The New Journal of Urology
eISSN 3023-6940
doi: https://doi.org/10.33719/nju1730282

RESEARCH ARTICLE

Benchmarking Artificial Intelligence Models for Clinical Guidance in Nocturia and Nocturnal Polyuria: A Comparative Evaluation of ChatGPT, Gemini, Copilot, and Perplexity

Gökhan Çeker [1], İsmail Ulus [2], İbrahim Hacıbey [1]

[1] Department of Urology, Başaksehir Çam and Sakura City Hospital, Istanbul, Türkiye

[2] Department of Urology, Bağcılar Training and Research Hospital, Istanbul, Türkiye

Submitted: *2025-06-30*
Accepted: *2025-08-11*

Corresponding Author;
Gökhan Çeker, MD

Address: Başakşehir Mahallesi G-434 Caddesi No: 2L, Başakşehir, Istanbul 34480, Türkiye

E-mail: drgokhanceker@gmail.com

ORCID

G.Ç. 0000-0002-7891-9450
İ.U. 0000-0002-2005-9734
İ.H. 0000-0002-2212-5504

Abstract

Objective: This study aimed to evaluate and compare the performance of four artificial intelligence (AI) models—ChatGPT-4.0, Gemini 1.5 Pro, Copilot, and Perplexity Pro—in answering clinical questions about nocturia and nocturnal polyuria.

Material and Methods: A total of 25 standardized clinical questions were developed across five thematic domains: general understanding, etiology and pathophysiology, diagnostic work-up, management strategies, and special populations. Responses from each AI model were scored by two blinded expert urologists using a five-point Likert scale across five quality domains: relevance, clarity, structure, utility, and factual accuracy. Mean scores were compared using repeated measures ANOVA or Friedman tests depending on data distribution. Inter-rater reliability was measured via the intraclass correlation coefficient (ICC).

Results: ChatGPT-4.0 and Perplexity Pro achieved the highest overall mean scores (4.61/5 and 4.52/5), significantly outperforming Gemini (4.35/5) and Copilot (3.63/5) ($p = 0.032$). ChatGPT scored highest in "general understanding" (4.86/5, $p = 0.018$), while Perplexity led in "management strategies" (4.74/5, $p = 0.021$). Copilot consistently scored lowest, particularly in "diagnostic work-up" (3.42/5, $p = 0.008$). In quality domain analysis, ChatGPT and Perplexity again outperformed others, especially in "factual accuracy" (4.48/5 and 4.44/5), with Copilot trailing (3.54/5, $p = 0.001$). Inter-rater reliability was excellent (ICC = 0.91).

Conclusion: ChatGPT and Perplexity Pro demonstrated strong performance in delivering clinically relevant and accurate information on nocturia and nocturnal polyuria. These findings suggest their potential as supportive tools for education and decision-making. Copilot's lower performance underscores the need for continued model refinement. AI integration in clinical contexts should remain guided by expert validation and alignment with current urological guidelines.

Keywords: artificial intelligence, large language models, nocturia, nocturnal polyuria

Cite; Ceker G, Ulus I, Hacıbey I. Benchmarking Artificial Intelligence Models for Clinical Guidance in Nocturia and Nocturnal Polyuria: A Comparative Evaluation of ChatGPT, Gemini, Copilot, and Perplexity. New J Urol. 2025;20(3):183-192. **doi:** https://doi.org/10.33719/nju1730282

INTRODUCTION

Nocturia and nocturnal polyuria are two of the most common and burdensome lower urinary tract symptoms, particularly in aging populations (1). Their clinical relevance extends beyond sleep disruption, with studies linking them to falls, depression, and cardiovascular morbidity (2,3). While nocturia is easily recognized as a symptom, identifying nocturnal polyuria as an underlying cause often requires quantitative assessment, and this distinction may not always receive adequate attention in routine clinical practice (4).

In parallel with advances in digital health technologies, artificial intelligence (AI)—particularly through large language models (LLMs) such as ChatGPT (5), Gemini (6), Perplexity (7), and Copilot (8) is gaining traction for its potential use in clinical education, patient interaction, and medical decision support. While these AI-powered models demonstrate linguistic fluency and contextual adaptability in general medical domains, their clinical reliability in specialty fields such as urology remains insufficiently evaluated.

