ORIGINAL RESEARCH
General Urology

Doi: 10.4274/jus.galenos.2025.2025-9-18
J Urol Surg

Consultation Potential of Artificial Intelligence Chatbots in Prostatitis Management: An Evaluation of Quality, Reliability and Readability

Halil Demirçakan[1], Burak Köseoğlu[2], Taha Numan Yıkılmaz[3]

[1]Çumra State Hospital, Clinic of Urology, Konya, Türkiye
[2]Polatlı Duatepe State Hospital, Clinic of Urology, Ankara, Türkiye
[3]Private Egekent Hospital Hospital, Clinic of Urology, Denizli, Türkiye

What's known on the subject? and What does the study add?

Prostatitis is a common condition with complex management, especially in chronic forms. Patients often seek online information beyond medical consultation. artificial intelligence chatbots are increasingly used in healthcare, but concerns exist regarding the quality, reliability, and readability of their responses. This study is the first to compare ChatGPT-4, Gemini Pro, and Llama 3.1 Large with respect to prostatitis. Gemini Pro provided higher-quality responses, Llama 3.1 Large offered more reliable answers, and ChatGPT-4 demonstrated the highest readability, although all exceeded the recommended reading level.

Abstract

Objective: This study aimed to assess the responses of three artificial intelligence (AI) chatbots (ChatGPT-4, Gemini Pro, Llama 3.1 Large) on prostatitis using quality, reliability, readability scales and examine their role in disease management.

Materials and Methods: Keywords related to prostatitis were identified using Google Trends and Semrush platforms. The search volume and regional distribution of these terms over the past five years were analyzed, leading to the selection of 25 questions. The core question set was categorized into six subgroups: general information, symptoms, diagnostic methods, treatment methods, complications, and myths. The responses generated by the AI chatbots were assessed for readability using the Flesch-Kincaid Grade Level (FKGL) and Flesch Reading Ease (FRES) scores. Quality and reliability were evaluated using the Educational Quality of Information for Patients (EQIP) score and the Modified DISCERN score.

Results: No significant difference was observed among AI chatbots in mean FKGL scores (p=0.354). However, ChatGPT-4 had a significantly higher mean FRES score than Gemini Pro and Llama 3.1 Large (p=0.016 and p=0.003, respectively). Gemini Pro had the highest mean EQIP score, significantly surpassing Llama 3.1 Large and ChatGPT-4 (p<0.001); Llama 3.1 Large had the highest median Modified DISCERN score (p<0.001). Across all subgroup analyses, Gemini Pro yielded the highest mean EQIP score, while Llama 3.1 Large had the highest median Modified DISCERN score.

Conclusion: The findings of this study suggest that Llama 3.1 Large provides more reliable responses to questions about prostatitis, whereas Gemini Pro delivers higher-quality responses. However, the readability levels of all AI chatbots exceeded the recommended 6th-grade level, indicating that their responses may be challenging for general audiences.

Keywords: Prostatitis, artificial intelligence, ChatGPT-4, Gemini Pro, Llama 3.1 large

Introduction

Prostatitis is a condition characterized by symptoms such as chronic pelvic pain, dysuria, and pain during ejaculation. It has a high prevalence, with an overall prevalence of 8% among adult men, making it a common reason for urological consultations (1). Unlike other prostate-related conditions such as benign prostatic hyperplasia and prostate cancer, which are more prevalent in older men, prostatitis affects men of all ages, with a higher incidence in middle-aged men (2). The treatment of prostatitis is determined by the disease type. Acute prostatitis is typically caused by a bacterial infection; since its etiological

Correspondence: Halil Demirçakan MD, Çumra State Hospital, Clinic of Urology, Konya, Türkiye
E-mail: drhalildemircakan@gmail.com **ORCID-ID:** orcid.org/0000-0003-2376-580X
Received: 03.10.2025 **Accepted:** 04.11.2025 **Epub:** 11.06.2026

Cite this article as: Demirçakan H, Köseoğlu B, Yıkılmaz TN. Consultation potential of artificial intelligence chatbots in prostatitis management: an evaluation of quality, reliability and readability. J Urol Surg. [Epub Ahead of Print]

well defined, its treatment follows a standardized, specific antibiotic therapy protocol. In contrast, chronic prostatitis is a multifactorial condition in which, besides infection, neurological, immunological, inflammatory, and psychosocial factors play a significant role. As a result, its treatment must be more individualized and multidisciplinary, making the management process more complex (3). Despite various treatment options, success rates for chronic prostatitis remain low, with a recurrence rate of approximately 50% (4,5). This leads patients to seek alternative sources of information outside conventional medical resources, particularly online sources that are readily accessible (6).

