurate or irrelevant responses were identified. Table 2 shows that Gemini provided comprehensive/correct responses to 6 (54.5%) questions on basic information, 4 (66.6%) on diagnosis/tests, 19 (65.5%) on management/treatment, and 7 (77.8%) on prognosis.

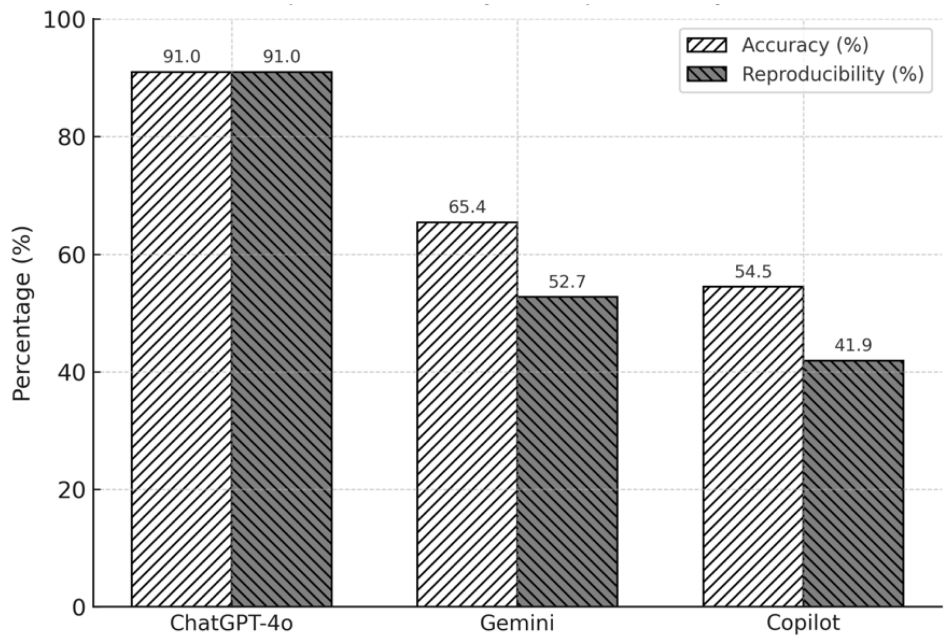

Figure 1. Comparative Accuracy and Reproducibility of ChatGPT-4o, Gemini, and Copilot.

Copilot provided 30 (54.5%) comprehensive/correct responses, 17 (31.0%) incomplete/partially correct responses, 6 (11.0%) inaccurate/misleading responses, and 2 (3.5%) completely inaccurate/irrelevant responses. Table 2 shows that Copilot provided comprehensive/correct responses to 5 (45.4%) questions on basic information, 3 (50.0%) on diagnosis/tests, 17 (58.6%) on management/treatment, and 5 (55.6%) on prognosis.

ChatGPT-4o answered questions correctly at the highest rate (91.0%) among all NLPCs. Data are presented as n (%) with 95% CIs shown in Table 2.

The chatbots produced reliably repeatable results. Reproducibility results are summarized in Table 3. ChatGPT-4o provided 50 (91.0%) reproducible responses overall. The subgroup reproducibility rates were 9 (81.8%) for basic information, 6 (100.0%) for diagnosis, 26 (89.7%) for management and treatment, and 9 (100.0%) for prognosis. Gemini provided 29 (52.7%) reproducible responses overall, with subgroup rates of 7 (63.6%) for basic information, 2 (33.3%) for diagnosis, 15 (51.7%) for management and treatment, and 5 (55.5%) for prognosis. Copilot provided 23 (41.9%) reproducible responses overall. Subgroup reproducibility rates were 5 (45.5%) for basic information, 3 (50.0%) for diagnosis, 12 (41.4%) for management and treatment, and 3 (33.3%) for prognosis. Data are presented as n (%) with 95% CIs shown in Table 3.

Inter-rater reliability was high for all chatbot evaluations. Cohen's κ values were 0.88 for ChatGPT-4o, 0.82 for Gemini, and 0.80 for Copilot, indicating substantial to almost perfect agreement between raters.

DISCUSSION

NLPCs are artificial intelligence (AI) systems derived from articles, books, and other internet-based sources. These chatbots aim to produce consistent and accurate text in accordance with the given instructions; however, they may sometimes produce inappropriate responses due to a lack of understanding of the context. It is very important to determine the performance of these language models because inaccurate or incomplete information can lead to misleading or harming patients. This study provides an exhaustive evaluation of ChatGPT-4o, Gemini, and Copilot in answering FAQs about childhood HT. The results showed that NLPCs have the potential to provide accurate answers to HT-related questions.

