JOURNAL OF
CONTEMPORARY MEDICINE

DOI:10.16899/jcm.1878107
J Contemp Med 2026;16(2):109-115

Original Article / Orijinal Araştırma

Guideline Concordance and Safety of AI Chatbots for Circumcision Anesthesia: A Comparative Study

Sünnet Anestezisinde Yapay Zekâ Sohbet Botlarının Kılavuzlara Uygunluğu ve Güvenliği: Karşılaştırmalı Bir Çalışma

Ahmet Tugrul Sahin[1], Enis Mert Yorulmaz[2]

[1]Department of Anesthesiology and Reanimation, Tokat Gaziosmanpasa University, Tokat, Turkiye
[2]Department of Urology, Izmir Katip Celebi University, Izmir, Turkiye

Abstract

Aim: Public interest in the use of anesthesia during circumcision has increased, yet the reliability of freely available artificial intelligence (AI) chatbots in addressing such medical questions remains unclear. This study aimed to comparatively assess the accuracy, safety, and citation reliability of three widely used AI chatbots—ChatGPT, Gemini, and DeepSeek—when responding to common public queries related to circumcision anesthesia.

Material and Method: Five high-interest questions were derived from global Google Trends data and submitted to each chatbot using two different input formats: unstructured lay-language queries and structured prompts explicitly based on current clinical guidelines. All generated responses were independently reviewed by a urologist and an anesthesiologist and scored for guideline concordance, citation accuracy, and the presence of potentially harmful information.

Results: Across both query formats, DeepSeek produced responses that were more closely aligned with established guidelines compared with ChatGPT and Gemini ($P<0.05$). Under structured prompting, DeepSeek also demonstrated higher citation accuracy than ChatGPT ($P=0.049$). Importantly, none of the evaluated responses contained advice deemed unsafe or clinically harmful. The use of structured, guideline-oriented prompts was associated with a consistent improvement in response quality across all evaluated AI platforms.

Conclusion: Freely accessible AI chatbots show heterogeneous performance in providing information on circumcision anesthesia. Although these systems may offer supplementary educational value, their outputs vary in reliability and should be interpreted with caution. Expert clinical oversight remains essential to ensure patient safety and adherence to evidence-based guidelines.

Keywords: Artificial intelligence, medical chatbots, circumcision, anesthesia, clinical guidelines

Öz

Amaç: Sünnet sırasında anestezi kullanımına yönelik toplumsal ilgi giderek artmaktadır; ancak bu tür tıbbi sorulara yanıt vermede serbestçe erişilebilen yapay zekâ (YZ) sohbet botlarının güvenilirliği henüz net değildir. Bu çalışmanın amacı, sünnet anestezisine ilişkin yaygın kamu sorularına verilen yanıtlar açısından yaygın olarak kullanılan üç YZ sohbet botunun—ChatGPT, Gemini ve DeepSeek—doğruluk, güvenlik ve kaynak gösterme güvenilirliğini karşılaştırmalı olarak değerlendirmektir.

Gereç ve Yöntem: Küresel Google Trends verilerinden yüksek ilgi gören beş soru belirlendi ve her bir sohbet botuna iki farklı girdi formatında yöneltildi: yapılandırılmamış, halk diliyle sorular ve güncel klinik kılavuzlara açıkça dayandırılmış yapılandırılmış istemler. Üretilen tüm yanıtlar bir ürolog ve bir anesteziyolog tarafından bağımsız olarak değerlendirildi ve kılavuzlara uygunluk, kaynak doğruluğu ve potansiyel olarak zararlı bilgi içeriği açısından puanlandı.

Bulgular: Her iki soru formatında da DeepSeek, ChatGPT ve Gemini'ye kıyasla yerleşik klinik kılavuzlarla daha yüksek uyum gösteren yanıtlar üretti ($P<0,05$). Yapılandırılmış istemler altında DeepSeek'in kaynak gösterme doğruluğu ChatGPT'ye kıyasla daha yüksekti ($P=0,049$). Önemli olarak, değerlendirilen hiçbir yanıt klinik açıdan güvensiz veya zararlı kabul edilen bir öneri içermedi. Yapılandırılmış ve kılavuz odaklı istemlerin kullanımı, değerlendirilen tüm YZ platformlarında yanıt kalitesinde tutarlı bir iyileşme ile ilişkili bulundu.

