Submitted: 29 October, 2025 Accepted: 31 December, 2025 Published: 30 June, 2026 DOI:10.22514/j.androl.2026.019

ORIGINAL RESEARCH

REV INT ANDROL
Revista Internacional de Andrología

Establishing a performance benchmark for artificial intelligence in pediatric urology: an expert evaluation of ChatGPT and Gemini on circumcision

Yasin Aktaş[1,*], Adem Tunçekin[1]

[1]Department of Urology, Faculty of Medicine, Uşak University, 64300 Uşak, Turkey

***Correspondence**
yasin.aktas@usak.edu.tr
(Yasin Aktaş)

Abstract

Background: Large language models (LLMs) based on artificial intelligence are increasingly being used for medical inquiries. However, the accuracy of these models regarding circumcision has not yet been evaluated. This study aimed to establish a performance benchmark by comparing the accuracy and quality of ChatGPT's and Gemini's responses to guideline-based and patient-focused circumcision questions. **Methods**: The comparative study was conducted from August to October 2025. A total of 50 questions were analyzed: 30 guideline-based/academic questions derived from the European Association of Urology and the American Urological Association guidelines and 20 patient-focused frequently asked questions (FAQs) from reputable sources. Two board-certified urologists evaluated ChatGPT-5 and Gemini 2.5 Flash responses using Binary Accuracy Scoring (BAS) and Detailed Accuracy Scoring (DAS). Inter-rater reliability was assessed using Cohen's Kappa coefficient and 95% confidence intervals (CI) were calculated via the Clopper-Pearson method. The Wilcoxon signed-rank test was used to compare paired ordinal data between models, and the Mann-Whitney U test was used to compare the different question categories (guideline-based *vs.* patient-focused). **Results**: Regarding BAS, both models achieved 100% accuracy, meaning no factually incorrect answers were found. In terms of DAS, for guideline-based questions, both ChatGPT and Gemini achieved a "completely correct" rate of 93.3% (95% CI: 77.9–99.2%) ($p = 0.705$). For patient-focused FAQs, Gemini scored 85% (95% CI: 62.1–96.8%) and ChatGPT scored 75% (95% CI: 50.9–91.3%), with no statistically significant difference between the two models ($p = 0.315$). There were no significant differences between the guideline-based and patient-focused question groups for either model (ChatGPT: $p = 0.68$; Gemini: $p = 0.322$). **Conclusions**: Both models demonstrated high reliability, providing a preliminary performance benchmark for this specific domain. While no significant performance difference was observed between the models in this dataset, qualitative limitations necessitate a "Physician-in-the-Loop" workflow, employing LLMs as drafting agents under expert supervision.

Keywords

Artificial intelligence; ChatGPT; Gemini; Circumcision; Pediatric urology; Large language models

Establecimiento de un punto de referencia de rendimiento para la inteligencia artificial en urología pediátrica: evaluación experta de ChatGPT y Gemini en la circuncisión

Resumen

Antecedentes: Los modelos lingüísticos grandes (LLM) basados en inteligencia artificialse utilizan cada vez más para consultas médicas. Sin embargo, no se ha evaluado su precisión en lo que respecta a la circuncisión. El objetivo de este estudio fue establecer un punto de referencia de rendimiento comparando la precisión y la calidad de las respuestas de ChatGPT y Gemini a preguntas sobre la circuncisión basadas en directrices y centradas en el paciente. **Métodos**: El estudio comparativo se llevó a cabo entre agosto y octubre de 2025. Se analizaron un total de 50 preguntas: 30 preguntas basadas en directrices académicas derivadas de las directrices de la European Association of Urology y la American Urological Association y 20 preguntas frecuentes (FAQ) centradas en el paciente procedentes de fuentes acreditadas. Dos urólogos certificados evaluaron las respuestas de ChatGPT-5 y Gemini 2.5 Flash utilizando la puntuación de precisión binaria (BAS) y la puntuación de precisión detallada (DAS). La fiabilidad entre evaluadores se evaluó utilizando el coeficiente Kappa de Cohen y se calcularon los intervalos de confianza (IC) del 95% mediante el método de Clopper-Pearson. Se utilizó la prueba de rangos con signo de Wilcoxon para comparar los datos ordinales emparejados entre los modelos y la prueba U de Mann-Whitney para comparar las diferentes categorías de preguntas (basadas en directrices frente a centradas en el paciente). **Resultados**: En cuanto al BAS, ambos modelos alcanzaron una precisión del 100%, lo que significa que no se encontraron respuestas incorrectas desde el punto de vista fáctico. En cuanto al DAS, en las preguntas basadas en directrices, tanto ChatGPT como Gemini alcanzaron una tasa de "totalmente correctas" del 93.3% (IC del 95%: 77.9–99.2%) (p = 0.705). En cuanto a las preguntas centradas en el paciente, Gemini obtuvo una puntuación del 85% (IC del 95%: 62.1–96.8%) y ChatGPT, del 75% (IC del 95%: 50.9–91.3%), sin diferencias estadísticamente significativas entre ambos modelos (p = 0.315). No hubo diferencias significativas entre los grupos de preguntas basadas en directrices y centradas en el paciente para ninguno de los dos modelos (ChatGPT: p = 0.68; Gemini: p = 0.322). **Conclusións**: Ambos modelos demostraron una alta confiabilidad, proporcionando un punto de referencia de rendimiento preliminar para este dominio específico.Aunque no se observaron diferencias significativas en el rendimiento entre los modelos de este conjunto de datos, las limitaciones cualitativas requieren un flujo de trabajo "Physician-in-the-Loop", en el que se emplean los LLM como agentes de redacción bajo la supervisión de expertos.

