Submitted: 25 January, 2025 Accepted: 28 March, 2025 Published: 30 September, 2025 DOI:10.22514/j.androl.2025.030

ORIGINAL RESEARCH

REV INT ANDROL
Revista Internacional de Andrología

Artificial intelligence applications as a source of information on penis-lengthening and penis girth enhancement: ChatGPT 4o Plus *vs.* Gemini advanced

Çağatay Özsoy[1], Erhan Ateş[1,*], Mücahit Gelmiş[2], Abdullah Akdağ[3]

[1]Department of Urology, Aydin Adnan Menderes University School of Medicine, 09100 Aydın, Türkiye
[2]Department of Urology, Gaziosmanpasa Training and Research Hospital, 34255 Istanbul, Türkiye
[3]Department of Urology, Basaksehir Cam and Sakura City Hospital, 34480 Istanbul, Türkiye

***Correspondence**
erhan.ates@adu.edu.tr
(Erhan Ateş)

Abstract

Background: To evaluate the quality, reliability, and readability of the information provided by artificial intelligence (AI) applications on penis-lengthening and penis girth enhancement, using ChatGPT 4o Plus and Gemini Advanced as examples for the first time. **Methods**: Frequently asked questions (FAQs) about penis-lengthening and penis girth enhancement were derived from the European Association of Urology (EAU) 2024 guidelines, review articles published in the last five years, Google Trends, health forums, YouTube, Instagram, Twitter and hospital websites. These questions were posed to the November 2024 versions of ChatGPT 4o Plus and Gemini Advanced. The reliability and quality of responses were assessed using the modified Global Quality Score (GQS), divided into reliability (GQS-R) and usefulness (GQS-U) subcategories. Readability was evaluated using the Flesch Reading Ease (FRE) and the Flesch-Kincaid Reading Grade Level (FKRGL) scales. Results were compared using appropriate statistical methods. **Results**: Both AI applications were asked 62 questions. ChatGPT 4o Plus scored significantly higher than Gemini Advanced on GQS-R and GQS-U ($p < 0.001$ for both). Gemini Advanced had significantly higher FKRGL scores than ChatGPT 4o Plus ($p < 0.001$), except for penis lengthening questions, where Gemini Advanced exhibited higher FRE scores. ChatGPT 4o Plus was significantly faster in responding ($p = 0.021$). **Conclusions**: Artificial intelligence applications provided high-quality, useful, and reliable information on penis lengthening and penis girth enhancement. However, their readability and comprehensibility remain challenging. ChatGPT 4o Plus delivers information more quickly, reliably and usefully, while Gemini Advanced is more readable and comprehensible.

Keywords

Artificial intelligence; ChatGPT; Gemini; Penis lengthening; Penis girth enhancement

Aplicaciones de inteligencia artificial como fuente de información sobre alargamiento de pene y aumento del grosor del pene: ChatGPT 4o Plus *vs.* Gemini advanced

Resumen

Antecedentes: Evaluar la calidad, fiabilidad y legibilidad de la información proporcionada por aplicaciones de inteligencia artificial (IA) sobre alargamiento de pene y aumento del grosor del pene, utilizando ChatGPT 4o Plus y Gemini Advanced como ejemplos por primera vez. **Métodos**: Las preguntas frecuentes (FAQs) sobre alargamiento de pene y aumento del grosor del pene se derivaron de las guías de la Asociación Europea de Urología (EAU) 2024, artículos de revisión publicados en los últimos cinco años, Google Trends, foros de salud, YouTube, Instagram, Twitter y sitios web de hospitales. Estas preguntas se plantearon a las versiones de noviembre de 2024 de ChatGPT 4o Plus y Gemini Advanced. La fiabilidad y calidad de las respuestas se evaluaron utilizando el puntaje de calidad global modificado (GQS), dividido en las subcategorías de fiabilidad (GQS-R) y utilidad (GQS-U). La legibilidad se evaluó utilizando las escalas de Facilidad de Lectura de Flesch (FRE) y el Nivel de Grado de Lectura de Flesch-Kincaid (FKRGL). Los resultados se compararon utilizando métodos estadísticos adecuados. **Resultados**: A ambas aplicaciones de IA se les realizaron 62 preguntas. ChatGPT 4o Plus obtuvo puntajes significativamente más altos que Gemini Advanced en GQS-R y GQS-U ($p < 0.001$ en ambos casos). Gemini Advanced tuvo puntajes FKRGL significativamente más altos que ChatGPT 4o Plus ($p < 0.001$), excepto en las preguntas sobre alargamiento del pene, donde Gemini Advanced mostró puntajes FRE más altos. ChatGPT 4o Plus fue significativamente más rápido en responder ($p = 0.021$). **Conclusiones**: Las aplicaciones de inteligencia artificial proporcionaron información de alta calidad, útil y confiable sobre el alargamiento del pene y el aumento del grosor del pene. Sin embargo, su legibilidad y comprensibilidad siguen siendo un desafío. ChatGPT 4o Plus ofrece información de manera más rápida, confiable y útil, mientras que Gemini Advanced es más legible y comprensible.