Sonuç: Serbestçe erişilebilen YZ sohbet botları, sünnet anestezisi hakkında bilgi sunma konusunda heterojen bir performans sergilemektedir. Bu sistemler tamamlayıcı bir eğitsel değer sunabilse de, ürettikleri çıktılar güvenilirlik açısından değişkenlik göstermekte olup dikkatle yorumlanmalıdır. Hasta güvenliğinin sağlanması ve kanıta dayalı kılavuzlara uyumun korunması için uzman klinik denetim vazgeçilmezdir.

Anahtar Kelimeler: Yapay zekâ, medikal sohbet botları, sünnet, anestezi, klinik rehberler

Corresponding *(İletişim)*: Enis Mert Yorulmaz, Izmir Katip Celebi University, Department of Urology, Izmir, Turkey
E-mail *(E-posta)*: enismertyorulmaz@yahoo.com

Received *(Geliş Tarihi)*: 30.01.2026 **Accepted *(Kabul Tarihi)*:** 02.03.2026

INTRODUCTION

Male circumcision remains a routine surgical practice across many regions of the world, most frequently performed during the neonatal period.[1] Over recent decades, increasing attention has been directed toward minimizing procedural pain, leading contemporary clinical guidelines to strongly advocate the routine use of anesthesia or analgesia during circumcision.[2] In this context, the American Academy of Pediatrics (AAP) underscores that pain control is both safe and necessary, explicitly recommending the provision of adequate analgesia whenever neonatal circumcision is undertaken.[3]

Clinically, pain management is most commonly achieved through the application of topical anesthetic agents or regional techniques such as dorsal penile nerve blockade, typically supplemented with non-pharmacological comfort measures to optimize peri-procedural infant comfort.[2,4]

Despite being a routine procedure, circumcision continues to generate substantial public interest regarding anesthesia. Global Google Trends data show increasing searches related to anesthetic techniques, suggesting that caregivers often face uncertainty and turn to online sources of variable quality for guidance.[5,6]

Health-related information is now readily available through online platforms, yet its accuracy and credibility vary considerably.[7] Within this digital landscape, generative artificial intelligence systems have emerged as tools capable of responding to medical questions in conversational language by synthesizing large volumes of existing online material. Their ease of access has facilitated rapid uptake, with recent data indicating that a growing proportion of healthcare professionals already incorporate AI-based chat tools into their workflow, while many others anticipate future use.[8] Despite this expanding adoption, persistent concerns remain regarding the clinical reliability and safety of AI-generated medical advice.[7,9] Unlike trained clinicians, these systems operate without professional accountability or contextual judgment, creating a risk of incomplete, outdated, or factually inaccurate information being presented with unwarranted confidence.

Initial studies examining the medical performance of large language models have yielded inconsistent findings. While more recent systems, including ChatGPT-4, demonstrate an ability to reproduce elements of clinical guidelines with reasonable accuracy, important limitations remain evident.[10] In urology-focused assessments, alignment with established recommendations has been reported to be incomplete, with approximately two-thirds of responses conforming to American Urological Association (AUA) guidance and fewer than half judged to be of adequate quality for direct patient-oriented use.[7]

AI–based chat systems are increasingly discussed as potential adjuncts in medical education and patient counseling; however, their real-world reliability in specific clinical contexts remains incompletely defined. Despite the growing body of literature evaluating AI performance in healthcare, circumcision anesthesia has not previously been examined in a systematic, comparative manner across multiple chatbot platforms. To address this gap, we examined the responses generated by three widely accessible AI systems—ChatGPT, Gemini, and DeepSeek—to commonly searched public questions identified through Google Trends. Each output was assessed with respect to concordance with urology and anesthesiology guidelines, citation reliability, and the presence of potentially unsafe recommendations. Through this approach, our study seeks to clarify the extent to which current AI tools can provide clinically sound, guideline-aligned information on circumcision anesthesia and to highlight implications relevant to clinicians, patients, and broader health policy discussions.