Palabras Clave

Inteligencia artificial; ChatGPT; Gemini; Circuncisión; Urología pediátrica; Modelos de lenguaje grandes

1. Introduction

Artificial intelligence-based large language models (LLMs) are attracting attention for their potential to answer patient and guideline-based clinical questions in medicine. Compared to competitors such as Gemini, ChatGPT offers higher accuracy and shorter responses in some medical fields; however, competitors may perform better in certain specialized areas, such as emergency situations [1]. In clinical cases, ChatGPT-4 can make diagnoses with a level of accuracy similar to that of specialist physicians [2]. In some medical licensing exams, ChatGPT-4 has achieved an accuracy rate of up to 81%, outperforming many medical students and surpassing previous models [3]. The consistency and repeatability of responses depend on the version of the model and whether the question is open- or closed-ended [4]. ChatGPT may sometimes provide incorrect or incomplete information, especially when it comes to citing references and sources [5]. The model's lack of specialized medical data in its training creates limitations in terms of accuracy and timeliness [6].

Circumcision is the surgical removal of the foreskin, the piece of skin that covers the tip of the penis. The practice dates back to ancient times. It is one of the most commonly performed surgical procedures worldwide for medical, religious, cultural, and social reasons [7]. Today, about one-third of men worldwide are circumcised. It is particularly common in Muslim and Jewish communities as a religious ritual. In some countries, such as the United States, Canada, and Australia, circumcision is often performed on newborns for medical or social reasons [8]. Circumcision is medically indicated for cases of phimosis (inability to retract the foreskin), recurrent infections, and certain urological conditions [7]. Medical research shows that circumcision can reduce the risk of urinary tract infections (UTIs), Human Immunodeficiency Virus (HIV), certain sexually transmitted infections, penile cancer [9]. But circumcision is not without risks. The most common problems that can happen are bleeding, infection, rarely injury to the penis. The complication rate is about 0.5% [10]. Therefore, circumcision should be performed by specialists when appropriate medical conditions exist.

The current literature does not contain a specific evaluation of the responses provided by ChatGPT or Gemini to patient or guideline-based questions directly related to circumcision. The primary aim of this study is not merely to compare the performance of two leading artificial intelligence (AI) models, but to provide a preliminary benchmark regarding the proficiency of state-of-the-art LLMs in the specific subdomain of circumcision. By evaluating their responses to a comprehensive set of 30 guideline-based and 20 patient-focused questions, we intend to establish a performance baseline for future AI

evaluations in pediatric urology and determine the readiness of these tools for clinical integration.

2. Material and methods

The comparative study was conducted from August to October of 2025. Because the study did not involve human participants or patient records and only used publicly available AI-generated data, it was exempt from formal Institutional Review Board approval. The research was conducted in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines [11]. We evaluated the performance of the ChatGPT-5 and Gemini 2.5 Flash models using two sets of questions related to circumcision.

Guideline-Based and Academic Questions (n = 30) (**Supplementary material**): Questions were derived by identifying strong recommendations within the European Association of Urology (EAU) and American Urological Association (AUA) guidelines and converting them into interrogative sentences to ensure content validity. Additionally, specific academically significant clinical questions regarding circumcision were included to cover areas where guidelines might be less explicit. These questions cover recommendations for indications of circumcision, management of complications, pain control, and patient education. Circumcision is not a main topic in the EAU guidelines. Rather, it is presented as a treatment recommendation under various headings, particularly in pediatric urology for the treatment of conditions such as phimosis and urinary tract infections. In oncology guidelines, circumcision is included as a treatment recommendation in the penile cancer section.