The accuracy of chatbots has been investigated in many areas in children. Naz et al[14] comprehensively evaluated the answers given by ChatGPT, Gemini, and Copilot to FAQs about chronic kidney disease (CKD) in children. They showed that chatbots have the potential to provide accurate answers to questions about CKD. Akyol Onder et al[12] evaluated the performance of ChatGPT in pediatric vesicoureteral reflux and showed that ChatGPT-4o provides high-quality, accessible information on vesicoureteral reflux (VUR). Qarajeh et al[15] found high accuracy rates in ChatGPT, Bard AI, and Bing chat in determining potassium and phosphorus content in dietary supplements for patients with CKD.

ChatGPT-4o answered most questions about pediatric HT accurately and reliably compared to Gemini and Copilot. Naz et al[14] showed that among 3 chatbots, Gemini was the most competent in providing correct answers to questions related to CKD, followed by ChatGPT and Copilot. In another study, in the study

Table 2. Comparative Accuracy of ChatGPT-4o, Gemini, and Copilot Responses by Topic

Category	Comprehensive/Correct, n (% [95% CI])	Incomplete/Partially Correct, n (% [95% CI])	Inaccurate/Misleading, n (% [95% CI])	Completely Inaccurate/ Irrelevant, n (% [95% CI])
ChatGPT-4o				
Overall (n = 55)	50 (91.0 [79.2-97.3])	5 (9.0 [2.7-20.8])	–	–
Basic information (n = 11)	10 (91.0 [71.0-99.0])	1 (9.0 [1.0-29.0])	–	–
Diagnosis and tests (n = 6)	5 (83.3 [36.5-99.1])	1 (16.7 [0.9-63.5])	–	–
Treatment (n = 29)	27 (93.1 [77.2-99.2])	2 (6.9 [0.8-22.8])	–	–
Prognosis (n = 9)	8 (88.8 [51.7-99.7])	1 (11.1 [0.3-48.3])	–	–
Gemini				
Overall (n = 55)	36 (65.4 [50.9-78.0])	15 (27.3 [15.0-42.8])	4 (7.3 [2.0-17.6])	–
Basic information (n = 11)	6 (54.5 [25.0-81.0])	4 (36.4 [10.9-69.2])	1 (9.1 [0.2-41.3])	–
Diagnosis and tests (n = 6)	4 (66.6 [22.3-95.7])	1 (16.7 [0.9-63.5])	1 (16.7 [0.9-63.5])	–
Treatment (n = 29)	19 (65.5 [45.7-82.1])	9 (31.0 [15.3-50.8])	1 (3.5 [0.1-17.8])	–
Prognosis (n = 9)	7 (77.8 [40.0-97.2])	1 (11.1 [0.3-48.3])	1 (11.1 [0.3-48.3])	–
Copilot				
Overall (n = 55)	30 (54.5 [40.0-68.4])	17 (31.0 [19.1-45.9])	6 (11.0 [4.1-22.6])	2 (3.5 [0.4-12.1])
Basic information (n = 11)	5 (45.4 [17.9-74.0])	4 (36.4 [10.9-69.2])	2 (18.2 [2.3-51.8])	–
Diagnosis and tests (n = 6)	3 (50.0 [11.8-88.2])	2 (33.3 [4.3-77.7])	1 (16.7 [0.9-63.5])	–
Treatment (n = 29)	17 (58.6 [38.9-76.5])	9 (31.0 [15.3-50.8])	2 (6.9 [0.8-22.8])	1 (3.5 [0.1-17.8])
Prognosis (n = 9)	5 (55.6 [21.2-86.3])	2 (22.2 [2.8-60.0])	1 (11.1 [0.3-48.3])	1 (11.1 [0.3-48.3])

Definitions of rating categories:
Comprehensive/correct (score 4): The response provides complete and accurate information without factual or contextual errors.
Incomplete/partially correct (score 3): The response is generally correct but lacks key details or contextual depth.
Inaccurate/misleading (score 2): The response contains both correct and incorrect information, including potentially misleading statements.
Completely inaccurate/irrelevant (score 1): The response includes incorrect or irrelevant information unrelated to the question.

of intraocular tumors, Bard showed a higher accuracy rate than ChatGPT-3.5 and Bing.[16] Pay et al[17] showed that ChatGPT had a higher accuracy rate compared to Bing and Gemini in coronary artery disease. These differences in results may be due to the complexity and effectiveness of the NLPC algorithms and the fact that they are trained on different data.

