J Pediatr Res 2026;13(3):294-302
DOI: 10.4274/jpr.galenos.2026.37039

Original Article

# Large Language Models in Pediatric Penile Conditions: Guideline Concordance, Readability, and Language-Based Performance

Bilge Türedi[1], Ali Sezer[2]

[1]University of Health Sciences Türkiye, Bakırköy Dr. Sadi Konuk Training and Research Hospital, Clinic of Pediatric Urology, İstanbul, Türkiye
[2]University of Health Sciences Türkiye, Kocaeli City Hospital, Clinic of Pediatric Urology, Kocaeli, Türkiye

**ABSTRACT**

**Aim:** Large language models (LLMs) are increasingly being used by patients and caregivers in order to obtain medical information. This study evaluated the guideline concordance and public readability of two contemporary LLMs in answering frequently asked questions regarding common pediatric penile conditions.

**Materials and Methods:** Five common pediatric penile conditions were evaluated: phimosis, hypospadias, congenital penile curvature, paraphimosis, and buried penis. Ten standardized questions were developed for each condition, covering diagnosis, treatment, surgical considerations, complications, and long-term outcomes. A total of 50 questions were submitted to two widely used LLMs in both English and Turkish. The responses were assessed according to the European Association of Urology Paediatric Urology Guidelines and graded as fully concordant (Grade 1), partially concordant (Grade 2), partially concordant with inaccuracies (Grade 3), or non-concordant (Grade 4). Public readability was evaluated using a five-point scale.

**Results:** ChatGPT (OpenAI, GPT-5.5 version; accessed June 2026) achieved Grade 1 concordance in 40 out of 50 responses (80%), with 49 responses (98%) considered acceptable (Grade 1 or 2). Gemini (Google DeepMind, Gemini 3.5 Pro version; accessed June 2026) achieved Grade 1 concordance in 41 of 50 responses (82%), and all responses (100%) were classified as acceptable. No Grade 4 responses were identified, and only one Grade 3 response was observed overall. Concordance rates did not differ significantly between the 2 models ($p>0.05$). Public readability was high for both systems (4.60±0.58 vs. 4.67±0.49; $p>0.05$). No significant differences were observed between the English and Turkish responses. Reduced concordance occurred in different questions, with only two overlapping areas identified.

**Conclusion:** Both LLMs demonstrated high guideline concordance and excellent public readability. No clinically significant misinformation was identified, and most deficiencies reflected omissions of specific guideline details. LLMs may serve as reliable educational resources but should complement, rather than replace, professional medical consultation.

**Keywords:** Artificial intelligence, clinical practice guidelines, hypospadias, large language models, patient education

**Corresponding Author**
Bilge Türedi, MD, University of Health Sciences Türkiye, Bakırköy Dr. Sadi Konuk Training and Research Hospital, Clinic of Pediatric Urology, İstanbul, Türkiye
**E-mail:** blgtrd@gmail.com **ORCID:** orcid.org/0000-0003-3532-0912
**Received:** 13.06.2026 **Accepted:** 09.07.2026 **Epub:** 21.08.2026 **Publication Date:** 18.09.2026

**Cite this article as:** Türedi B, Sezer A. Large language models in pediatric penile conditions: guideline concordance, readability, and language-based performance. J Pediatr Res. 2026;13(3):294-302

Copyright 2026 The Author(s). Published by Galenos Publishing House on behalf of Ege University Faculty of Medicine, Department of Pediatrics and Ege Children's Foundation. Licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0)

## Introduction

Pediatric penile conditions constitute a common source of concern for parents and caregivers. Conditions such as phimosis, hypospadias, congenital penile curvature (CPC), paraphimosis, and buried penis frequently prompt families to seek medical information long before specialist consultation. Although many of these conditions are benign or treatable, misconceptions and misinformation may lead to unnecessary anxiety, delayed presentation, inappropriate treatment attempts, or unrealistic expectations regarding outcomes (1).

In recent years, large language models (LLMs) have emerged as increasingly popular sources of health information. Millions of users now consult conversational artificial intelligence platforms for medical advice, explanations of diagnoses, and treatment recommendations. Consequently, the quality, accuracy, and readability of the information generated by these systems have become important topics of investigation (2). Early studies evaluating LLM-generated medical information reported variable performance, with concerns regarding factual inaccuracies, hallucinations, and incomplete recommendations. However, rapid model development and continuous training have resulted in substantial improvements in both medical accuracy and user-oriented communication (3).