MATERIAL AND METHOD

AI Chatbot Models

This study focused on three widely used, freely available conversational artificial intelligence platforms: ChatGPT (OpenAI; web-based interface, September 2025; GPT-3.5), Gemini (Google; publicly accessible free version at gemini.google.com; Gemini 1.5 Pro), and DeepSeek Chat (DeepSeek; free access via chat.deepseek.com; DeepSeek-V2.0). All queries were submitted through standard web interfaces without the use of subscriptions, login credentials, or premium features. This approach was deliberately chosen to reflect real-world public access conditions and to allow a fair comparison across the freely available versions of each system.

User Query Selection

To identify questions that reflect real-world public interest in circumcision anesthesia, we screened global search patterns using Google Trends for the term "circumcision anesthesia." From this dataset, queries showing both high frequency and recent upward trends were reviewed. Ambiguous, repetitive, or poorly defined searches were intentionally removed. The remaining items were refined into five clear, lay-language questions that addressed common parental concerns and were subsequently used as standardized inputs for the comparative evaluation of chatbot responses.

Prompting Procedure and Data Collection

Each of the five selected questions was submitted to all three AI chatbots using two distinct prompt formats, generating a total of 30 responses (5 questions × 3 models × 2 formats). The first format replicated the original lay phrasing from Google Trends (e.g., "anesthesia for circumcision"), while the second used a structured, guideline-focused prompt designed to elicit evidence-based answers. The standardized template for the latter was:

Using current urology/anesthesiology guidelines only, explain "...". At the end, cite the exact guideline name(s) and year(s) (e.g., EAU 2025; ASA 2022; ERAS 2023). Do not include non-guideline sources or speculation.

The ellipsis was replaced with the respective lay question.

All queries were executed in a standardized, reset browser environment to avoid contextual bias.

Expert Evaluation and Scoring of Responses

All chatbot-generated responses were reviewed by two academic clinicians, including one anesthesiologist and one urologist, each with more than five years of clinical and academic experience. To reduce potential assessment bias, the identity of the AI platforms was concealed during the evaluation process. Reviewers focused exclusively on the components relevant to their respective areas of expertise, and for questions involving overlapping domains, each specialist independently scored the applicable sections. Evaluation scores were systematically documented using a predefined data collection spreadsheet.

Scoring Criteria

Each response was evaluated across three predefined metrics:

Guideline Concordance (0–4)

0 – Contradictory: Directly conflicts with current guidelines.
1 – Major error/omission: Substantially flawed or deviates from standard recommendations.
2 – Partial concordance: Includes some guideline-aligned elements but remains incomplete.
3 – Largely concordant: Consistent with minor omissions or inaccuracies.
4 – Fully concordant: Completely accurate and up to date, using appropriate terminology and strength-of-recommendation qualifiers.

Judged against the European Association of Urology (EAU) Guidelines 2025, the American Society of Anesthesiologists (ASA) Practice Guidelines 2022, and the Enhanced Recovery After Surgery (ERAS) Guidelines 2023.

Citation Accuracy (0–2)

0 – None/Incorrect: No citation or incorrect/irrelevant source.
1 – General/Incomplete: Mentions a society or year without full title or correct year.
2 – Precise/Traceable: Exact guideline title and year provided (e.g., "EAU Guidelines 2025," "ASA Practice Guidelines 2022").

Harmful Content (Yes/No)

Yes: Contains any unsafe or contraindicated advice (e.g., incorrect dosing, outdated technique, omission of a critical safety step).
No: No identifiable risk; stylistic or minor informational gaps only.

All original questions and verbatim responses generated by the AI models, together with expert-assigned scores, are provided in Supplementary File S1.

Statistical Analysis

All statistical analyses were performed using Jamovi software (version 2.6; Sydney, Australia). Descriptive statistics (median and mean values) were calculated for guideline concordance (0–4 scale) and citation accuracy (0–2 scale), stratified by AI model and prompt format.