Artificial intelligence (AI) refers to computer systems designed to mimic human intelligence and perform complex tasks such as problem-solving, learning, decision-making, and language processing. AI systems are designed to acquire information from their environment, process the collected data, and make informed decisions (7). AI chatbots have rapidly emerged as valuable tools in healthcare, offering innovative applications ranging from diagnosis and treatment to patient management and optimization of operational processes. AI-based systems stand out due to their ability to provide rapid access to information, analyze patient data, and offer clinical guidance. However, concerns remain regarding the quality, accuracy, and readability of the content generated by AI chatbots (8).

To the best of our knowledge, no study has compared the responses provided by AI chatbots regarding prostatitis. This study aims to evaluate and compare the quality and readability of information on prostatitis provided by different AI chatbots.

Materials and Methods

In this study, responses provided by three AI chatbots (ChatGPT-4, Gemini Pro, and Llama 3.1 Large) to frequently asked questions about prostatitis were evaluated for quality, reliability, and readability. Since the study did not involve human participants and the data were obtained from publicly available sources, ethics committee approval was not required.

Evaluation Parameters

Reliability refers to the degree of accuracy, impartiality, and trustworthiness of the information provided by the chatbots. In this context, the evaluation assessed whether the purpose of the information was clearly stated, whether it was supported by reliable sources, whether the content was presented in a balanced and unbiased manner, whether additional sources of information were provided, and whether the scientific uncertainties or limitations of the available evidence related to the topic were explicitly acknowledged (9).

Quality refers to the educational value, comprehensiveness, and presentation of the content. Rather than focusing solely on factual accuracy, this parameter evaluates how clearly, objectively, and instructively information is presented to patients, and how up to date that information is. It includes elements such as whether the content states its purpose clearly, follows a logical structure, explains treatment options with associated risks and benefits, and provides references to additional sources of information (10).

Readability refers to how easily a text can be understood and followed by the reader. It is measured based on linguistic features such as sentence structure, word length, syntactic complexity, and overall grammatical composition (11).

Data Collection and Analysis

In the first stage, online search trends related to "prostatitis" were determined using Google Trends (http://trends.google.com/) and Semrush (http://www.semrush.com/). The search volume and regional distribution of terms over the past five years were analyzed, and 25 questions were formulated. The core set of questions was categorized into four subgroups: general information (six questions), symptoms and diagnostic methods (eight questions), treatment methods (seven questions), and complications and myths (four questions). The comprehensive list of these 25 queries directed to the AI chatbots is presented in Table 1, while representative chatbot responses are provided in Supplementary Appendix 1. In the second stage, the 25 selected questions were presented to three AI chatbots (ChatGPT-4, Gemini Pro, and Llama 3.1 Large), and responses from each were recorded in Microsoft Word. All prompts were entered on the same date and under identical default configurations for each AI chatbot to ensure methodological consistency and fairness across models.

The collected data were analyzed for word count (WS), sentence count, and syllable count (SYC). To assess text readability, the average Flesch Reading Ease (FRES) score and average Flesch-Kincaid Grade Level (FKGL) score were calculated. The readability level based on the FRES score was classified as follows: "easy" for scores above 70, "moderate" for scores between 60 and 70, and "difficult" for scores below 60 (12). The FRES score is calculated using the following formula:

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

The FKGL score indicates the readability level of a text, with a score of <6 corresponding to an elementary school level and a score of >13 indicating a graduate-level comprehension (13). The FKGL formula is as follows:

