*Androl Bul 2025;27:8–12*
https://doi.org/10.24898/tandro.2025.33349

Erkek Cinsel Sağlığı

# Comparing ChatGPT and Google Gemini in urology: Which ai model provides superior patient education on penile prosthesis?

## ChatGPT ve Google Gemini'nin ürolojide karşılaştırılması: Penil protez hasta eğitiminde hangi yapay zeka modeli üstün?

Mücahit Gelmiş[1], Ali Ayten[1], Çağatay Özsoy[2], Berk Bulut[1], Mustafa Gökhan Köse[1]

## ABSTRACT

**OBJECTIVE:** This study aimed to compare the performance of AI-powered chatbots ChatGPT-4 and Google Gemini in patient education on penile prostheses. Specifically, the evaluation focused on the accuracy, speed, and reproducibility of their responses to patient questions. Access to accurate and comprehensive information about penile prosthesis surgeries directly impacts patients' decision-making processes and treatment outcomes. Therefore, examining the effectiveness of AI-powered platforms in this domain is of significant importance.

**MATERIAL and METHODS:** Fifty questions were sourced from the "People also ask" section of Google search results. These questions were separately submitted to ChatGPT-4 and Google Gemini. The responses were independently evaluated by two experienced urologists using the Global Quality Score (GQS). Misleading information was classified as misinformation. Response times and reproducibility rates were statistically analyzed, with a significance level set at p <0.05.

**RESULTS:** ChatGPT-4 demonstrated a higher overall GQS average compared to Google Gemini (4.9±0.31 vs. 3.45±0.94, p <0.001) and provided faster response times (12.3±2.1 seconds vs. 18.7±3.4 seconds, p <0.001). No statistically significant difference was observed in reproducibility rates between the two platforms (ChatGPT: 94%, Google Gemini: 90%, p=0.20).

**CONCLUSION:** ChatGPT-4 outperformed Google Gemini by providing both faster and more accurate responses. These findings highlight the potential of AI-powered chatbots in patient education. However, the necessity of human oversight to ensure the accuracy of the information provided by these platforms should not be overlooked.

**Keywords:** artificial intelligence, chatbot, patient education, penile prosthesis

## ÖZ

**AMAÇ:** Bu çalışmanın amacı, yapay zekâ destekli sohbet botları ChatGPT-4 ve Google Gemini'nin penil protezle ilgili hasta eğitimi alanındaki performanslarını karşılaştırmaktır. Özellikle bu iki platformun hasta sorularına verdikleri yanıtların doğruluğu, hız ve tekrarlanabilirlik açısından değerlendirilmesi hedeflenmiştir. Penil protez ameliyatları hakkında bilgi arayan hastaların doğru ve kapsamlı bilgiye erişimi, karar verme süreçlerini ve tedavi sonuçlarını doğrudan etkileyebilir. Bu nedenle, yapay zekâ destekli platformların bu alandaki etkinliğini incelemek önemlidir.

**GEREÇ ve YÖNTEMLER:** Çalışmada, Google arama sonuçlarının "People also ask" bölümünden alınan toplam 50 soru kullanılmıştır. Bu sorular, ChatGPT-4 ve Google Gemini'ye ayrı ayrı yönlendirilmiştir. Yanıtlar, iki deneyimli ürolog tarafından bağımsız olarak Global Quality Score (GQS) kullanılarak değerlendirilmiştir. Yanıltıcı bilgiler yanlış bilgi olarak sınıflandırılmıştır. Yanıt süreleri ve tekrarlanabilirlik oranları istatistiksel olarak analiz edilmiştir. Analizlerde p <0,05 anlamlılık düzeyi olarak kabul edilmiştir.

**BULGULAR:** ChatGPT-4, Google Gemini'ye kıyasla daha yüksek bir genel GQS ortalaması (4,9±0,31 vs. 3,45±0,94, p <0,001) ve daha hızlı yanıt süresi (12,3±2,1 saniye vs. 18,7±3,4 saniye, p <0,001) göstermiştir. Tekrarlanabilirlik oranlarında ise anlamlı bir fark bulunmamıştır (ChatGPT: %94, Google Gemini: %90, p=0,20).

