ORIGINAL RESEARCH
Endourology

Doi: 10.4274/jus.galenos.2025.2025-9-5
J Urol Surg 2026;13(2):90-95

# LLM-based Chatbots for Kidney Stones: A Readability and Quality Assessment

**Ahmet Burak Yılmaz[1], Derya Bakır[2]**

*[1]Sincan Training and Research Hospital, Clinic of Urology, Ankara, Türkiye*
*[2]Atılım University Faculty of Medicine, Department of Internal Medicine, Ankara, Türkiye*

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

Kidney stone disease is highly prevalent, and patients frequently use online resources to seek information. Artificial intelligence–driven chatbots, including large language models (LLMs), are increasingly used as digital health tools. Previous studies in urology have shown that chatbot outputs may align with guidelines in certain contexts but often lack accuracy, completeness, and readability. This study is the first systematic comparison of three widely used LLM-based chatbots (ChatGPT GPT-4o, Google Gemini 2.5 Pro, and DeepSeek R1) for kidney stone-related patient queries. It demonstrates that while quality scores were similar and generally limited to acceptable, readability significantly differed, with ChatGPT requiring higher health literacy than Gemini or DeepSeek. Findings highlight both the potential utility and the current limitations of chatbots in patient education, emphasizing the need for expert oversight and domain-specific refinement.

## Abstract

**Objective:** Kidney stone disease is among the most common urological disorders worldwide. Patients frequently search online for information regarding etiology, management, and prevention; however, the quality and readability of available resources are variable. This study aimed to evaluate and compare the quality and readability of responses generated by three large language model (LLM)-based chatbots—OpenAI GPT-4, Google Gemini 2.5 Pro, and DeepSeek R1—for common patient-oriented kidney stone queries.

**Materials and Methods:** A set of 15 frequently asked questions was curated from online search trends and categorized into three domains: definitions and epidemiology, medical and surgical management, and lifestyle or behavioral aspects. Readability was assessed using Flesch Reading Ease Score (FRES) and Flesch-Kincaid Grade Level (FKGL). Response quality was evaluated with the Ensuring Quality Information for Patients (EQIP) tool and the modified DISCERN instrument. Statistical analyses were performed using the Kruskal-Wallis test with Dunn's post-hoc comparisons.

**Results:** Mean DISCERN and EQIP scores did not significantly differ among platforms, with overall ratings falling in the "limited to acceptable" range. FRES scores were comparable across groups, whereas FKGL revealed significant differences: Gemini responses required a lower educational level than those of ChatGPT ($p<0.016$) and DeepSeek (adjusted $p<0.02$). No differences were observed in word count, sentence count, or total text length.

**Conclusion:** Although all three LLMs generated structured, patient-centered outputs, quality remained modest and readability varied. Some ChatGPT responses demand higher health literacy, potentially limiting accessibility. These findings underscore the need for expert oversight and domain-specific refinement before widespread clinical adoption.

**Keywords:** Endourology, general urology, radiology

**Correspondence:** Ahmet Burak Yılmaz MD, Sincan Training and Research Hospital, Clinic of Urology, Ankara, Türkiye
**E-mail:** abyilmaz05@gmail.com **ORCID-ID:** orcid.org/0000-0001-7269-445X
**Received:** 09.09.2025 **Accepted:** 22.09.2025 **Epub:** 16.03.2026 **Publication Date:** 01.06.2026

**Cite this article as:** Yılmaz AB, Bakır D. LLM-based chatbots for kidney stones: a readability and quality assessment. J Urol Surg. 2026;13(2):90-95.

## Introduction

Kidney stone disease represents one of the most prevalent urological conditions, with a global lifetime risk ranging from 3% to 15% depending on geographic and demographic factors (1). Its recurrent nature and potential complications, including obstruction, infection, and renal impairment, make it a significant public health burden. Patients frequently search online for information about causes, treatment options, prevention strategies, and prognosis. However, online resources often vary in quality, reliability, and readability, raising concerns about misinformation and patient safety (2).

Artificial intelligence (AI)-powered chatbots, especially large language models (LLMs), have emerged as valuable tools for healthcare communication, delivering quick, tailored, and confidential responses to patient inquiries, which may enhance access to health information and alleviate barriers like stigma or embarrassment (3). In urology, studies have demonstrated that chatbot outputs may align with guideline-based recommendations in some contexts, such as urolithiasis, yet inconsistencies and inaccuracies remain (4,5).