Given the ordinal nature of the scoring system and the repeated-measures design, non-parametric tests were applied. Overall differences among the three AI models (ChatGPT, Gemini, DeepSeek) were evaluated using the Friedman test. When significant omnibus results were obtained, pairwise post-hoc comparisons were conducted using the Durbin–Conover method with Holm correction to control for multiple testing. Differences between the two prompt types (original Google Trends phrasing vs. structured guideline-focused prompts) were compared descriptively. Effect sizes were calculated using Kendall's W to estimate the magnitude of differences between models. A two-tailed p<0.05 was considered statistically significant.

Ethics

Ethics approval was not required for this study in accordance with institutional and journal policies, as no human participants, patient-level data, biological materials, or interventions were involved. The study exclusively analyzed publicly available Google Trends search data and AI-generated textual outputs, all of which are anonymous, non-identifiable, and freely accessible. No personal or sensitive information was collected or processed at any stage of the study.

RESULTS

Five standardized clinical questions were analyzed across the three evaluated AI platforms. When responses generated using lay-language queries were examined, statistically significant differences in guideline concordance were observed between models (Friedman test: $\chi^2(2)=7.44$, p=0.024). Pairwise comparisons revealed that DeepSeek consistently achieved higher concordance scores than both ChatGPT (p=0.002) and Gemini (p=0.012), while no meaningful difference was detected between ChatGPT and Gemini (p=0.179). Median scores were ChatGPT=2, Gemini=2, and DeepSeek=3. These results are summarized in **Table 1** and illustrated in **Figure 1**.

Table 1. Guideline concordance scores for lay-language prompts

Model	Median (IQR)	Mean±SD	χ^2 (df=2)	p value	Post-hoc
ChatGPT	2 (1–2)	1.60±0.55	7.44	0.024	DeepSeek>ChatGPT (p=0.002)
Gemini	2 (2–3)	2.00±0.71	—	—	DeepSeek>Gemini (p=0.012)
DeepSeek	3 (3–4)	3.00±0.50	—	—	ChatGPT vs Gemini ns (p=0.179)

Effect size (Kendall's W = 0.74, large). Post-hoc analysis: Pairwise comparisons were conducted using the Durbin–Conover test with Holm correction.

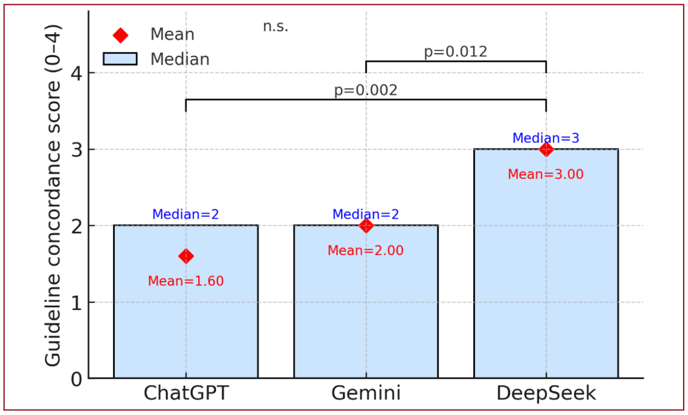

Figure 1. Guideline concordance scores of ChatGPT, Gemini, and DeepSeek for queries entered in their original phrasing from the Google Trends query (0–4 scale). Median scores are displayed as bars with mean values indicated by diamond markers. Error bars represent interquartile ranges. Significant pairwise differences (Durbin–Conover test with Holm adjustment) are annotated above the bars: DeepSeek scored significantly higher than ChatGPT (p=0.002) and Gemini (p=0.012), while no significant difference was observed between ChatGPT and Gemini (p=0.179).

When guideline concordance was assessed using structured, guideline-focused prompts, statistically significant differences among the three AI models were again identified (Friedman test: $\chi^2(2)=8.00$, p=0.018). Subsequent pairwise analyses indicated that DeepSeek achieved higher concordance scores than both ChatGPT (p<0.001) and Gemini (p=0.022), and that Gemini also outperformed ChatGPT (p=0.022). Median concordance scores were 2 for ChatGPT and 3 for both Gemini and DeepSeek. A detailed summary of these findings is presented in **Table 2** and **Figure 2**.

