ORIGINAL ARTICLE ÖZGÜN ARAŞTIRMA

Evaluating the Quality of Responses by ChatGPT-3.5, Google Gemini, and Microsoft Copilot to Common Pediatric Questions: A Content-Based Assessment

ChatGPT-3.5, Google Gemini ve Microsoft Copilot'un Yaygın Pediatrik Sorulara Verdiği Yanıtların Kalitesinin Değerlendirilmesi: İçerik Tabanlı Bir Değerlendirme

Abdulkerim Elmas (0009-0002-3788-8325), Mustafa Akçam (0000-0002-4635-7633)

Süleyman Demirel University Faculty of Medicine, Department of Pediatrics, Division of Pediatric Gastroenterology, Hepatology and Nutrition, Isparta, Türkiye

Cite this article as: Elmas A, Akçam M. Evaluating the quality of responses by ChatGPT-3.5, Google Gemini, and Microsoft copilot to common pediatric questions: a content-based assessment. J Curr Pediatr. 2025;23(3):164-171

Abstract

Introduction: The objective of this study was to compile a list of the most frequently asked questions by parents during their first visits to the pediatrician and to evaluate the reliability and success of responses provided by artificial intelligence-supported chatbots against these questions.

Materials and Methods: The 20 most frequently asked questions by parents of infants during their pediatrician outpatient visits were posed to ChatGPT3.5, Google Gemini, and Microsoft Copilot applications. The responses provided by the applications were evaluated on a Likert scale from 1 (least adequate) to 5 (most adequate) by a pediatric gastroenterologist, pediatrician, and pediatric assistant, all of whom were physicians.

Results: Upon scoring the responses provided by artificial intelligence (AI) applications to the 20 questions posed, Google Gemini was found to have received the highest score (286) and was statistically significant ($p < 0.001$). No significant difference was observed when Copilot and ChatGPT were compared. Upon evaluation of responses generated by AI applications, pediatricians were found to have assigned the highest ratings.

Conclusion: The Gemini AI application demonstrated greater success than ChatGPT3.5 and Copilot in responding. While AI chatbots demonstrate the capability to deliver information, advice, and guidance regarding health and diseases, it is imperative that the responses generated by these systems undergo rigorous evaluation by healthcare professionals.

Öz

Giriş: Bu çalışmanın amacı, ebeveynlerin çocuk doktoruna yaptıkları ilk başvurular sırasında en sık sordukları soruların bir listesini derlemek ve yapay zekâ destekli sohbet robotları tarafından bu sorulara verilen yanıtların güvenilirliğini ve başarısını değerlendirmektir.

Gereç ve Yöntem: Bebeklerin poliklinik ziyaretleri sırasında ebeveynleri tarafından en sık sorulan 20 soru, ChatGPT3.5, Google Gemini ve Microsoft Copilot uygulamalarına yöneltilmiştir. Uygulamalar tarafından verilen yanıtlar; bir çocuk gastroenteroloğu, bir çocuk sağlığı ve hastalıkları uzmanı ve bir pediatri asistanı olmak üzere üç hekim tarafından 1 (en yetersiz) ile 5 (en yeterli) arasında derecelendirilen Likert ölçeği kullanılarak değerlendirilmiştir.

Keywords
Artificial intelligence, pediatrics, chatbot, parental counseling, content evaluation

Anahtar kelimeler
Yapay zeka, pediatri, chatbot, ebeveyn danışmanlığı, içerik değerlendirme

Received/Geliş Tarihi : 01.05.2025
Accepted/Kabul Tarihi : 29.07.2025
Published Date/ Yayınlanma Tarihi : 29.12.2025

DOI:10.4274/jcp.2025.40370

Address for Correspondence/Yazışma Adresi:
Abdulkerim Elmas, Süleyman Demirel University Faculty of Medicine, Department of Pediatrics, Division of Pediatric Gastroenterology, Hepatology and Nutrition, Isparta, Türkiye
E-mail: akelmas@gmail.com

Bulgular: Yapay zekâ (YZ) uygulamalarının yöneltilen 20 soruya verdikleri yanıtların puanlanması sonucunda, en yüksek puanı Google Gemini almış (286) ve bu sonuç istatistiksel olarak anlamlı bulunmuştur (p < 0,001). Copilot ve ChatGPT karşılaştırıldığında ise anlamlı bir fark saptanmamıştır. Yapay zekâ uygulamaları tarafından üretilen yanıtların değerlendirilmesi sonucunda, en yüksek puanlamayı çocuk sağlığı ve hastalıkları uzmanlarının yaptığı belirlenmiştir.