Several studies have evaluated the quality of online health information using content obtained from search engines, social media platforms, video-sharing websites, or frequently asked questions (FAQs) collected from public sources (4). However, standardized evaluations of LLM performance within narrowly defined pediatric urological domains remain limited. Furthermore, comparisons between different languages and between contemporary LLM platforms are scarce (5).

Pediatric penile conditions represent an ideal model for such an evaluation as they are frequently encountered in clinical practice, generate substantial parental concern, and are associated with numerous recurring questions regarding diagnosis, treatment, surgery, long-term outcomes, and fertility (6).

The aim of this study was to assess the guideline concordance and public readability of LLM-generated responses regarding common pediatric penile conditions and to compare the performance of two contemporary LLM platforms across different languages.

## Materials and Methods

### Study Design

This cross-sectional comparative study evaluated the quality and guideline concordance of responses generated by two LLMs regarding common pediatric penile conditions. This assessment focused on both the medical accuracy and the public readability of the responses provided in both English and Turkish.

### Selection of Clinical Topics and Questions

Five commonly encountered pediatric penile conditions were selected based on their clinical relevance and frequency in pediatric urology practice:

1. Phimosis,
2. Hypospadias,
3. CPC,
4. Paraphimosis,
5. Buried penis.

In order to ensure a systematic and comprehensive assessment across all conditions, questions were developed according to predefined thematic domains commonly addressed by patients and caregivers seeking health information. These domains included:

- Definition,
- Epidemiology,
- Etiology,
- Symptoms,
- Diagnosis,
- Classification (when applicable),
- Treatment,
- Surgical considerations,
- Complications,
- Long-term outcomes.

For each disease, ten FAQs were generated, resulting in a total of 50 questions. Equivalent question sets were prepared in both English and Turkish.

### LLM Response Generation

Each question was submitted separately to the evaluated LLMs in both English and Turkish. All responses were generated during the same study period using the default model settings without additional prompting, refinement, or follow-up interactions. The responses were collected and archived for subsequent analysis.

### Guideline Concordance Assessment

The medical accuracy of each response was evaluated using the current European Association of Urology (EAU) Guidelines on Paediatric Urology-2026 as the reference standard (7). Guideline concordance assessments were performed independently by two pediatric urologists certified by the European Board of Paediatric Urology. Any discrepancies in grading were resolved through discussion and consensus.

Each response was independently reviewed and assigned one of four concordance grades:

- **Grade 1:** Fully concordant with the guideline recommendations,
- **Grade 2:** Generally concordant but containing minor omissions or incomplete details,
- **Grade 3:** Partially concordant but including inaccurate or potentially misleading information,
- **Grade 4:** Non-concordant with the guideline recommendations.

For secondary analyses, responses were categorized as:

- **Successful:** Grade 1.
- **Partially successful:** Grade 2.
- **Failed:** Grade 3 or Grade 4.

Additionally, acceptable responses were defined as those being Grade 1 or Grade 2.

### Public Readability Assessment

In order to evaluate the usefulness of LLM-generated information for non-medical users, all responses were assessed for public readability. Readability assessments were performed independently by two non-medical reviewers representing the target lay audience. One reviewer was a native English speaker who evaluated the English responses, whereas the second reviewer was a bilingual Turkish-English speaker who evaluated the Turkish responses. Neither reviewer had a formal medical education or healthcare-related professional experience, thereby reflecting the perspective of the typical information-seeking patients and caregivers.

Readability was evaluated using a five-point Likert-type scale based on the overall comprehensibility, the clarity of language, the use of medical terminology, sentence complexity, and the suitability of the information given for a lay audience.

Scores were assigned as follows:

- **5 points:** Excellent readability; easily understood by the general public.
- **4 points:** Good readability with minor use of medical terminology.
- **3 points:** Moderate readability requiring some health literacy.
- **2 points:** Poor readability with substantial technical content.
- **1 point:** Very poor readability; unlikely to be understood by a non-medical audience.

Disease-specific and overall readability scores were calculated separately for the English and Turkish responses.

