Aksaray Journal of Medical Sciences
Aksaray Üniversitesi Tıp Bilimleri Dergisi

asujms 2025; 5(3) :93-97

Research Article

AI-Assisted Knowledge Assessment: Comparison of ChatGPT and Gemini on Undescended Testicle in Children

Yapay Zeka Destekli Karşılaştırma: Çocuklarda İnmemiş Testis Bilgisi Konusunda ChatGPT ve Gemini Karşılaştırması

Emine Özdemir Kaçer [*1], Mustafa Tuşat [*2], Murat Kılıçaslan [*3], Sebahattin Memiş [*3]

[*1] Aksaray University, Faculty of Medicine, Department of Pediatrics, Aksaray/TÜRKİYE

[*2] Aksaray University, Faculty of Medicine, Department of Pediatrics Surgery, Aksaray/TÜRKİYE

[*3] Aksaray Training and Research Hospital, Department of Pediatrics, Aksaray/TÜRKİYE

Aim: This study aimed to evaluate the accuracy and completeness of ChatGPT-4 and Google Gemini in answering questions about undescended testis (UDT), as these AI tools can sometimes provide seemingly accurate but incorrect information, raising caution in medical applications.

Material and Method: Researchers created 20 identical questions independently and submitted them to both ChatGPT-4 and Google Gemini.A pediatrician and a pediatric surgeon evaluated the responses for accuracy, using the Johnson et al. scale (accuracy rated from 1 to 6 and completeness from 1 to 3). Responses that lacked content received a score of 0. Statistical analyses were performed using R Software (version 4.3.1) to assess differences in accuracy and consistency between the tools.

Results: Both chatbots answered all questions, with ChatGPT achieving a median accuracy score of 5.5 and a mean score of 5.35, while Google Gemini had a median score of 6 and a mean of 5.5. Completeness was similar, with ChatGPT scoring a median of 3 and Google Gemini showing comparable performance.

Conclusion: ChatGPT and Google Gemini showed comparable accuracy and completeness; however, inconsistencies between accuracy and completeness suggest these AI tools require refinement. Regular updates are essential to improve the reliability of AI-generated medical information on UDT and ensure up-to-date, accurate responses.

Keywords: ChatGPT, Gemini, Children, Undescended Testicle.

Amaç: Bu çalışma, ChatGPT-4 ve Google Gemini'nin inmemiş testisle ilgili soruları yanıtlamadaki doğruluğunu ve eksiksizliğini değerlendirmeyi amaçlamıştır. Çünkü bu yapay zeka araçları bazen görünüşte doğru ama yanlış bilgiler sağlayabilmektedir ve bu da tıbbi uygulamalarda dikkatli olunmasını gerektirmektedir.

Gereç ve Yöntem: Araştırmacılar, 20 özdeş soruyu bağımsız olarak oluşturup hem ChatGPT-4 hem de Google Gemini'ye göndermişlerdir. Bir çocuk doktoru ve bir çocuk cerrahı, yanıtları doğruluk açısından Johnson ve ark. ölçeğini (doğruluk 1 ile 6 arasında, eksiksizlik ise 1 ile 3 arasında derecelendirilmiştir) kullanarak değerlendirmiştir. İçerik içermeyen yanıtlar 0 puan almıştır. Araçlar arasındaki doğruluk ve tutarlılık farklılıklarını değerlendirmek için istatistiksel analizler R Yazılımı (sürüm 4.3.1) kullanılarak gerçekleştirilmiştir.

Bulgular: Her iki sohbet robotu da tüm soruları yanıtlamış; ChatGPT'nin ortanca doğruluk puanı 5,5 ve ortalama puanı 5,35 iken, Google Gemini'nin ortanca puanı 6 ve ortalama puanı 5,5 olmuştur. Tamlık benzerdi; ChatGPT'nin ortalama puanı 3 iken, Google Gemini benzer bir performans gösterdi.

Sonuç: ChatGPT ve Google Gemini benzer doğruluk ve tamlık gösterdi; ancak doğruluk ve tamlık arasındaki tutarsızlıklar, bu yapay zeka araçlarının iyileştirilmesi gerektiğini gösteriyor. UDT'de yapay zeka tarafından oluşturulan tıbbi bilgilerin güvenilirliğini artırmak ve güncel, doğru yanıtlar sağlamak için düzenli güncellemeler şarttır.