Palabras Clave

Inteligencia artificial; ChatGPT; Gemini; Alargamiento del pene; Aumento del grosor del pene

1. Introduction

The internet has become a frequently used tool for accessing health-related information, due to ease of access. Patients, particularly those who feel embarrassed to consult a doctor, often turn to online sources. Platforms, such as YouTube, Facebook, Instagram and Twitter, are commonly used as health information sources. Numerous studies have explored the accuracy of information provided by these platforms [1–3].

Today, artificial intelligence (AI) and large language models (LLMs) have rapidly transformed the healthcare sector, introducing both promising opportunities and notable challenges, particularly in the realms of patient education and early diagnosis [4]. Among the most widely used LLM systems is Chat Generative Pre-training Transformer (ChatGPT), launched by OpenAI (San Francisco, CA, USA) in November 2022. ChatGPT is a deep learning-based AI chatbot built on natural language processing models and designed to provide logical informative answers instantly by mimicking human language patterns [5]. Released in March 2023, Gemini AI, formerly known as Google Bard and based on the Pathways Language Model 2, is another chatbot similar to ChatGPT but distinguished by its real-time internet access and LaMDA (Language Model for Dialogue Applications) communication model [6].

The reliability and accuracy of AI-generated medical information play a crucial role in shaping patients' decisions and their engagement with healthcare providers [7]. Advanced language models like ChatGPT and Gemini have the potential to offer valuable guidance in the management of urological disease [8–10], yet their effectiveness requires thorough evaluation. Assessing AI's role in patient guidance, comparing its responses to established medical knowledge, and ensuring the dissemination of accurate information before clinical consultations are essential steps. Nevertheless, the growing integration of AI in medicine raises concerns regarding the credibility of its outputs and the influence these automated recommendations may have on patient expectations and clinical outcomes [11].

Although male sexual dysfunction, including hypogonadism, low sexual desire, hypoactive sexual desire disorder, erectile dysfunction, ejaculatory disorders, penile curvature, and priapism have also been investigated using ChatGPT [10, 12–14], penis-lengthening and penis girth enhancement have not been studied with AI. YouTube and social media videos addressing these topics show that the quality and reliability of such information are extremely poor [3, 15, 16].

This study aimed to determine the quality, reliability and readability of information provided by AI applications, specifically ChatGPT 4o Plus and Gemini Advanced, on penis-lengthening and penis girth enhancement for the first time in the literature.

2. Materials and methods

2.1 Question development and AI application assessment

Questions were designed based on the 2024 European Association of Urology (EAU) guidelines and review articles published in the past 5 years [17–19]. Additional questions were included based on frequently asked questions (FAQs) on Google Trends, health forums, social media and hospital web-

sites regarding penis-lengthening and penis girth enhancement (**Supplementary material**). Care was taken to ensure that the questions covered all psychiatric, surgical and non-surgical aspects of penis lengthening and penis girth enhancement.

The questions were categorized into general questions about penis size, penis-lengthening and penis girth enhancement. The questions were presented to ChatGPT 4o Plus (November 2024 version, https://chatgpt.com/) and Gemini Advanced (November 2024 version, https://gemini.google.com/), and their responses were recorded for further analysis. To evaluate AI's ability to respond to instant queries without being influenced by previous questions, the AI session was terminated before each new question, the browser cache was cleared, and a new session was started on a newly opened page before proceeding to the next question. The time between posing the question and receiving the response was measured using a stopwatch. For this purpose, a stopwatch was started when the question was presented to the AI, and it was stopped when the AI's response was completed. To evaluate reproducibility, all questions were repeated on newly opened pages with the browser cache cleared, and the consistency of the responses was assessed. In the reproducibility evaluation, the focus was not on the AI's response being identical to the previous one but rather on the coherence of the conveyed information and whether any information was missing compared to the previous response. An all-or-none approach was used: responses that were fully consistent with the previous answer were recorded as "Yes", while those with any missing information were marked as "No". All interactions with AI bots were conducted by a single researcher (ÇÖ). As no patient data were used, informed consent and ethical committee approval were not required.