### Statistical Analysis

Categorical variables are presented as frequencies and percentages, whereas continuous variables are reported as mean±standard deviation. Fisher's exact test was used for pairwise comparisons between the 2 models and the 2 languages, while chi-square tests were used to compare concordance outcomes among the disease categories. Readability scores are summarized descriptively and compared between the 2 languages and the 2 models. A two-sided p value $<0.05$ was considered statistically significant.

### Outcomes

The primary outcome was guideline concordance of LLM-generated responses.

The secondary outcomes included:

- The public readability scores,
- A comparison of the English and Turkish responses,
- A comparison between the LLMs,
- The identification of disease-specific and model-specific knowledge gaps,
- The distribution of successful, partially successful, and failed responses.

Importantly, beyond overall concordance rates, the specific questions associated with reduced concordance were compared between the 2 models in order to identify model-specific knowledge-gap profiles.

## Results

A total of 50 questions covering five common pediatric penile conditions (phimosis, hypospadias, CPC, paraphimosis, and buried penis) were evaluated for each LLM in both English and Turkish. The responses were assessed for guideline concordance according to the EAU Paediatric Urology Guidelines and for public readability.

### Guideline Concordance

#### ChatGPT

Among the 50 evaluated responses, 40 (80%) were classified as Grade 1 (fully concordant), 9 (18%) as Grade 2 (partially concordant), and 1 (2%) as Grade 3. No Grade 4 responses were identified. Consequently, 49 of 50 responses (98%) were considered acceptable (either Grade 1 or Grade 2) (Table I).

Disease-specific Grade 1 concordance rates were 70% for phimosis, 90% for hypospadias, 80% for CPC, 100% for paraphimosis, and 60% for buried penis. When acceptable responses (Grade 1 or 2) were considered, concordance rates reached 90% for phimosis and 100% for all of the remaining conditions (Figure 1).

No significant differences were observed among the disease categories regarding successful responses (p=0.37) or for the acceptable responses (p=0.91).

#### Gemini

Among the 50 evaluated responses, 41 (82%) were classified as Grade 1 and 9 (18%) as Grade 2. No Grade 3 or Grade 4 responses were identified. Therefore, all responses (50/50, 100%) were considered acceptable according to the predefined criteria (Table II).

Disease-specific Grade 1 concordance rates were 80% for phimosis, 80% for hypospadias, 80% for CPC, 90% for paraphimosis, and 80% for buried penis. Acceptable response rates (Grade 1 or 2) reached 100% across all disease categories (Figure 2).

No significant differences were detected among the disease categories regarding successful responses (p=0.989) or for the acceptable responses (p=1.000).

### Comparison Between Models

Overall guideline concordance was comparable between the two LLMs. Grade 1 concordance rates were 80% for ChatGPT and 82% for the Gemini (p=1.000). Acceptable response rates (Grade 1 or 2) were 98% and 100%, respectively (p=1.000). The frequency of failed responses (Grade 3-4) did not differ significantly between the two models (2% vs. 0%, p=1.000).

Although overall performance was similar, the specific questions associated with reduced concordance differed substantially between the two models. Only a limited overlap was observed in non-Grade 1 responses, suggesting model-specific knowledge-gap profiles rather than disease-specific weaknesses.

**Table I.** Guideline concordance outcomes according to disease category and language

| Disease | Language | Successful (Grade 1) n (%) | Partially successful (Grade 1, 2) n (%) | Failed (Grade 3, 4) n (%) | p value* |
|---|---|---|---|---|---|
| **Phimosis** | English | 7/10 (70%) | 9/10 (90%) | 1/10 (10%) | |
| | Turkish | 7/10 (70%) | 9/10 (90%) | 1/10 (10%) | |
| **Hypospadias** | English | 9/10 (90%) | 10/10 (100%) | 0/10 (0%) | |
| | Turkish | 9/10 (90%) | 10/10 (100%) | 0/10 (0%) | |
| **CPC** | English | 8/10 (80%) | 10/10 (100%) | 0/10 (0%) | |
| | Turkish | 8/10 (80%) | 10/10 (100%) | 0/10 (0%) | |
| **Paraphimosis** | English | 10/10 (100%) | 10/10 (100%) | 0/10 (0%) | |
| | Turkish | 10/10 (100%) | 10/10 (100%) | 0/10 (0%) | |
| **Buried penis** | English | 6/10 (60%) | 10/10 (100%) | 0/10 (0%) | |
| | Turkish | 6/10 (60%) | 10/10 (100%) | 0/10 (0%) | |
| **Overall** | **English** | **40/50 (80%)** | **49/50 (98%)** | **1/50 (2%)** | **1.000†** |
| | **Turkish** | **40/50 (80%)** | **49/50 (98%)** | **1/50 (2%)** | |
| **p value (among disease categories)** | **English** | **0.37‡** | **0.91§** | - | |
| | **Turkish** | **0.37‡** | **0.91§** | - | |

