

# The potential of AI applications to provide clinical counseling to parents: a pilot study

©Nilsun Kuas<sup>\*1</sup>, ©Özgür Aktaş<sup>2</sup>, ©Ergün Bahadırhan Karkın<sup>3</sup>, ©Berkay Tekkanat<sup>3</sup>

<sup>1</sup>Department of Pediatric Urology, Eskişehir City Hospital, Eskişehir, Türkiye

<sup>2</sup>Department of Pediatric Surgery, Eskişehir City Hospital, Eskişehir, Türkiye

<sup>3</sup>Department of Pediatric Surgery, Faculty of Medicine, Eskişehir Osmangazi University, Eskişehir, Türkiye

Received: 08/09/2026

Accepted: 09/10/2026

Published: 20/10/2026

Cite this article: Kuas N, Aktaş Ö, Karkın EB, Tekkanat B. The potential of AI applications to provide clinical counseling to parents: a pilot study. *Surg Child.* 2026;3(4):133-138. doi:10.51271/SOC-0086

\*Corresponding Author: Nilsun Kuas, dr.nilsun@hotmail.com, dr.nilsunnn@gmail.com

## ABSTRACT

**Aims:** This study aimed to evaluate the responses of current large language model-based Artificial Intelligence (AI) applications to pediatric surgery and urology clinical scenarios using expert physician-defined criteria, and to investigate their potential for providing clinical guidance to parents.

**Methods:** Ten clinical questions were submitted to three AI applications (ChatGPT-5.2, DeepSeek-V3.1, Gemini 3). Responses were scored blindly by four specialist physicians using ten criteria across four domains (maximum, 20 points per question). A safety veto criterion (E1) was included as an independent safeguard; any response failing to recognize a clinical emergency or containing potentially harmful information was rendered invalid regardless of other scores.

**Results:** No safety related veto was applied. The overall mean scores were as follows: ChatGPT-5.2: 19.8/20, Gemini 3: 19.8/20 and DeepSeek-V3.1: 19.6/20. The lowest single-rater scores were observed in DeepSeek-V3.1 responses for varicocele (S6, min: 16/20) and intermittent hernias (S7, min: 16/20), although question-level mean scores for these items remained high (18,3/20 and 18,8/20, respectively).

**Conclusion:** In this pilot evaluation, current LLM-based AI applications demonstrated promising performance in responding to parental questions in pediatric urology and general pediatrics, though these findings should be interpreted within the scope of the limited number of clinical scenarios evaluated. They may have potential as a source of supplementary information but cannot replace physician consultations.

**Keywords:** Artificial Intelligence, urology, child, pediatrics, generative Artificial Intelligence, caregivers

## INTRODUCTION

The use of Artificial Intelligence (AI) applications in the healthcare field has been rapidly increasing in recent years. Large Language Model (LLM) based applications are increasing accessibility to medical information; parents are increasingly turning to these tools to access quick and reliable information about their children's health problems.<sup>1,2</sup>

Parents whose children are sick experience high levels of anxiety and uncertainty due to the nature of the disease. This uneasiness; It can cause even non-emergency situations to be brought to the emergency room and outpatient clinics to be burdened with unnecessary application.<sup>3,4</sup> Studies in Türkiye have shown that the vast majority of admissions to pediatric emergency departments are not truly emergencies, placing a significant burden on the healthcare system.<sup>5</sup> In addition, the majority of parents turn to social media and internet resources for information about their children's health

problems.<sup>6,7</sup> Content with questionable medical accuracy and lacking any regulatory mechanisms is widely circulated on Instagram, YouTube, and similar platforms; this can create an environment where parents are misled.<sup>8</sup> Having a reliable digital advisory resource has the potential to reduce unnecessary emergency calls while also decreasing the risk of genuine emergencies going unnoticed.

Pediatric and urological complaints frequently create uncertainty and anxiety for parents. Life-threatening emergencies such as testicular torsion, incarcerated hernia, or neonatal sepsis, along with elective conditions requiring evaluation such as hydronephrosis, varicocele, or phimosis, constitute a wide clinical spectrum. Accurate classification of this spectrum and guidance of parents on appropriate steps are critical from both a medical and ethical perspective.