2.2 Evaluation of the responses

The responses provided by the AI were stripped of any identifiers such as AI name, font style, numbering or any other distinguishing markers. They were then converted into plain text without any indicators and presented to the evaluator in a completely random order. The evaluation was conducted blindly by an andrologist (EA) with more than 10 years of clinical experience and expertise in andrology, serving as secretary-general of the National Andrology Association and editor of its peer-reviewed journal. After the evaluation of all responses, another researcher (MG) recorded the evaluation results under main headings (Fig. 1).

2.3 Evaluation of reliability and usefulness

The reliability and usefulness of the responses were assessed based on the "Penile Size Abnormalities and Dysmorphophobia" section of the 2024 EAU Guidelines. The Global Quality Score (GQS) assessment scale was utilized, with a Likert-type scale ranging from 1 to 5 based on the quality, flow, and ease of use of information found online, as defined by Bernard *et al.* [20] This scale was modified into two categories of reliability (GQS-R) and usefulness (GQS-U) (Table 1). Reliability was assessed based on the completeness, accuracy, and presence of misleading or incorrect information in the responses. Scores ranged from 1 (poor reliability, significant missing information) to 5 (excellent reliability, comprehensive and accurate responses). Usefulness was evaluated based on the practicality and relevance of the responses to patient inquiries. A score of 1 indicated information that was not useful at all, whereas a score of 5 represented highly useful responses that provided actionable insights for patients.

2.4 Assessment of readability

The Flesch formula, most commonly used readability metric for evaluating the readability of written health information materials, was used. The Flesch Reading Ease (FRE) and the Flesch-Kincaid Reading Grade Level (FKRGL) calculate readability based on average sentence length (in terms of word count) and average word length (in terms of syllable count) (Fig. 2) [21]. Readability scores were calculated using https://readable.com, an online tool that allows the computation of FRE and FKGL scores. The AI-generated responses were stripped of headings and unnecessary spaces, converting them into plain text before the analysis. Scores obtained from the FRE range from 0 (unreadable) to 100 (very easy to read). The FKRGL provides a score corresponding to a grade level. In our study, the interpretation of the FRE and FKRGL scores was made according to Table 2. A lower FKRGL level and a higher FRE score were considered better readability. To reach individuals with low literacy levels, the standard difficulty level was determined as FRE scores of 61–70 points and FKRGL grade levels 8–9, following the literature [22].

2.5 Statistical analysis

Statistical analysis was conducted to compare the GQS, FKRGL, FRE, reproducibility, and response times for penis-related question types between ChatGPT 4o Plus and Gemini Advanced. Descriptive statistics are presented as mean ± standard deviation if the distribution was normal and as median (range) values if it was not, and with a percentage (%) for categorical variables. The normality assumption was checked with the Shapiro-Wilk test. The Mann-Whitney U-test was used for non-parametric comparisons of continuous variables between two independent groups, and Student's *t*-test was used for parametric comparisons. Pearson's chi-square test was used to analyze the relationships between categorical variables. Reproducibility was assessed by repeating each question multiple times (n = 62 questions in total) and calculating the number and percentage of consistent answers for each model. Response times were recorded in seconds for each question and compared between the models. A p-value < 0.05 was considered significant, and all analyses were performed using SPSS version 29 software (IBM Corp., Armonk, NY, USA). Results are presented with 95% confidence intervals.

3. Results

A total of 62 questions were asked of both AI applications. Of these, 17 were general questions about penis size, 24 were related to penis lengthening, and 21 were about penis girth enhancement. As a result, the GQS-R (reliability) and GQS-U (usefulness) scores of ChatGPT 4o Plus were significantly