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

Table 1. Comprehensive list of queries directed to ChatGPT-4, Gemini Pro, and Llama 3.1 large (by category)
General information
What is prostatitis and who is more likely to have it?
What are the different types of prostatitis and how are they classified?
What are the differences between acute and chronic prostatitis?
What are the causes of chronic prostatitis and how can it be prevented?
Can prostatitis be associated with stress?
What are the current studies on the relationship between the microbiome and prostatitis?
Symptoms and diagnostic methods
What tests are performed for the diagnosis of prostatitis?
What is the importance of microbial analyses in the treatment of chronic prostatitis?
Is the PSA test a reliable method for diagnosing prostatitis?
What imaging methods are used in the diagnosis of prostatitis?
What are the differences between chronic and acute prostatitis, and which symptoms help distinguish them?
Can prostatitis cause sexual dysfunction?
What is the relationship between prostatitis and urinary tract infections?
How does chronic prostatitis affect work and social life?
Treatment methods
What medications are used to treat prostatitis and how do they work?
Which natural supplements may help prevent prostate inflammation?
Is there scientific evidence supporting the use of herbal supplements such as saw palmetto for chronic prostatitis symptoms?
Is prostate massage truly beneficial in the treatment of chronic prostatitis, or is it a weakly supported practice?
Can stress management, meditation, or cognitive behavioral therapy help control chronic prostatitis pain?
How effective are pelvic floor muscle exercises in managing chronic prostatitis?
When is surgical intervention necessary for prostatitis?
Complications and myths
What are the ways to improve quality of life in patients with chronic prostatitis?
Can prostate inflammation turn into cancer?
What urological complications can occur if chronic prostatitis treatment is delayed?
Is chronic prostatitis likely to recur, and what measures can be taken to prevent recurrence in the long term?

The quality, reliability, and adequacy of the content were evaluated using the Patient Education Quality Information (EQIP) score (0-100 points) and the Modified DISCERN score (0-5 points). The EQIP scoring system is based on 20 questions, each of which can be answered as "yes," "partially," "no," or "not applicable." A "yes" response scores 1 point, a "partially" response scores 0.5 points, and a "no" response scores 0 points. The average score is calculated by dividing the sum of points by the number of questions answered. "Not applicable" responses are excluded from this calculation. The resulting value is then multiplied by 100, providing an EQIP score expressed as a percentage (14). An EQIP score of 0-25% indicates "serious quality issues," 26-50% indicates "significant quality issues," 51-75% indicates "good quality with minor issues," and 76-100% indicates "high quality" (15). The Modified DISCERN score comprises five questions, each scored 0 or 1 depending on whether its criterion is met. The scores for both scales were obtained by averaging the evaluations of two independent clinicians who were highly knowledgeable about prostatitis.

Statistical Analysis

The dataset was imported into IBM SPSS Statistics 27. Descriptive statistics were presented as the median [interquartile range (IQR)] for non-parametric data and as the mean ± standard deviation for parametric data. The normality of data distribution was assessed using the Shapiro-Wilk test. One-way ANOVA was used for normally distributed variables, while the Kruskal-Wallis test was applied to non-parametric or ordinal data. In cases where a significant difference was identified, pairwise comparisons were conducted using the Bonferroni correction. Non-parametric methods were prioritized when subgroup sample sizes were small. A p-value of <0.05 was considered statistically significant.

Results

The mean number of sentences per response was 12.72±6.31 for ChatGPT-4, 19.40±5.70 for Gemini Pro, and 11.96±4.92 for Llama 3.1 Large. A statistically significant difference was observed among the AI chatbots regarding the average number of sentences per response (p<0.001). Pairwise comparisons revealed that Gemini Pro had a significantly higher number of sentences than both ChatGPT-4 and Llama 3.1 Large (p<0.001 for both). However, no statistically significant difference was found between ChatGPT-4 and Llama 3.1 Large (p>0.05).

A statistically significant difference was also observed among the AI chatbots in the number of words per response (p<0.001). Pairwise comparisons demonstrated that Gemini Pro had a significantly higher WS than both ChatGPT-4 and Llama 3.1 Large (p<0.001 and p=0.001, respectively). Additionally, Llama 3.1 Large had a significantly higher WS than ChatGPT-4 (p=0.002).

The mean FKGL scores were 22.9±2.9 for ChatGPT-4, 23.7±2.5 for Gemini Pro, and 24.1±2.6 for Llama 3.1 Large, with no statistically significant differences observed among the AI chatbots (p=0.354). The mean FRES scores were 36.2±11.6 for ChatGPT-4, 27.4±11.1 for Gemini Pro, and 25.6±9.4 for Llama 3.1 Large, indicating a statistically significant difference among the AI chatbots (p=0.002). Pairwise comparisons showed that ChatGPT-4 had a significantly higher mean FRES score than Gemini Pro and Llama 3.1 Large (p=0.016 and p=0.003, respectively), while no significant difference was found between Gemini Pro and Llama 3.1 Large (p>0.05).