**SONUÇ:** ChatGPT-4, hem daha hızlı hem de daha doğru yanıtlar sunarak Google Gemini'ye üstünlük sağlamıştır. Bu bulgular, yapay zekâ destekli sohbet botlarının hasta eğitimi alanında önemli bir potansiyel sunduğunu ortaya koymaktadır. Ancak, bu platformların bilgi doğruluğunu sağlamak için insan denetimine ihtiyaç duyduğu unutulmamalıdır.

**Anahtar Kelimeler:** hasta eğitimi, penil protez, sohbet botu, yapay zeka

## INTRODUCTION

In the field of urology, technological advancements continue to enhance both patient care and surgical outcomes, particularly in the realm of penile prosthesis implantation. This procedure, which is performed to treat erectile dysfunction that is unresponsive to medical therapy, requires precise patient education to ensure informed decision-making and optimal postoperative satisfaction.[1] Traditionally, urologists have been the primary source of this vital information, guiding

[1]*Gaziosmanpaşa Eğitim ve Araştırma Hastanesi, Gaziosmanpaşa, İstanbul, Türkiye*
[2]
*Aydın Adnan Menderes Üniversitesi Tıp Fakültesi, Aydın, Türkiye*

**Yazışma Adresi/** *Correspondence:*
Uzm. Dr. Mücahit Gelmiş
Karayolları, Osmanbey Cd. 621 Sokak, 34255 Gaziosmanpaşa 34255 İstanbul - Türkiye
Tel: +90 212 945 30 00
E-mail: mucahitgelmis@gmail.com

**Geliş/** *Received:* 09.12.202
**Kabul/** *Accepted:* 16.12.2024

patients through the complexities of surgical options, risks, benefits, and post-surgical care.

However, the digital age has introduced new tools that patients increasingly turn to for information. Most notably, artificial intelligence (AI) chatbots.[2] Among these, ChatGPT (OpenAI, San Francisco, CA, USA) and Google Gemini which was previously named as Google Bard (Google LLC, Mountain View, CA, USA) stand out for their accessibility and advanced natural language processing (NLP) capabilities.[3] While ChatGPT is known for its extensive training data up to October 2023, offering a robust knowledge base, Google Gemini's ability to retrieve real-time information from the internet provides potentially more up-to-date insights. This distinction is particularly critical in a field like urology, where ongoing research and new surgical techniques can directly impact patient outcomes.

As AI models become more integrated into patient education, understanding their accuracy, reliability, and potential impact on clinical decision-making becomes increasingly important.[4] In the context of penile prosthesis, where patients may have specific concerns about surgical procedures, risks, and long-term satisfaction, the quality of information provided by these AI models could significantly influence their choices.

This article aims to evaluate and compare ChatGPT and Google Gemini as sources of patient education on penile prosthesis. By analyzing the accuracy, depth, and relevance of their responses to frequently asked questions in this specialized area of urology, we hope to provide insights into the strengths and limitations of each platform, ultimately guiding both patients and healthcare providers in their use of these emerging technologies.

## MATERIALS and METHODS

### Question Selection Process

To comprehensively evaluate patient education on penile prosthesis, questions were selected based on common concerns encountered during the penile prosthesis placement process. A Google search was conducted using the keyword "penile prosthesis" in an incognito mode with a previously cleared search history. Questions were extracted from the "People also ask" section, where the first 50 questions were collected. To refine this list, duplicate questions with the same or similar meanings were removed. Finally, these questions were distributed evenly into three domains of interest:

1. Penile Prosthesis Types and Surgical Process (17 questions)
2. Postoperative Period and Complications (17 questions)
3. Long-term Use and Patient Satisfaction (16 questions)

### AI Platform Querying Process

These 50 questions were sent separately to ChatGPT-4.0 (OpenAI, San Francisco, CA, USA) and Google Gemini (Google LLC, Mountain View, CA, USA). To ensure consistency, a new user account was created for this study, and chat histories were cleared before each question was submitted to minimize potential bias stemming from memory retention features of the AI platforms.

### Evaluation of AI Responses

Responses generated by the AI platforms were independently assessed by two urological surgeons with expertise in andrology. Each response was evaluated using the Global Quality Score (GQS), a validated scoring system ranging from 1 (poor quality) to 5 (excellent quality).[5] The raters were blinded to the sources of the responses to mitigate bias, and their evaluations were conducted independently. The dissemination of erroneous, misleading, or false information was categorized as misinformation.