Anahtar Kelimeler: ChatGPT, İnmemiş Testis., Gemini, Çocuklar

Corresponding Author: Emine Özdemir Kaçer
Aksaray University, Faculty of Medicine, Department of Pediatrics, Aksaray/TÜRKİYE
dremineozdemir@hotmail.com

Recieved : 10.07.2025
Accepted : 11.08.2025

Cite: Kacer EO, Tusat M, Kılıcaslan M, Memis S.AI-Assisted Knowledge Assessment: Comparison of ChatGPT and Gemini on Undescended Testicle in Children. asujms 2025; 5(3):93-97

INTRODUCTION

The adoption of AI chatbots in healthcare is expanding rapidly, with medical applications becoming a key area of research. Among these tools, OpenAI's ChatGPT has emerged as one of the most popular, while Google's Gemini has also gained attention for its advanced capabilities and innovative features (1). Chatbots are increasingly utilized by both patients and healthcare professionals to access medical information, often replacing traditional search engines.

Undescended testicle (UDT) refers to the absence of one or both testicles from the lower part of the scrotum, with the testicles instead located in the groin or abdominal cavity (2). The prevalence of this condition varies; studies suggest that it affects 3% to 5% of newborns and 1-2% during the first year (3). Testicular exams are essential components of routine pediatric assessments to ensure timely intervention, reducing risks such as infertility and testicular cancer (4). Diagnosis typically relies on clinical evaluation, with a detailed pediatric genital exam often sufficient (5).

Cryptorchidism poses significant risks, including an elevated chance of malignancy and compromised fertility (6). Although spontaneous resolution is seen in about 70% of cases within the first three to four months, further descent beyond six months is rare, requiring therapeutic intervention (7). Given its high prevalence, raising awareness is crucial, as reflected in the results of national screening programs (8).

The Internet has become one of the primary sources of public health information, particularly for caregivers seeking accurate and understandable advice (9,10). Chatbots, designed to consolidate information from multiple reliable sources, will likely play an increasingly significant role in healthcare in the future, especially when paired with effective data management systems (11).
This study aims to compare ChatGPT and Google Gemini by evaluating their responses to questions about UDT. As far as we know, this research is the first of its kind in the literature.

MATERIALS AND METHODS

Ethics

Since no patient data were used, ethics committee approval was not required for this study.

Study Design

The study was conducted between August 2 and September 2, 2024, at the Pediatrics and Pediatric Surgery Departments of Aksaray Training and Research Hospital. A pediatrician (E.O.K.) and a pediatric surgeon (M.T.) collaborated to create a list of 20 questions focusing on UDT. During this process, online health resources such as the 2022 European Association of Urology (EAU) Guidelines were reviewed. Frequently asked questions from patients and caregivers were selected and adapted to a clinical context.

Responses were generated between September 4 and 18, 2024, using the free versions of ChatGPT 4 (OpenAI) and Google Gemini (Google LLC). To avoid bias from previous sessions, conversations were reset after each query. Evaluations were conducted in two rounds on different days, with a 24-hour gap to reduce redundancy.

Accuracy Assessment

Response quality was evaluated using two predefined scales following Johnson et al.'s framework (12). Accuracy Scale (6-point Likert):1 = Completely incorrect, 2 = More false than true, 3 = Balanced (equal true and false), 4 = More true than false, 5 = Almost all true, 6 = Completely true. Completeness Scale (3-point Likert):1 = Incomplete (some key aspects missing), 2 = Adequate (minimum essential information provided), 3 = Comprehensive (provides additional context beyond expectations).

Two independent reviewers (E.Ö.K. and M.T.) assessed the responses for accuracy and completeness. Disagreements were resolved through discussion and consensus to minimize bias and ensure reliability.

Statistical Analysis

Data were transferred to Microsoft Excel for further processing. Continuous variables were summarized as mean ± SD and median (25th–75th percentile), while categorical data were presented as frequencies and percentages. For group comparisons, Student's t-test or Mann-Whitney U test was used, depending on the data distribution. The Intraclass Correlation Coefficient (ICC) was calculated using the Two-Way Random Model with Absolute Agreement to assess agreement between the chatbots. p-values <0.05 were considered statistically significant.

