Anatolian Journal of Emergency Medicine 2026;9(1):39-43

ORIGINAL ARTICLE / ÖZGÜN ARAŞTIRMA MAKALESİ

The Role of Large Language Models in Pediatric Emergency Medicine: Accuracy and Decision-Support Potential of ChatGPT

Pediatrik Acil Tıpta Büyük Dil Modellerinin Rolü: ChatGPT'nin Doğruluğu ve Karar Destek Potansiyeli

Çağlar Kuas[1], Ertuğ Günsoy[2], Mustafa Emin Çanakçı[1], Murat Çetin[3], Volkan Ercan[1], Engin Özakın[1], Nurdan Acar[1]

ABSTRACT

Aim: To assess the accuracy and decision-support potential of ChatGPT in pediatric emergency practice by comparing its performance with human responses to structured multiple-choice questions.

Material and Methods: This cross-sectional study used 100 randomly selected questions from Pediatric Emergency Medicine: Just the Facts, Second Edition. The GPT-4o model was tested without prior prompts, and its answers were compared with the reference solutions and human accuracy rates reported in the source. Accuracy rates were calculated and compared using z-tests. Correlation between ChatGPT and human performance was analyzed with Spearman's test.

Results: ChatGPT answered 85 of 100 questions correctly, achieving an accuracy of 85% (95% CI: 78.0–92.0), which was significantly higher than the mean human accuracy of 54% (95% CI: 50.8–57.4) ($p < 0.001$). Topic-based analysis showed that ChatGPT's accuracy ranged from 75% to 100%, while human accuracy ranged from 30% to 65%, with higher variance. Among the 15 questions answered incorrectly by ChatGPT, 60% were case-based; the average correct human response rate for these was 35 ± 17%. A moderate positive correlation was observed between human and ChatGPT performance ($\rho = 0.40$, $p < 0.001$).

Conclusion: ChatGPT demonstrated high accuracy on structured pediatric emergency questions, suggesting potential as a supportive tool in decision-making and education. While its strengths lie in knowledge-based tasks, limitations remain in complex case-based reasoning. These findings indicate that LLMs could complement, but not replace, human expertise. Prospective studies are warranted to evaluate real-world integration in pediatric emergency care.

Keywords: ChatGPT, large language models, pediatric emergency medicine, decision support, artificial intelligence

ÖZ

Amaç: Bu çalışmada, yapılandırılmış çoktan seçmeli sorularda ChatGPT'nin doğruluğu insan yanıtlarıyla karşılaştırılarak, çocuk acil pratiğinde karar destek potansiyeli değerlendirildi.

Gereç ve Yöntemler: Kesitsel çalışmada Pediatric Emergency Medicine: Just the Facts, Second Edition kitabından rastgele seçilen 100 soru kullanıldı. GPT-4o modeli ek komut verilmeden test edildi. Yanıtlar kaynak cevaplarla ve literatürde bildirilen insan doğruluk oranlarıyla karşılaştırıldı. Doğruluk oranları hesaplandı; z-testi ile karşılaştırma, Spearman testi ile korelasyon analizi yapıldı.

Bulgular: ChatGPT 100 sorunun 85'ine doğru yanıt verdi (%85; GA %78,0–92,0). İnsan doğruluk oranı ortalama %54 (GA %50,8–57,4) olup fark istatistiksel olarak anlamlıydı ($p < 0,001$). Konu bazlı analizde ChatGPT'nin doğruluğu %75–100, insan yanıtlarının doğruluğu %30–65 arasında değişti. ChatGPT'nin yanlış yanıtladığı 15 sorunun %60'ı vaka bazlıydı; bu sorularda insan doğruluk oranı %35±17 idi. İnsan doğruluk oranları ile ChatGPT performansı arasında orta düzeyde pozitif korelasyon bulundu ($\rho = 0,40$; $p < 0,001$).

Sonuç: ChatGPT, bilgi yoğunluğu yüksek pediatrik acil sorularında yüksek doğruluk göstermiştir. Bulgular, modelin eğitim ve karar desteğinde tamamlayıcı bir araç olabileceğini göstermektedir. Bununla birlikte, vaka bazlı muhakemede sınırlılıkları devam etmektedir. ChatGPT ve benzeri modellerin klinik kullanıma entegrasyonu için ileriye dönük çalışmalara ihtiyaç vardır.

Anahtar Kelimeler: ChatGPT, büyük dil modelleri, pediatrik acil tıp, karar destek, yapay zekâ

Received: 2 September 2025 Accepted: 1 December 2025

[1]Eskişehir Osmangazi University Medical School, Department of Emergency Medicine, Eskişehir, Türkiye.