Despite the growing interest in AI integration into clinical practice, few investigations have specifically evaluated chatbot performance on kidney stone–related patient queries. Existing evidence has focused on general urological conditions or cancer care, but a systematic analysis of chatbot-generated responses to kidney stone questions remains limited (6). Addressing this gap is clinically relevant given the high prevalence of stone disease and the reliance of patients on digital health platforms for guidance.

## Materials and Methods

### Study Purpose and Design

This cross-sectional, descriptive study was conducted to evaluate and compare the readability and quality of responses generated by three LLM-based chatbots—OpenAI GPT-4, Google Gemini 2.5 Pro, and DeepSeek R1—in addressing common patient-oriented questions about kidney stone disease. As the data consisted solely of machine-generated text with no human subjects, ethical approval or informed consent was not required.

### Construction of Question Set

A curated set of patient-relevant questions was developed through a systematic review of online search behaviors. Search volume data for terms such as "kidney stone," "urolithiasis," and related synonyms were analyzed using Google Trends (Google LLC, USA) and keyword tools, including Semrush (Semrush Inc., USA) and Ahrefs (Ahrefs Pte. Ltd., Singapore), over a 10-year period. Redundant, ambiguous, or overly specific queries were excluded. The resulting set included 15 questions, grouped into three thematic areas:

- Definitions and epidemiology (5 questions)
- Medical and surgical management (5 questions)
- Lifestyle, behavioral, and psychological factors (5 questions)

### Data Collection Process

Questions were submitted in English to the interfaces of the three chatbots (gpt-4o-2024-11-20, gemini-2.5-pro, deepseek-r1), using newly created accounts to eliminate potential bias from prior interactions. Full responses were collected and systematically stored in a standardized "Question-Model-Response" format as plain-text files for analysis.

### Assessment of Readability

For each response, metrics including word count (WC), sentence count (SC), and syllable count (SYC) were calculated using custom Python scripts (Python Software Foundation) and verified in the R statistical environment (R Foundation for Statistical Computing). Two established readability metrics were applied. Flesch Reading Ease Score (FRES) is a 0-100 scale, where higher scores indicate greater ease of understanding. Flesch-Kincaid Grade Level (FKGL) reflects the U.S. educational grade level needed to comprehend the text. These metrics were computed using the formulae (7):

$$FRES = 206.835 - (1.015 \times WC/SC) - (84.6 \times SYC/WC)$$

$$FKGL = (0.39 \times WC/SC) + (11.8 \times SYC/WC) - 15.59$$

### Evaluation of Response Quality

The reliability and accuracy of chatbot outputs were assessed using two validated tools: Ensuring Quality Information for Patients (EQIP): A 20-item scale, with items scored as yes (1), partly (0.5), or no (0). Four quality groups were defined: 0-25% = very poor, 26-50% = limited quality, 51-75% = acceptable but improvable, 76-100% = high quality. The percentage score was calculated as (8):

$$EQIP\ (\%) = [(yes \times 1) + (partly \times 0.5) + (no \times 0)] \div (20 - non\text{-}applicable\ items) \times 100.$$

Modified DISCERN (9,10): A 5-item binary scale (range 0-5), with higher scores denoting greater reliability. Lower scores (0-1) indicate poor or misleading information, 2 = incomplete, 3 = fair, 4 = good, 5 = excellent, and comprehensive.

### Statistical Analysis

SPSS v20 was employed for statistical analysis. Normality was assessed via the Kolmogorov-Smirnov test. Group comparisons and subgroup analyses (Background & Epidemiology, Clinical Management, Lifestyle, & Patient Factors) were performed

using the Kruskal-Wallis test, and followed by Dunn's post-hoc pairwise comparisons with Bonferroni correction to adjust for multiple testing. Two researchers independently evaluated all responses. Inter-rater reliability was measured using weighted kappa statistics for ordinal data and intraclass correlation coefficients (ICC) for continuous data. A p-value <0.05 was considered statistically significant.

## Results

A total of 45 AI-generated responses were evaluated across three platforms (Gemini, DeepSeek, and ChatGPT) for quality and readability metrics (Table 1).

### Quality Metrics

The mean modified DISCERN score was comparable among the groups (Gemini: 1.93±0.8; DeepSeek: 2.0±0.65; ChatGPT: 2.1±0.8), with no statistically significant overall difference (p=0.85). Similarly, EQIP scores showed no significant variation (Gemini: 56±4.6; DeepSeek: 57.7±4.7; ChatGPT: 57.5±4.1), p=0.55. Inter-rater agreement was substantial for both the modified DISCERN score [weighted $\kappa$ = 0.8, 95% confidence interval (CI): 0.750-0.852, p<0.001] and EQIP score (ICC: 0.870, 95% CI: 0.721-0.952, p<0.001).