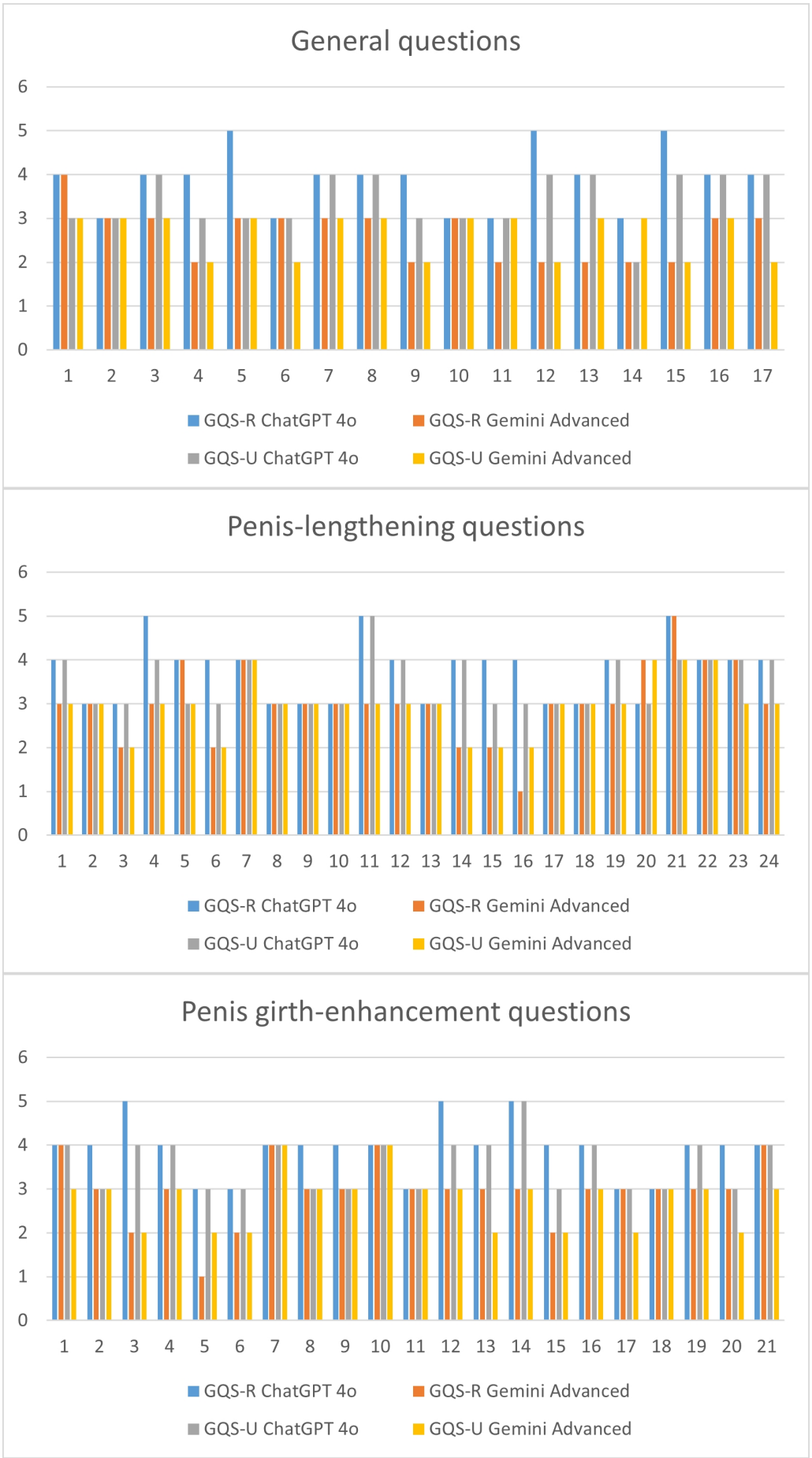

FIGURE 1. The GQS-R and GQS-U scores assigned to the responses provided by ChatGPT-4o Plus and Gemini advanced for questions related to penis size. GQS-R: Global Quality Score reliability; GQS-U: Global Quality Score usefulness.

TABLE 1. Global quality scale reliability and usefulness.

	1	2	3	4	5
GQS-R	Poor reliability, most information missing	Generally poor reliability, some information listed but many important topics missing	Moderate reliability, some important information is adequately discussed but others poorly discussed	Good reliability, most of the relevant information is listed, but some topics not covered	Excellent reliability
GQS-U	Poor usefulness, not at all useful for patients	Generally poor usefulness, very limited use to patients	Moderate usefulness, somewhat useful for patients	Good usefulness, useful for patients	Excellent usefulness, very useful for patients

GQS-R: Global Quality Scale reliability; GQS-U: Global Quality Scale usefulness.

Flesch Reading Ease (FRE)

$$206.835 - 1.015\left(\frac{total\ words}{total\ sentences}\right) - 84.6\left(\frac{total\ syllabes}{total\ words}\right)$$

Flesch–Kincaid Reading Grade Level (FKRGL)

$$0.39\left(\frac{total\ words}{total\ sentences}\right) + 11.8\left(\frac{total\ syllabes}{total\ words}\right) - 15.59$$

FIGURE 2. Calculation of the Flesch Reading Ease (FRE) and the Flesch-Kincaid Reading Grade Level (FKRGL).

TABLE 2. Interpretation of flesch reading ease scores.