Patient-Focused FAQs (Frequently Asked Questions) (n = 20) (**Supplementary material**): Questions most frequently asked by parents and patients were compiled from reputable international urology associations and patient education portals. The questions are written in plain language and reflect common concerns encountered in clinical practice.

The ChatGPT-5 (OpenAI, San Francisco, CA, USA) and Gemini 2.5 Flash (Google DeepMind, Mountain View, CA, USA) models were presented with all questions in English. A clean chat session was opened for each question to eliminate the influence of previous responses. Responses were recorded in their original form without any intervention. To ensure the evaluation reflected typical user behavior, a "zero-shot" prompting strategy was used. All questions were entered into the models verbatim in English without prior examples, role-playing instructions, or specific formatting constraints. Additionally, the prompts did not explicitly direct the models to cite literature or provide a bibliography. This approach allowed us to assess the models' intrinsic propensity to offer evidence-based support. Consequently, any references provided were generated spontaneously as part of the models' default response mechanisms.

Two independent reviewers selected guideline- and evidence-based questions, matching each one with a clear guideline recommendation or evidence-based statement from the scientific literature. Two board-certified urologists evaluated the responses generated by the models independently. During the evaluation, the reviewers were blinded to the specific AI model (ChatGPT or Gemini) generating the response to prevent assessment bias. Any conflicting scores were resolved through direct discussion until full consensus was achieved.

The evaluation utilized two scoring systems as follows [12]:

Binary Accuracy Scoring (BAS): 1 = Correct; 0 = Incorrect.

Detailed Accuracy Scoring (DAS): 1 = Completely correct; 2 = Correct but inadequate; 3 = Mixed correct and misleading information; 4 = Completely incorrect.

Statistical analysis was performed using IBM SPSS Statistics 26.0 (IBM Corp., Armonk, NY, USA). The Wilcoxon signed-rank test was used to compare paired ordinal data (detailed accuracy scores) between ChatGPT and Gemini for the same set of questions. The Mann-Whitney U test was used to compare the distributions of detailed accuracy scores between the guideline-based and patient-focused question groups. Descriptive statistics were presented as frequencies and percentages. The results are reported as z and p values, with a p value of < 0.05 considered statistically significant. Inter-rater reliability was assessed using Cohen's Kappa coefficient to quantify the consistency between the two independent urologists. A formal a priori power analysis was not conducted. Instead, the sample size was determined by the availability of relevant content. This reflects the limited number of high-level recommendations in the EAU and AUA guidelines, as well as specific, academically significant clinical questions regarding circumcision. To provide an estimate of the precision of the accuracy proportions, 95% confidence intervals (CI) were calculated using the Clopper-Pearson exact method.

3. Results

A total of 50 questions were analyzed: 30 guideline-based/academic and 20 patient-focused FAQs. For the guideline-based questions, both ChatGPT and Gemini received DAS of 93.3% (95% CI: 77.9–99.2%), indicating that they provided completely correct responses. One ChatGPT response and two Gemini responses were rated as "correct but inadequate", and one ChatGPT response contained mixed or misleading information. No completely incorrect answers were identified. There was no statistically significant difference between ChatGPT and Gemini (p = 0.705, z = −0.378; Wilcoxon signed-rank test) (Table 1).

For the patient-focused FAQs, Gemini had a slightly higher proportion of correct answers (85%) (95% CI: 62.1–96.8%) than ChatGPT (75%) (95% CI: 50.9–91.3%). ChatGPT provided more mixed responses (15% *vs.* 5%), but this difference was not statistically significant (p = 0.315, z = −0.933; Wilcoxon signed-rank test) (Table 2).

A qualitative assessment of the responses revealed distinct patterns in the limitations of the AI models. Responses classified as "Correct but Inadequate" (DAS 2) provided accurate general information but lacked specific clinical details. For example, they omitted particular reduction maneuvers for paraphimosis or failed to specify preferred anesthetic agents. Conversely, responses labeled as "Mixed/Misleading" (DAS 3) were predominantly observed in patient-focused questions about subjective or controversial

TABLE 1. Detailed accuracy scoring of ChatGPT and Gemini responses to guideline-based and academic questions (n = 30).