### Readability Metrics

The FRES indicated slightly easier readability for DeepSeek (42.1±6.4) compared to Gemini (42.2±9.3) and ChatGPT (40.9±8.4), though the overall difference was not significant (p=0.9). By contrast, the FKGL demonstrated significant variability across platforms. Gemini produced texts requiring the lowest educational level (18.2±1.6 grade level), while ChatGPT responses were significantly higher (20.8±1.1 grade level). Post-hoc pairwise analyses revealed a significant difference between Gemini and ChatGPT (p<0.016) and a significant difference between Gemini and DeepSeek (adjusted p<0.02), but there was not a significant difference between DeepSeek and ChatGPT (p=0.66).

### Subgroup Analysis

Subgroup analyses stratified by clinical category revealed no statistically significant differences among the three LLMs for modified DISCERN, EQIP, FKGL, or FRES scores within any category. The relative performance of the models remained consistent across different types of patient-oriented questions, and no category-specific patterns of superiority or inferiority were observed.

### Text Length Metrics

Regarding output length, no significant differences were observed among platforms. The total WCs were 343±101 (Gemini), 358±104 (DeepSeek), and 349±106 (ChatGPT) (p=0.93). Similarly, the total SC was consistent across groups (Gemini: 29±11; DeepSeek: 27±9; ChatGPT: 29±8; p=0.82). Finally, total SYC did not differ significantly (Gemini: 1007±305; DeepSeek: 991±259; ChatGPT: 980±242; p=0.93).

While overall quality scores (modified DISCERN and EQIP) and readability ease (FRES) did not vary significantly, the FKGL revealed that ChatGPT responses required a significantly higher educational level than those of Gemini. No differences were observed in the quantity of generated text across platforms. A comprehensive summary of all findings is presented in Table 2.

**Table 1. List of questions on kidney stones by category**

| Category | Questions |
|---|---|
| General definitions and epidemiology | What is a kidney stone?<br>How common are kidney stones in the general population?<br>What are the main risk factors for kidney stone formation?<br>Do kidney stones have a genetic basis?<br>What is the natural course of kidney stones? |
| Medical and surgical treatments | Which medications are used to relieve pain in kidney stone disease?<br>What is medical expulsive therapy?<br>When is surgical treatment necessary for kidney stones?<br>How effective is shock wave lithotripsy?<br>What are the risks of ureteroscopy and percutaneous nephrolithotomy? |
| Behavioral, lifestyle, and psychological aspects | How much water should I drink to prevent kidney stones?<br>Which foods increase the risk of kidney stones?<br>How high is the risk of recurrence after a kidney stone episode?<br>Do chronic diseases increase the risk of kidney stones?<br>What lifestyle changes can help prevent kidney stones? |

**Table 2. Comparison of quality and readability metrics among AI-generated texts**

| | Gemini | DeepSeek | ChatGPT | p |
|---|---|---|---|---|
| Modified DISCERN score | 1.93±0.8 | 2±0.65 | 2.1±0.8 | 0.850* |
| Ensuring quality information for patients (EQIP) score | 56±4.6 | 57.7±4.7 | 57.5±4.1 | 0.550* |
| Flesch reading ease score (FRES) | 42.2±9.3 | 42.1±6.4 | 40.9±8.4 | 0.900* |
| Flesch-Kincaid grade level (FKGL) | 18.2±1.6 | 20.6±1.4 | 20.8±1.1 | Gemini vs DeepSeek: <0.020*<br>Gemini vs ChatGPT: <0.016*<br>DeepSeek vs ChatGPT: 0.660* |
| Total word count | 343±101 | 358±104 | 349±106 | 0.930* |
| Total sentence count | 29±11 | 27±9 | 29±8 | 0.820* |
| Total syllable count | 1007±305 | 991±259 | 980±242 | 0.930* |

*: Kruskal-Wallis. Variables are presented as mean ± standard deviation. For non-parametric variables, overall group comparisons were performed using the Kruskal-Wallis test followed by Dunn's post-hoc tests with Bonferroni correction for pairwise contrasts. In the table, p-values are presented for both the overall group effect and the relevant pairwise contrasts (Gemini vs. DeepSeek, Gemini vs. ChatGPT, DeepSeek vs. ChatGPT). Bold p-values indicate statistical significance

## Discussion

This study provides one of the first systematic comparisons of LLM-based chatbots in addressing patient-centered questions about kidney stone disease. By analyzing both readability and quality metrics across ChatGPT (GPT-4), Google Gemini 2.5 Pro, and DeepSeek R1, we aimed to evaluate the extent to which these emerging tools can provide accurate, accessible, and clinically useful information for patients.