Evidence suggests that although LLMs can produce grammatically coherent and context-aware responses, their outputs often vary in factual accuracy and alignment with clinical guidelines. In a comprehensive review, Abd-alrazaq et al. highlighted that current LLMs, despite their pedagogical potential, may propagate misinformation or provide inconsistent recommendations—especially when used without professional oversight in educational or clinical contexts (9). These findings emphasize the importance of careful model evaluation and contextual validation when implementing LLMs in specialty-specific healthcare environments. Given the high prevalence, diagnostic challenges, and clinical significance of nocturia and nocturnal polyuria, these conditions are ideal targets for assessing the performance and practical value of large language models in clinical urology.

While previous benchmarking studies have evaluated LLMs in other urological and medical domains, to our knowledge (10-13), this is the first study to systematically benchmark multiple state-of-the-art LLMs specifically on the clinical topics of nocturia and nocturnal polyuria. Our methodological approach is distinguished by the use of a guideline-driven, thematically structured question set, as well as blinded, domain-expert evaluation, providing new insights into the strengths and limitations of AI models within this under-explored area of urological practice.

The present study aims to systematically evaluate and compare the performance of four state-of-the-art LLMs—ChatGPT-4.0, Gemini 1.5 Pro, Copilot, and Perplexity Pro—on a structured set of questions related to nocturia and nocturnal polyuria. By doing so, we seek to assess their accuracy, consistency, and potential role in clinical urology.

MATERIALS AND METHODS

Study Design

This study was designed as a cross-sectional evaluation of the performance of four LLMs—ChatGPT-4.0 (OpenAI), Gemini 1.5 Pro (Google), Copilot based on GPT-4 (Microsoft), and Perplexity Pro (Perplexity AI)—in providing medical information about nocturia and nocturnal polyuria.

Questionnaire Development

A set of 25 standardized clinical questions was developed based on established international guidelines, including those from the European Association of Urology (EAU) and the International Continence Society (ICS), as well as expert input from urologists and commonly encountered patient queries. For instance, the first two questions were: (1) What is the standard International Continence Society (ICS) definition of nocturia? In addition, (2) How is nocturnal polyuria defined according to the International Continence Society (ICS)? The full list of questions is provided in the Supplementary Material. This approach ensured that the questions comprehensively and accurately reflect current evidence-based practices in the diagnosis and management of nocturia and nocturnal polyuria. The sample size of 25 questions was selected to comprehensively cover all major clinical domains relevant to nocturia and nocturnal polyuria while ensuring the evaluation process remained feasible and manageable for expert reviewers. Although no formal power calculation was performed, this number is consistent with similar benchmarking studies in the literature (14, 15). Two independent urologists were chosen as evaluators to maximize inter-rater reliability. We acknowledge that the sample size and number of raters may limit the statistical power and generalizability of the findings. The questions were systematically divided into five thematic categories:

1. General Understanding
2. Etiology and Pathophysiology
3. Diagnostic Work-Up
4. Management Strategies
5. Special Populations and Research

These categories were selected to encompass both foundational and advanced aspects of the topic, ensuring a broad and structured evaluation of LLMs' performance.

Prompting Methodology

Each of the 25 questions was submitted to the four LLMs (ChatGPT-4.0, Gemini 1.5 Pro, Copilot, and Perplexity Pro) using a standardized prompt format. All questions were entered in English, exactly as worded in the Supplementary Material, with no additional context or preamble. For each model, default settings were used (e.g., temperature, maximum tokens, and model-specific parameters were left at their platform defaults; browsing or enhanced real-time data retrieval was not enabled). Each response was generated in a single turn, and no follow-up clarifications or edits were made to the model output. This approach ensured consistent, unbiased, and reproducible input conditions across all AI platforms.

All large language models were accessed via their official platforms in April 2025, using the latest versions available at that time. For each model, default settings were applied, and features such as web browsing or real-time data retrieval were turned off to ensure standardization across all platforms. Nevertheless, we acknowledge that inherent differences in the models' functionalities and potential platform updates may serve as confounding factors in comparative performance analyses.

Data Collection

Each question was individually submitted to the four selected LLMs during April 2025. For consistency, default settings were used for each model without enabling additional features such as browsing or enhanced real-time data retrieval. All responses were collected in their original form without any modifications.