*p values compare distributions of guideline concordance outcomes.
†Overall English versus Turkish comparison (Fisher's exact test).
‡Comparison among the disease categories for successful responses (Grade 1) using Fisher-Freeman-Halton exact test.
§Comparison among the disease categories for partially successful responses (Grade 1, 2) using Fisher-Freeman-Halton exact test.
CPC: Congenital penile curvature

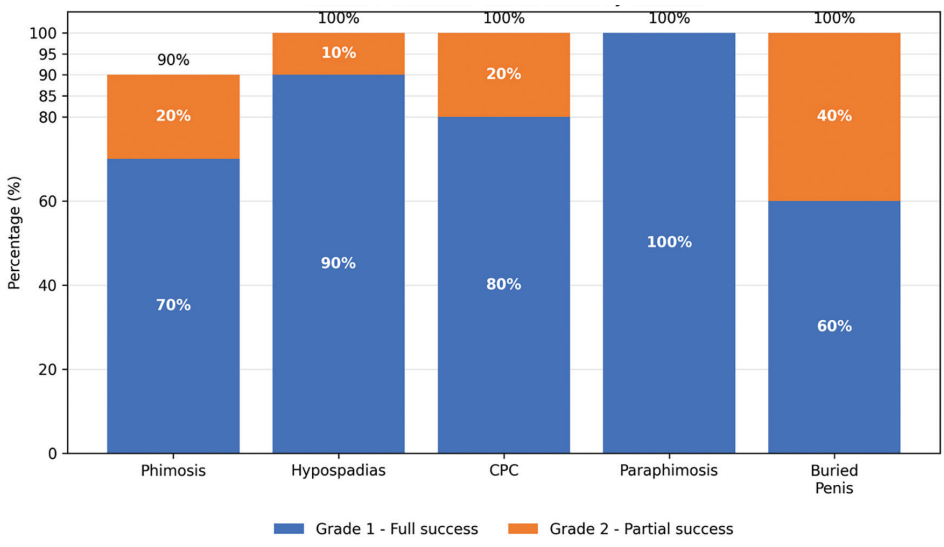


**Figure 1.** ChatGPT guideline concordance by disease

**Table II.** Guideline concordance outcomes according to disease category and language for Gemini

| Disease | Language | Grade 1 n (%) | Grade 2 n (%) | Grade 3, 4 n (%) | Acceptable responses (Grade 1, 2) n (%) |
|---|---|---|---|---|---|
| **Phimosis** | English | 8 (80%) | 2 (20%) | 0 (0) | 10 (100%) |
| | Turkish | 8 (80%) | 2 (20%) | 0 (0) | 10 (100%) |
| **Hypospadias** | English | 8 (80%) | 2 (20%) | 0 (0) | 10 (100%) |
| | Turkish | 8 (80%) | 2 (20%) | 0 (0) | 10 (100%) |
| **CPC** | English | 8 (80%) | 2 (20%) | 0 (0) | 10 (100%) |
| | Turkish | 8 (80%) | 2 (20%) | 0 (0) | 10 (100%) |
| **Paraphimosis** | English | 9 (90%) | 1 (10%) | 0 (0) | 10 (100%) |
| | Turkish | 9 (90%) | 1 (10%) | 0 (0) | 10 (100%) |
| **Buried penis** | English | 8 (80) | 2 (20) | 0 (0) | 10 (100) |
| | Turkish | 8 (80) | 2 (20) | 0 (0) | 10 (100) |
| **Overall** | **English** | **41 (82)** | **9 (18)** | **0 (0)** | **50 (100)** |
| | **Turkish** | **41 (82)** | **9 (18)** | **0 (0)** | **50 (100)** |
| **p value (among disease categories)** | **English** | **0.989** | - | - | **1.000** |
| | **Turkish** | **0.989** | - | - | **1.000** |
| **p value (English vs. Turkish)** | - | **1.000** | **1.000** | **1.000** | **1.000** |