Detailed Accuracy Scoring	ChatGPT (n, %)	Gemini (n, %)
1 = Completely correct	28 (93.3%)	28 (93.3%)
2 = Correct but inadequate	1 (3.3%)	2 (6.7%)
3 = Mixed correct and misleading information	1 (3.3%)	0 (0%)
4 = Completely incorrect	0 (0%)	0 (0%)

$p = 0.705$; $z = -0.378$ (Wilcoxon Signed Rank test).

TABLE 2. Detailed accuracy scoring of ChatGPT and Gemini responses to patient-focused frequently asked questions (n = 20).

Detailed Accuracy Scoring	ChatGPT (n, %)	Gemini (n, %)
1 = Completely correct	15 (75%)	17 (85%)
2 = Correct but inadequate	2 (10%)	2 (10%)
3 = Mixed correct and misleading information	3 (15%)	1 (5%)
4 = Completely incorrect	0 (0%)	0 (0%)

$p = 0.315$; $z = -0.933$ (Wilcoxon Signed Rank test).

topics, such as the potential for psychological trauma or long-term regret. In these instances, the models attempted to balance conflicting viewpoints, occasionally resulting in ambiguous answers that lacked clear guidance for lay readers.

In BAS, both models correctly answered all guideline-based and patient-focused questions.

The inter-rater reliability was assessed using Cohen's Kappa coefficient. For DAS, the reviewers demonstrated almost perfect agreement for ChatGPT (Cohen's $\kappa = 0.850$) and substantial agreement for Gemini (Cohen's $\kappa = 0.782$). Scoring discrepancies were observed in only 2 out of 50 questions for both ChatGPT and Gemini. These differences were resolved through direct discussion between the reviewers until full consensus was reached. For BAS, there was 100% absolute agreement between the raters.

Fig. 1 shows comparisons of DAS distributions between guideline-based and patient-focused question groups for ChatGPT (Panel A) and Gemini (Panel B). According to the Mann-Whitney U test, there was no significant difference between the two groups for either model ($p = 0.68$ for ChatGPT and $p = 0.322$ for Gemini).

4. Discussion

According to EAU guidelines, circumcision reduces the incidence of penile cancer and is recommended as an additional treatment for superficial noninvasive disease (PeIN) and invasive disease limited to the glans (categories T1 and T2) [13, 14]. Secondary phimosis is the sole absolute indication; however, it is also indicated for primary phimosis refractory to medical treatment, recurrent balanoposthitis, and UTIs associated with urinary tract abnormalities [15]. Circumcision is further proposed as a preventive measure due to its role in reducing the incidence of sexually transmitted infections (STIs), including HIV, syphilis, and herpes simplex virus type 2 [16], and its inverse relationship with genital Human Papillomavirus (HPV) prevalence [17]. The AUA guidelines affirm that neonatal circumcision is generally safe when performed by an experienced surgeon, noting a minor complication rate of approximately 3%. The well-established medical benefits and low morbidity of circumcision may encourage anxious parents and health-conscious individuals to seek more information through AI-driven resources. Therefore, the accuracy of medical advice obtained from these platforms is clinically relevant.

AI and chatbots, like ChatGPT, are being used more and more to provide patients and physicians with information about urological diseases. Studies in recent years have shown these systems have advantages and limitations regarding response quality, accuracy, and patient safety. Importantly, the absence of a statistically significant difference between ChatGPT and Gemini in this dataset does not prove equivalence. Rather, it demonstrates the considerable proficiency of current generative AI technologies in this domain. Our findings suggest that both models have reached a level of competence that allows them to serve as useful adjuncts for patient education with expert supervision. Therefore, this study provides a preliminary performance benchmark, shifting the focus of future research from determining "superiority" to optimising the safe integration of these tools into clinical practice.

To effectively integrate these tools into daily urology practice, we propose a "physician-in-the-loop" workflow [18]. In this framework, LLMs can rapidly draft responses to common patient inquiries or generate discharge instructions. However, these responses must be rigorously vetted by the urologist. This supervision is critical not only for identifying obvious errors but also for recognizing "correct but inadequate" advice, a risk that our study highlighted when models occasionally omitted specific drug dosages or procedural nuances. We hypothesize that the frequent occurrence of "correct but inadequate" responses to clinical management questions stems from an LLM optimization strategy. Models prioritize general, globally safe information while omitting specific, high-