Mean EQIP scores were 68.2±1.1 for Gemini Pro, 56.7±2.5 for ChatGPT-4, and 57.2±2.0 for Llama 3.1 Large. A statistically significant difference in mean EQIP score was observed among the AI chatbots (p<0.001). Pairwise comparisons indicated that Gemini Pro had a significantly higher mean EQIP score than both Llama 3.1 Large and ChatGPT-4 (p<0.001 for both), while no significant difference was observed between Llama 3.1 Large and ChatGPT-4 (p>0.05).

The median (IQR) Modified DISCERN scores were 2 (1) for Gemini Pro, 2 (1) for ChatGPT-4, and 4 (0) for Llama 3.1 Large (p<0.001). The median Modified DISCERN score was significantly higher for Llama 3.1 Large than for Gemini Pro and ChatGPT-4 (p<0.001 for both comparisons). However, no significant differences were observed among the other AI chatbots (p=0.608; Table 2).

In all subgroup analyses, Gemini Pro had a significantly higher mean EQIP score than both Llama 3.1 Large and ChatGPT-4 (Figure 1), while the median Modified DISCERN score was significantly higher for Llama 3.1 Large than for Gemini Pro and ChatGPT-4 (Figure 2).

Discussion

Chronic prostatitis is a recurrent condition that remains difficult to manage despite the availability of various treatment options (4). This challenge often leads patients to seek alternative sources of health information outside traditional medical channels, such as easily accessible online platforms (6). Therefore, evaluating how accurately and clearly AI chatbots convey health-related information is of clinical importance. To the best of our knowledge, this is the first study to assess the readability, reliability, and quality of AI chatbot responses to frequently asked questions about prostatitis. The findings revealed notable differences among the evaluated models. Specifically, Gemini Pro provided more comprehensive and higher-quality information; Llama 3.1 Large generated the most reliable responses; and ChatGPT-4 achieved the highest readability scores.

Readability is defined in the literature as a fundamental component of health literacy, ensuring the comprehensibility of documents (16). Gül et al. (17) assessed the content provided by three AI chatbots (ChatGPT-4, Bard, and Perplexity) about subdural hematoma and reported that the responses did not meet recommended reading levels and that the average reader would not be able to fully benefit from these sources. In the same study, it was noted that while Google Bard provided more

Table 2. Comparison of the FKRE, FKGL, EQIP and Modified DISCERN scores of the three different AI chatbots

	ChatGPT	Gemini	Llama	p-value
FKRE Mean ± SD	36.2±11.6	27.4±11.1	25.6±9.4	0.002[a]
FKGL Mean ± SD	22.9±2.9	23.7±2.5	24.1±2.6	0.354
EQIP Score Mean ± SD	56.7±2.5	68.2±1.1	57.2±2.0	< 0.001[b]
Modified DISCERN score Median (IQR)	2 (1)	2 (1)	4 (0)	< 0.001[c]

[a]: Difference between ChatGPT and other, [b]: Difference between Gemini and other, [c]: Difference between Llama and other, FKRE: FLesch kincaid reading ease, FKGL: Flesch kincaid grade level, EQIP: Educational quality of information for patients, AI: Artificial intelligence, SD: Standard deviation, IQR: Interquartile range