RESULTS

When analyzing the 20 questions, neither chatbot received the lowest accuracy score (1). ChatGPT scored the highest possible accuracy (6) in 11 questions (55%), while Gemini achieved full scores in 13 questions (65%). Both platforms performed similarly in terms of completeness, each scoring 55% on the top level.

ChatGPT had a score of 4 for 8% (n=4) of the answers, while the lowest completeness score (2) was observed in 45% (n=9) of the responses. Comparing their answers, both chatbots differed in three questions (n=3). Detailed distributions of Likert scores are shown in Table 1.

Table 1: Questions about undescended testis and ChatGPT and Google Gemini responses score

Question	Accuracyscore	Completenessscore	Accuracyscore	Completenesscore
	ChatGPT 4		Gemini	
1. What is an undescended testicle?	5	2	6	2
2. How is undescended testicle diagnosed?	6	3	6	3
3. What causes undescended testicles?	5	2	6	3
4. What happens if undescended testicle is not operated on?	6	3	6	3
5. Is it possible to have children with undescended testicles?	6	3	4	2
6. When does the undescended testicle descend?	6	3	6	2
7. Until what age do testicles develop?	6	3	5	3
8. Can undescended testicles occur at birth?	4	2	5	2
9. Will my child's future sexual health be affected by undescended testicles?	6	3	6	3
10. Is undescended testicle condition genetic? Can it be seen in other members of the family?	5	3	5	2
11. Can my other children have undescended testicles?	5	2	6	3
12. Why is the location of the testicle important and why might the testicle not fully descend?	6	3	6	2
13. Are there any other treatment methods other than surgery?	5	2	6	3
14. What is the cost of treatment?	4	2	6	3
15. What is the postoperative process? How long does it take for my child to recover?	6	3	4	2
16. Is there a risk of undescended testicle causing infertility in the future if left untreated?	5	3	6	3
17.Does undescended testicle increase the risk of testicular cancer in the future?	6	3	6	3
18. If one testicle has not descended, will the other testicle be affected as well?	5	2	5	2
19. Will there be recurrences after surgery?	6	2	4	2
20. Can undescended testicles be treated with medication?	4	2	6	3

The median accuracy score for ChatGPT was 5.5 (mean: 5.35, SD: 0.75), while Gemini had a median of 6 (mean: 5.5, SD: 0.76). Both platforms achieved a median completeness score of 3 (mean: 2.55, SD: 0.51). Comparison of completeness and accuracy scores between ChatGPT and Google Gemini responses on UDT is shown in Table 2.

Table 2: Comparison of completeness and accuracy scores between ChatGPT and Google Gemini responses on undecended testis

	ChatGPT		Gemini		*p*
	Mean±SD	Median (Q1-Q3),	Mean±SD	Median (Q1-Q3),	
Accuracy	5,35±0,75	5,5 (5-6)	5,50±0,76	6 (5-6)	0,437
Completeness	2,55±0,51	3 (2-3)	2,55±0,51	3 (2-3)	1,000

SD: standard deviation, (Q1-Q3): 25th–75th percentile. p-value <0.05 was considered statistically significant

Statistical analysis showed no significant correlation between accuracy and completeness scores for either chatbot. ChatGPT's p-value was 0.437, while Gemini's was 1.00, indicating no significant relationship between the two measures. The ICC results also showed no statistically significant agreement between accuracy and completeness for either platform.

DISCUSSION

Large language models (LLMs) like ChatGPT and Google Gemini are becoming increasingly essential in healthcare due to their ability to provide fast information retrieval and support decision-making algorithms. These tools not only offer general health information but also assist medical staff by enabling them to respond quickly to electronic patient inquiries in a reliable, user-friendly way, with the flexibility to adapt to different clinical needs (13).

The present study compared the accuracy and completeness of responses generated by ChatGPT and Gemini on the topic of UDT. Similar research evaluating the reliability of AI chatbots in pediatric orthopedics reported comparable agreement levels for ChatGPT and Gemini, with 67% and 69% agreement, respectively (14). Our study also observed that both chatbots produced relevant responses consistently on the first attempt for all questions.