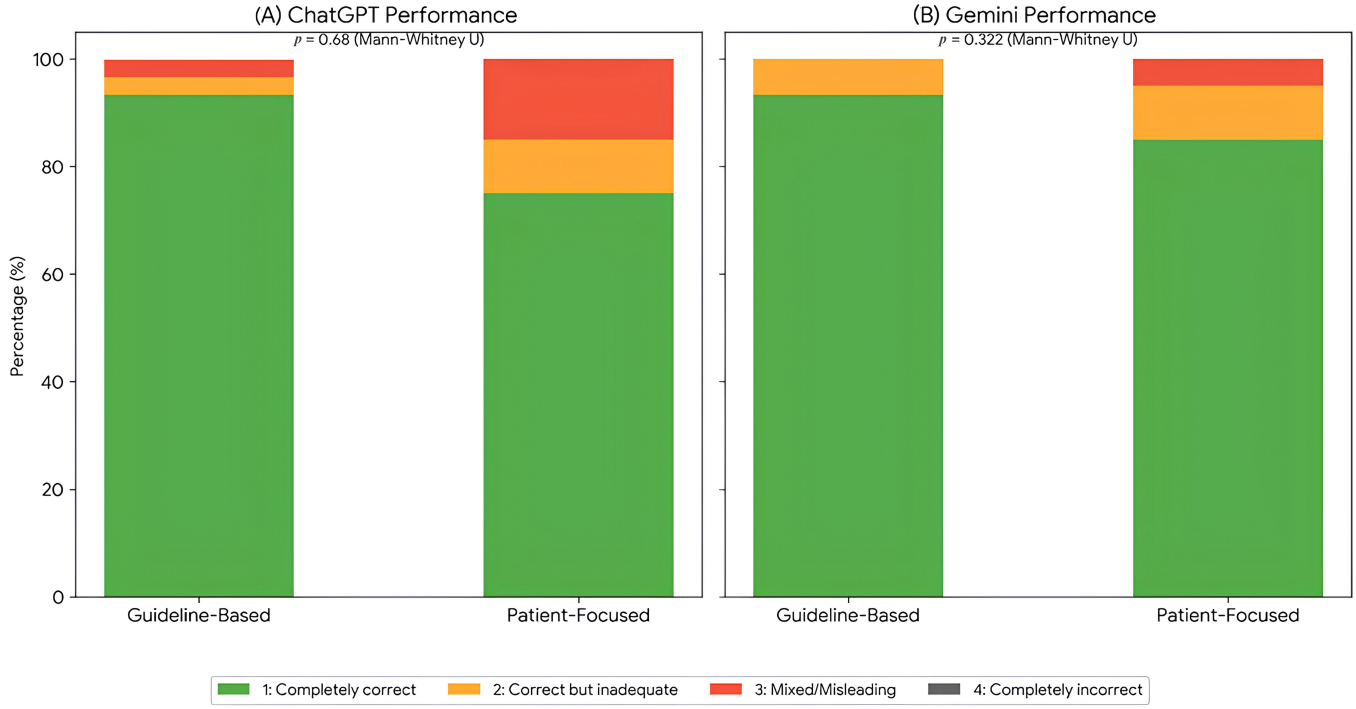

FIGURE 1. Detailed accuracy score comparison for guideline-based and patient-focused frequently asked questions provided by ChatGPT (Panel A) and Gemini (Panel B) ($p = 0.68$ & $p = 0.322$).

stakes, context-dependent details (*e.g.*, drug dosages or procedural steps) to mitigate the risk of providing unsafe or region-specific misinformation. Treating LLMs as "drafting engines" rather than "decision-makers" allows clinicians to leverage AI's speed while mitigating the risks of contextual misunderstanding and incomplete guidance.

Our results regarding the accuracy of LLMs in circumcision align with broader trends observed in pediatric urology. In a comparable study evaluating pediatric urology guidelines and FAQs using ChatGPT-3.5, the responses were found to be completely accurate 92% of the time, with high consistency [19]. AI-based models can accurately diagnose, treat, and predict the prognosis of urinary tract infections, kidney stones, and urological cancers. For instance, artificial neural networks have demonstrated an accuracy rate of 98.3% in diagnosing UTIs [20]. Similarly, our study found high accuracy rates for both models. However, distinctions emerge when compared to complex oncological topics. While a study of GPT-3.5 reported accurate answers regarding urological cancers (prostate, bladder, and kidney), previous studies have noted its moderate-to-low applicability and comprehensibility issues [21]. In contrast, our study suggests that for a specific, widely understood procedure like circumcision, the models generate information that is generally more accurate and actionable.