Current AI language models such as ChatGPT, Gemini, and DeepSeek have become systems capable of interpreting complex medical questions and generating context-aware responses.<sup>9,10</sup> However, the extent to which these responses are adequate in terms of clinical accuracy, safety, and comprehensibility by the patient's family has not been sufficiently investigated. In this study, ten commonly encountered pediatric/urological questions representing different levels of urgency and clinical profiles were created; AI responses were evaluated independently and blindly by four specialist physicians.

## METHODS

This study is a descriptive and comparative pilot study in which the responses of AI language models to pediatric clinical questions were examined through expert physician evaluation. This study was approved by the Ethics Committee for Scientific Researches at Eskişehir City Hospital (Date: 10.04.2026, Decision No: ESH/BAEK 2026/367). All procedures were carried out in accordance with the ethical rules and the principles of the Declaration of Helsinki.

Ten realistic clinical questions were prepared by the research team. The questions, which included emergency (S1-S3), semi-emergency (S4-S7), and elective (S8-S10) categories, were presented in a natural language format that parents would use to direct their complaints, which are frequently encountered in clinical practice, to an AI application (Table 1).

In this study, free new membership login web-based applications ChatGPT-5.2 (OpenAI), DeepSeek-V3.1 (DeepSeek AI), and Gemini 3 (Google DeepMind) were used. Each clinical question was asked in a separate, newly opened chat. Each questions was presented to each application individually in independent sessions using the same prompt format. All queries were completed in April 2026.

Responses were independently and blindly evaluated by four specialist physicians (Assessor 1–4). The four assessors were the study authors, all of whom are specialists in pediatric surgery or pediatric urology. Each assessor independently completed a separate evaluation form for each AI application and submitted their scores without access to the ratings of other assessors. As all assessors were active users of the evaluated AI applications, complete blinding to the identity of the AI models cannot be guaranteed; however, inter-assessor independence was maintained throughout the scoring process. The evaluation form consisted of four domains and ten sub-criteria (Table 2). Each criterion was scored from 0 (inappropriate)–1 (partially appropriate)–2 (appropriate); the maximum answer score was 20. The evaluation criteria were developed by the authors with the primary aim of assessing the feasibility and accuracy of AI applications as a counseling resource for parents with low health literacy. Criteria were organized under four domains: urgency optimization (A), correct referral unit (B), clinical knowledge adequacy (C), and health literacy and comprehensibility (D). This framework was informed by common clinical scenarios encountered in daily pediatric and urological practice, typical parental concerns, and the core question of whether AI-

**Table 1.** Clinical questions used in the study

| No  | Category    | Clinical question                                                                      | Expected diagnosis          | Referral            |
|-----|-------------|----------------------------------------------------------------------------------------|-----------------------------|---------------------|
| S1  | Emergency   | Sudden-onset severe testicular pain and nausea in a 13-year-old boy                    | Testicular torsion          | Emergency           |
| S2  | Emergency   | Persistent inguinal swelling and redness in a 3-month-old infant                       | Incarcerated hernia         | Emergency           |
| S3  | Emergency   | 38.5°C fever, poor feeding, and foul-smelling urine in a 6-week-old infant             | Febrile UTI/neonatal sepsis | Emergency           |
| S4  | Semi-urgent | Antenatally detected renal swelling in a 10-day-old infant                             | Antenatal hydronephrosis    | Elective specialist |
| S5  | Semi-urgent | 39°C fever and crying during urination in a 2-year-old boy                             | Febrile UTI                 | Same-day physician  |
| S6  | Semi-urgent | Increasing testicular pain when standing and visible veins in a 15-year-old adolescent | Varicocele                  | Elective specialist |
| S7  | Semi-urgent | Inguinal swelling that appears when crying in a 4-year-old child                       | Inguinal hernia             | Elective specialist |
| S8  | Elective    | Non-retractable foreskin and white material in a 3-year-old boy                        | Phimosis/smegma             | Routine check-up    |
| S9  | Elective    | Intermittently non-palpable testis in a 5-year-old child                               | Retractile testis           | Elective specialist |
| S10 | Elective    | Swelling, redness, and crusting on day 5 after circumcision                            | Normal healing              | Routine follow-up   |