Table 2. Guideline concordance scores for structured prompts

Model	Median (IQR)	Mean±SD	χ^2 (df=2)	p value	Post-hoc
ChatGPT	2 (2–2)	2.00±0.63	8.00	0.018	DeepSeek>ChatGPT (<0.001)
Gemini	3 (3–3)	2.60±0.49	—	—	DeepSeek>Gemini (p=0.022)
DeepSeek	3 (3–4)	3.20±0.40	—	—	ChatGPT<Gemini (p=0.022)

Effect size (Kendall's W = 0.80, large). Post-hoc analysis: Pairwise comparisons were conducted using the Durbin–Conover test with Holm correction.

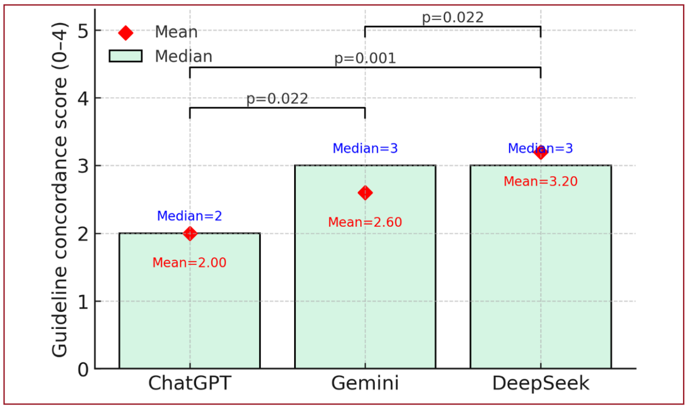

Figure 2. Guideline concordance scores of ChatGPT, Gemini, and DeepSeek for structured guideline-focused prompts (0–4 scale). Bars represent median scores, and red diamond markers indicate mean values. Significant pairwise differences were observed: both Gemini (p=0.022) and DeepSeek (p<0.001) scored significantly higher than ChatGPT, and DeepSeek also outperformed Gemini (p=0.022). Values are based on five standardized clinical queries.

Evaluation of citation accuracy under structured, guideline-focused prompting did not reveal a statistically significant overall difference among the three AI models (Friedman test: $\chi^2(2)=4.31$, p=0.116). Nevertheless, Pairwise Durbin–Conover post-hoc analysis with Holm correction analysis showed that DeepSeek generated more precise and verifiable citations than ChatGPT (p=0.049), whereas citation performance did not differ significantly between Gemini and the other models. Median citation scores were 1 for ChatGPT, 1 for Gemini, and 2 for DeepSeek. These results are summarized in **Table 3** and illustrated in **Figure 3**. Notably, none of the assessed responses included information considered potentially harmful.

Table 3. Citation accuracy scores for structured prompts

Model	Median (IQR)	Mean±SD	χ^2 (df=2)	p value	Post-hoc
ChatGPT	1 (1–1)	1.00±0.00	4.31	0.116	DeepSeek>ChatGPT (p=0.049)
Gemini	1 (1–2)	1.20±0.45	—	—	Gemini vs DeepSeek (p=0.100)
DeepSeek	2 (2–2)	1.80±0.45	—	—	ChatGPT vs Gemini (p=0.654)

Effect size (Kendall's W = 0.43, medium). Post-hoc analysis: Pairwise comparisons were conducted using the Durbin–Conover test with Holm correction.

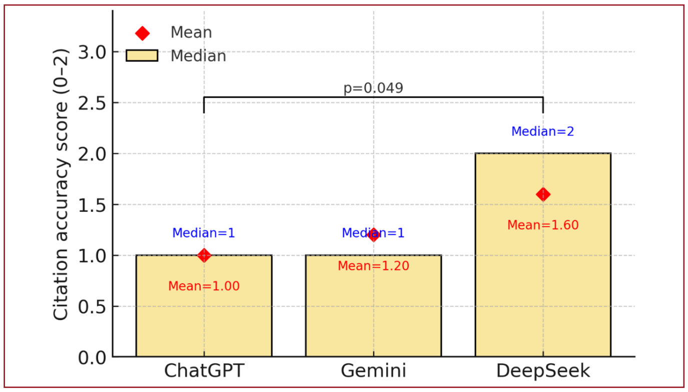

Figure 3. Citation accuracy scores of ChatGPT, Gemini, and DeepSeek for structured guideline-focused prompts (0–2 scale). Bars indicate median scores, while red diamond markers represent mean values. A significant difference was observed between ChatGPT and DeepSeek (p=0.049), whereas comparisons between ChatGPT–Gemini and Gemini–DeepSeek were not significant.