Although models can occasionally match human performance, they consistently underperform in terms of clinical reasoning and visual interpretation. Success rates vary significantly across medical specialties [22]. The comparative performance of AI models varies across urological subspecialties. Previous research has shown that Gemini and ChatGPT have similar accuracy rates for bladder-related diseases [23], while ChatGPT has been found to be more comprehensive for bladder cancer [24]. On the contrary, ChatGPT had a lower rate of poor-quality responses than Gemini for erectile dysfunction [25]. Our study contributes to this comparative landscape by demonstrating that, in the specific domain of circumcision, the performance gap narrows significantly, with no statistically significant difference observed between the two models. These findings suggest that, for standard surgical procedures, the choice of model may be less critical than the implementation of expert oversight.

ChatGPT has the potential to transform urology in the areas of education, research, and clinical practice. Although it functions as a powerful complement to human expertise, deploying this AI technology requires a steadfast commitment to ethical and responsible standards [26]. The data sets used to train the models may be outdated and therefore may not reflect new medical developments [27]. LLMs can sometimes generate information that is not based in reality, which is sometimes referred to as a "hallucination" [27, 28]. Despite the high accuracy observed in our dataset, the potential for misinformation remains a critical concern for clinical implementation. A survey of urologists revealed that 74.9% believe LLMs can disseminate misinformation. The most commonly expressed ethical concerns were artificial hallucinations (46.1%) and plagiarism (51.9%). Since nearly half of users identified inaccuracy as a primary limitation, urologists must be vigilant in verifying AI outputs to prevent the dissemination of misleading information [29]. Although ChatGPT improves access to information, the risk of incomplete responses underscores the importance of human expertise [30]. While we did not identify hallucinations in our dataset, the prevalence of "correct but inadequate" answers confirms the need for caution and validates the conclusion that AI cannot yet replace clinical judgment in

diagnostic and therapeutic decision-making [31].

It is important to note that this study used English-language prompts. Since LLM performance can vary significantly across languages, future research should prioritize multilingual testing to determine if these high levels of accuracy are maintained for non-English-speaking patient populations. Additionally, given the rapid pace of AI development, longitudinal studies are necessary to monitor the influence of future model updates on the consistency and reliability of medical advice over time. The accuracy rate drops significantly in languages other than English; success in multilingual medical exams can decrease by up to 10% [27]. Since LLMs are typically trained on data collected within a specific timeframe, they may be inadequate for current events or evolving information. Studies have shown that models trained with timestamps not only have better recall of information from the training period but also update more efficiently with new data [32]. Recent studies have revealed that variations in prompts significantly impact model performance, and existing methods are insufficient for predicting this sensitivity. Additionally, interactive and visual prompt engineering tools can help users develop more effective prompts [33].

Models generally do not cite sources in their responses and lack transparency [28]. For instance, only 43% of citations provided by ChatGPT-4 have been verified as accurate, with 57% being false or entirely hallucinated [34]. ChatGPT-4o achieved a higher accuracy rate in answering medical questions than its predecessors, ChatGPT-3.5 and 4, reaching an average score of 4.46 out of 5 while lowering the false response rate to 3.7%. However, its reproducibility, or consistency in responding to repeated queries, is reportedly lower [35]. Overall, newer iterations of ChatGPT (specifically versions 4 and 4o) demonstrate a significant improvement in the accuracy of the information they provide. While the AI models were not instructed to cite sources in our study, ChatGPT-5 provided references when answering questions about the cost-effectiveness of circumcision and penile cancer. Upon investigation, we confirmed the accuracy of these citations; they were not the result of hallucinations.

This study has several limitations. First, the data were collected through simulations and do not reflect real patient interactions. The questions were predefined and modeled only in English, which limits the study's generalizability to different languages or cultural contexts. Additionally, AI models' knowledge base may not fully cover the current literature, and some responses may lack verifiable sources. Since the study focused solely on questions related to circumcision, the results cannot be generalized directly to other urological or medical fields.

5. Conclusions

In general, AI chatbots can provide accurate and balanced information about urological diseases, often performing at near-human levels. However, they have significant limitations, such as hallucinations, poor reasoning, a lack of citations, and issues with timeliness. In this study, ChatGPT and Gemini demonstrated high accuracy in answering 30 guideline-based and 20 patient-focused questions; their performance showed no statistically significant difference within this specific question set. These results provide a preliminary performance benchmark, suggesting that current LLMs have the potential to support clinical practice when used under expert supervision. To safely integrate this potential into daily workflows, we recommend a "physician-in-the-loop" model that utilizes LLMs as drafting assistants while ensuring rigorous oversight to mitigate risks such as "correct but inadequate" advice. Ultimately, with expert supervision, these models can serve as effective adjuncts for patient education and clinical efficiency.