UTI: Urinary tract infection

**Table 2.** Evaluation criteria

| Group | Criterion                                        | Description                                                                               |
|-------|--------------------------------------------------|-------------------------------------------------------------------------------------------|
| A     | A1–No unnecessary emergency referral             | Does not make an unnecessary emergency referral in elective/semi-urgent situations        |
| A     | A2–Does not cause delay in emergencies           | Recommends waiting is not advised in emergency scenarios; emergency is not missed         |
| B     | B1–Correct referral unit suggestion              | Suggests the correct specialty, such as pediatric surgery, urology, or pediatrics         |
| B     | B2–Consistent with the Turkish healthcare system | Provides referral guidance consistent with Turkish healthcare practices                   |
| C     | C1–Warning signs stated                          | Red-flag symptoms are clearly stated                                                      |
| C     | C2–Clear conditions for re-presentation          | States when the patient should see a doctor again                                         |
| C     | C3–Diagnosis/condition consistency               | The response is sufficiently consistent with the probable diagnosis or clinical condition |
| D     | D1–Language understandable for parents           | Medical terms are conveyed in language appropriate for parents                            |
| D     | D2–Clear action plan provided                    | Clearly states what the parent should do                                                  |
| D     | D3–Balanced, reassuring tone                     | Demonstrates a reassuring approach without creating unnecessary panic                     |

generated responses contain sufficient, correctly directed, and understandable information for a non-medical caregiver. The E1 safety veto was defined as any response that failed to identify a clinical emergency—specifically, a condition in which delayed diagnosis or treatment could result in loss of life or organ function—or that contained information with potential to cause harm. Activation of the E1 veto rendered the entire response invalid, regardless of scores on other criteria.

## RESULTS

### Safety Assessment

All three AI applications produced safe responses in all questions; no E1 safety veto was applied in any assessment.

### Question-Based Overall Scores

The question-based averages of the four evaluators are presented in **Table 3**. ChatGPT-5.2 received an overall average score of 19.8/20, Gemini 3 19.8/20, and DeepSeek-V3.1 19.6/20. The lowest question averages were observed in DeepSeek-V3.1's S6 (varicocele: 18.3/20) and S7 (inguinal hernia: 18.8/20) responses.

**Table 3. Question-based mean scores—average of 4 raters (max: 20)**

| Question                   | ChatGPT-5.2 mean | DeepSeek-V3.1 mean | Gemini 3 mean |
|----------------------------|------------------|--------------------|---------------|
| S1–Testicular torsion      | 19.3             | 20.0               | 20.0          |
| S2–Incarcerated Hernia     | 20.0             | 19.8               | 19.8          |
| S3–Febrile UTI <3 months   | 20.0             | 20.0               | 20.0          |
| S4–Prenatal hydronephrosis | 20.0             | 20.0               | 19.8          |
| S5–Febrile UTI >3 months   | 19.8             | 19.5               | 20.0          |
| S6–Varicocele              | 19.8             | 18.3               | 19.5          |
| S7–Intermittent hernia     | 19.5             | 18.8               | 20.0          |
| S8–Phimosis/smegma         | 19.8             | 19.8               | 20.0          |
| S9–Retractile testis       | 20.0             | 20.0               | 19.3          |
| S10–Post-circumcision      | 20.0             | 20.0               | 20.0          |
| Overall mean               | 19.8             | 19.6               | 19.8          |

20=Full score, 18–19=Partially deficient, ≤17=Insufficient, D1=Rater 1 (Kuas N.), D2=Rater 2 (Aktaş Ö.), D3=Rater 3 (Karkin EB.), D4=Rater 4 (Tekkanat B.), Max: Maximum, UTI: Urinary tract infection

### Criteria-Based Assessment

The criterion-based averages of the four evaluators are presented in **Table 4**. All applications demonstrated high performance in criteria groups A and C. The most significant shortcomings were; in DeepSeek-V3.1, it was found to be in the B1 (correct referral unit recommendation: 18.5/20) and D group communication criteria (D1–D3 average: 19.0/20), and in ChatGPT-5.2, it was found to be in the D3 (balanced tone: 19.0/20) criterion.