Sonuç: Gemini yapay zekâ uygulaması, ChatGPT3.5 ve Copilot'a kıyasla sorulara yanıt verme konusunda daha başarılı bulunmuştur. Yapay zekâ destekli sohbet robotları; sağlık ve hastalıklarla ilgili bilgi, öneri ve rehberlik sunma potansiyeline sahip olmakla birlikte, bu sistemler tarafından üretilen yanıtların sağlık profesyonelleri tarafından titiz bir değerlendirmeye tabi tutulması zorunludur.

Introduction

In recent years, the application of artificial intelligence (AI) has become increasingly prevalent in all areas of our daily lives. Despite the perception that AI will not replace human doctors in the health sector, it is anticipated that it will assist in diagnosing and treating patients through algorithms (1). Experimental studies on AI are being conducted in some hospitals worldwide. Artificial intelligence-supported chat robots (AISR), which can interact with users using natural language, have started to replace conventional search engines with the widespread use of smart devices (2). It is anticipated that AI will be extensively employed in the future to address individuals' health concerns (3). There is a growing trend towards the utilization of AI-powered online platforms that offer health-related advice. Nevertheless, concerns about the reliability of these platforms persist (4,5).

ChatGPT3.5 is an OpenAI-developed AISR with the most commonly known natural language processing and machine learning capabilities. According to analyst data, the application, which was released in November 2022, achieved a remarkable milestone of over 100 million users within a mere two months. This exponential growth trajectory established it as an unparalleled phenomenon in the realm of consumer applications (6). Despite generating highly detailed and persuasive responses to a wide range of health-related inquiries, from general patient questions to complex scientific queries posed by medical professionals, these systems frequently produce inaccurate and contradictory information (7). Google Gemini was developed by Alphabet and DeepMind, one of Google's parent companies, in the final months of 2023 and was made available to users in 2024. Microsoft Copilot is an AI-based chatbot developed by Microsoft. The present landscape is characterized by a multitude of AI models designed to execute a broad spectrum of tasks, encompassing image and sound processing, creative generation, computational operations, and statistical analysis. Our research focused on three specific AI chatbots due to their accessibility at no cost, prevalence in real-world applications, and the substantial backing provided by large corporations for their underlying infrastructure and ongoing development.

The lack of experience, limited knowledge, fear of making mistakes, and protective instinct are among the factors that contribute to stress and depression in first-time parents (8). Parental self-efficacy is the self-confidence that parents possess to have children and fulfill child-rearing tasks. A study found that parental self-efficacy was low in families with a nuclear family structure that did not receive support from family elders (9). It is not uncommon for parents to be unable to reach a pediatrician after the birth of their child and to search for information on the internet instead. Subsequent to the COVID-19 pandemic, the exponential growth in telehealth services has positioned AI-driven chatbots as indispensable tools for patient engagement and remote healthcare delivery (10). The growing reliance of parents on electronic resources to alleviate medical concerns and obtain expert opinions has yielded several potential advantages. One such advantage is the 24/7 accessibility of AISRs, providing a convenient resource for parents, particularly during off-hours (11-13). This technology can be particularly beneficial in reducing the burden on healthcare providers in developing countries where access to care is often limited, especially for rural populations and the uninsured by offering an alternative means of delivering healthcare services and contributing to the reduction of disparities in access and quality (14).

Previous research has not yielded any articles that assessed the sufficiency and trustworthiness of responses generated by AISRs to the frequently asked questions of parents regarding pediatric care. The role of AISRs in healthcare delivery is a subject of considerable debate. While proponents extol their potential to address individual health concerns and reduce the workload of healthcare professionals, critics caution against their limitations, such as the accuracy of AI-generated diagnoses, the unique nature of individual patient presentations, and the lack of human oversight. The ongoing debate underscores the need for further research to determine the most appropriate and reliable methods for integrating AISRs into clinical practice (6,15,16). Recent systematic reviews have highlighted the potential and limitations of AI chatbots in pediatric settings, particularly regarding the accuracy of medical information and parental satisfaction (4,17). These findings underscore the need for real-world evaluations, as addressed in our study.

This study aims to evaluate the accuracy and reliability of AI-generated responses in addressing parents' questions with a focus on content. The evaluation will be conducted using ChatGPT-3.5, Google Gemini, and Microsoft Copilot.