Reading Difficulty	FRE score	Estimated FKRGL reading grade level
Very easy	91–100	Grade 5 or 11 year old
Easy	81–90	Grade 6
Fairly easy	71–80	Grade 7
Standard	61–70	Grade 8–9 or 13–15 year old
Fairly difficult	51–60	Grade 10–12
Difficult	31–50	Grade 13–16
Very difficult	0–30	College graduate

FRE: Flesch Reading Ease; FKRGL: Flesch-Kincaid Reading Grade Level.

higher in all three categories of general penis size, penis-lengthening, and penis girth enhancement compared to Gemini Advanced (GQS-R, $p < 0.001$; GQS-U, $p < 0.001$) (Table 3). Additionally, ChatGPT 4o Plus tended to display greater consistency in providing the same answers when the questions were repeated, compared to Gemini Advanced ($p = 0.044$). A Comparison of the FKGRL, the FRE, reproducibility and, response times for penis-related question types between ChatGPT 4o Plus and Gemini Advanced is presented in Table 4. Regarding response times, ChatGPT 4o Plus was faster than Gemini Advanced, and this difference was significant ($p = 0.021$). Gemini Advanced had higher FKGRL scores than ChatGPT 4o Plus. This difference was significant for penis-lengthening questions and across all questions ($p < 0.001$). However, Gemini Advanced tended to show higher FRE scores in categories other than penis-lengthening, although this difference was not significant. In conclusion, while the readability of responses provided by both AI bots was generally challenging, Gemini Advanced had a slightly better readability level than ChatGPT 4o Plus. Both applications require a reading level equivalent to someone above the age of 15. The response time of ChatGPT 4o Plus was significantly shorter than that of Gemini Advanced ($p = 0.021$).

TABLE 3. Comparison of global quality scale reliability and usefulness scores for penis-related question types between ChatGPT 4o and Gemini advanced.

	ChatGPT 4o	Gemini Advanced	*p* value
Generel questions about penis size (n = 17)			
GQS-R	4 (3–5)	3 (2–4)	<0.001
GQS-U	3 (2–4)	3 (2–3)	<0.001
Penis-lengthening questions (n = 24)			
GQS-R	4 (3–5)	3 (1–5)	0.003
GQS-U	3 (3–5)	3 (2–4)	0.006
Penis girth enhancement questions (n = 21)			
GQS-R	4 (3–5)	3 (2–4)	<0.001
GQS-U	4 (3–5)	3 (2–4)	<0.001
All questions (n = 62)			
GQS-R	4 (3–5)	3 (1–5)	<0.001
GQS-U	4 (3–5)	3 (2–4)	<0.001

The findings are presented as median (minimum–maximum). The statistical analysis was performed using the Mann-Whitney U test. GQS-R: Global Quality Scale reliability; GQS-U: Global Quality Scale usefulness.

TABLE 4. Comparison of the flesch-kincaid grade level, the flesch reading ease, reproducibility and response times for penis-related question types between ChatGPT 4o and Gemini advanced.

	ChatGPT 4o	Gemini Advanced	*p* value
FKRGL			
Generel questions about penis size (n = 17)	10.80 (6.8–13)	11.4 (5.8–13.9)	0.306
Penis-lengthening questions (n = 24)	10.25 (8.3–12.9)	12.45 (9.3–34.5)	<0.001
Penis girth enhancement questions (n = 21)	9.8 (8.4–14.1)	10.8 (8.4–16.2)	0.131
All questions (n = 62)	10.4 (6.8–14.1)	11.70 (5.8–30.5)	<0.001
FRE			
Generel questions about penis size (n = 17)	32.40 ± 9.98	40.33 ± 10.66	0.858
Penis-lengthening questions (n = 24)	35.97 ± 9.49	30.67 ± 9.28	0.511
Penis girth enhancement questions (n = 21)	37.87 ± 9.34	38.93 ± 12.76	0.486
All questions (n = 62)	32.40 ± 9.98	35.97 ± 9.49	0.200
Reproducibility, n (%)	54 (87%)	45 (72.5%)	0.044
Response Time (s)	3 (1–6)	4 (2–6)	0.021

The findings are presented as median (minimum–maximum). mean ± SD or as percentages (n %). Statistical analyses included Mann-Whitney U test, Student T test and Chi-square test. SD: standard deviation; FKRGL: The Flesch-Kincaid Reading Grade Level; FRE: The Flesch Reading Ease.

4. Discussion

The present study is the first to demonstrate that AI provided high-quality and useful information on penis-lengthening and penis girth enhancement, yet the readability and comprehensibility of this information remained challenging. Furthermore, this study highlights the superiority of ChatGPT 4o Plus over Gemini Advanced in terms of reliability, usefulness, reproducibility and response speed, while Gemini Advanced outperformed ChatGPT 4o Plus in readability and comprehensibility.