The results indicate that Gemini achieved higher accuracy, with a median score of 6, compared to ChatGPT's 5.5. In 13 of the 20 questions, Gemini provided fully accurate answers, whereas ChatGPT did so for 11 questions. These findings align with previous studies where Bard outperformed ChatGPT in responding to medical questions about pregnancy (15). Both chatbots demonstrated similar performance in terms of completeness, with median scores of 3 for both platforms. This level of consistency was also reported in a study by Mediboina et al., who assessed ChatGPT's responses across different datasets and found comparable completeness scores (16).

While the performance across various contexts suggests robustness, previous research has highlighted that some AI chatbots, despite high accuracy, lack perfect consistency in their responses (17). For instance, inconsistencies in chatbot-generated answers to vascular surgery questions were documented, although the overall information remained valid (18).

In our study, statistical tests did not reveal significant correlations between accuracy and completeness scores for either chatbot. This aligns with Mediboina et al.'s findings, where minor inaccuracies were detected but did not correlate with completeness, particularly in responses related to abortion care (16). The low correlation in our results could stem from the small sample size used in this evaluation.

Question-specific performance also varied between the chatbots. Gemini performed better in answering questions about treatment costs and medication options, while ChatGPT excelled in explaining postoperative recovery and the potential for recurrence after surgery. This variability mirrors findings from other studies comparing AI chatbots' responses in areas like bone health and skeletal biology (19). Similarly, research on glucocorticoid-induced osteoporosis found that different chatbots demonstrated strengths in distinct areas, reflecting their specialized capacities (20).

Studies in other fields, such as dentistry, have also identified limitations in AI-generated responses. Although these chatbots often provide informative answers, they sometimes deliver ambiguous or partially inaccurate content (8). Similarly, the lack of significant agreement between accuracy and completeness in our study, measured by the Intraclass Correlation Coefficient (ICC), points to challenges in evaluating chatbot performance reliably. These discrepancies highlight the need for more rigorous assessments to ensure that chatbots meet the expectations of healthcare providers and users.

Limitations

A key limitation of our study is the small dataset, which restricted the ability to generalize findings across broader medical contexts. Additionally, as the data collection occurred at a specific point in time, the results may not reflect future updates or improvements in chatbot functionality. Future studies should involve larger datasets and incorporate longitudinal assessments to monitor performance changes over time.

CONCLUSION

Artificial intelligence platforms like ChatGPT and Google Gemini have become essential tools for improving access to healthcare information. The comparison of these chatbots highlights their potential to enhance learning experiences and knowledge dissemination in medical fields. As these models continue to advance, their utility as reliable sources of information will likely expand. However, to ensure the delivery of accurate and dependable medical advice, comprehensive evaluations and regular audits of these platforms are required, especially when addressing specific health conditions like UDT.

The role of expert oversight remains indispensable in using AI-generated health data. Healthcare professionals must remain actively involved in validating and interpreting chatbot-generated information to safeguard patient care. Additionally, collaboration between medical practitioners and AI developers is crucial to optimize these tools and improve the quality and safety of their responses. Establishing such partnerships will help AI platforms become more effective and trustworthy resources in healthcare.

Declarations

Since no patient data were used, ethics committee approval was not required for this study
Authors' Contributions
EMK: Designed study, collected data, wrote the article, wrote and revised article, supervised study.
MT: Analyzed data
MK: Collected data
SM: Collected data
Both authors have read and approved the final version of the manuscript. Each author meets the ICMJE authorship criteria and accepts responsibility for the integrity and accuracy of the work.
Conflict of Interest: No conflict of interest was declared by the authors.
Financial Disclosure: The authors declared that this study received no financial support.