Materials and Methods

The study was conducted at Suleyman Demirel University, Department of Pediatrics, in accordance with the Declaration of Helsinki. Given that the study did not entail the use of personal data, human participants, or medical records, it was concluded that review by an ethics committee was not required. The 20 most frequent questions were selected based on a combined approach: (1) a systematic review of online sources, (2) the authors' own clinical experience, and (3) a nationwide consultation via messaging apps with actively practicing pediatricians to validate the representativeness of these questions (Table 1).

Table 1. Most frequently asked questions to pediatricians
Can I give my baby a pacifier?
When can the baby be bathed after birth?
How can I tell if breast milk is enough for my baby?
Is it recommended to use a baby walker for babies?
What foods should not be given to the baby before the age of one year?
Does my baby have gas pains, and how can I help them?
How often should I change my baby's diaper, and how I do it?
How can I establish my baby's sleep patterns?
How can I strengthen my baby's immune system and protect them from diseases?
Should I worry if my baby hiccups or sneezes frequently?
Which products should I use for my baby's skincare, and which products should I avoid?
What activities can I do for my baby's emotional and mental development?
Can I let my baby watch television?
What should I pay attention to for my baby's ear care, and how should ear cleaning be done?
My baby has hair loss. Is it normal?
Does the temperature rise after vaccination in babies? What should I do if they have a fever?
When and how long should I allow my baby to be exposed to sunlight, and how should sun protection be provided?
Is my baby's breathing normal, and what should I pay attention to regarding breathing problems?
Is my baby's appetite normal, and is it getting enough food? How can I assess this?
Is my baby's sweating normal, and what should I do to prevent excessive sweating?

Questions were asked to AISRs in Turkish. A comprehensive literature review was conducted using the search term 'most common questions asked of pediatricians.' Additionally, pediatricians practicing in diverse regions of Turkey were contacted via messaging apps to gather firsthand information on the most frequently asked questions by new parents. The collected data, combined with our own clinical expertise, was analyzed to identify the top 20 most recurrent questions. These questions formed the foundation of our research. The questions were posed to ChatGPT3.5, Google Gemini, and Microsoft Copilot AI software in the same format in May 2024 in a new chat window to minimize the influence of previous posts, and the responses were recorded without data loss. The study was completed with the active versions of all three AI chatbot tools in May 2024. The questions were then scored by a pediatric gastroenterology specialist (PGS), a pediatrician (P), and a pediatric assistant (PA) using a 5-point Likert scale. Three clinicians, PGS, P, and PA, possessing 35, 15, and 3 years of clinical experience respectively, were tasked with independently assessing the accuracy and reliability of responses generated by the AISR. To ensure objectivity, each clinician evaluated the responses across five predefined categories without knowledge of the others' assessments. In accordance with the aforementioned criteria, the following responses were recorded: (1) AI provided an incorrect answer; (2) AI was unable to provide an adequate response and could not identify the correct source; (3) AI was unable to provide an adequate response but suggested the correct source; (4) AI provided an adequate response but not an optimal one; (5) AI provided an optimal response. According to this scaling, the lowest rating was given to (1), while the highest rating was given to (5). The scoring system yielded total scores ranging from 20 to 100. The absence of a validity and reliability assessment for this system was recognized as a limitation of the present study.

Statistical Analysis

IBM SPSS Statistics 27 (Corp. I. IBM SPSS Statistics for Windows. Version 270. Armonk: NY: IBM Corp; Released 2020). package program was used for statistical analysis. Since each of the three clinicians evaluated the same 20 questions across three different AI applications using a 5-point Likert scale, the data represent repeated measures with related (dependent) samples. Additionally, as the Likert scale provides ordinal data and the assumption of normality was not met, non-parametric methods were preferred. The Friedman test was used to compare the differences in scores across the three AI systems, which is the appropriate non-parametric alternative to repeated-measures ANOVA. Pairwise comparisons were conducted using the Wilcoxon signed-rank test. Statistically, $p<0.05$ was considered statistically significant.

Results

A comparison of the scores obtained by the physicians of the AISR revealed that Gemini received the highest score, while ChatGPT3.5 and Copilot received the lowest score (Table 2).