ChatGPT, trained on extensive textual data, effectively captures the nuances and complexities of human language. It provides instant, logical, informative, contextually relevant and appropriate responses to queries [5]. Given the ever-increasing body of medical knowledge, it is inevitable that the healthcare field will embrace this technology, as it is characterized by large amounts of textual data and complex clinical applications. Indeed, physicians and patients have shown interest in ChatGPT due to its ability to deliver accurate and prompt answers on a wide range of topics [23]. Gemini AI, formerly known as Google Bard, is based on the Pathways Language Model 2 and similarly engages in conversational interactions in response to human input [6].

The quality and reliability of medical information provided by Gemini AI and ChatGPT have been investigated, and the two systems have been compared [24–26]. Andrological disorders, encompassing male reproductive and sexual health, are

among the topics that have been explored. Şahin *et al.* [14] entered the most frequently searched terms related to premature ejaculation (PE) into ChatGPT, evaluating the generated responses for readability. They reported that the text produced by ChatGPT was of questionable quality, significantly difficult to read and comprehend and of a literary level that could only be understood by highly educated individuals. Similarly, in our study, the readability and comprehensibility of the ChatGPT responses were challenging. However, our findings differed by highlighting the high quality and reliability of the responses.

In terms of readability, Gemini Advanced had higher FKRGL scores, making its responses more complex and suited for highly literate users, it also showed better FRE scores in some areas, improving accessibility. In contrast, ChatGPT 4o Plus had lower FKRGL scores, making its responses easier to understand for a broader audience but potentially reducing comprehensiveness. These findings highlight the need for AI models to balance readability with medical depth to serve diverse user populations effectively. The fact that readability differences were significant in some categories but not in others suggests that the generalizability of these findings may be limited. The variations across different types of questions indicate that AI models do not perform uniformly across all topics, which should be considered when interpreting the results.

Yigman *et al.* [10] identified the four most searched keywords for sexual dysfunction, erectile dysfunction and PE on Google and grouped them under headings, such as definitions, causes, pharmacological treatments and general treatments. They reported that ChatGPT was an acceptably useful and reliable source for acquiring information about sexual dysfunction and its most common causes, erectile dysfunction and PE. The scales used to characterize the usefulness and reliability of information provided by ChatGPT were consistent and reliable. However, they noted questionable consistency among evaluators regarding the usefulness of PE information. Ergin *et al.* [13] recorded frequently asked questions on health websites, urology association websites, and social media platforms regarding topics, such as male hypogonadism, erectile dysfunction (ED), ejaculatory disorders, penile curvature, penile size abnormalities, priapism and male infertility. They converted the strongly recommended guidelines of the EAU sexual and reproductive health section into questions and presented them to ChatGPT-3.5 and 4o. The accuracy rates for the responses were 85.2% for frequently asked questions in ChatGPT-3.5 and 88.8% in ChatGPT 4o. The accuracy rates for questions derived from the guidelines were 81.5% for ChatGPT-3.5 and 88.9% for ChatGPT 4o. However, the evaluation of penile size abnormalities was limited to questions like "What is the normal length and girth of the penis?" and "Is penis size important to women?". Our study went further by asking 14 questions about penile size, allowing for a more in-depth assessment. Similar findings were reported by Caglar *et al.* [12], who entered frequently asked questions from hospital websites, YouTube, and Instagram into ChatGPT using the EAU guideline recommendations on andrological topics, such as male hypogonadism and ED. They reported that ChatGPT answered 87.9% of questions accurately and sufficiently, 9.3% accurately but insufficiently, and no questions incorrectly. The highest accuracy rates were observed for ejaculatory disorders, penile curvature, and male hypogonadism.

From ancient times, the penis has been associated with power and viewed as a symbol of masculinity. The societal idealization of larger penises can lead to significant psychological distress for patients. Numerous surgical and non-surgical methods have been developed to improve penile length and girth [17–19]. While patients often research this topic on social media, studies evaluating the quality and reliability of these platforms are limited. Seranio *et al.* [3] assessed the top 100 most-viewed YouTube videos on penis-lengthening, and most videos discussed non-surgical methods, such as penile traction devices (19.2%) and surgical methods (65.1%). However, the overall quality and reliability of the videos were poor. Videos created by physicians were rated higher in quality. Another study investigating the relationship between the quality of YouTube videos on penis-lengthening surgery and the academic profiles of surgeons reported a significant correlation between video quality and the surgeon's h-index and total publication count [27]. Videos containing academic knowledge were of higher quality. In our study, we tested the grasp of AI for academic knowledge by asking questions derived from the EAU guidelines and review articles. Çağlayan *et al.* [28] analyzed the reliability of Instagram posts with the hashtag #penislengthening and the impact of such posts on young adult men's perception of their penis size. As results, only six of 1000 posts (1%) contained reliable information. During the survey phase of the study, participants reported a significant decline in genital self-image ($p < 0.001$) and a significant increase in media exposure ($p < 0.001$).