### Performance by Question Categories

Emergency questions (S1–S3): In the testicular torsion, incarcerated hernia, and febrile UTI <3 months questions, all applications received full or near-full scores from all four evaluators. In terms of safety, complete performance was demonstrated in all emergency questions. Semi-emergency questions (S4–S7): In the varicocele (S6) and intermittent hernia (S7) questions, DeepSeek-V3.1 received relatively low scores. Evaluator 1 gave 16/20 for S6; Evaluator 2 gave 17/20 for S6 and 16/20 for S7. The possible reasons for these

**Table 4. Criterion-based mean scores—average of 4 raters (max: 20)**

| Evaluation criterion                          | ChatGPT-5.2 mean | DeepSeek-V3.1 mean | Gemini 3 mean |
|-----------------------------------------------|------------------|--------------------|---------------|
| A1–No unnecessary emergency referral          | 20.0             | 19.8               | 20.0          |
| A2–Does not cause delay in emergencies        | 19.5             | 19.8               | 20.0          |
| B1–Correct referral unit suggestion           | 19.8             | 18.5               | 19.8          |
| B2–Consistency with Turkish healthcare system | 20.0             | 20.0               | 20.0          |
| C1–Warning signs stated                       | 20.0             | 19.8               | 19.3          |
| C2–Clear conditions for re-presentation       | 20.0             | 20.0               | 20.0          |
| C3–Diagnosis/condition consistency            | 20.0             | 19.8               | 20.0          |
| D1–Language understandable for parents        | 20.0             | 19.5               | 20.0          |
| D2–Clear action plan provided                 | 19.5             | 19.5               | 19.3          |
| D3–Balanced, reassuring tone                  | 19.0             | 19.3               | 20.0          |
| Overall criterion mean                        | 19.8             | 19.6               | 19.8          |

20=Full score, 18–19=Partially deficient, ≤17=Insufficient, D1=Rater 1 (Kuas N.), D2=Rater 2 (Aktaş Ö.), D3=Rater 3 (Karkin EB.), D4=Rater 4 (Tekkanat B.), Max: Maximum

declines are deficiencies in the criteria of correct referral unit recommendation (B1) and parent communication (group D). ChatGPT-5.2 showed the most consistent performance in this group. Elective questions (S8–S10): All interventions received high scores in the phimosis/smegma, retractile testis, and post-circumcision care questions. ChatGPT-5.2's score of 18/20 from Assessor 4 in question S7 is due to a partial deficiency in criterion D2 (clear action plan).

No statistically significant differences were found among the three models. Wilcoxon post-hoc comparisons, when performed with Bonferroni correction, also showed no significant pairwise comparisons. Since the Shapiro-Wilk test rejected normal distribution in all models ( $p < 0.01$ ), a non-parametric approach was used instead of parametric tests.

Inter-rater reliability was assessed using the intraclass correlation coefficient, ICC(2,k) (two-way random-effects, average-measures, absolute agreement), based on the total score (0–20) for each question. The overall ICC(2,k) was 0.35 (95% CI: –0.11 to 0.65), with model-specific values of 0.08 for ChatGPT-5.2, 0.40 for DeepSeek-V3.1, and 0.45 for Gemini 3, all corresponding to poor reliability according to the criteria of Koo and Li.<sup>11</sup> Fleiss' kappa, calculated on the criterion-level 0/1/2 ratings, likewise yielded an overall value of 0.08 (slight agreement); no kappa could be computed for criteria B2 and C2, as all raters consistently assigned the maximum score (2) on these items, resulting in zero variance. These low reliability coefficients should be interpreted in light of the uniformly high scores observed across raters (**Tables 3, 4**): the clustering of nearly all scores within the 18–20/20 range (a ceiling effect) indicates not genuine disagreement among raters, but rather a limited capacity of the scale to discriminate between raters at this performance level. Effect sizes (Cohen's d) for pairwise comparisons of mean question-level scores were moderate for ChatGPT vs. DeepSeek ( $d = 0.48$ ) and for DeepSeek vs. Gemini ( $d = -0.53$ ), and negligible for ChatGPT vs. Gemini ( $d = -0.10$ );



however, none of these practical differences reached statistical significance ( $n=10$ , power-limited).