### Ethical Approval

This study did not involve human participants or the use of patient data. The data analyzed were generated solely by artificial intelligence platforms (ChatGPT and Gemini) in response to standardized questions.

### Statistical Analysis

The statistical analysis was performed using IBM Statistical Package for Social Sciences (SPSS) program version software version 27 (IBM, Chicago, IL, USA). Independent-sample t-tests were employed to identify statistically significant differences between the average GQS scores of ChatGPT-4.0 and Google Gemini. Reproducibility of responses, defined as the proportion of consistent and reliable answers, was analyzed using frequency metrics, presented as n (%). Differences in response time (seconds) were compared using independent-sample t-tests. A p-value <0.05 was considered statistically significant for all comparisons.

## RESULTS

### Global Quality Score (GQS) Evaluation

The average GQS of ChatGPT-4.0 was significantly higher across all three domains compared to Google Gemini (Table 1). ChatGPT achieved an overall average score of 4.9±0.31, whereas Google Gemini scored 3.45±0.94 (p

**Table 1.** Comparative analysis of AI platforms in penile prosthesis patient education

| | *ChatGPT 4* | *Gemini* | *P value* |
|---|---|---|---|
| GQS* | | | |
| Penile Prosthesis Types & Process (17 Questions) | 4.8 ± 0.2 | 3.7 ± 0.8 | <0.001 |
| Postoperative Period & Complications (17 Questions) | 5.0 ± 0.0 | 3.4 ± 0.9 | <0.001 |
| Long-term Use & Satisfaction (16 Questions) | 4.9 ± 0.1 | 3.3 ± 1.0 | <0.001 |
| Reproducibility, n (%) | 47/50 (94%) | 45/50 (90%) | 0.20 |
| Response Time (second) | 12.3 ± 2.1 | 18.7 ± 3.4 | <0.001 |

*Global Quality Score, ** mean ± standard deviation

**Table 2.** Comparative analysis of AI platforms in penile prosthesis patient education

| | *ChatGPT 4* | *Gemini* | *P value* |
|---|---|---|---|
| GQS* | | | |
| Penile Prosthesis Types & Process (17 Questions) | 4.8 ± 0.2 | 3.7 ± 0.8 | <0.001 |
| Postoperative Period & Complications (17 Questions) | 5.0 ± 0.0 | 3.4 ± 0.9 | <0.001 |
| Long-term Use & Satisfaction (16 Questions) | 4.9 ± 0.1 | 3.3 ± 1.0 | <0.001 |
| Reproducibility, n (%) | 47/50 (94%) | 45/50 (90%) | 0.20 |

*Global Quality Score, ** mean ± standard deviation

<0.001). Both AI platforms demonstrated a capacity to provide good-quality answers; however, ChatGPT consistently delivered more detailed and accurate responses.

### Domain-specific Performance

- **Penile Prosthesis Types and Surgical Process**: ChatGPT outperformed Google Gemini (4.8±0.2 vs. 3.7±0.8, p <0.001).
- **Postoperative Period and Complications**: ChatGPT scored a perfect 5.0±0.0, while Google Gemini scored 3.4±0.9 (p <0.001).
- **Long-term Use and Patient Satisfaction**: ChatGPT scored 4.9±0.1, and Google Gemini scored 3.3±1.0 (p <0.001).

### Reproducibility and Response Time

Reproducibility, measured as consistent and accurate answers, was similar between platforms, with ChatGPT achieving 47/50 (94%) and Google Gemini 45/50 (90%) (p=0.20). In contrast, response time differed significantly; ChatGPT was faster, averaging 12.3±2.1 seconds per response, compared to Google Gemini's 18.7±3.4 seconds (p <0.001) (Table 1).

## DISCUSSION

Artificial intelligence (AI) technology has been increasingly applied across various fields, with significant adoption in healthcare.[6] Among the most commonly utilized platforms in this area are ChatGPT, developed by OpenAI, and Gemini, created by Google. Patients considering penile prosthesis implantation due to erectile dysfunction often approach this decision with hesitation and seek information from diverse online sources. This study aimed to evaluate the reliability of responses provided by ChatGPT and Gemini by comparing their answers to 50 frequently asked questions about penile prostheses.