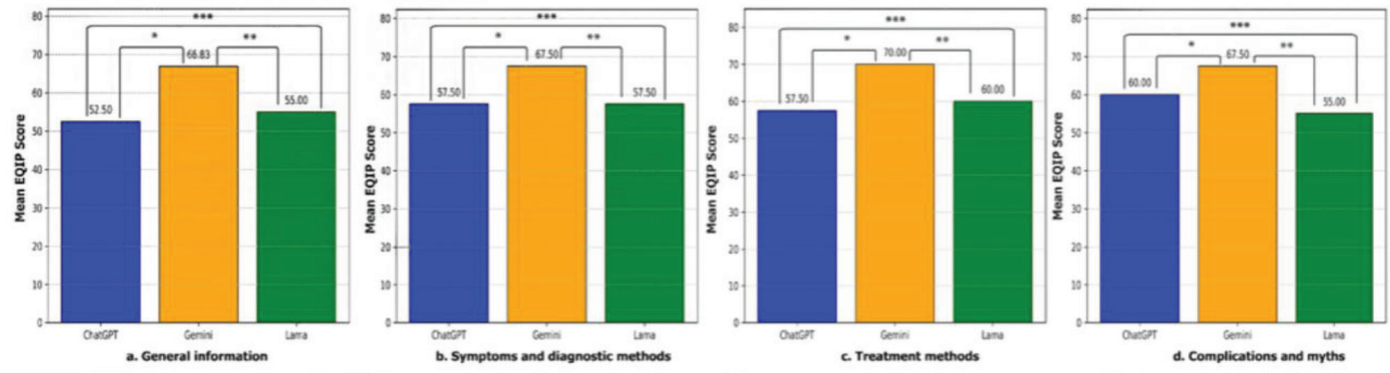

Figure 1. Comparison of the EQIP scores of three different AI chatbots for all subgroups

a. *p<0.001, **p<0.001, ***p=0.189

b. *p<0.001, **p<0.001, ***p>0.05

c. *p<0.001, **p<0.001, ***p<0.001

d. *p=0.005, **p<0.001, ***p=0.046

EQIP: Educational quality of information for patients, AI: Artificial intelligence

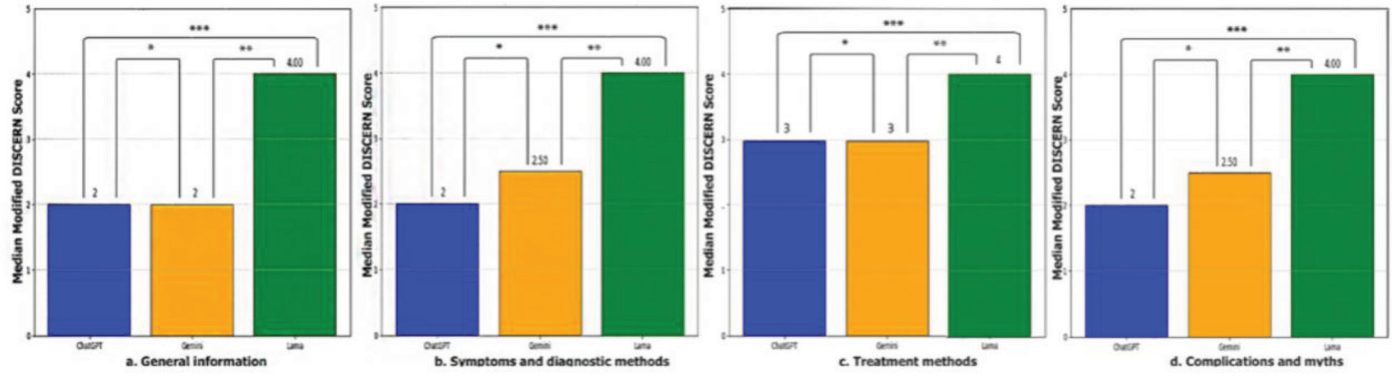

Figure 2. Comparison of the modified DISCERN scores of three different artificial intelligence chatbots for all subgroups

a. *p=0.725, **p=0.001, ***p=0.003

b. *p=0.765, **p=0.001, ***p<0.001

c. *p=0.75, **p=0.001, ***p<0.001

d. *p=0.392, **p=0.032, ***p=0.003

readable responses, ChatGPT-4's responses were more difficult to understand than those of the other chatbots. Similarly, Şahin et al. (18) demonstrated that AI chatbots' responses to questions about erectile dysfunction did not meet readability requirements. In this study, ChatGPT-4 was reported to generate content that was more difficult to understand than content produced by other AI chatbots. Another study examining the readability and quality of responses from five AI chatbots regarding palliative care information found that all five provided content exceeding a sixth-grade reading level. The study also indicated that Gemini Pro generated content that was significantly more difficult to read than content generated by ChatGPT-4 (19). In our study, although the three AI chatbots had similar FKGL scores, ChatGPT-4 had a significantly higher FRES score than the other two, indicating superior readability and comprehensibility. However, all three AI models had FRES scores below 60 and FKGL scores above 6, indicating that their responses exceeded the 6th-grade reading level recommended by the National Institutes of Health and the American Medical Association for health information, which makes them difficult to read (20,21). Poor readability of AI chatbot responses may limit their usability for patients. Additionally, the greater complexity of medical terminology relative to lay language may have contributed to the evaluated AI chatbots' inability to generate comprehensible responses for the general public.