[2]University of Health Sciences, Van Training and Research Hospital, Emergency Department, Van, Türkiye.

[3]University of Health Sciences, Dr. Behçet Uz Pediatric Diseases And Surgery Education And Research Hospital, İzmir, Türkiye.

Corresponding Author: Çağlar Kuas, MD. **Address:** Büyükdere, Osmangazi University Meşelik Campus, Osmangazi University Health Practice and Research Hospital 26040 Odunpazarı, Eskişehir, Türkiye. **Telephone:** +905061768178 **E-mail**: dr.ckuas@gmail.com.

Atıf için/Cited as: Kuas Ç, Günsoy E, Çanakçı ME, et al. The Role of Large Language Models in Pediatric Emergency Medicine: Accuracy and Decision-Support Potential of ChatGPT. Anatolian J Emerg Med 2026;9(1): 39-43. https://doi.org/10.54996/anatolianjem.1776853.

Introduction

Emergency medicine is one of the most dynamic fields of medicine, characterized by rapid decision-making, time management, and high patient volume. This dynamic structure becomes even more pronounced in pediatric emergency departments. The clinical presentations of pediatric patients differ from those of adults; physical examination findings may be limited, medical history is often obtained through parents, and diagnostic and therapeutic protocols vary by age group. Therefore, it is critically important for physicians working in pediatric emergency departments to apply their broad knowledge base quickly and accurately. However, for physicians with limited clinical experience, coping with uncertainty, making decisions, and accessing reliable information remain significant challenges. At this point, the role of large language models (LLMs) in clinical applications comes into focus. Recent advances in artificial intelligence (AI), particularly in natural language processing (NLP), have created new opportunities in healthcare delivery (1). ChatGPT (Chat Generative Pre-trained Transformer), one of these models, has been investigated in fields such as medical education, patient communication, and decision support due to its capacity to provide human-like, guideline-consistent, and rapid responses based on a vast knowledge base (2).

The majority of studies in the literature assessing ChatGPT's capacity to provide decision support in emergency medicine have focused on adult patient scenarios (3,4). In contrast, the limited number of studies conducted in pediatric emergency medicine have reported that large language models can achieve accuracy rates comparable to or even higher than those of pediatric emergency physicians, even in complex clinical situations (5,6). Although these studies provide preliminary insights into the potential use of ChatGPT in pediatric emergency practice, scientific evidence in this area remains insufficient.

The aim of this study is to evaluate the usability of ChatGPT as a clinical decision support tool in pediatric emergency practice. By examining the accuracy of the model in response to structured questions specific to pediatric patients, this study seeks to shed light on the potential role of AI-based systems in pediatric emergency care. To this end, a comparative analysis with human responses was conducted to determine the reliability and consistency of ChatGPT in knowledge-based decision-making processes.

Material and Methods

This cross-sectional study was conducted using multiple-choice questions and employed the freely accessible version of ChatGPT. A total of 100 multiple-choice questions were randomly selected from the 1,003 pediatric emergency questions available in the digital textbook Pediatric Emergency Medicine: Just the Facts, Second Edition hosted on www.accessemergencymedicine.mhmedical.com. AccessEmergency Medicine is a regularly updated, comprehensive subscription-based online emergency medicine resource designed for emergency physicians and medical students (7). All questions were used in accordance with the AccessEmergencyMedicine content-use policy, and no copyrighted question text or proprietary material was reproduced in this manuscript. This study did not involve human participants, patient data, or biological samples. Therefore, ethical approval was not required.

The selected questions were individually entered into the GPT-4o model, developed by OpenAI and offered as the default version for free users as of 2024. No prompts or instructions were provided before each query; the model was simply asked to respond directly to the question. The responses generated by ChatGPT were compared with the reference answers provided in the source textbook and recorded. We used the accuracy rates reported in the source digital textbook, which reflect the percentage of correct answers submitted by platform users (medical students, residents, and physicians) who attempted each question. These published accuracy rates were extracted directly from the platform and were used as the measure of human performance in our study.

The included questions were classified into case-based and theoretical knowledge categories. ChatGPT's responses were categorized as correct or incorrect, and the proportions of correct and incorrect answers were calculated within each predefined question group.