Our findings demonstrated no significant difference among the three models in terms of quality, as assessed by the modified DISCERN and EQIP tools. Mean scores across platforms were in the "limited to acceptable" range, highlighting that while chatbots are capable of providing structured answers, they often lack the depth and reliability required for complex medical decision-making. This aligns with Cocci et al. (2), who reported that only 52% of ChatGPT's urology-related responses were deemed appropriate, with particularly poor performance in emergency scenarios. Similarly, McCarter et al. (6) found that while ChatGPT outperformed Perplexity and Bing in terms of overall quality and reduced misinformation, completeness of responses remained suboptimal. Collectively, these findings emphasize that although chatbot-generated responses can provide an initial framework of information, they require expert oversight to ensure accuracy and clinical safety.

In terms of readability, our analysis revealed that while FRES scores did not differ significantly, FKGL levels varied considerably, with ChatGPT responses requiring a higher educational level compared to Gemini. This finding is clinically relevant, as health literacy is a major determinant of patient adherence and outcomes. Although our analysis indicated relatively high FKGL scores, it should be emphasized that this index is calculated only from sentence length, WC, and SYC. Because of this formula-based structure, FKGL may sometimes overestimate the true difficulty of chatbot-generated texts. A higher score does not automatically mean that patients would be unable to understand the content; rather, it often reflects longer sentences or more complex word forms. Even so, adopting strategies such as plain-language principles, readability adaptation, and layered summarization could further improve the accessibility of these responses. Prior analyses similarly noted that ChatGPT responses in urology often correspond to a college-level reading difficulty, which may limit accessibility for the average patient (2). Huang and Scotland (11), in their evaluation of UroGPT™, reported that ease of use and patient satisfaction were high, suggesting that domain-specific tailoring of chatbots can help bridge this gap. Future developments should prioritize adaptive outputs that can adjust complexity according to user literacy.

Another important consideration is the degree of alignment with established clinical guidelines. Our study did not specifically assess guideline adherence, but previous literature provides insight. Talyshinskii et al. (4) demonstrated that while GPT-4's outputs often aligned with European Association of Urology guidelines in urolithiasis, critical omissions and inaccuracies were frequent, particularly in metaphylaxis and surgical planning. Likewise, Cil and Dogan (12) reported that although ChatGPT achieved high accuracy in straightforward kidney stone diagnostic scenarios, completeness of responses to more complex queries remained inadequate. These observations highlight the dual reality of LLMs: They can approximate guideline-based answers but are not yet robust enough for unsupervised use in clinical contexts.

From a patient perspective, LLM-driven chatbots offer anonymity, immediacy, and accessibility; features that may reduce stigma and enhance engagement. Studies have shown that patients with kidney stones are interested in interactive technologies, such as apps or chatbots, for improving dietary adherence and fluid intake (11). In fact, AI-based dietary counseling tools have already been piloted for oxalate management, with varying accuracy across platforms-Bard performing best, while GPT

models were less reliable (13). These findings suggest that condition-specific chatbot development may be more effective than relying solely on general-purpose LLMs.

Beyond performance metrics, ethical and practical concerns warrant attention. Chatbots may disseminate misinformation with confidence, potentially undermining patient trust or delaying medical care. Editorial commentary in Translational Andrology and Urology cautioned against over-reliance on LLMs, emphasizing their tendency to produce "confidently incorrect" outputs in urological oncology contexts (14). Moreover, issues of data privacy, informed consent, and medico-legal responsibility remain unresolved. As discussed by Ogbodo et al. (15), integrating AI into the consent process must be balanced with safeguards to ensure comprehension and mitigate liability.

### Future Directions

While our findings highlight current limitations, the trajectory of LLM development is promising. Recent reviews indicate that LLMs are increasingly capable of assisting in diagnosis, patient counseling, and education in urolithiasis, but should currently be regarded as adjuncts rather than replacements for physician input (5). Domain-specific training, multimodal integration (e.g., combining chatbots with imaging or electronic health records), and iterative user feedback may improve performance. Additionally, stratified evaluation frameworks are needed to ensure outputs meet both clinical accuracy and readability thresholds. Another important direction for future research will be to systematically evaluate the concordance of chatbot outputs with established clinical guidelines, such as the European Association of Urology urolithiasis guidelines. Such an approach would provide stronger clinical relevance and allow a more comprehensive assessment of the accuracy and reliability of LLM-generated content.