In the evaluation, PGS ranked Gemini as the highest performing model and ChatGPT 3.5 as the lowest. Similarly, P ranked Gemini highest and ChatGPT lowest. PA's evaluation indicated Gemini as the top-performing model while Copilot was ranked the lowest. The results demonstrated that Gemini scored statistically significantly higher than the other AIs ($p<0.001$). No statistically significant difference was observed when Copilot and ChatGPT were compared.

Table 2. Comparison of the scores of AI applications

	ChatGPT3.5	Gemini	Copilot	p*
PGS median total	4 (1-5) 73	5 (4-5) 95	4 (2-5) 74	<0.001
P median total	4 (3-5) 75	5 (4-5) 96	4 (2-5) 79	<0.001
PA median total	4 (3-5) 76	5 (3-5) 95	3,5 (2-5) 71	<0.001
Total score	224	286	224	

Descriptive statistics are given as median (min.-max.) and total score.
*Friedman test, **Wilcoxon test
PGS: Pediatric gastroenterology specialist, P: Pediatrician, PA: Pediatric assistant

Upon evaluation of responses generated by AI applications, pediatrician were found to have assigned the highest ratings (Table 3).

A statistically significant difference was identified between the scores assigned by P and PA to Copilot ($p = 0.033$).

The items that all AIs most successfully answered were questions 12 and 14. These were: 'What kind of activities can I do for my baby's emotional and mental development?' and 'What should I pay attention to for my baby's ear care, and how should ear cleaning be done?'. The questions numbered 5 and 6, which inquired about the foods that should not be given to babies before the age of one year and about the causes and treatment of infant gas pains, respectively, were the least successfully answered. A detailed breakdown of the scores assigned by the evaluators is provided in Table 4.

The analysis revealed that the AI model exhibited significant inaccuracies when responding to questions concerning nutrition and colic. These errors may be attributed to the specific phrasing of the questions or to inherent limitations within the AI model, such as the generation of hallucinated content. While the overall evaluation suggests satisfactory performance in terms of accuracy and reliability, the identified shortcomings in the context of health-related inquiries warrant further investigation.

Discussion

Our study represents the inaugural investigation of the utilization of AISR in our country's pediatrics domain. The results of this study highlight the promising potential of AI technologies in the healthcare sector. However, given the nascent stage of these technologies, ongoing evaluation by domain experts is crucial to ensure their reliability and safety. Rather than focusing on technological differences between AI platforms, this study aimed to determine the reliability of

Table 3. Comparison of the scores given by the physicians to the applications

	PGS score median (min-max)	P score median (min-max)	PA score median (min-max)	p*	PGS-P**	PGS-PA**	P-PA**
ChatGPT-3.5	4 (1-5)	4 (3-5)	4 (3-5)	0.828	0.763	0.405	0.705
Gemini	5 (4-5)	5 (4-5)	5 (3-5)	0.819	0.564	1	0.564
Copilot	4 (2-5)	4 (2-5)	3,5 (2-5)	0.08	0.132	0.366	0.033
Total score	242	250	242				

Descriptive statistics are given as median (min.-max.) and total score.
*Friedman test, **Wilcoxon test
PGS: Pediatric gastroenterology specialist, P: Pediatrician, PA: Pediatric assistant

Table 4. Median scores (min-max) given to each AI model by question

Question No	Question Topic	ChatGPT3.5	Gemini	Copilot	Best Scoring Model
Q1	Use of pacifiers	4 (4-4)	5 (5-5)	3 (2-3)	Gemini
Q2	Bathing after birth	3 (4-4)	4 (3-4)	5 (5-5)	Copilot
Q3	Breastfeeding adequacy	3 (3-4)	5 (5-5)	3 (3-4)	Gemini
Q4	Baby walker use	4 (3-4)	5 (5-5)	4 (3-4)	Gemini
Q5	Unsafe foods before 1 year	3 (1-4)	5 (4-5)	3 (2-3)	Gemini
Q6	Infantile colic / gas pain	4 (3-4)	4 (4-4)	2 (2-3)	ChatGPT and Gemini
Q7	Diaper change	4 (4-4)	5 (5-5)	3 (3-4)	Gemini
Q8	Sleep patterns	4 (4-4)	5 (5-5)	4 (3-5)	Gemini
Q9	Strengthening immunity	5 (4-5)	5 (4-5)	4 (3-4)	ChatGPT and Gemini
Q10	Hiccups and sneezing	3 (3-3)	5 (5-5)	3 (3-4)	Gemini
Q11	Skincare products	4 (4-4)	5 (5-5)	4 (4-4)	Gemini
Q12	Emotional and mental development	4 (4-5)	5 (5-5)	5 (5-5)	Gemini and Copilot
Q13	Screen time	4 (4-5)	5 (5-5)	5 (4-5)	Gemini and Copilot
Q14	Ear care	5 (4-5)	5 (5-5)	5 (4-5)	Equal
Q15	Hair loss	3 (3-4)	5 (5-5)	5 (4-5)	Gemini and Copilot
Q16	Post-vaccine fever	4 (3-4)	5 (5-5)	4 (4-5)	Gemini
Q17	Sun exposure and protection	4 (4-4)	5 (5-5)	4 (3-4)	Gemini
Q18	Breathing patterns	4 (3-4)	5 (4-5)	2 (2-3)	Gemini
Q19	Appetite and feeding assessment	3 (3-3)	5 (4-5)	4 (4-4)	Gemini
Q20	Sweating	4 (3-4)	4 (4-4)	3 (3-4)	ChatGPT and Gemini