In today's digital landscape, reliable online health information is critically important for patients (22). The inclusion of citations and references is essential for assessing the credibility of health-related information. In a study by Dursun and Bilici Geçer (12) that evaluated responses generated by ChatGPT-3.5, ChatGPT-4, Gemini Pro, and Copilot regarding orthodontic clear aligners, all AI chatbots were reported to produce responses of moderate reliability. The study found that Copilot provided the most reliable responses, followed by Gemini Pro, ChatGPT-4, and ChatGPT-3.5. The lower reliability of ChatGPT-3.5 and ChatGPT-4 compared to other chatbots was attributed to their inability to provide references or citations for the information they generated. In contrast, a study evaluating the recommendations of Bard, BingAI, and ChatGPT-4 regarding melanoma management found that ChatGPT-4 provided more reliable, evidence-based clinical recommendations than Bard and BingAI, achieving higher scores across all reliability scales (23). In our study, responses from Llama 3.1 Large were the most reliable. This was attributed to Llama 3.1 Large's ability to cite scientific publications and its superior organization of textual content. In contrast, Gemini Pro and ChatGPT-4 lacked reference materials and had unclear response origins, resulting in lower reliability compared with Llama 3.1 Large. The variability in reliability findings across studies may be explained by differences in question content and scoring systems.

The EQIP score, developed by healthcare professionals and patient information specialists, is used to assess the quality of health information sources such as websites and patient brochures (14). In a study evaluating the quality of responses provided by ChatGPT-4 on osteoporosis, the mean EQIP score was 48.7, and significant concerns were raised regarding the quality of the generated responses (24). Another study examining ChatGPT-4's responses to frequently asked questions about spinal cord injuries found that the EQIP scores averaged 43.02±6.37, raising concerns about the quality of the responses (25). A study investigating responses from five different chatbots to questions about erectile dysfunction reported EQIP scores as follows: ChatGPT-4 (40.0±4.2), Bing (39.5±3.1), Bard (32.1±30.4), Ernie Bot (53.1±20.6), and Copilot (63.5±12.7). The study concluded that ChatGPT-4, Bing Chat, and Copilot provided "acceptable quality with minor issues," whereas Ernie Bot and Bard exhibited "significant quality issues" (18). In our study, the mean EQIP scores for Gemini Pro, ChatGPT-4, and Llama 3.1 Large were found to be 68.2±1.1, 56.7±2.5, and 57.2±2.0, respectively. The responses provided by all chatbots were categorized as "good quality with minor issues." The variations in mean EQIP scores across studies can be attributed to differences in methodological approaches and evaluators' expertise.

In a study by Durmaz Engin et al. (26), responses of AI models (ChatGPT-4, Bing AI, and Gemini Pro) to categorized questions about retinopathy of prematurity (ROP) were evaluated using the DISCERN and EQIP instruments and three readability scales. All responses from each AI chatbot were analyzed collectively to obtain a single composite score for each model. The study found no significant differences in median scores among the three AI chatbots in the "general information" category, which can be attributed to the standardized and well-established nature of disease definitions. In contrast, ChatGPT-4 outperformed other AI chatbots in the screening, diagnosis, treatment, and prognosis subcategories. The variability in diagnostic criteria and treatment methods, and the impact of these factors on prognosis, were cited as explanations for differences among AI chatbots. In our study, across all subgroups, Llama 3.1 Large provided more reliable responses, while Gemini Pro generated higher-quality responses. The higher reliability scores for Llama 3.1 Large can be explained by its inclusion of citations to scientific publications and the use of evidence-based statements, such as clinical classifications and prevalence data. Gemini Pro, in contrast, presented information more simply, more comprehensibly, and in a patient-oriented manner, avoiding unnecessary technical details. Moreover, by organizing the content in a logical sequence that allows readers to follow the topic step by step, Gemini Pro demonstrated superior content coherence and educational value compared to Llama 3.1 Large. We believe that these differences may also be attributable to structural variations in the models' training data sources and text-generation strategies.