### Study Limitations

Our study is not without limitations. First, only three chatbots were evaluated, and future iterations of these models may yield different results. Second, we assessed responses in English only, whereas linguistic variability may influence performance. Third, the evaluation of quality and readability, though based on validated tools, cannot fully capture nuanced aspects such as empathy, cultural appropriateness, or dynamic adaptability in conversational settings.

## Conclusion

In conclusion, this study highlights the potential and limitations of LLM-based chatbots in addressing patient-oriented kidney stone disease queries. While all three models produced responses of comparable quality and readability, with no significant differences in modified DISCERN, EQIP, or FRES metrics, ChatGPT required a higher educational level (FKGL) than Gemini, potentially limiting accessibility for patients with lower health literacy. The findings underscore that while these chatbots offer accessible, rapid responses, their outputs often lack the depth and reliability needed for complex clinical scenarios, necessitating expert oversight. Future advancements in domain-specific training, adaptive readability, and guideline alignment are essential to enhance their utility as reliable adjuncts in patient education and urological care.

### Ethics

**Ethics Committee Approval:** Not necessary.

**Informed Consent:** Not necessary.

### Footnotes

### Authorship Contributions

Concept: A.B.Y., D.B., Design: A.B.Y., D.B., Data Collection or Processing: A.B.Y., D.B., Analysis or Interpretation: A.B.Y., D.B., Literature Search: A.B.Y., D.B., Writing: A.B.Y., D.B.

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

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

## References

1. Thongprayoon C, Krambeck AE, Rule AD. Determining the true burden of kidney stone disease. Nat Rev Nephrol. 2020;16:736-746. [Crossref]
2. Cocci A, Pezzoli M, Lo Re M, Russo GI, Asmundo MG, Fode M, Cacciamani G, Cimino S, Minervini A, Durukan E. Quality of information and appropriateness of ChatGPT outputs for urology patients. Prostate Cancer Prostatic Dis. 2024;27:103-108. [Crossref]
3. Shool S, Adimi S, Saboori Amleshi R, Bitaraf E, Golpira R, Tara M. A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Med Inform Decis Mak. 2025;25:117. [Crossref]
4. Talyshinskii A, Juliebø-Jones P, Zeeshan Hameed BM, Naik N, Adhikari K, Zhanbyrbekuly U, Tzelves L, Somani BK. ChatGPT as a clinical decision maker for urolithiasis: compliance with the current European Association of Urology Guidelines. Eur Urol Open Sci. 2024;69:51-62. [Crossref]
5. Ates T, Tamkac N, Sukur IH, Ok F. What is the role of large language models in the management of urolithiasis?: a review. Urolithiasis. 2025;53:92. [Crossref]
6. McCarter J, Applewhite J, Desai S, Hinojosa-Gonzalez D, Badal J. Assessment of artificial intelligence chatbot responses to the top search queries related to kidney stones: a cross-sectional survey study. Journal of Medical Artificial Intelligence. 2025;8. [Crossref]
7. Brewer JC. Measuring text readability using reading level. In book: Advanced Methodologies and Technologies in Modern Education Delivery (pp. 93-103). [Crossref]
8. 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]

9. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health. 1999;53:105-111. [Crossref]

10. Singh AG, Singh S, Singh PP. YouTube for information on rheumatoid arthritis--a wakeup call? J Rheumatol. 2012;39:899-903. [Crossref]

11. Huang M, Scotland K. Usefulness and ease of use of an artificial intelligence chatbot for kidney stone formers. Cureus. 2025;17:e77398. [Crossref]

12. Cil G, Dogan K. The efficacy of artificial intelligence in urology: a detailed analysis of kidney stone-related queries. World J Urol. 2024;42:158. [Crossref]

13. Aiumtrakul N, Thongprayoon C, Arayangkool C, Vo KB, Wannaphut C, Suppadungsuk S, Krisanapan P, Garcia Valencia OA, Qureshi F, Miao J, Cheungpasitporn W. Personalized medicine in urolithiasis: AI chatbot-assisted dietary management of oxalate for kidney stone prevention. J Pers Med. 2024;14:107. [Crossref]

14. Simon BD, Gelikman DG, Turkbey B. Evaluating the efficacy of artificial intelligence chatbots in urological health: insights for urologists on patient interactions with large language models. Transl Androl Urol. 2024;13:879-883. [Crossref]

15. Ogbodo E, Talyshinskii A, Moen CA, Emiliani E, Somani BK, Tzelves L, Beisland C, Juliebø-Jones P. Patient consent in the modern era: novel tools and practical considerations in urology. Curr Urol. 2025;19:235-240. [Crossref]