Descriptive statistics are given as median (min-max). N/A: Not applicable

chatbot responses from the perspective of pediatric care. The findings have direct implications for how parents interact with AI tools when pediatric consultation is not immediately accessible.

Conversational tools that establish dialogue with the user by mimicking human interaction through written, verbal, and visual communication are referred to as AISR. With the increasing use of technological devices (e.g., smartphones and computers) and access to the Internet, AISR is becoming accessible and interesting. They offer the potential to provide health-related information and autonomous services, which could be promising for technology-assisted interventions. Moreover, these chatbots have the potential to alleviate current healthcare resource burdens by automating functions that previously required face-to-face interaction (18). Gonsard et al. (19) aimed to assess the acceptability of AI-powered home monitoring systems among pediatric asthma patients. Their findings revealed a notable generational gap, with adolescents expressing a more positive attitude towards self-management using AI-driven tools than their parents. Nevertheless, at this juncture, healthcare professionals must validate the veracity of the information provided by AI. The application of AI to analyze vast datasets and medical records has yielded remarkable results in the diagnosis of complex and intricate diseases (20,21). Ying et al. (22) demonstrated that while ChatGPT performed reasonably well in providing responses to queries related to the diagnosis and screening of pediatric endocrine disorders, it exhibited limitations in its ability to account for nuances within disease subgroups. Furthermore, the study highlighted the inconsistency of responses across different languages, suggesting a lower level of reliability.

In our study, the questions were posed in the same format to the AISRs, who responded with varying content and length. In similar studies conducted by Taşkın et al. (23) and Perez-pino et al. (24), it was observed that the AISRs provided different responses to the same questions. This discrepancy affects the dependability of the received output. The discrepancy may be attributed to the sources from which the AISR is derived. In our study, it was observed that the answers given by AI were long. In fact, if we had designed a study comparing the answers given by clinicians and the answers given by AI to the same questions, we could have obtained different results. As a matter of fact, in a study conducted on this subject, the answers given by clinicians were found to be shorter and more superficial than the answers given by AI (25). However, we do not know how AISRs will perform when responding to patient questions in a real clinical setting. We hope that research on this subject will encourage future studies for the routine use of AI in the healthcare.

Although studies examining the answers given by asking health-related questions to AI are increasing today, they are still few. In the ophthalmology clinic, the AISR, was employed for diagnostic and triage purposes. The ChatGPT4 achieved the highest accuracy rate (3). In our study, Gemini was the application that received the most successful responses to the questions. The pediatrician assigned the highest scores overall, which may reflect greater familiarity with AI interfaces or a more forgiving evaluation approach compared to the pediatric gastroenterologist or assistant. This evaluator effect underlines the subjectivity inherent in expert scoring, despite efforts to standardize the evaluation categories. The pediatric assistant tended to give more conservative scores compared to the specialist physicians. This may reflect a more cautious approach due to limited clinical authority or less familiarity with AI-generated content. Understanding such evaluator variability is crucial for interpreting subjective rating-based research. Future studies may benefit from including more raters and inter-rater reliability testing to strengthen the generalizability of findings. ChatGPT4 is a more recent and paid version than the previous version, ChatGPT3.5. We used the free version instead of ChatGPT4 because we prefer AI applications that are easily and freely accessible to the general public.