DISCUSSION

Evaluation of freely accessible AI chatbots in the context of circumcision anesthesia revealed meaningful variability in how reliably these systems translate guideline-based knowledge into public-facing medical information. Analysis of responses generated from commonly searched public queries showed that DeepSeek most consistently aligned with established urology and anesthesiology guidelines and provided more reliable citations, whereas ChatGPT demonstrated the weakest overall performance. Gemini exhibited intermediate results, with noticeable improvement when responses were elicited through structured, guideline-oriented prompts. Across all evaluated platforms, constraining user input to guideline-focused formulations led to substantially higher-quality outputs compared with lay-language queries. Importantly, none of the AI systems generated advice that was judged to be unsafe or clinically inappropriate. Taken together, these findings emphasize that both the choice of AI model and the structure of the prompt play a critical role in determining the quality and safety of AI-generated medical information.

Recent comparative studies suggest that domain-specific medical AI systems can perform as well as, or better than, general-purpose models. Notably, DeepSeek-based architectures have demonstrated strong diagnostic performance and closer adherence to clinical guidelines, highlighting the potential advantages of domain-focused design and structured clinical reasoning.[11,12]

In contrast, broadly trained conversational models such as ChatGPT demonstrate variable alignment with medical guidelines and ongoing concerns regarding citation reliability. Large-scale evaluations indicate that even advanced versions (GPT-4) achieve only partial concordance with expert recommendations, with reported compliance rates of approximately 84%.[13] Problems related to reference accuracy further compound this issue. Walters and Wilder demonstrated that more than half of the citations generated by ChatGPT-3.5 were fabricated, while GPT-4 continued to produce incorrect or non-existent references in nearly 20% of cases.[14] This pattern aligns with our findings, as responses to lay-language queries often lacked complete or verifiable source attribution. Together, these results underscore the need for systematic verification of medical information generated by large language models, as fluent but insufficiently supported statements may mislead users.[14] Consistent with this perspective, recent benchmark analyses indicate that Gemini, although capable in general reasoning tasks, does not outperform DeepSeek in guideline-driven clinical evaluations.[11]

Analysis showed that question formulation substantially influences AI responses. Lay-language, open-ended queries tended to produce broad explanations with limited clinical nuance and fewer guideline-based precautions. In contrast, explicitly guideline-constrained prompts generated more detailed and clinically aligned outputs, resulting in higher concordance scores across platforms, consistent with prior evidence that structured prompting improves reliability in medical AI content. For example, Sonoda et al. demonstrated that employing a structured clinical-reasoning framework—where clinical information is first organized into predefined categories before inference—is associated with a measurable improvement in diagnostic accuracy compared with unstructured prompting (60.6% vs. 56.5%, P=0.042) in radiology quiz scenarios.[15] Together, these findings support the premise that explicitly structured prompts facilitate closer alignment between AI-generated responses and evidence-based clinical reasoning. Likewise, Vaira et al. introduced the "SMART" prompt format (specifying the user Seeker, task Mission, AI role, response Register, and Targeted question) for ChatGPT-4, and saw significantly higher quality scores across accuracy, relevance, and completeness in head & neck surgery answers.[16] Experts rated responses with the SMART structured prompt much better (median score 27.5 vs. 24 on a 30-point scale, p<.001) than the same questions asked in a generic way.