CPC: Congenital penile curvature

### Public Readability

#### ChatGPT

The mean readability scores were high in both languages. English responses achieved an overall score of 236/250 (4.72±0.48), while Turkish responses achieved 224/250 (4.48±0.66). Although English responses demonstrated numerically higher readability scores, the difference did not reach statistical significance (p=0.081).

Disease-specific readability scores for English responses ranged from 4.6 to 4.8, whereas Turkish scores ranged from 4.2 to 4.7. No significant differences were observed among the disease categories for the English (p=0.812) or the Turkish responses (p=0.091).

#### Gemini

Gemini also demonstrated high readability in both languages. English responses achieved an overall readability

score of 236/250 (4.72±0.45), while Turkish responses achieved 231/250 (4.62±0.53). No significant difference was observed between the two languages (p=0.365) (Table III)

Disease-specific readability scores ranged from 4.6 to 4.8 for the English responses and from 4.4 to 4.8 for the Turkish responses. No significant variations among the disease categories were detected for the English (p=0.851) or the Turkish responses (p=0.634).

### Comparison Between Models

English readability scores were identical between ChatGPT and Gemini (4.72 vs. 4.72, p=1.000). Turkish readability scores were numerically higher for Gemini (4.62 vs. 4.48), although this difference was not statistically significant (p=0.31). Overall readability scores did not differ significantly between the two models (4.60±0.58 vs. 4.67±0.49, p=0.54).

### Analysis of Non-Grade 1 Responses

Although overall guideline concordance rates were comparable between ChatGPT and Gemini, the specific questions associated with reduced concordance differed considerably between the two models. ChatGPT deficiencies were mainly related to epidemiological, etiological, and counseling-related topics, whereas Gemini more frequently demonstrated omissions involving diagnostic evaluation, surgical indications, treatment details, and the timing of interventions.

Only two overlapping areas of reduced concordance were identified across the 50 evaluated questions (Phimosis Q4 and Buried Penis Q7). Overall, 15 of 17 non-Grade 1 responses (88.2%) were model-specific, suggesting distinct knowledge-gap profiles despite similar overall performance metrics (Table IV).

### Overall Findings

Both LLMs demonstrated high levels of guideline concordance and public readability across all of the evaluated pediatric penile conditions. No statistically significant differences were identified between the English and Turkish responses or between the two models evaluated.

Gemini achieved a slightly higher rate of fully concordant responses and demonstrated no guideline-discordant answers. However, these numerical differences did not reach statistical significance. Importantly, despite similar overall performance metrics, the specific domains associated with reduced guideline concordance differed between the two models, indicating distinct patterns of knowledge limitations.