In our study, Google Gemini received significantly higher scores than ChatGPT3.5 and Copilot across all evaluators. This result may be attributable to Gemini's underlying model infrastructure, which was observed to provide more structured, concise, and medically relevant responses. Notably, Gemini performed especially well in questions related to infant care routines and developmental advice, such as emotional development and hygiene practices, whereas it underperformed—along with other models—in addressing nutrition-related concerns like gas pain or food restrictions. These findings suggest that current AI systems may be more reliable for behavioral and general care topics than for complex, medically nuanced issues requiring clinical judgement. This demonstrated that parents can obtain accurate responses to certain queries through AI applications without consulting a physician. However, AI applications that lack a robust infrastructure comprising health professionals may provide erroneous and inadequate responses, potentially posing significant legal and vital risks. In this context, Rokhshad et al. (26) asserted that chatbots are valuable tools for training and disseminating patient information. Still, they are not yet equipped to replace physicians in making diagnostic decisions.

AISRs can support the simple questions of patients with messages during the busy shifts of clinicians or allied health personnel. However, it should be reviewed and evaluated by the healthcare personnel that correct and consistent answers are given to the questions by AI. Thus, in countries with limited health personnel and clinicians, time savings and the ability to assign personnel to more critical units can be achieved. Especially out-of-hours patients who have problems in reaching the health centre and who cannot take time off from their workplace can get answers to their health-related questions quickly and unnecessary clinic visits and loss of labour force can be prevented. If more patients' questions are answered quickly, empathically and to a high standard, unnecessary clinic visits may decrease and resources may be freed up for those in need (27).

There is a pressing need for comprehensive studies on the use of AI, which has become a popular source of health information in recent years. Given that AI constantly evolves, further studies utilizing the latest AI versions may be advisable. It is important to recognize that the results of such studies may differ significantly over time. Future research could involve comparative analyses of responses provided by advanced AI applications and human healthcare professionals to fundamental health-related queries.

Study Limitations

The main limitation of our study was that it compared fixed answers to specific questions. Since the patient's previous health records were not analysed here, personalised disease states may have been omitted because the patient-physician relationship and related conditions vary. As it is known, diseases can be personalised and may occur with different symptoms instead of the same symptoms in every situation and in every individual. In addition, the directions made by the AI's answers to the questions in the clinicians' opinions were not taken into account. The absence of a validity and reliability assessment for this system was recognized as a limitation of the present study. If the study had evaluated the answers given by clinicians and AI with unbiased physicians, different results may have emerged. Due to the evolving nature of AI platforms, the same question may yield different answers over time. This temporal variability limits reproducibility and generalizability. Although the evaluators of our study were blinded to each other, they were also co-authors of the study and this may have biased the study.

Conclusion

The potential exists for AI applications to alleviate the burden on healthcare systems in developing countries. In our study, Gemini, Copilot, and ChatGPT-3.5 demonstrated satisfactory performance in general and exhibited considerable potential for patient information and education. Nevertheless, it is necessary to conduct an evaluation and preliminary examination by experts in the before recommending the use of AI in healthcare. Although our study shows promising results, it needs to be studied for a long time due to its limitations and ethical issues related to AI-supported healthcare.

Ethics

Ethical Approval: The study was conducted at Süleyman Demirel University, Department of Pediatrics, in accordance with the Declaration of Helsinki. Given that the study did not entail the use of personal data, human participants, or medical records, it was concluded that review by an ethics committee was not required.

Footnotes

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The authors declared that this study received no financial support.