ChatGPT-4o obtained the best accuracy, delivering thorough and accurate answers to 91.0% of inquiries, exceeding Gemini (65.4%) and Copilot (54.5%). This noticeable diversity indicates that ChatGPT-4o possesses a superior capacity to provide reliable medical information on pediatric HT compared to other chatbots. ChatGPT-4o demonstrated consistent accuracy across subcategories, achieving the highest performance in management/treatment at 93.1% and the lowest in diagnosis/tests at 83.3%. Conversely, both Gemini and Copilot exhibited difficulties in sustaining a consistent level of performance across categories, with Copilot's responses including a portion of entirely erroneous or irrelevant information (3.5%).

Although ChatGPT-4o exhibited the highest overall accuracy and consistency, Gemini's relatively stronger performance in the prognosis category deserves attention. This advantage may not necessarily reflect superior medical reasoning, but rather the model's broader exposure to open-access, outcome-oriented health materials during training. Gemini's corpus likely incorporates educational and patient-facing resources such as MedlinePlus or Mayo Clinic, which emphasize disease course and long-term outlook. Consequently, prognosis-related questions may have aligned more closely with the linguistic and semantic structures present in Gemini's training data. Similar observations have been reported in previous evaluations showing that models trained on heterogeneous internet-based health content perform better in general and outcome-focused queries, while those optimized for concise factual responses, such as Copilot, may lack contextual elaboration.[18] These findings suggest that domain-specific variability in training data composition and contextual framing can significantly influence chatbot performance across medical subdomains, particularly for topics like prognosis where interpretive and patient-centered information is essential.

The combination of accurate and somewhat accurate answers from Gemini suggests sensible trustworthiness; nonetheless, the

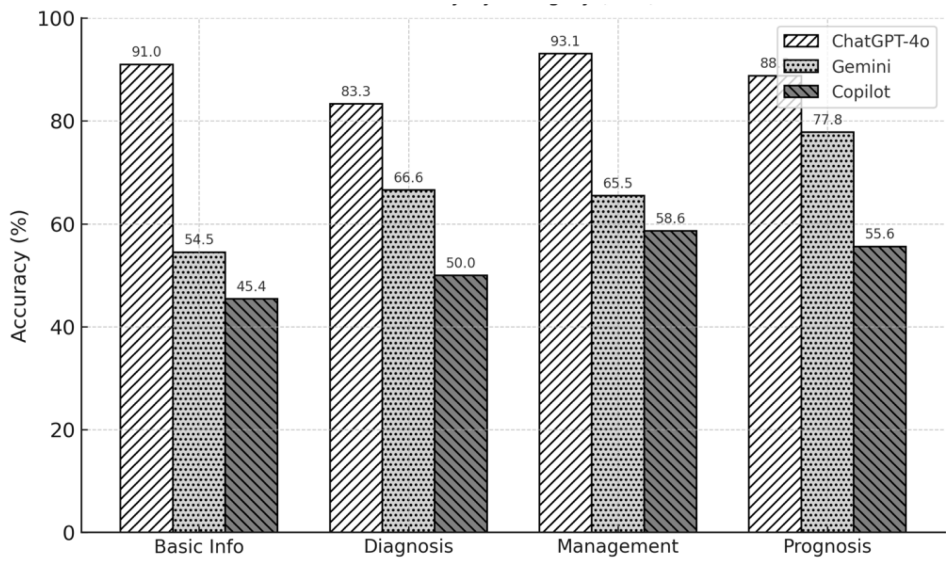

Figure 2. Accuracy by Subcategory Among ChatGPT-4o, Gemini, and Copilot.

presence of false information (7.3%) raises concerns, particularly for severe health problems. Copilot's efficacy was further undermined by a mix of incomplete, misleading, and entirely erroneous responses, which constituted 45.5% of its answers. This underscores the necessity of meticulous vetting and quality assurance when employing AI-driven solutions for patient education.

In the present study, incomplete or partially correct responses generally reflected omissions of important details, insufficient contextual explanation, or lack of guideline citation. These responses may limit the educational value of chatbot-generated information but are unlikely to cause direct harm. Conversely, clinically misleading responses included factual inaccuracies or inappropriate clinical guidance, which could potentially influence caregiver decisions or delay professional consultation. Differentiating between these 2 error types is essential, as incomplete information primarily affects patient education quality, while misleading information poses a more serious risk to patient safety. Similar concerns have been highlighted in recent evaluations of AI-based medical information systems, which emphasize that incomplete content mainly affects comprehension, whereas misleading content may jeopardize clinical safety.[18,19]

Reproducibility is essential for guaranteeing the credibility of chatbot responses, especially for health problems. ChatGPT-4o outperformed other chatbots with a 91.0% repeatability rate. This high level of consistency highlights the capacity to provide stable responses across multiple interactions. In contrast, Gemini and Copilot attained repeatability rates of 52.7% and 41.9%, respectively, prompting concerns regarding their reliability for coherent patient education.