## DISCUSSION

This pilot study suggests that current LLM-based AI applications may produce clinically appropriate responses to urological and pediatric clinical questions in children, based on expert physician evaluations of a limited set of clinical scenarios. ChatGPT-5.2 (19.8/20) and Gemini 3 (19.8/20) achieved near-perfect scores, while DeepSeek-V3.1 (19.6/20) showed relative deficiencies, particularly in varicocele and inguinal hernia questions. These preliminary findings are consistent with the possibility that AI applications may hold potential for parental clinical guidance, warranting further investigation with a larger and more diverse set of scenarios.<sup>12,13</sup>

In Türkiye, pediatric outpatient and emergency department admissions are significantly higher than the global average in both number and intensity. Studies show that a large proportion of pediatric emergency department admissions do not truly constitute an emergency; most admissions consist of problems that can be resolved at the primary and secondary care levels.<sup>3-5</sup> One study reported that in Türkiye, pediatricians have to work with 60-100 patients a day in outpatient clinics, and the consultation time per patient is only 3-5 minutes.<sup>14</sup> This table clearly demonstrates the need for reliable tools that can provide parents with accurate preliminary information without requiring them to visit a hospital.

High levels of anxiety exhibited by parents regarding childhood illnesses are a significant predictor of unnecessary emergency room visits.<sup>3</sup> Fever phobia is the best-documented example of this; a significant number of parents perceive even low fevers as a potential danger and seek emergency medical attention.<sup>15</sup> Similar anxiety patterns can be observed in urological complaints such as groin swelling or urinary problems; this can lead to both unnecessary burden on emergency services and, sometimes, the risk of overlooking real emergencies.

Today's parents are increasingly turning to social media for health information, rather than relying on advice from neighbors or older family members.<sup>6</sup> Platforms such as Instagram, YouTube and Facebook have become parents' main sources of information about child health, and it is known that most of the content on these platforms is produced by individuals without medical education.<sup>7</sup> Systematic reviews of the prevalence of health misinformation on social media highlight that YouTube is a particularly undesirable source of health information.<sup>8</sup> In this context, AI applications have the capacity to provide more structured, coherent, and clinically based information, in contrast to static and uncontrolled social media content. Indeed, in our study, no response was blocked by a security veto; all applications correctly identified emergency questions and highlighted the need for immediate response.<sup>12</sup>

In a recent study that tested the performance of ChatGPT-4o, Gemini, and DeepSeek-V3 regarding metabolic and bariatric

surgery using 30 questions and evaluating the responses of 93 specialists on a scale of 0-3, DeepSeek-V3 was found to be significantly superior to other applications in generating guideline-compliant answers. This indicates the need for field-specific studies before actively implementing these applications in the healthcare field.<sup>16</sup>

The performance of DeepSeek-R1, ChatGPT (GPT-o3-mini), and Gemini 2.0 Flash was evaluated in German using 200 interdisciplinary multiple-choice questions drawn from the German Progress Test Medicine, spanning both preclinical and clinical domains. All three models achieved high accuracy (DeepSeek-R1: 96%, Gemini 2.0 Flash: 94%, ChatGPT: 92.5%), and pairwise comparisons using the Cochran Q and McNemar tests with Bonferroni correction revealed no statistically significant differences between the models, which parallels our findings. Notably, despite this high accuracy, performance for both DeepSeek-R1 and ChatGPT declined as question length increased, indicating that item-level consistency was not uniform even among high-performing models; this pattern is conceptually similar to the limited inter-rater reliability observed among our own evaluators, in that high overall accuracy did not guarantee stable consistency at the item level. This demonstrates that high individual accuracy does not necessarily translate into consistent performance across models or between human evaluators. With different questioning formats, whether using multiple-choice questions or a conversational format, it can be argued that these models have significant potential as clinical decision support tools.<sup>17</sup>

In their 2024 study, Tortum<sup>18</sup> compared the decisions of triage nurses with AI recommendations using ChatGPT, Gemini, and Pi. Unlike this study, although ChatGPT provided the closest triage decisions to doctors, they concluded that the rate of incomplete triage was 26% for the yellow zone and 42% for the red zone. Of course, it is expected that the decisions of the research team within the same clinic would be more consistent internally. In light of today's findings, it is also true that AI is not completely and reliably accurate. However, it could at least replace other fixed-source platforms for parents with low health literacy.