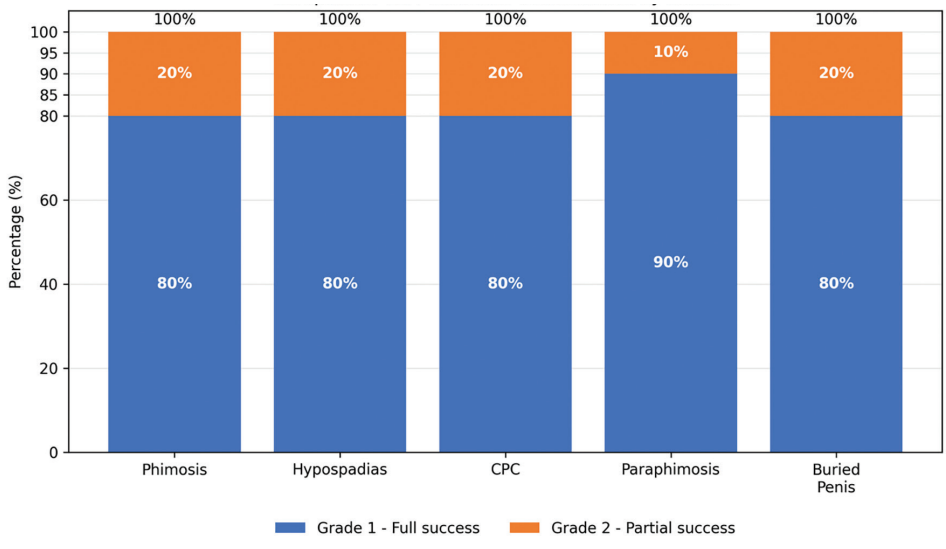


**Figure 2.** Gemini guideline concordance by disease

**Table III.** Public readability scores according to disease category and language

| Disease | ChatGPT English | ChatGPT Turkish | p value | Gemini English | Gemini Turkish | p value |
|---|---|---|---|---|---|---|
| **Phimosis** | 4.8±0.4 | 4.7±0.5 | 0.640 | 4.8±0.4 | 4.7±0.5 | 0.651 |
| **Hypospadias** | 4.8±0.4 | 4.5±0.7 | 0.322 | 4.7±0.5 | 4.6±0.5 | 0.681 |
| **CPC** | 4.6±0.5 | 4.2±0.8 | 0.170 | 4.6±0.5 | 4.4±0.7 | 0.576 |
| **Paraphimosis** | 4.7±0.5 | 4.7±0.5 | 1.000 | 4.8±0.4 | 4.8±0.4 | 1.000 |
| **Buried Penis** | 4.7±0.5 | 4.3±0.7 | 0.146 | 4.7±0.5 | 4.6±0.5 | 0.681 |
| **Overall** | **4.72±0.48** | **4.48±0.66** | **0.081** | **4.72±0.45** | **4.62±0.53** | **0.365** |
| **p value (among disease categories)** | **0.812** | **0.091** | - | **0.851** | **0.634** | - |

CPC: Congenital penile curvature

**Table IV.** Comparison of non-Grade 1 responses between ChatGPT and Gemini

| Disease | ChatGPT (Grade 2-3 responses) | Gemini (Grade 2 responses) | Overlap |
|---|---|---|---|
| **Phimosis** | Q2 (Etiology)<br>Q3 (Epidemiology)<br>Q4 (Symptoms) | Q4 (Symptoms)<br>Q7 (Non-surgical treatment) | Q4 |
| **Hypospadias** | Q7 (Need for surgery) | Q5 (Diagnosis)<br>Q8 (Ideal age for repair) | None |
| **CPC** | Q2 (Etiology)<br>Q8 (Sexual function) | Q6 (Surgical indication)<br>Q7 (Timing of treatment) | None |
| **Paraphimosis** | None | Q6 (Treatment) | None |
| **Buried penis** | Q2 (Etiology)<br>Q3 (Epidemiology)<br>Q7 (Surgical indication)<br>Q8 (Complications) | Q7 (Surgical indication)<br>Q9 (Surgical outcomes) | Q7 |
| **Overall** | **10 non-Grade responses** | **9 non-Grade 1 responses** | **2 overlapping questions** |

CPC: Congenital penile curvature

## Discussion

The present study demonstrated that both of the LLMs evaluated provided highly accurate and readable information regarding common pediatric penile conditions. Guideline concordance rates exceeded 98% for ChatGPT and reached 100% for Gemini when acceptable responses (Grade 1 or 2) were considered. Importantly, no Grade 4 responses were identified and only a single Grade 3 response was observed across all of the ChatGPT responses evaluated. These findings suggest that current-generation LLMs are capable of delivering information which is largely consistent with contemporary pediatric urology guidelines.

The performance observed in the present study appears superior to that reported in many early investigations evaluating previous generations of LLMs. Initial studies frequently highlighted concerns regarding hallucinated information, factual inaccuracies, and inconsistent recommendations. As model architectures and training datasets have evolved, however, recent evidence has demonstrated progressive improvements in medical accuracy and clinical reasoning (8). Our findings support this trend and suggest that current LLM versions may provide useful supplementary educational information for both healthcare professionals and the general public. Nevertheless, clinical decisions should continue to rely on professional expertise, guideline-based recommendations, and individualized patient assessments.