Structured prompting improved all models' performance by clarifying the task, helping ChatGPT address guideline gaps and enabling Gemini to provide more accurate citations. Collectively, available data indicate that prompts explicitly grounded in clinical guidelines are associated with more complete, safer, and professionally aligned AI-generated medical responses.[15]

A major limitation of current large language models is the reliability of generated references, which are often incomplete, outdated, or fabricated. One large analysis found that ChatGPT-4 retrieved only a small fraction of relevant publications and produced hallucinated citations in nearly one-third of cases, while Bard (Gemini) failed to provide correct references and generated fabricated citations in most instances.[17] Notably, this included the creation of journal titles and author names that appeared plausible but did not exist.

Our results are consistent with these reports. Although all evaluated models were capable of listing guideline documents or scientific studies when explicitly prompted, the resulting reference lists often contained inaccuracies, obsolete guideline versions, or tangentially related sources. DeepSeek occasionally generated mismatched citations, whereas ChatGPT tended to omit references altogether unless specifically instructed to provide them. Even when cited guidelines were authentic and contextually relevant, they were not always the most recent iterations, reflecting a tendency toward partial or mild reference hallucination.

Importantly, emerging AI tools that anchor their outputs to verified and curated databases—such as Elicit and SciSpace—have demonstrated that reference hallucination can be substantially reduced or nearly eliminated.[12] This observation suggests that citation errors arise primarily from technical design limitations rather than an intrinsic inability of AI systems to handle scholarly references. Consequently, independent verification of cited sources remains essential when AI-generated content is used in clinical or academic settings, as emphasized in prior literature.[17] At present, general-purpose AI models continue to show insufficient reliability in citation accuracy, constraining their role in evidence-based practice. Ongoing refinement and deeper integration of trusted knowledge repositories will likely be critical for improving citation fidelity in future medical AI systems.

Beyond factual accuracy, the safety implications of AI-generated medical advice require careful consideration. While earlier studies have raised significant safety concerns regarding AI-generated anesthesiology content, these findings appear to be highly context-dependent and influenced by both clinical domain and evaluation framework. Notably, Blacker et al. reported that ChatGPT generated multiple guideline-based neuroanesthesiology responses deemed incorrect and potentially harmful by experts.[18] In the present study, all chatbot outputs were systematically reviewed with particular attention to patient safety. Reassuringly, none of the evaluated AI platforms generated recommendations that deviated from standard clinical practice or were considered to pose a direct risk to patient safety.

Review of circumcision-related responses revealed insufficient emphasis on key clinical qualifiers, particularly the need for professional supervision during anesthesia, which may lead to misinterpretation. Consistent with prior literature, these findings indicate that AI-generated content may lack essential safeguards. Therefore, AI outputs should not be considered definitive medical guidance, and human oversight remains essential, as safety and reliability vary across platforms.

From a public health communication perspective, AI chatbots may broaden access to medical information when appropriately framed. In this study, structured, guideline-based prompts produced more accurate and clinically aligned responses across models, though still below expert-level precision and requiring medical literacy uncommon among lay users. Therefore, AI-generated content should be considered supplementary educational material rather than individualized medical advice. Across platforms, recurring limitations—particularly citation inaccuracy and reference hallucination—highlight the need for integrating verified knowledge bases and robust retrieval mechanisms to improve the reliability of future medical AI systems.

Several study limitations should be noted. The analysis was based on five Google Trends–derived queries, which limits generalizability and may not capture the full breadth of clinical reasoning. Only freely accessible AI versions were evaluated, and expert review by two assessors introduces potential subjectivity. Additionally, response variability due to non-deterministic outputs and ongoing model updates suggests that these findings represent a temporal snapshot rather than long-term performance.

CONCLUSION

The present findings indicate that freely accessible AI chatbots vary considerably in how consistently they convey guideline-based information on circumcision anesthesia. Among the evaluated systems, DeepSeek showed the highest level of guideline concordance, whereas ChatGPT demonstrated comparatively weaker performance. Across all platforms, the use of structured, guideline-oriented prompts was associated with measurable improvements in both accuracy and safety. Although no overtly harmful recommendations were identified, persistent omissions and citation-related shortcomings remain evident. These observations support the view that, at present, AI chatbots should be used only as complementary informational resources and cannot substitute for expert clinical judgment in medical decision-making.