Eraybar<sup>19</sup> compared the triage decisions of 86 healthcare professionals trained in triage and AI applications in 50 virtual patient scenarios. They examined the responses of Google Bard and OpenAI applications. The evaluation showed that healthcare professionals gave significantly more correct answers compared to Google Bard and ChatGPT ( $n=30.7$  vs.  $n=25.5$ ). In this study, no significant difference was found between ChatGPT and Bard for low and high severity scenarios ( $p=0.821$ ). AI applications may not be as successful in determining triage color codes as they are in distinguishing between urgent and non-urgent cases. However, this result may also be due to the fact that the answer key was still generated by healthcare professionals. Although urgency level performance was not formally evaluated in our study, the evaluated AI responses did not fail to flag emergency conditions within the ten scenarios tested. This pilot-level observation may be of interest for future studies examining AI utility for parents with low health literacy.



Forty questions aimed at evaluating the reliability, quality, and readability of responses related to urolithiasis in children were administered to the ChatGPT-4o twice, one week apart. The ChatGPT-4o was found to provide reliable and high-quality information about urinary stones in children; however, the advanced reading level of its responses posed a significant hindrance to comprehensibility for patients and caregivers. In our study, all models scored highly in terms of comprehensibility. This may be due to the questions being phrased in everyday parenting language.<sup>20</sup>

One of the key differences between AI applications and traditional internet resources is their dialogical nature. Parents can engage in the conversation by providing additional information, clarifying vague concerns, and receiving more detailed, situation-specific responses. This interactive feature aligns with research findings suggesting that AI agents can better inform individuals compared to static web pages.<sup>21</sup> Thus, AI applications can help parents more accurately assess whether their complaint is truly urgent.

Despite all these findings, it is clear that AI applications cannot replace medical examinations.<sup>9,10</sup> In situations requiring physical examination, diagnostic testing, and clinical judgment, AI responses may only serve as preliminary information. However, within the scope of this pilot study, no safety-critical failures were identified, as no EI safety veto was applied by any evaluator across the ten scenarios tested. Whether this finding extends to broader real-world clinical practice requires evaluation with a larger and more representative scenario set.

### Limitations

This study has some limitations; the number of questions is relatively small, and the fact that all evaluators received similar training may affect the scoring results. Indeed, the calculated inter-rater reliability coefficients (ICC and Fleiss' kappa) were found to be weak-slight, largely due to the limited variance of the scores because of the ceiling effect; it is recommended that more distinctive scales with wider scoring ranges be used in future studies. Since the questions were based on possible clinical situations, they may not reflect real-world parental use.

### CONCLUSION

This pilot study suggests that current AI language models (ChatGPT-5.2, DeepSeek-V3.1, and Gemini 3) produced clinically appropriate responses to urological and pediatric clinical questions in children within the evaluated scenarios. ChatGPT-5.2 achieved the highest overall average (19.8/20), Gemini 3 showed similar performance (19.8/20), while DeepSeek-V3.1 performed relatively worse in semi-emergency scenarios (19.6/20). No safety vetoes were applied across the ten tested scenarios. These findings are preliminary and should be interpreted cautiously; AI tools may serve as a supportive preliminary information source for parents where physician access is limited, but cannot replace physician consultation. Larger and more diverse studies are needed to confirm these findings.

## ETHICAL DECLARATIONS

### Ethics Committee Approval

This study was approved by the Ethics Committee for Scientific Researches at Eskişehir City Hospital (Date: 10.04.2026, Decision No: ESH/BAEK 2026/367).

### Informed Consent

As no human participants or patient-level data were included, informed consent was not required.

### Peer Review Process

This manuscript was subject to external peer review.

### Conflict of Interest

The authors declare no conflicts of interest related to this study.

### Financial Disclosure

The authors received no financial support for the conduct or publication of this research.

### Author Contributions

Concept: NK, ÖA; Design: NK; Consultation: ÖA, BT; Data Collection and/or Processing: NK, ÖA, EBK, BT; Analysis and/or Interpretation: NK, ÖA; Literature Review: BT, EBK; Manuscript Preparation: NK; Critical Review: All Authors.