For statistical analysis, descriptive statistics were first computed. Accuracy rates of ChatGPT and human responses were determined, and their overall means were compared. The significance of differences between the two groups was assessed using the z-test for two dependent proportions, with z values and corresponding p values reported. Spearman correlation analysis was performed to evaluate the relationship between ChatGPT and human accuracy. The correlation coefficient (r) and p value were calculated to determine the existence and direction of a linear relationship between the two variables. Additionally, the distribution of correct and incorrect responses by ChatGPT and humans was visualized using boxplots and scatterplots. Frequency distributions of categorical variables (e.g., question type, topic) were presented in frequency tables. Data were compiled in Microsoft Excel, and statistical analyses were performed using IBM SPSS Statistics for Windows, Version 21.0 (IBM Corp., Armonk, NY). A p value < 0.05 was considered statistically significant for all analyses.

Results

In this study, a total of 100 questions representing 18 clinical categories frequently encountered in pediatric emergency medicine were evaluated. The included questions were equally divided between case-based and theoretical formats. Content analysis revealed that the most common topic was infectious emergencies (n = 17), followed by pediatric trauma questions (n = 9). Among the subtopics, central nervous system infections and head trauma were the most frequently addressed areas (n = 4 each).

ChatGPT answered 85 out of 100 questions correctly, achieving an accuracy rate of 85% (95% CI: 78.0–92.0). The average accuracy rate of human responses was 54% (95% CI: 50.8–57.4). The difference between ChatGPT and human accuracy rates was statistically significant ($p < 0.001$). The z-test yielded a z value of 4.75 ($p < 0.001$).

When analyzed by topic, ChatGPT's accuracy was consistently higher than that of humans across all categories. ChatGPT's category-specific accuracy rates ranged from 75% to 100%, while human accuracy rates

ranged from 30% to 65%. Furthermore, the variance in human accuracy was observed to be higher across categories (Table 1). The diagnostic performance according to question type (case-based vs. theoretical) is presented in Table 2.

For the 15 questions that ChatGPT answered incorrectly, the mean percentage of correct human responses was 39 ± 14%. Of these 15 incorrectly answered questions, 9 (60%) were case-based questions. For these case-based questions, the mean percentage of correct human responses was 35 ± 17%. To assess the relationship between human accuracy and ChatGPT performance, a correlation analysis was conducted. A moderate positive and statistically significant correlation was observed between human accuracy rates and ChatGPT's accuracy (Spearman's ρ = 0.40, p < 0.001) (Figure 1).

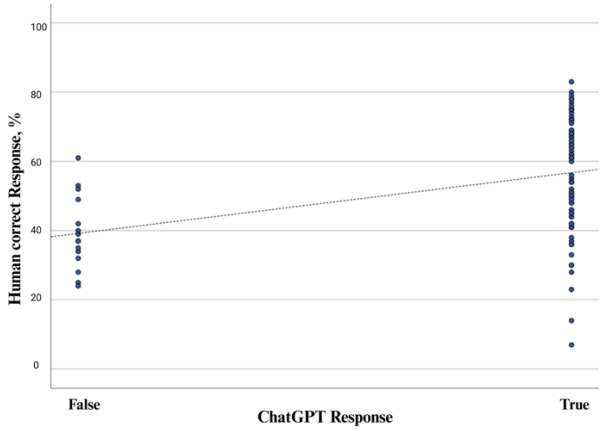

Figure 1. Correlation Between Human Accuracy and ChatGPT Correctness

Scatterplot demonstrating the relationship between ChatGPT's response accuracy (True/False) and human correct response rates. Each point represents a single question. A trend toward higher human accuracy is observed in items that ChatGPT answered correctly. Spearman correlation: ρ = 0.40, p < 0.001.

- ***X-axis (ChatGPT):*** *ChatGPT response (False = incorrect, True = correct)*
- ***Y-axis (Human Correct Response %):*** *Proportion of human respondents who answered each question correctly*

Discussion

In this study, the accuracy of responses provided by the large language model ChatGPT to theoretical and case-based multiple-choice questions specific to pediatric emergency medicine was evaluated in comparison with human responses. The findings demonstrated that ChatGPT's overall accuracy rate was significantly higher than that of human respondents. This result suggests that ChatGPT can exhibit high cognitive performance when faced with knowledge-based structured questions and holds potential as a decision-support tool in pediatric emergency settings.

In our study, the particularly high accuracy achieved by ChatGPT on theoretical knowledge questions indicates the model's strong capacity to access up-to-date medical content included in its training data. Consistent with this finding, the literature highlights that LLMs, owing to their high accuracy rates, may support clinical decision-making in pediatric emergency cases that involve diagnostic challenges (5). On the other hand, the relatively lower performance observed in case-based questions underscores the model's limitations in domains requiring higher cognitive skills such as clinical reasoning, contextual analysis, and causal inference. Since ChatGPT was not specifically designed to answer medical questions, it cannot be expected to fully comprehend the complex relationships among different clinical scenarios and treatment options. Although the model relies on vast and diverse data sources, issues such as the recency, contextual appropriateness, and lack of integration with clinical experience pose important constraints in generating accurate and reliable clinical decisions (8). This highlights that while ChatGPT is highly capable in knowledge-based tasks, it is not yet at a level to replace human experts in clinical decision-making processes.