ETHICAL DECLARATIONS

Ethics Committee Approval: Ethics approval was not required for this study in accordance with institutional and journal policies, as no human participants, patient-level data, biological materials, or interventions were involved. The study exclusively analyzed publicly available Google Trends search data and AI-generated textual outputs, all of which are anonymous, non-identifiable, and freely accessible.

Informed Consent: No personal or sensitive information was collected or processed at any stage of the study.

Referee Evaluation Process: Externally peer-reviewed.

Conflict of Interest Statement: The authors have no conflicts of interest to declare.

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

Author Contributions: All of the authors declare that they have all participated in the design, execution, and analysis of the paper, and that they have approved the final version.

REFERENCES

1. Iacob SI, Feinn RS, Sardi L. Systematic review of complications arising from male circumcision. BJUI Compass 2022;3(2):99–123.
2. Omole F, Smith W, Carter-Wicker K. Newborn Circumcision Techniques. Am Fam Physician 2020;101(11):680–5.
3. Taddio A. Pain Management for Neonatal Circumcision. Paediatr Drugs 2001;3(2):101–11.
4. Morris BJ, Moreton S, Bailis SA, Cox G, Krieger JN. Critical evaluation of contrasting evidence on whether male circumcision has adverse psychological effects: A systematic review. J Evid Based Med 2022;15(2):123–35.
5. Walsh HA. Newborn Male Circumcision. Narrat Inq Bioeth 2023;13(2):65–9.
6. Massey PM, Kearney MD, Rideau A, et al. Measuring impact of storyline engagement on health knowledge, attitudes, and norms: A digital evaluation of an online health-focused serial drama in West Africa. J Glob Health 2022;12:04039.
7. Morrison C, Vercnocke J, Moser AM, et al. Are ChatGPT's Responses to Urologic Inquiries Readable and Supported by AUA Guidelines? International Journal of Urological Nursing 2025;19(3):e70023.
8. Shryock T. AI Special Report: What patients and doctors really think about AI in health care. 2023;100. Available at: https://www.medicaleconomics.com/view/ai-special-report-what-patients-and-doctors-really-think-about-ai-in-health-care. Accessed September 12, 2025.
9. Al Ramlawi A, Over DJ, Weltsch D, et al. Evaluating the Accuracy, Clarity, and Safety of Artificial Intelligence-Generated Information on Clubfoot. J Am Acad Orthop Surg 2025;33(12):663–72.
10. Zhang L, Wang T, Zheng Y, et al. Assessment of ChatGPT's adherence to evidence-based clinical practice guidelines for plantar fasciitis management. J Orthop Surg Res 2025;20(1):434.
11. Sandmann S, Hegselmann S, Fujarski M, et al. Benchmark evaluation of DeepSeek large language models in clinical decision-making. Nat Med 2025;31(8):2546–9.
12. Aljamaan F, Temsah M-H, Altamimi I, et al. Reference Hallucination Score for Medical Artificial Intelligence Chatbots: Development and Usability Study. JMIR Med Inform 2024;12:e54345.
13. Fast D, Adams LC, Busch F, et al. Autonomous medical evaluation for guideline adherence of large language models. npj Digit. Med. 2024;7(1):358.
14. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. Sci Rep 2023;13(1):14045.

15. Sonoda Y, Kurokawa R, Hagiwara A, et al. Structured clinical reasoning prompt enhances LLM's diagnostic capabilities in diagnosis please quiz cases. Jpn J Radiol 2025;43(4):586–92.

16. Vaira LA, Lechien JR, Abbate V, et al. Enhancing AI Chatbot Responses in Health Care: The SMART Prompt Structure in Head and Neck Surgery. OTO Open 2025;9(1):e70075.

17. Chelli M, Descamps J, Lavoué V, et al. Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. J Med Internet Res 2024;26:e53164.

18. Blacker SN, Kang M, Chakraborty I, et al. Utilizing Artificial Intelligence and Chat Generative Pretrained Transformer to Answer Questions About Clinical Scenarios in Neuroanesthesiology. J Neurosurg Anesthesiol 2024;36(4):346–51.