## REFERENCES

1. Topol EJ. High-performance medicine: the convergence of human and Artificial Intelligence. *Nat Med*. 2019;25(1):44-56. doi:10.1038/s41591-018-0300-7
2. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)*. 2023;11(6):887. doi:10.3390/healthcare11060887
3. Akbayram HT, Coskun E. Paediatric emergency department visits for non-urgent conditions: can family medicine prevent this? *Eur J Gen Pract*. 2020;26(1):134-139. doi:10.1080/13814788.2020.1825676
4. Yüksel A. Characteristics of pediatric emergency patients in a Turkish tertiary-level hospital: a 1-year cross-sectional study. *SAGE Open Med*. 2022;10:20503121221135401. doi:10.1177/20503121221135401
5. Saygili S, Demirbas KC, Bağcı B, et al. Underserved and undervalued: the alarming collapse in pediatric residency preference in Türkiye. *Turk Arch Pediatr*. 2025;60(4):352-354. doi:10.5152/TurkArchPediatr.2025.1112254
6. Frey E, Bonfiglioli C, Brunner M, Frawley J. Parents' use of social media as a health information source for their children: a scoping review. *Acad Pediatr*. 2022;22(4):526-539. doi:10.1016/j.acap.2021.12.006
7. Smailhodzic E, Hooijsma W, Boonstra A, Langley DJ. Social media use in healthcare: a systematic review of effects on patients and on their relationship with healthcare professionals. *BMC Health Serv Res*. 2016;16(1):442. doi:10.1186/s12913-016-1691-0
8. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. *J Med Internet Res*. 2021;23(1):e17187. doi:10.2196/17187
9. 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. doi:10.1001/jamainternmed.2023.1838
10. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930-1940. doi:10.1038/s41591-023-02448-8
11. Koo TK, Li MY. A Guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155-163. doi:10.1016/j.jcm.2016.02.012
12. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172-180. doi:10.1038/s41586-023-06291-2



13. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198
14. Çıraklı Ü. 18 OECD ülkesinde 2002-2018 yılları arası kişi başı ortalama hasta muayene sürelerinin karşılaştırılması. *Sağlıkta Kalite ve Akreditasyon Derg*. 2020;3(1):43-54.
15. Crocetti M, Moghbeli N, Serwint J. Fever phobia revisited: have parental misconceptions about fever changed in 20 years? *Pediatrics*. 2001;107(6):1241-1246. doi:10.1542/peds.107.6.1241
16. Hany M, Zidan MH, Parmar C, et al. Accuracy and knowledge base evaluation of ChatGPT-4o, Gemini-2.0-Flash, and DeepSeek-V3 in metabolic and bariatric surgery: an expert-rated blinded study. *Obes Surg*. 2026;36(5):2160-2171. doi:10.1007/s11695-026-08562-z
17. Meyer A, Karay Y, Steinbicker AU, Streichert T, Overbeek R. Performance of DeepSeek-R1, ChatGPT (GPT-o3-mini), and Gemini 2.0 Flash on German medical multiple-choice questions: comparative evaluation. *JMIR Form Res*. 2025;9:e77357. doi:10.2196/77357
18. Tortum F, Kasali K. Exploring the potential of Artificial Intelligence models for triage in the emergency department. *Postgrad Med*. 2024; 136(8):841-846. doi:10.1080/00325481.2024.2418806
19. Eraybar S, Dal E, Aydın MO, Begenen M. Transforming emergency triage: a preliminary, scenario-based cross-sectional study comparing Artificial Intelligence models and clinical expertise for enhanced accuracy. *Bratisl Lek Listy*. 2024;125(11):738-743. doi:10.4149/BLL\_2024\_114
20. Ensari E, Boztas AE. Evaluation of ChatGPT-4o' responses on pediatric urolithiasis: is it useful? *Urolithiasis*. 2025;53(1):202. doi:10.1007/s00240-025-01880-4
21. Laranjo L, Dunn AG, Tong HL, et al. Conversational agents in healthcare: a systematic review. *J Am Med Inform Assoc*. 2018;25(9): 1248-1258. doi:10.1093/jamia/ocy072