References

1. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2:230-43.
2. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence-based chatbots for promoting health behavioral changes: systematic review. J Med Internet Res. 2023;25:e40789.
3. Lyons RJ, Arepalli SR, Fromal O, Choi JD, Jain N. Artificial intelligence chatbot performance in triage of ophthalmic conditions. Can J Ophthalmol. 2024;59:e301-8.
4. Nov O, Singh N, Mann D. Putting ChatGPT's medical advice to the (turing) test: survey study. JMIR Med Educ. 2023;9:e46939.
5. Altamimi I, Altamimi A, Alhumimidi AS, Altamimi A, Temsah MH. Snakebite advice and counseling from artificial intelligence: an acute venomous snakebite consultation with ChatGPT. Cureus. 2023;15:e40351.
6. Le M, Davis M. ChatGPT yields a passing score on a pediatric board preparatory exam but raises red flags. Glob Pediatr Health. 2024;11:2333794X241240327.
7. Thirunavukarasu AJ, Hassan R, Mahmood S, Sanghera R, Barzangi K, El Mukashfi M, et al. Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. JMIR Med Educ. 2023;9:e46599.
8. Serhan N, Ege E, Ayrancı U, Kosgeroglu N. Prevalence of postpartum depression in mothers and fathers and its correlates. J Clin Nurs. 2013;22:279-84.
9. Montigny F, Lacharité C. Perceived parental efficacy: concept analysis. J Adv Nurs. 2005;49:387-96.
10. Bhaskar S, Nurtazina A, Mittoo S, Banach M, Weissert R. Editorial: telemedicine during and beyond COVID-19. Front Public Health. 2021;9:662617.
11. Doraiswamy S, Abraham A, Mamtani R, Cheema S. Use of telehealth during the COVID-19 pandemic: scoping review. J Med Internet Res. 2020;22:e24087.
12. Wong J, Foussat AC, Ting S, Acerbi E, van Elburg RM, Mei Chien C. A Chatbot to engage parents of preterm and term infants on parental stress, parental sleep, and infant feeding: usability and feasibility study. JMIR Pediatr Parent. 2021;4:e30169.
13. Kadariya D, Venkataramanan R, Yip HY, Kalra M, Thirunarayanan K, Sheth A. kBot: Knowledge-enabled personalized chatbot for asthma self-management. proc int conf smart comput SMARTCOMP. 2019;2019:138-43.
14. West CP, Dyrbye LN, Shanafelt TD. Physician burnout: contributors, consequences and solutions. J Intern Med. 2018;283:516-29.
15. Dash D, Thapa R, Banda JM. Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare delivery. arXiv preprint. arXiv:230413714. 2023.
16. Harskamp RE, De Clercq L. Performance of ChatGPT as an AI-assisted decision support tool in medicine: a proof-of-concept study for interpreting symptoms and management of common cardiac conditions (AMSTELHEART-2). Acta Cardiol. 2024;79:358-66.
17. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence-based chatbots for promoting health behavioral changes: systematic review. J Med Internet Res. 2023;25:e40789.
18. de Cock C, Milne-Ives M, van Velthoven MH, Alturkistani A, Lam C, Meinert E. Effectiveness of conversational agents (virtual assistants) in health care: protocol for a systematic review. JMIR Res Protoc. 2020;9:e16934.
19. Gonsard A, AbouTaam R, Prévost B, Roy C, Hadchouel A, Nathan N, et al. Children's views on artificial intelligence and digital twins

for the daily management of their asthma: a mixed-method study. Eur J Pediatr. 2023;182:877-88.

20. Yuan C, Adeloye D, Luk TT, Huang L, He Y, Xu Y, et al. The global prevalence of and factors associated with Helicobacter pylori infection in children: a systematic review and meta-analysis. Lancet Child Adolesc Health. 2022;6:185-94.
21. Ren Y, Loftus TJ, Datta S, Ruppert MM, Guan Z, Miao S, et al. Performance of a machine learning algorithm using electronic health record data to predict postoperative complications and report on a mobile platform. JAMA Netw Open. 2022;5:e2211973.
22. Ying L, Li S, Chen C, Yang F, Li X, Chen Y, et al. Screening/diagnosis of pediatric endocrine disorders through the artificial intelligence model in different language settings. Eur J Pediatr. 2024;183:2655-61.
23. Taşkın S, Geçgelen Cesur M, Uzun M. Yapay zekâ destekli sohbet robotlarinin yaygin ortodontik sorulari cevaplama başarisinin değerlendirilmesi. Med J Sdu. 2023;30:680-6.
24. Perez-Pino A, Yadav S, Upadhyay M, Cardarelli L, Tadinada A. The accuracy of artificial intelligence-based virtual assistants in responding to routinely asked questions about orthodontics. Angle Orthod. 2023;93:427-32.
25. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, Faix DJ, Goodman AM, Longhurst CA, Hogarth M, Smith DM. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183:589-96.
26. Rokhshad R, Zhang P, Mohammad-Rahimi H, Pitchika V, Entezari N, Schwendicke F. Accuracy and consistency of chatbots versus clinicians for answering pediatric dentistry questions: A pilot study. J Dent. 2024;144:104938.
27. Rasu RS, Bawa WA, Suminski R, Snella K, Warady B. Health literacy impact on national healthcare utilization and expenditure. Int J Health Policy Manag. 2015;4:747-55.