AVAILABILITY OF DATA AND MATERIALS

The data supporting the findings of this study are ethically restricted and are available from the corresponding author (YA) upon reasonable request.

AUTHOR CONTRIBUTIONS

YA—writing, data curation, original draft. AT—data curation, conceptualization. Both authors read and approved the final manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable. Because the study did not involve human participants or patient records and only used publicly available AI-generated data, it was exempt from formal Institutional Review Board approval.

ACKNOWLEDGMENT

The authors would like to express their gratitude to all the dedicated urology professionals for their hard work.

FUNDING

This research received no external funding.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

SUPPLEMENTARY MATERIAL

Supplementary material associated with this article can be found, in the online version, at https://files.intandro.com/files/article/2071860267858182144/attachment/Supplementary%20material.docx.

REFERENCES

[1] Fattah FH, Salih AM, Salih AM, Asaad SK, Ghafour AK, Bapir R, *et al.* Comparative analysis of ChatGPT and Gemini (Bard) in medical inquiry: a scoping review. Frontiers in Digital Health. 2025; 7: 1482712.

[2] Hirosawa T, Kawamura R, Harada Y, Mizuta K, Tokumasu K, Kaji Y, *et al*. ChatGPT-generated differential diagnosis lists for complex case-derived clinical vignettes: diagnostic accuracy evaluation. JMIR Medical Informatics. 2023; 11: e48808.

[3] Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, *et al.* Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. Journal of Medical Internet Research. 2024; 26: e60807.

[4] Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis. Journal of Biomedical Informatics. 2024; 151: 104620.

[5] Garg RK, Urs VL, Agarwal AA, Chaudhary SK, Paliwal V, Kar SK. Exploring the role of ChatGPT in patient care (diagnosis and treatment) and medical research: a systematic review. Health Promotion Perspectives. 2023; 13: 183–191.

[6] Ruksakulpiwat S, Kumar A, Ajibade A. Using ChatGPT in medical research: current status and future directions. Journal of Multidisciplinary Healthcare. 2023; 16: 1513–1520.

[7] Prabhakaran S, Ljuhar D, Coleman R, Nataraja RM. Circumcision in the paediatric patient: a review of indications, technique and complications. Journal of Paediatrics and Child Health. 2018; 54: 1299–1307.

[8] American Academy of Pediatrics Task Force on Circumcision. Male circumcision. Pediatrics. 2012; 130: e756–785.

[9] Friedman B, Khoury J, Petersiel N, Yahalomi T, Paul M, Neuberger A. Pros and cons of circumcision: an evidence-based overview. Clinical Microbiology and Infection. 2016; 22: 768–774.

[10] Simpson M. Urologic conditions in infants and children: circumcision. FP Essentials. 2020; 488: 11–15.

[11] von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. The Lancet. 2007; 370: 1453–1457.

[12] Beyatlı M, Güngör HS, İnkaya A, Sobay R, Tahra A, Küçük EV. Expert evaluation of ChatGPT-4 responses to upper tract urothelial carcinoma questions: a prospective comparative study with guideline-based and patient-focused queries. Journal of Clinical Medicine. 2025; 14: 6353.

[13] Tsen HF, Morgenstern H, Mack T, Peters RK. Risk factors for penile cancer: results of a population-based case-control study in Los Angeles County (United States). Cancer Causes & Control. 2001; 12: 267–277.

[14] Philippou P, Shabbir M, Malone P, Nigam R, Muneer A, Ralph DJ, *et al*. Conservative surgery for squamous cell carcinoma of the penis: resection margins and long-term oncological control. Journal of Urology. 2012; 188: 803–808.

[15] Ladenhauf HN, Ardelean MA, Schimke C, Yankovic F, Schimpl G. Reduced bacterial colonisation of the glans penis after male circumcision in children—a prospective study. Journal of Pediatric Urology. 2013; 9: 1137–1144.

[16] Tobian AA, Serwadda D, Quinn TC, Kigozi G, Gravitt PE, Laeyendecker O, *et al.* Male circumcision for the prevention of HSV-2 and HPV infections and syphilis. The New England Journal of Medicine. 2009; 360: 1298–1309.

[17] Albero G, Castellsagué X, Giuliano AR, Bosch FX. Male circumcision and genital human papillomavirus: a systematic review and meta-analysis. Sexually Transmitted Diseases. 2012; 39: 104–113.