The growing presence of AI in scientific research has sparked concerns about its potential to generate inaccurate or misleading information. While AI offers numerous advantages, the risk of producing incorrect diagnoses or recommendations must be carefully considered [11]. The performance of chatbots is influenced by various factors, including data availability, linguistic structures, and contextual differences [4]. To ensure accuracy and reliability, multilingual chatbots must adapt to these variations when processing and delivering medical information.

If AI models are trained with outdated or incomplete datasets, they may generate incorrect medical advice and fail to identify rare diseases, leading to significant health risks. AI-powered chatbots cannot replace professional medical evaluations, as they rely on user-input symptoms rather than comprehensive clinical assessments. A major limitation is their inability to account for individual health conditions such as chronic illnesses, allergies, or psychiatric disorders [7].

To enhance the reliability of AI-driven healthcare tools, continuous updates with the latest medical literature and clinical research are essential. Additionally, establishing oversight mechanisms that verify the accuracy of medical information in chatbot responses can significantly improve their effectiveness. Implementing such safeguards will help address existing quality concerns and ensure that AI applications in healthcare remain a valuable and trustworthy resource.

This study had some limitations. The findings were based solely on data available during the study period. As AI continues to learn and evolve, responses to the same questions

may vary over time, due to updates and changes in training data. The evaluation of responses from ChatGPT inherently introduced subjectivity to the study, as it relied on human judgment. Additionally, questions were asked only in one language, and the quality of responses may have differed in other languages.

5. Conclusions

The findings of this study suggest that LLMs have the potential to provide reliable information in the field of penile lengthening and penile girth enhancement. Artificial intelligence models can serve as an easily accessible source of information for patients, yet the balance between the accuracy, reliability, and readability of their responses must be carefully assessed. Our study provides valuable insights for healthcare professionals, patients and AI developers regarding the evaluation of AI-generated health information.

Notably, ChatGPT 4o Plus demonstrated superiority in reproducibility, quality, reliability, usefulness and response speed, making it a potential tool for clinical information support. In contrast, Gemini Advanced exhibited better readability and comprehensibility, suggesting its potential use in patient education materials. However, both models need further development to improve the comprehensibility of their responses. For AI systems to be more widely adopted in the healthcare field, their content must be optimized for better readability. Future research should focus on integrating AI models into medical information dissemination processes and ensuring that patients can easily understand the information provided.

AVAILABILITY OF DATA AND MATERIALS

The data supporting the findings of the present study may be found at https://doi.org/10.5281/zenodo.14645247.

AUTHOR CONTRIBUTIONS

ÇÖ and EA—designed the research study; wrote the manuscript. ÇÖ—performed the research. AA—provided help and advice on methodology and resources. MG—analyzed the data. All authors read and approved the final manuscript. All authors contributed to editorial changes in the manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

ACKNOWLEDGMENT

Not applicable.

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/1972591395783753728/attachment/Supplementary%20material.docx.

REFERENCES

[1] Baydilli N, Selvi I. Is social media reliable as a source of information on Peyronie's disease treatment? International Journal of Impotence Research. 2022; 34: 295–301.

[2] Xu AJ, Myrie A, Taylor JI, Matulewicz R, Gao T, Pérez-Rosas V, *et al*. Instagram and prostate cancer: using validated instruments to assess the quality of information on social media. Prostate Cancer and Prostatic Diseases. 2022; 25: 791–793.

[3] Seranio N, Muncey W, Cox S, Belladelli F, Del Giudice F, Glover F, *et al*. Size matters: characterizing penile augmentation content from the 100 most popular YouTube videos. International Journal of Impotence Research. 2024; 36: 493–497.

[4] Walker HL, Ghani S, Kuemmerli C, Nebiker CA, Müller BP, Raptis DA, *et al*. Reliability of medical information provided by ChatGPT: assessment against clinical guidelines and patient information quality instrument. Journal of Medical Internet Research. 2023; 25: e47479.

[5] Roumeliotis KI, Tselikas ND. ChatGPT and Open-AI models: a preliminary review. Future Internet. 2023; 15: 192.

[6] Singh SK, Kumar S, Mehra PS. Chat GPT & Google Bard AI: a review. 2023 International Conference on IoT, Communication and Automation Technology (ICICAT). 23–24 June 2023. IEEE: India. 2023.

[7] Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, *et al*. ChatGPT and other large language models are double-edged swords. Radiology. 2023; 307: e230163.