A noteworthy finding was that questions incorrectly answered by ChatGPT also exhibited low average accuracy among human respondents, suggesting that these questions represent a high level of cognitive difficulty for both humans and AI models. Such questions are likely to involve uncertain clinical presentations, rare disease scenarios, or multi-step decision processes. This indicates that advanced clinical reasoning is especially critical in pediatric emergencies, where decision-support systems may play an important role. However, current LLMs do not yet possess sufficient contextual understanding and predictive capacity to fully meet this need (6). Therefore, combining the clinical expertise of healthcare professionals with the rapid information retrieval and analytical abilities of LLMs such as ChatGPT could pave the way for more comprehensive and personalized decision-support systems (9).

The positive and statistically significant correlation identified between human and model accuracy supports the notion that ChatGPT's responses are not random but systematically grounded in knowledge. This finding is important as it demonstrates that the model goes beyond simple memorization and can perform knowledge-based inference. Nevertheless, this performance should be carefully validated before integration into real-time patient management and clinical decision-making, and must be addressed within frameworks of ethics, safety, and verification.

The high accuracy rate achieved by ChatGPT in pediatric emergency-specific content suggests that such AI tools may serve a supportive role in areas such as medical education, exam preparation, and rapid access to information (6). For junior physicians and pediatric residents in particular, structured access to knowledge and support from decision-support systems may contribute both to patient safety and to clinical learning.

However, the scope of this study was limited to structured multiple-choice questions, which do not directly reflect real-life clinical scenarios. Thus, prospective studies based on patient cases are needed to evaluate the usability of models like ChatGPT for decision-support in pediatric emergency practice. It must also be emphasized that such tools should serve only an assistive and guiding role in medical decision-making, with ultimate responsibility remaining with the physician (10).

Several domain-specific AI systems have been developed in recent years to assist clinicians with narrowly defined diagnostic tasks—such as radiology-focused convolutional neural networks, dermatology image classifiers, or rule-based triage algorithms—each of which is trained on specialized datasets optimized for a single clinical problem (11). In contrast, GPT represents a general-purpose large

Topics	Total (n:100)	ChatGPT Correct Response n (%)	Human Correct Response %, (SD)
Infectious Emergencies	17	15 (88)	58±9
Trauma	9	8 (89)	45±11
Respiratory Emergencies	8	6 (75)	52±8
Cardiovascular Emergencies	8	7 (88)	61±11
Cardinal Presentations	7	7 (100)	63±10
Immunologic Emergencies	7	6 (86)	65±12
Toxicologic Emergencies	6	5 (83)	39±12
Gastrointestinal Emergencies	6	5 (83)	65±17
Genitourinary Emergencies	6	5 (83)	59±18
Sedation, Analgesia, and Imaging	5	3 (60)	49±18
Endocrine Emergencies	5	5 (100)	44±10
Emergency Medical Services and Mass Casualty Incidents	5	4 (80)	55±9
Hematologic and Oncologic Emergencies	5	4 (80)	58±7
Environmental Emergencies	2	1 (50)	33±11
Psychosocial Emergencies	2	2 (100)	49±11
Ophthalmologic Emergencies	1	1 (100)	49±17
Neurologic Emergencies	1	1 (100)	7±3
Total	**100**	**85 (85)**	**54**±8

Table 1. Performance of ChatGPT and Human Respondents in Pediatric Emergency Topics
Data are presented as **n (%)**. Percentages are calculated over non-missing observations.
Definition of percentages: **ChatGPT response (%)** = 100×(number of correctly answered items ÷ total items in the category). **Human response (%)** = the mean of item-level correct-response rates for the same items (for each item: 100×[# participants correct ÷ # participants responding], then averaged across items).

Question Category	ChatGPT Correct Response n (%)	Human Correct Response %, (SD)
Case-based (n:50)	36 (80)	52±17
Theoretical Knowledge (n:50)	39 (87)	56±19

Table 2: Diagnostic Performance by Question Type (Case-based vs. Theoretical)
Data are presented as **n (%)**. Percentages are calculated over non-missing observations.
Definition of percentages: **ChatGPT response (%)** = 100×(number of correctly answered items ÷ total items in the category). **Human response (%)** = the mean of item-level correct-response rates for the same items (for each item: 100×[# participants correct ÷ # participants responding], then averaged across items).

language model trained on broad and heterogeneous multimodal data, enabling flexible reasoning across a wide range of pediatric emergency scenarios rather than a single diagnosis category. This distinction may explain why GPT demonstrated strong alignment with human diagnostic reasoning despite not being trained specifically for pediatric emergency medicine. Our findings, therefore, complement the growing literature suggesting that general LLMs can approximate or augment the performance of domain-specific tools, while offering substantially wider applicability in educational and decision-support contexts (10).