[18] Sezgin E. Artificial intelligence in healthcare: complementing, not replacing, doctors and healthcare providers. Digital Health. 2023; 9: 20552076231186520.

[19] Caglar U, Yildiz O, Meric A, Ayranci A, Gelmis M, Sarilar O, *et al*. Evaluating the performance of ChatGPT in answering questions related to pediatric urology. Journal of Pediatric Urology. 2024; 20: 26.e21–26.e25.

[20] Ozkan IA, Koklu M, Sert IU. Diagnosis of urinary tract infection based on artificial intelligence methods. Computer Methods and Programs in Biomedicine. 2018; 166: 51–59.

[21] Musheyev D, Pan A, Loeb S, Kabarriti AE. How well do artificial intelligence chatbots respond to the top search queries about urological malignancies? European Urology. 2024; 85: 13–16.

[22] Lin SY, Hsu YY, Ju SW, Yeh PC, Hsu WH, Kao CH. Assessing AI efficacy in medical knowledge tests: a study using Taiwan's internal medicine exam questions from 2020 to 2023. Digital Health. 2024; 10: 20552076241291404.

[23] Azizoğlu M, Klyuev S. A comparative study on the question-answering proficiency of artificial intelligence models in bladder-related conditions: an evaluation of Gemini and ChatGPT 4.o. Medical Records. 2025; 7: 201–205.

[24] Alasker A, Alshathri N, Alsalamah S, Almansour N, Alsalamah F, Alghafees M, *et al.* ChatGPT *vs.* Gemini: which provides better information on bladder cancer? International Society of Urology Journal. 2025; 6: 34.

[25] Barlas İ, Tunç L. Quality of chatbot responses to the most popular questions regarding erectile dysfunction. Urology Research and Practice. 2025; 50: 253–260.

[26] Solano C, Tarazona N, Angarita GP, Medina AA, Ruiz S, Pedroza VM, *et al*. ChatGPT in urology: bridging knowledge and practice for tomorrow's healthcare, a comprehensive review. Journal of Endourology. 2024; 38: 763–777.

[27] Alonso I, Oronoz M, Agerri R. MedExpQA: multilingual benchmarking of large language models for medical question answering. Artificial Intelligence in Medicine. 2024; 155: 102938.

[28] Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M, *et al.* Toward expert-level medical question answering with large language models. Nature Medicine. 2025; 31: 943–950.

[29] Eppler M, Ganjavi C, Ramacciotti LS, Piazza P, Rodler S, Checcucci E, *et al*. Awareness and use of ChatGPT and large language models: a prospective cross-sectional global survey in urology. European Urology. 2024; 85: 146–153.

[30] Braga A, Nunes NC, Santos EN, Veiga ML, Braga AANM, de Abreu GE, *et al.* Use of ChatGPT in urology and its relevance in clinical practice: is it useful? International Brazilian Journal of Urology. 2024; 50: 192–198.

[31] Yu QX, Feng DC, Wu RC, Li DX. Auxiliary use of ChatGPT in surgical diagnosis and treatment—correspondence. International Journal of Surgery. 2024; 110: 617–618.

[32] Dhingra B, Cole JR, Eisenschlos JM, Gillick D, Eisenstein J, Cohen WW. Time-aware language models as temporal knowledge bases. Transactions of the Association for Computational Linguistics. 2022; 10: 257–273.

[33] Strobelt H, Webson A, Sanh V, Hoover B, Beyer J, Pfister H, *et al*. Interactive and visual prompt engineering for *ad-hoc* task adaptation with large language models. IEEE Transactions on Visualization and Computer Graphics. 2023; 29: 1146–1156.

[34] Ghanem D, Zhu AR, Kagabo W, Osgood G, Shafiq B. ChatGPT-4 knows its A B C D E but cannot cite its source. JBJS Open Access. 2024; 9: e24.00099.

[35] Ye Y, Zheng ED, Lan QL, Wu LC, Sun HY, Xu BB, *et al.* Comparative evaluation of the accuracy and reliability of ChatGPT versions in providing information on *Helicobacter pylori* infection. Frontiers in Public Health. 2025; 13: 1566982.

How to cite this article: Yasin Aktaş, Adem Tunçekin. Establishing a performance benchmark for artificial intelligence in pediatric urology: an expert evaluation of ChatGPT and Gemini on circumcision. Revista Internacional de Andrología. 2026; 24(2): 58-64. doi: 10.22514/j.androl.2026.019.