[8] Szczesniewski JJ, Tellez Fouz C, Ramos Alba A, Diaz Goizueta FJ, García Tello A, Llanes González L. ChatGPT and most frequent urological diseases: analysing the quality of information and potential risks for patients. World Journal of Urology. 2023; 41: 3149–3153.

[9] 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.e1–26.e5.

[10] Yigman M, Untan I, Dogan AE. ChatGPT: a new hope for sexual dysfunction sufferers? Journal of Men's Health. 2024; 20: 135–140.

[11] Sedaghat S. Early applications of ChatGPT in medical practice, education and research. Clinical Medicine. 2023; 23: 278–279.

[12] Caglar U, Yildiz O, Ozervarli MF, Aydin R, Sarilar O, Ozgor F, *et al*. Assessing the performance of chat generative pretrained transformer (ChatGPT) in answering andrology-related questions. Urology Research and Practice. 2023; 49: 365–369.

[13] Ergin İE, Sancı A. Can ChatGPT help patients understand their andrological diseases. Revista Internacional de Andrología. 2024; 22: 14–20.

[14] Şahin MF, Keleş A, Özcan R, Doğan Ç, Topkaç EC, Akgül M, *et al*. Evaluation of information accuracy and clarity: ChatGPT responses to the most frequently asked questions about premature ejaculation. Sexual Medicine. 2024; 12: qfae036.

[15] Aktas B, Demirel D, Celikkaleli F, Bulut S, Ozgur EG, Kizilkan Y, *et al*. YouTube™ as a source of information on prostatitis: a quality and reliability analysis. International Journal of Impotence Research. 2024; 36: 242–247.

[16] Babar M, Loloi J, Patel RD, Singh S, Azhar U, Maria P, *et al*. Cross-sectional and comparative analysis of videos on erectile dysfunction treatment on YouTube and TikTok. Andrologia. 2022; 54: e14392.

[17] Romero-Otero J, Manfredi C, Ralph D, Osmonov D, Verze P, Castiglione F, *et al*. Non-invasive and surgical penile enhancement interventions for

aesthetic or therapeutic purposes: a systematic review. BJU International. 2021; 127: 269–291.

[18] Marra G, Drury A, Tran L, Veale D, Muir GH. Systematic review of surgical and nonsurgical interventions in normal men complaining of small penis size. Sexual Medicine Reviews. 2020; 8: 158–180.

[19] Hehemann MC, Towe M, Huynh LM, El-Khatib FM, Yafi FA. Penile girth enlargement strategies: what's the evidence? Sexual Medicine Reviews. 2019; 7: 535–547.

[20] 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. The American Journal of Gastroenterology. 2007; 102: 2070–2077.

[21] Jindal P, MacDermid JC. Assessing reading levels of health information: uses and limitations of flesch formula. Education for Health. 2017; 30: 84–88.

[22] Kutner M, Greenburg E, Jin Y, Paulsen C. The health literacy of America's adults: results from the 2003 national assessment of adult literacy. (NCES 2006-483). Washington, DC: U.S. Department of Education, National Center for Education Statistics; September 2006. 2006.

[23] Liu S, Wright AP, Patterson BL, Wanderer JP, Turer RW, Nelson SD, *et al*. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. Journal of the American Medical Informatics Association. 2023; 30: 1237–1245.

[24] Miyake Y, Retrosi G, Keijzer R. Artificial intelligence and pediatric surgery: where are we? Pediatric Surgery International. 2025; 41: 19.

[25] Pirkle S, Yang J, Blumberg TJ. Do ChatGPT and Gemini provide apparopriate recommendations for pediatric orthopaedic conditions? Journal of Pediatric Orthopaedics. 2025; 45: e66–e71.

[26] Malak A, Şahin MF. How useful are current chatbots regarding urology patient information? Comparison of the ten most popular chatbots' responses about female urinary incontinence. Journal of Medical Systems. 2024; 48: 102.

[27] Bülbül E, İlki FY. Assessment of the relationship between the quality of YouTube videos on penile enlargement surgery and scholarly profiles of surgeons. Journal of Urological Surgery. 2024; 11: 105–110.

[28] Çağlayan A, Gül M. #Penisenlargement on Instagram: a mixed-methods study. International Journal of Impotence Research. 2024; 36: 218–222.

How to cite this article: Çağatay Özsoy, Erhan Ateş, Mücahit Gelmiş, Abdullah Akdağ. Artificial intelligence applications as a source of information on penis-lengthening and penis girth enhancement: ChatGPT 4o plus *vs.* Gemini advanced. Revista Internacional de Andrología. 2025; 23(3): 51-59. doi: 10.22514/j.androl.2025.030.