ChatGPT (4.9±0.31) and Google Gemini (3.45±0.94) demonstrated comparable average scores, with ChatGPT significantly outperforming Gemini in all categories (p <0.001). Previous studies, such as those by Caglar et al., have shown ChatGPT's high accuracy rates (92%) and excellent reproducibility in answering pediatric urology questions.[7] In our study, feedback from two expert urologists specializing in andrology confirmed that ChatGPT's responses were clearer and more concise. These findings suggest that ChatGPT 4.0 provides more effective and reliable information for patients seeking guidance on penile prosthesis implantation, particularly regarding patient education and comprehension. This is an encouraging development, especially given the potential for AI-generated responses to appear misleadingly comprehensive.

While both chatbots provided moderately good quality responses, Google Gemini received a "2" Global Quality Score (GQS) on two questions: "How long do penile prostheses last?" and "How is sexual intercourse with a penile prosthesis?" This low score indicates significant gaps in the detail and accuracy of information provided on these critical topics. The complexity of factors influencing penile prosthesis surgery timing, including patient age, symptoms, and severity, requires detailed understanding. Insufficient responses may lead to patient confusion and incorrect decisions. Giorgino et al. found similar deficiencies in chatbots concerning complex topics like flatfoot in pediatric orthopedics, which could also lead to confusion or incorrect decisions by patients or parents.[8]

Similarly, as reported in Silbergleit et al., Google Gemini also demonstrated lower performance in our study.[9] This could be attributed to gaps in the information provided, insufficient depth of explanations, or less intuitive and user-friendly response formats. These limitations hinder its effectiveness in conveying critical health information, reducing its utility as a resource for patient education and communication.

Overall, both experts agreed that the two chatbots show promising potential as advanced educational tools for patients. However, the general evaluation revealed occasional deficiencies in the coverage of important topics. Ideally, the responses should achieve at least a “Good quality, key topics addressed, useful for patients” GQS rating of 4.[5] The “Postoperative Period and Complications” category is particularly critical, as it directly impacts patient satisfaction. ChatGPT’s perfect score in this category (5.0±0.0) highlights its ability to provide comprehensive and reliable information on this sensitive topic. For instance, addressing issues such as infection risk management and device malfunction preparedness helps patients approach the postoperative period with greater confidence. Conversely, Google Gemini’s lower score in this category suggests that inadequate depth of information could negatively influence patient outcomes. Given the complexity of postoperative care, insufficient guidance in this area could lead to confusion and poor decision-making among patients.

There was no statistically significant difference in reproducibility between the two platforms (p=0.20), with ChatGPT achieving a slightly higher consistency rate (94%) compared to Gemini (90%). This suggests that both platforms can serve as reliable sources of information. However, ChatGPT demonstrated a statistically significant advantage in response time (12.3±2.1 seconds vs. Gemini’s 18.7±3.4 seconds; p <0.001), making it a more appealing option in scenarios where time is a critical factor in patient education. In contrast to the findings of Silbergleit et al., where ChatGPT-3.5 was the fastest platform but provided the least accurate responses compared to ChatGPT-4 and Gemini, our study showed that ChatGPT-4 not only provided faster responses than Gemini but also delivered more accurate results, establishing its superiority in both speed and accuracy.[9]

Patients often rely on search engines for medical answers and tend to trust the information they find.[10] Unlike social media platforms, AI platforms integrate information from multiple reliable sources. The conversational nature of these platforms enhances comprehension by allowing ongoing dialogue. Consequently, AI programs like ChatGPT and Gemini are considered more accessible and user-friendly than other forms of social media-based information. However, users must critically evaluate the content of the responses to ensure accuracy. AI also exhibits “AI hallucination,” where chatbots generate inaccurate responses that appear convincing. This phenomenon poses a risk, particularly when users lack sufficient knowledge, leading them to accept false information as accurate.[11]