Interestingly, despite similar overall concordance rates, the specific domains associated with reduced concordance differed considerably between the two models. Only two overlapping areas of reduced concordance were identified among the fifty evaluated questions. Most non-Grade 1 responses were model-specific, indicating distinct knowledge-gap profiles despite comparable overall performances. This observation suggests that future evaluations should not focus solely on overall accuracy rates but should also investigate the nature and distribution of model-specific deficiencies.

Another notable finding was the consistently high readability of the responses. Both LLMs achieved mean readability scores above 4.5/5 in both English and Turkish, indicating that the information generated was generally understandable to individuals without formal medical training. Although the English responses tended to achieve slightly higher scores, no statistically significant differences were observed between the two languages.

Similarly, no significant language-related differences were identified in guideline concordance. Taken together, these findings suggest that contemporary LLMs can provide relatively consistent health information across different linguistic settings. Such multilingual consistency may be particularly important for those patients and caregivers in non-English-speaking regions where access to specialist pediatric urology resources may be more limited. The ability of modern LLMs to maintain a comparable performance across languages may therefore contribute to reducing language-related disparities in access to health information (9).

The observed variation among the disease categories may partly reflect differences in the amount of the available training data. Conditions such as hypospadias have been extensively studied and discussed within both scientific and public domains for decades. A simple search of the biomedical literature demonstrates that hypospadias has generated substantially more publications than many other pediatric penile conditions (10). Consequently, it is not surprising that the LLMs demonstrated particularly strong performances in this area. In contrast, less frequently discussed conditions may provide fewer opportunities for model training and refinement, potentially contributing to occasional omissions or incomplete responses.

Compared with previous investigations evaluating online health information, our study employed a more standardized methodology. Earlier studies commonly relied on questions collected from search engines, YouTube, Facebook, Google Trends, or publicly available FAQ databases (5,11). While these approaches reflect real-world information-seeking behavior, they may introduce substantial variability in topic selection and question complexity. In order to minimize this variability, we developed a structured question set based on predefined clinical domains, ensuring consistent coverage of definitions, epidemiology, etiology, symptoms, diagnosis, treatment, complications, surgical issues, and long-term outcomes across all of the disease categories.

**Study Limitations**

Several limitations should be acknowledged. First, LLMs frequently generate information beyond the specific question asked. While this may enhance educational value, it complicates objective scoring because responses often contain content not directly addressed by the evaluation criteria. Second, only those pediatric penile conditions specifically addressed within the EAU Paediatric Urology Guidelines were evaluated. Therefore, our findings may not be generalizable to conditions not covered by these guideline recommendations. Third, readability assessments were performed using a limited number of evaluators and may not fully represent the perceptions of all patient populations. In addition, readability assessment inherently involves a degree of subjectivity. Fourth, only two contemporary LLM platforms were evaluated, and other models may demonstrate different performance characteristics. Finally, this study focused on general educational questions commonly asked by families. Complex clinical scenarios, individualized treatment decisions, and unusual presentations were not assessed. Consequently, these findings support the reliability of LLMs for general patient education but should not be extrapolated to complex clinical decision-making (12). Visual information was not evaluated because corresponding illustrations were not included within the reference guideline used for concordance assessment. Another important limitation is the dynamic nature of LLM development. Model architectures, training datasets, and system updates evolve continuously over time. Therefore, the present findings should be interpreted as reflecting model performance during the study period rather than the permanent characteristics of the evaluated systems.

Taken together, contemporary LLMs appear capable of providing accurate and understandable information regarding common pediatric penile conditions. Although they cannot replace professional medical consultations, they may serve as useful educational resources. Continued reassessment remains necessary as both clinical guidelines and LLM technologies evolve over time (13,14).

## Conclusion

Contemporary LLMs demonstrated high guideline concordance and excellent public readability when answering FAQs regarding common pediatric penile conditions. Importantly, no clinically significant misinformation was identified in either of the models evaluated, and most observed deficiencies reflected minor omissions of specific guideline details rather than inaccurate recommendations.

No significant differences were observed between the two models or between the English and Turkish responses. Although overall performance was similar, the specific areas of reduced concordance differed between the two models, suggesting distinct knowledge-gap profiles.

These findings suggest that the current-generation of LLMs may serve as reliable educational resources for families and healthcare professionals. However, they should complement rather than replace professional medical consultations, particularly in complex clinical situations requiring individualized decision-making. Further studies evaluating additional pediatric urological conditions, visual content, and future versions of LLMs are warranted as both clinical guidelines and artificial intelligence technologies continue to evolve over time.