Nevertheless, the classification of the set of questions used in our study into four main categories — general information, symptoms and diagnostic methods, treatment methods, and complications and myths — demonstrates that AI chatbots may play roles at different stages of healthcare delivery. AI chatbots can perform functions such as providing general health information via clear, accessible explanations and supporting patients in recognizing their symptoms and seeking appropriate medical care. Our findings indicate that, at present, AI should be used only as a complementary tool in the education, awareness, diagnosis, and treatment phases of healthcare and cannot substitute for professional medical evaluation or physician examination.

Study Limitations

One of the strengths of our study is that the AI chatbot's responses were evaluated by two clinicians with extensive expertise in prostatitis. However, our study has some notable limitations. First, the study was conducted in English, and AI search results were limited to English. This limitation prevented an evaluation of the quality and readability of responses in other languages. Additionally, each question was posed to the AI chatbots only once, meaning that different responses might have been obtained if the same questions had been posed at different times or repeated. Furthermore, the scores were

obtained by averaging the evaluations of two independent clinicians, but since individual scores were not recorded separately, statistical analysis of interrater agreement could not be performed. Finally, the scoring systems used in this study were not specifically designed for AI chatbots, which may have introduced certain limitations in the evaluations.

Conclusion

AI chatbot models have generally provided high-quality, moderately reliable responses to questions about prostatitis. Among these models, the Gemini Pro model has generated the highest-quality responses, whereas the Llama 3.1 Large model has offered the most reliable responses. A significant limitation of ChatGPT-4 and Gemini Pro is their inability to provide references or citations to support their answers. Additionally, all chatbot models have produced content that exceeds the recommended sixth-grade reading level, which may limit their accessibility to a broader audience. These findings indicate that AI models should be enhanced with more evidence-based information and improved readability. Future research should therefore focus on increasing the accessibility of AI-based information sources and optimizing the role of these models in patient education and disease management.

Ethics

Ethics Committee Approval: Not necessary.

Informed Consent: Not necessary.

Footnotes

Authorship Contributions

Surgical and Medical Practices: H.D., B.K., T.N.Y., Concept: H.D., B.K., T.N.Y., Design: H.D., B.K., T.N.Y., Data Collection or Processing: H.D., B.K., T.N.Y., Analysis or Interpretation: H.D., B.K., T.N.Y., Literature Search: H.D., B.K., T.N.Y., Writing: H.D., B.K., T.N.Y.

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

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

Supplementary Appendix 1. https://d2v96fxpocvxx.cloudfront.net/bb2eeae3-0e60-42a4-acea-81e4a349912c/content-images/d2a8f682-3776-4ec2-83af-17a5b4cc247c.pdf