The surge in the usage of NLPCs in medical practice exposes the need to develop appropriate and repeatable systems. The high effectiveness of ChatGPT-4o shows that such technology is a rather helpful instrument in response to questions concerning childhood HT. Its exceptional accuracy and consistency make it extremely useful for patient education, a key element of good HT management. The difference noted in Gemini and Copilot indicates the serious threat of deceiving information or lopsided advice that can cause patients to experience long delays in competent medical intervention or inaccurate management of their problems.

The study reinforces the importance of lifestyle alterations in the control of childhood HT and also raises the possibilities of

Table 3. Reproducibility of ChatGPT-4o, Gemini, and Copilot Responses to Childhood Hypertension-Related Questions by Subgroup

Category	ChatGPT-4o, n (% [95% CI])	Gemini, n (% [95% CI])	Copilot, n (% [95% CI])
Basic Information (n = 11)	9 (81.8 [48.2-97.7])	7 (63.6 [30.8-89.1])	5 (45.5 [16.7-76.6])
Diagnosis and tests (n = 6)	6 (100.0 [54.1-100.0])	2 (33.3 [4.3-77.7])	3 (50.0 [11.8-88.2])
Treatment (n = 29)	26 (89.7 [72.6-97.8])	15 (51.7 [32.5-70.6])	12 (41.4 [23.5-61.1])
Prognosis (n = 9)	9 (100.0 [66.4-100.0])	5 (55.5 [21.2-86.3])	3 (33.3 [7.5-70.1])
Total (n = 55)	50 (91.0 [79.2-97.3])	29 (52.7 [38.6-66.4])	23 (41.9 [28.0-56.7])

the use of chatbots to teach families about dieting, exercising, and sticking to prescriptions. Patients have to be very careful whenever they seek AI-based directions in their treatment and recognize the finite capacity of chatbots as healthcare tools.

The limitations of this study can be outlined as detailed below. First, ChatGPT-4o does not elaborate on the methodologies it uses when answering questions. In fact, citation information is omitted for output results in ChatGPT-4o. The algorithm of ChatGPT-4o is complex and interactive, relying heavily on human feedback, and it is expected that future versions will feature wider knowledge bases.

Second, the study design focused exclusively on structured FAQs, which may not fully capture the variability, nuance, and contextual ambiguity of real-world caregiver or patient questions. In clinical settings, questions often reflect differences in health literacy, emotional tone, and prior knowledge, which structured formats cannot completely reproduce. Therefore, the accuracy and reproducibility observed here may differ from chatbot performance in spontaneous or open-ended interactions.

Third, this study only conducted an analysis of English questions and answers, which restricts the scope of this results being applicable for other languages. Fourth, the subjective evaluation of the models may be influenced by assessor bias, as the ratings were performed by 2 specialists from similar academic backgrounds.

From a clinical perspective, pediatricians should guide families on the responsible use of chatbots in the management of childhood HT. Chatbots may serve as supplementary educational tools, providing general information about BP monitoring, healthy lifestyle practices, and treatment adherence. However, they should not replace professional evaluation or guideline-based management. Families should be encouraged to verify chatbot-generated information with their physician, avoid using AI tools for medication adjustments or diagnostic decisions, and rely on validated clinical sources such as AAP or ESH guidelines. Emphasizing this distinction between educational use and clinical decision-making is crucial to ensure patient safety and prevent misinformation-related risks.

In conclusion, all 3 NLPCs could answer questions on childhood HT to varying extents, but ChatGPT-4o significantly outperformed the others in both correctness and consistency. Nevertheless, these tools should not replace professional advice or evidence-based guidelines. For the safe use of NLPCs in pediatric HT education, better integration of clinical practice guidelines and content validation are important steps.

Ethics approval this article does not contain any studies with human participants or animals performed by any of the authors. Institutional review board approval was not be obtained because all responses collected from chatbots were publicly available. The study was conducted in accordance with the standards of the Declaration of Helsinki.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author.

Artificial Intelligence Usage Statement: The authors declared that no Artificial Intelligence tool was used in the preparation of the manuscript.