Both platforms were evaluated using a transparent and impartial methodology, with efforts made to eliminate potential biases, such as clearing search histories and anonymizing responses from andrology experts. However, several limitations must be considered in interpreting the results of this study. First, restricting the analysis to 50 questions inherently limits the scope, as it only addresses a portion of the potential concerns or issues patients may face regarding penile prosthesis implantation. Future studies could provide a more comprehensive assessment of AI chatbots’ capabilities by incorporating a broader range of clinical scenarios. Additionally, the evaluation was limited to two chatbots, ChatGPT 4.0 and Google Gemini, which may not fully represent the overall performance of AI-assisted chatbots. Excluding other chatbots may reduce the generalizability of the findings. Another limiting factor is the absence of patient feedback analysis. Patient perceptions and needs may differ from expert evaluations, and neglecting this perspective could diminish the practical value of the study’s findings for individuals seeking medical information. Nevertheless, we believe that preliminary analysis by professionals is crucial. Furthermore, the accuracy and reliability of responses generated by these chatbots are heavily dependent on the currency of their training data. Outdated fundamental data could lead to incorrect or less reliable responses, particularly in the context of health information. Although the study was conducted under ideal conditions (using incognito mode and new user profiles), real-world applications may yield different results. Factors such as search history, personal preferences, and evolving search terms in practical scenarios may influence chatbot responses, potentially affecting their effectiveness and reliability in patient education.

Encouragingly, continued application development and collaboration with expert healthcare teams may enhance their reliability. Identifying areas where AI platforms fall short and emphasizing the importance of human expertise and validation in the context of medical information remain essential. Finally, the single-center nature of this study and the evaluation of responses by only two reviewers also represent limitations.

## CONCLUSION

This study demonstrates that ChatGPT 4.0 is a more reliable source of information for patients seeking guidance on penile prostheses compared to Google Gemini. However, it is crucial to critically evaluate the information provided by AI-supported chatbots in healthcare and ensure its accuracy through validation by a healthcare professional. Future research could explore the impact of utilizing more up-to-date AI models and expand the scope of investigations to encompass a broader range of clinical scenarios. Additionally, efforts to enhance training data and provide access to reliable sources may further improve the utility of AI-based chatbots in urological patient education.

**Peer-review**
Externally peer-reviewed.

**Conflict of Interest**
No conflict of interest was declared by the authors.

**Financial Disclosure**
No financial support has been received.

## REFERENCES

1. Krzastek SC, Smith R. An update on the best approaches to prevent complications in penile prosthesis recipients. Ther Adv Urol. 2019;11:1756287218818076. [CrossRef]
2. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595. [CrossRef]
3. Clark M, Bailey S. Chatbots in health care: connecting patients to information: emerging health technologies. Canadian Agency for Drugs and Technologies in Health. 2024;4(1):1–22. [CrossRef]
4. Kaneda Y, Takita M, Hamaki T, Ozaki A, Tanimoto T. ChatGPT's potential in enhancing physician efficiency: a Japanese case study. Cureus. 2023;15(11):e48235. [CrossRef]
5. Bernard A, Langille M, Hughes S, Rose C, Leddin D, Veldhuyzen van Zanten S. A systematic review of patient inflammatory bowel disease information resources on the World Wide Web. Am J Gastroenterol. 2007;102(9):2070–7. [CrossRef]
6. Li W, Fu M, Liu S, Yu H. Revolutionizing neurosurgery with GPT-4: a leap forward or ethical conundrum?. Ann Biomed Eng. 2023;51(10):2105–12. [CrossRef]
7. Caglar U, Yildiz O, Meric A, Ayranci A, Gelmis M, Sarilar O, Ozgor F. Evaluating the performance of ChatGPT in answering questions related to pediatric urology. J Pediatr Urol. 2024;20(1):26.e1–26.e5. [CrossRef]
8. Giorgino R, Alessandri-Bonetti M, Del Re M, Verdoni F, Peretti GM, Mangiavini L. Google Bard and ChatGPT in orthopedics: which is the better doctor in sports medicine and pediatric orthopedics? The role of AI in patient education. Diagnostics (Basel, Switzerland). 2024;14(12):1253. [CrossRef]
9. Silbergleit M, Tóth A, Chamberlin JH, Hamouda M, Baruah D, Derrick S, et al. ChatGPT vs Gemini: comparative accuracy and efficiency in CAD-RADS score assignment from radiology reports. J Imaging Inform Med. 2024;10.1007/s10278-024-01328–y. Online ahead of print. [CrossRef]
10. Diaz JA, Griffith RA, Ng JJ, Reinert SE, Friedmann PD, Moulton AW. Patients' use of the Internet for medical information. J Gen Intern Med. 2002;17(3):180–5. [CrossRef]
11. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. 2023;388(13):1233–9. [CrossRef]