REFERENCES

1. Patil NS, Huang RS, van der Pol CB, Larocque N. Comparative Performance of ChatGPT and Bard in a Text-Based Radiology Knowledge Assessment. Can Assoc Radiol J. 2024;75(2):344-50.
2. Haid B, Rein P, Oswald J. Undescended testes: Diagnostic Algorithm and Treatment. Eur Urol Focus. 2017;3(2-3):155-7.
3. Bradshaw CJ, Corbet-Burcher G, Hitchcock R. Age at orchidopexy in the UK: has new evidence changed practice? J Pediatr Urol. 2014;10(4):758-62.
4. Kolon TF, Herndon CD, Baker LA, Baskin LS, Baxter CG, Cheng EY, et al. Evaluation and treatment of cryptorchidism: AUA guideline. J Urol. 2014;192(2):337-45.
5. Promm M, Dittrich A, Brandstetter S, Fill-Malfertheiner S, Melter M, Seelbach-Göbel B, et al. Evaluation of Undescended Testes in Newborns: It Is Really Simple, Just Not Easy. Urol Int. 2021;105(11-12):1034-8.
6. Holland AJ, Nassar N, Schneuer FJ. Undescended testes: an update. Curr Opin Pediatr. 2016;28(3):388-94.
7. Batra NV, DeMarco RT, Bayne CE. A narrative review of the history and evidence-base for the timing of orchidopexy for cryptorchidism. J Pediatr Urol. 2021;17(2):239-45.
8. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the Performance of Generative AI Large Language Models ChatGPT, Google Bard, and Microsoft Bing Chat in Supporting Evidence-Based Dentistry: Comparative Mixed Methods Study. J Med Internet Res. 2023;25:e51580.
9. McMullan M. Patients using the Internet to obtain health information: how this affects the patient-health professional relationship. Patient Educ Couns. 2006;63(1-2):24-8.
10. Ozdemir Kacer E, Kacer I. Evaluating the quality and reliability of YouTube videos on scabies in children: A cross-sectional study. PloS one. 2024;19(10):e0310508.
11. Wong ZSY, Zhou J, Zhang Q. Artificial Intelligence for infectious disease Big Data Analytics. Infect Dis Health. 2019;24(1):44-8.
12. Johnson D, Goodman R, Patrinely J, Stone C, Zimmerman E, Donald R, et al. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. Res Sq. 2023.
13. Durmaz Engin C, Karatas E, Ozturk T. Exploring the Role of ChatGPT-4, BingAI, and Gemini as Virtual Consultants to Educate Families about Retinopathy of Prematurity. Children (Basel). 2024;11(6).
14. Pirkle S, Yang J, Blumberg TJ. Do ChatGPT and Gemini Provide Appropriate Recommendations for Pediatric Orthopaedic Conditions? J Pediatr Orthop. 2024.
15. Khromchenko K, Shaikh S, Singh M, Vurture G, Rana RA, Baum JD. ChatGPT-3.5 Versus Google Bard: Which Large Language Model Responds Best to Commonly Asked Pregnancy Questions? Cureus. 2024;16(7):e65543.
16. Mediboina A, Badam RK, Chodavarapu S. Assessing the Accuracy of Information on Medication Abortion: A Comparative Analysis of ChatGPT and Google Bard AI. Cureus. 2024;16(1):e51544.
17. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI Responds to Common Lung Cancer Questions: ChatGPT vs Google Bard. Radiology. 2023;307(5):e230922.
18. Chervonski E, Harish KB, Rockman CB, Sadek M, Teter KA, Jacobowitz GR, et al. Generative artificial intelligence chatbots may provide appropriate informational responses to common vascular surgery questions by patients. Vascular. 2024:17085381241240550.
19. Cung M, Sosa B, Yang HS, McDonald MM, Matthews BG, Vlug AG, et al. The performance of artificial intelligence chatbot large language models to address skeletal biology and bone health queries. J Bone Miner Res. 2024;39(2):106-15.
20. Tong L, Zhang C, Liu R, Yang J, Sun Z. Comparative performance analysis of large language models: ChatGPT-3.5, ChatGPT-4 and Google Gemini in glucocorticoid-induced osteoporosis. J Orthop Surg Res. 2024;19(1):574.1.

Corresponding Author: Emine Özdemir Kaçer
ebrusglm55@gmail.com
Orcid: 0000-0002-0111-1672

Author: Mustafa Tuşat
Orcid: 0000-0003-2327-4250

Author: Murat Kılıçaslan
Orcid: 0000-0003-1243-9830

Author: Sebahattin Memiş
Orcid: 0000-0002-3829-9218