Limitations

This study has several limitations. First, the evaluation was restricted to structured multiple-choice questions, which do not directly represent real-life clinical practice. ChatGPT's performance may vary when considering clinical uncertainties, incomplete patient histories, and interpretation of physical examination findings. Additionally, the questions were selected from a single textbook, which limits the generalizability of the findings to the broader spectrum of pediatric emergency medicine. The version of ChatGPT used in this study is a continuously updated model, and accuracy rates may differ across versions. Moreover, human responses were evaluated only through the example user results provided in the textbook, without assessing the real-time performance of practicing clinicians.

Conclusion

This study demonstrated that ChatGPT achieved high accuracy when answering knowledge-intensive structured questions in pediatric emergency medicine. These findings suggest that AI-based language models, when appropriately trained and supervised, may serve as complementary tools in clinical environments. However, further research is needed to assess their reliability, limitations, and impact on clinical outcomes.

Conflict of Interest: The authors declare that there is no conflict of interest.

Financial Support: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' Contribution CK: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Writing - original draft preparation, Writing - review & editing **EG:** Conceptualization, Investigation, Writing - review & editing **MEÇ:** Writing - original draft,

Writing - review & editing **MÇ:** Conceptualization, Formal analysis, Validation, Writing - review & editing VE: Writing - original draft, Writing - review & editing **EÖ:** Conceptualization, Methodology, Writing - original draft **NA:** Conceptualization, Formal analysis, Methodology, Project administration, Supervision, Validation, Writing - original draft preparation, Writing - review & editing

All authors read and approved the final submitted version of the manuscript. All authors have agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature.

Ethical Approval: This study did not involve human participants, patient data, or biological samples. The research was conducted exclusively using publicly available multiple-choice questions from a published textbook (Pediatric Emergency Medicine: Just the Facts, Second Edition) and the freely accessible version of ChatGPT. Since no patient information or clinical interventions were included, institutional review board (IRB) or ethics committee approval was not required.

References

1. Berg H ten, van Bakel B, van de Wouw L, et al. ChatGPT and Generating a Differential Diagnosis Early in an Emergency Department Presentation. Ann Emerg Med. 2023 Sep;S019606442300642X.
2. Preiksaitis C, Ashenburg N, Bunney G, et al. The Role of Large Language Models in Transforming Emergency Medicine: Scoping Review. JMIR Med Inform. 2024 May 10;12:e53787.
3. Hoppe JM, Auer MK, Strüven A, Massberg S, Stremmel C. ChatGPT With GPT-4 Outperforms Emergency Department Physicians in Diagnostic Accuracy: Retrospective Analysis. J Med Internet Res. 2024 Jul 8;26:e56110.
4. Kaboudi N, Firouzbakht S, Shahir Eftekhar M, et al. Diagnostic Accuracy of ChatGPT for Patients' Triage; a Systematic Review and Meta-Analysis. Arch Acad Emerg Med. 2024 Jul 30;12(1):e60.
5. Del Monte F, Barolo R, Circhetta M, et al. Diagnostic efficacy of large language models in the pediatric emergency department: a pilot study. Front Digit Health. 2025;7:1624786.
6. Ramgopal S, Varma S, Gorski JK, Kester KM, Shieh A, Suresh S. Evaluation of a Large Language Model on the American Academy of Pediatrics' PREP Emergency Medicine Question Bank. Pediatr Emerg Care. 2024 Dec;40(12):871–5.
7. AccessEmergency Medicine. About AccessEmergency Medicine. McGraw Hill. Web site. Available at: https://accessemergencymedicine.mhmedical.com/ss/about.aspx. Accessed 20 July 2025.
8. 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 Mar 4;47(1):33.
9. Liu J, Wang C, Liu S. Utility of ChatGPT in Clinical Practice. J Med Internet Res. 2023 Jun 28;25:e48568.
10. Fatima A, Shafique MA, Alam K, Fadlalla Ahmed TK, Mustafa MS. ChatGPT in medicine: A cross-disciplinary systematic review of ChatGPT's (artificial intelligence) role in research, clinical practice, education, and patient interaction. Medicine (Baltimore). 2024 Aug 9;103(32):e39250.
11. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56.