References

1. Guimaraes CTS, Sauer LJ, Romano RFT, Pacheco EO, Bittencourt LK. Prostate benign diseases. Top Magn Reson Imaging. 2020;29:1-16. [Crossref]
2. Nickel JC, Downey J, Hunter D, Clark J. Prevalence of prostatitis-like symptoms in a population based study using the National Institutes of Health chronic prostatitis symptom index. J Urol. 2001;165:842-845. [Crossref]
3. Magri V, Boltri M, Cai T, Colombo R, Cuzzocrea S, De Visschere P, Giuberti R, Granatieri CM, Latino MA, Larganà G, Leli C, Maierna G, Marchese V, Massa E, Matteelli A, Montanari E, Morgia G, Naber KG, Papadouli V, Perletti G, Rekleiti N, Russo GI, Sensini A, Stamatiou K, Trinchieri A, Wagenlehner FME. Multidisciplinary approach to prostatitis. Arch Ital Urol Androl. 2019;90:227-248. [Crossref]
4. Nickel JC, Shoskes DA, Wagenlehner FM. Management of chronic prostatitis/chronic pelvic pain syndrome (CP/CPPS): the studies, the evidence, and the impact. World J Urol. 2013;31:747-753. [Crossref]
5. Resim S. Akut ve bakteriyel prostatit olgularında tedavi yaklaşımları. Androl Bul. 2020;22:113-123. [Crossref]
6. Aktas BK, Demirel D, Celikkaleli F, et al. YouTube™ as a source of information on prostatitis: a quality and reliability analysis. Int J Impot Res. 2024;36:242-247. [Crossref]
7. Deng J, Lin Y. The benefits and challenges of ChatGPT: an overview. Front Comput Intell Syst. 2023;2:81-83. [Crossref]
8. Xu L, Sanders L, Li K, Chow JCL. Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. JMIR Cancer. 2021;7:e27850. [Crossref]
9. Uzun O. Assessment of reliability and quality of videos on medial epicondylitis shared on YouTube. Cureus. 2023;15:e37250. [Crossref]
10. Hu M, Zou P, Li T, Wang Y. Evaluating the quality of ChatGPT-generated medical information on major ophthalmic conditions: a comparative assessment against the EQIP tool and guidelines. PLoS One. 2025;20:e0334250. [Crossref]
11. Erkin Y, Hanci V, Ozduran E. Evaluating the readability, quality and reliability of online patient education materials on transcutaneuous electrical nerve stimulation (TENS). Medicine (Baltimore). 2023;102:e33529. [Crossref]
12. Dursun D, Bilici Geçer R. Can artificial intelligence models serve as patient information consultants in orthodontics? BMC Med Inform Decis Mak. 2024;24:211. [Crossref]
13. Kincaid JP, Fishburne RP Jr, Rogers RL, Chissom BS. Derivation of new readability formulas: automated readability index, fog count and Flesch reading ease formula. Orlando (FL): Institute for Simulation and Training; 1975. Accessed November 1, 2023. [Crossref]
14. Moult B, Franck LS, Brady H. Ensuring quality information for patients: development and preliminary validation of a new instrument to improve the quality of written health care information. Health Expect. 2004;7:165-175. [Crossref]
15. Hain T. Improving the quality of health information: the contribution of C-H-i-Q. Health Expect. 2002;5:270-273. [Crossref]
16. Meade MJ, Dreyer CW. Orthodontic treatment consent forms: a readability analysis. J Orthod. 2022;49:32-38. [Crossref]
17. Gül Ş, Erdemir İ, Hanci V, Aydoğmuş E, Erkoç YS. How artificial intelligence can provide information about subdural hematoma: assessment of readability, reliability, and quality of ChatGPT, BARD, and perplexity responses. Medicine (Baltimore). 2024;103:e38009. [Crossref]
18. Şahin MF, Ateş H, Keleş A, Özcan R, Doğan Ç, Akgül M, Yazıcı CM. Responses of five different artificial ıntelligence chatbots to the top searched queries about erectile dysfunction: a comparative analysis. J Med Syst. 2024;48:38. [Crossref]
19. Hancı V, Ergün B, Gül Ş, Uzun Ö, Erdemir İ, Hancı FB. Assessment of readability, reliability, and quality of ChatGPT®, BARD®, Gemini®, Copilot®, Perplexity® responses on palliative care. Medicine (Baltimore). 2024;103:e39305. [Crossref]
20. Weis BD. Health literacy: a manual for clinicians. American Medical Association Foundation and American Medical Association; 2003. [Crossref]

21. National Institute of Health. How to write easy-to-read health materials. [Crossref]
22. Whiles BB, Bird VG, Canales BK, DiBianco JM, Terry RS. Caution! AI bot has entered the patient chat: ChatGPT has limitations in providing accurate urologic healthcare advice. Urology. 2023;180:278-284. [Crossref]
23. Mu X, Lim B, Seth I, Xie Y, Cevik J, Sofiadellis F, Hunter-Smith DJ, Rozen WM. Comparison of large language models in management advice for melanoma: Google's AI BARD, BingAI and ChatGPT. Skin Health Dis. 2023;4:e313. [Crossref]
24. Erden Y, Temel MH, Bağcıer F. Artificial intelligence insights into osteoporosis: assessing ChatGPT's information quality and readability. Arch Osteoporos. 2024;19:17. [Crossref]
25. Temel MH, Erden Y, Bağcıer F. Information quality and readability: ChatGPT's responses to the most common questions about spinal cord injury. World Neurosurg. 2024;181:e1138-e1144. [Crossref]
26. 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;11:750. [Crossref]