Ethics Committee Approval: N/A.

Informed Consent: N/A.

Peer-review: Externally peer-reviewed.

Author Contributions: Concept – E.E.; Design – E.N.A.Ö.; Supervision – P.E.; Resources – E.N.A.Ö.; Materials – E.E.; Data Collection and/or Processing – E.E.; Analysis and/or Interpretation – E.N.A.Ö.; Literature Search – E.E.; Writing Manuscript – E.E.; Critical Review – P.E.

Declaration of Interests: The authors have no conflict of interest to declare.

Funding: The authors declared that this study has received no financial support.

REFERENCES

1. Flynn JT, Kaelber DC, Baker-Smith CM, et al. Clinical practice guideline for screening and management of high blood pressure in children and adolescents. *Pediatrics*. 2017;140(3):e20171904. [CrossRef]
2. Hardy ST, Sakhuja S, Jaeger BC, et al. Trends in blood pressure and hypertension among US children and adolescents, 1999-2018. *JAMA Netw Open*. 2021;4(4):e213917. [CrossRef]
3. Song P, Zhang Y, Yu J, et al. Global prevalence of hypertension in children: a systematic review and meta-analysis. *JAMA Pediatr*. 2019;173(12):1154-1163. [CrossRef]
4. National High Blood Pressure Education Program Working Group on High Blood Pressure in Children and Adolescents. The fourth report on the diagnosis, evaluation, and treatment of high blood pressure in children and adolescents. *Pediatrics*. 2004;114(2) (suppl 4th Report:555-576. [CrossRef]
5. Lurbe E, Agabiti-Rosei E, Cruickshank JK, et al. 2016 European Society of Hypertension guidelines for the management of high blood pressure in children and adolescents. *J Hypertens*. 2016;34(10):1887-1920. [CrossRef]
6. United States Food and Drug Administration. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. *AI/ML-Enabled Medical Devices* [Internet]; 2022. Available at: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device. Accessed February 24, 2025.
7. Pazoki R, Dehghan A, Evangelou E, et al. Genetic predisposition to high blood pressure and lifestyle factors: associations with midlife blood pressure levels and cardiovascular events. *Circulation*. 2018;137(7):653-661. [CrossRef]

8. Huang L, Trieu K, Yoshimura S, et al. Effect of dose and duration of reduction in dietary sodium on blood pressure levels: systematic review and meta-analysis of randomised trials. *BMJ*. 2020;368:m315. [CrossRef]
9. Siegel D, Lopez J, Meier J. Antihypertensive medication adherence in the Department of Veterans Affairs. *Am J Med*. 2007;120(1): 26-32. [CrossRef]
10. Munoz-Zuluaga C, Zhao Z, Wang F, Greenblatt MB, Yang HS. Assessing the accuracy and clinical utility of ChatGPT in laboratory medicine. *Clin Chem*. 2023;69(8):939-940. [CrossRef]
11. The Lancet Digital Health. ChatGPT: friend or foe? *Lancet Digit Health*. 2023;5(3):e102. [CrossRef]
12. Akyol Onder EN, Ensari E, Ertan P. ChatGPT-4o's performance on pediatric vesicoureteral reflux. *J Pediatr Urol*. 2024;21(2):504-509. (S1477-5131(24)00619-3). [CrossRef]
13. Flynn JT, Urbina EM, Brady TM, et al. Ambulatory blood pressure monitoring in children and adolescents: 2022 update: A scientific statement from the American Heart Association. *Hypertension*. 2022;79(7):e114-e124. [CrossRef]
14. Naz R, Akacı O, Erdoğan H, Açıkgöz A. Can large language models provide accurate and quality information to parents regarding chronic kidney diseases? *J Eval Clin Pract*. 2024;30(8):1556-1564. [CrossRef]
15. Qarajeh A, Tangpanithandee S, Thongprayoon C, et al. AI-powered renal diet support: performance of ChatGPT, Bard AI, and Bing Chat. *Clin Pract*. 2023;13(5):1160-1172. [CrossRef]
16. Lim ZW, Pushpanathan K, Yew SME, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *EBiomedicine*. 2023;95:104770. [CrossRef]
17. Pay L, Yumurtaş AÇ, Çetin T, Çınar T, Hayıroğlu Mİ. Comparative evaluation of chatbot responses on coronary artery disease. *Turk Kardiyol Dern Ars*. 2025;53(1):35-43. [CrossRef]
18. Ayers JW, Poliak A, Dredze M, 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. [CrossRef]
19. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *J Med Syst*. 2023;47(1):33. [CrossRef]