### Ethics

**Ethics Committee Approval:** Ethics committee approval was not required because the study did not involve human participants, animal subjects, or identifiable personal data.

**Informed Consent:** Not applicable. This study did not involve human participants or patient data. The analyses were based on responses generated by large language models to standardized guideline-based questions; therefore, informed consent was not required.

### Acknowledgements

The authors thank Dr. Emre Bülbül for performing the statistical analyses of this study. The authors also thank Ivan Pointon, a native English-speaking non-medical reviewer, and Onur Can Türedi, a bilingual Turkish-English non-medical reviewer, for their valuable contributions to the public readability assessment of the LLM-generated responses.

### Footnotes

### Authorship Contributions

Concept: B.T., A.S., Design: B.T., A.S., Data Collection or Processing: B.T., A.S., Analysis or Interpretation: B.T., A.S., Literature Search: B.T., A.S., Writing: B.T., A.S.

**Conflict of Interest:** The authors declare no conflicts of interest.

**Financial Disclosure:** The authors received no financial support for the conduct, authorship, or publication of this study.

## References

1. Ratnagandhi JA, Godavarthy P, Gnaneswaran M, Lim B, Vittalraj R. Enhancing anesthetic patient education through the utilization of large language models for improved communication and understanding. Anesth Res. 2025; 2:4.
2. Tulgar S, Aksu C, Selvi O, et al. A comparative evaluation of the quality of responses provided by different large language model chatbots to frequently asked questions regarding nerve blocks. BMC Anesthesiol. 2026; 26:98.
3. Green Z, Ashton JJ, Beattie RM. Learning from the past, structuring the future: using large language models to unlock a century of paediatric research in Archives of Disease in Childhood. Arch Dis Child. 2026; 111:418-22.
4. Kocaoğlu MH, Demirel A, Kaya İ. Accuracy, quality, and readability analyses of responses from large language models to questions on pediatric dental sedation. BMC Oral Health. 2026; 26:492.
5. Caglar U, Yildiz O, Meric A, et al. Evaluating the performance of ChatGPT in answering questions related to pediatric urology. J Pediatr Urol. 2024; 20:26.e1-5.
6. Spinoit AF, Waterschoot M, Sinatti C, et al. Fertility and sexuality issues in congenital lifelong urology patients: male aspects. World J Urol. 2021; 39:1013-9.
7. Radmayr C, van Uitert A, Kennedy U, et al. EAU Guidelines on Paediatric Urology. In: EAU Guidelines. Arnhem, The Netherlands, European Association of Urology, 2026.
8. Cheng Y, Zhu L. A review of ChatGPT in medical education: exploring advantages and limitations. Int J Surg. 2025; 111:4586-602.
9. Tan J, Wang L, Wang G, et al. Safety and user perception of general-purpose large language models in pediatric healthcare: evaluations of ChatGPT by doctors and parents. Digit Health. 2026; 12:20552076261427505.
10. Faraj S, Clermidi P, Irtan S, Madec FX. Artificial intelligence chatbots vs. YPUC pediatric urologists: performance on a Campbell Walsh urology hypospadiology questionnaire. World J Urol. 2025; 43:719.
11. Gumus K, Yilmaz AB, Cinbek AF, Ozdemir HB, Kutman KG. Evaluating AI chatbots in enuresis nocturna information: a comparative analysis of readability, reliability and quality. Fr J Urol. 2026; 36:103062.
12. Adriansyah IA, Wahyudi I, Vallasciani S, et al. Evaluating the utility of ChatGPT in enhancing parental education and clinical support in hypospadias care. J Pediatr Urol. 2025; 21:639-43.
13. MacNevin W, Dawe N, Harkness L, Salman B, Keefe DT. Evaluation of ChatGPT's performance on answering pediatric urology questions based on association guidelines. Can Urol Assoc J. 2025; 19:E362-7.
14. Kucuker K, Akinci A, Duran MB, et al. Awareness, use, and perceived barriers to artificial intelligence in pediatric urology: a multicenter survey. Ther Adv Urol. 2026; 18:17562872261422939.