Evaluation Process

Two independent expert urologists, each with at least five years of clinical experience in managing lower urinary tract symptoms, served as evaluators. For structured evaluation, each LLM-generated response was assessed using a standardized 5-point Likert scale (16) adapted to clinical quality assessment across five quality domains:

- **Relevance:** The extent to which the answer directly addressed the question.
- **Clarity:** The readability and ease of understanding of the response.
- **Structure:** The logical organization and coherence of the information.
- **Utility:** The practical usefulness of the information for clinical or educational purposes.
- **Factual Accuracy:** The accuracy of the information is based on current evidence and clinical guidelines.

The Likert scale was defined as follows:
1 = Poor (inaccurate or irrelevant),
2 = Fair (partially correct but lacking key information),
3 = Satisfactory (generally correct but not well-supported by evidence),
4 = Good (mostly accurate with minor omissions),
5 = Excellent (fully accurate, comprehensive, and aligned with scientific literature).

Scores from both evaluators were averaged to calculate a final domain score per response. To reduce potential bias, evaluators were blinded to each other's ratings and to the identity of the LLM that generated the response.

Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics version 27.0 (IBM Corp., Armonk, NY, USA). Descriptive statistics (mean, standard deviation, and range) were calculated to summarize the Likert scale scores for each evaluation domain across the four LLMs. The normality of data distribution was assessed using the Shapiro–Wilk test. The assumption of normality was met for the General Understanding and Special Populations and Research categories ($p > 0.05$). In contrast, the data for Etiology & Pathophysiology, Diagnostic Work-Up, and Management Strategies significantly deviated from a normal distribution ($p < 0.05$). Accordingly, repeated measures ANOVA was applied to normally distributed data, while the Friedman test was used as a non-parametric alternative for domains that violated

the normality assumption. Post-hoc pairwise comparisons were performed using the Bonferroni correction to control for multiple testing where applicable. Inter-rater reliability between the two expert evaluators was assessed using the intraclass correlation coefficient (ICC). An ICC above 0.75 was interpreted as indicating good agreement, while values above 0.90 were considered excellent (17). All statistical tests were two-sided, and a p-value < 0.05 was considered statistically significant.

Ethical Considerations

This study did not involve human participants, animal subjects and patient data. Therefore, ethical approval was not required in accordance with institutional and national research committee standards. All AI models were accessed through publicly available platforms under their respective terms of use.

RESULTS

Inter-rater reliability between the two expert urologists was excellent, with an ICC of 0.91, indicating strong agreement in scoring.

Overall Performance Across All Questions

Among the four LLMs, ChatGPT achieved the highest overall mean score (4.61 ± 0.32), followed by Perplexity Pro (4.52 ± 0.30) and Gemini (4.35 ± 0.28), while Copilot scored the lowest (3.63 ± 0.45). These differences were statistically significant (p = 0.032) (Table 1, Fig. 1). As an example, in response to the question "At what age-related thresholds is nocturnal urine output considered excessive?", ChatGPT provided a guideline-concordant answer: *"For individuals over 65 years, nocturnal urine output is considered excessive when it exceeds 33% of the total 24-hour urine output. For younger adults, the threshold is 20%."* This response received high scores in relevance, clarity, and factual accuracy. In contrast, Copilot answered: *"For adults over 65 years, nocturnal urine output exceeding 20-33% of the total 24-hour output is considered excessive."* This response was assigned lower scores, as it reflects guideline ambiguity and lacks precise cut-off values.

Performance Across Thematic Categories

LLM performance was further analyzed across five thematic subcategories:

- General Understanding: ChatGPT (4.86 ± 0.21) and Perplexity (4.52 ± 0.29) significantly outperformed Gemini (3.62 ± 0.38) and Copilot (3.58 ± 0.36) (p = 0.018).
- Etiology & Pathophysiology: All models except Copilot performed comparably (ChatGPT: 4.28, Gemini: 4.44, Perplexity: 4.30), while Copilot lagged behind (3.82 ± 0.41) (p = 0.047).
- Diagnostic Work-Up: ChatGPT (4.80 ± 0.27) had the highest performance, followed by Gemini and Perplexity, with Copilot again trailing (3.42 ± 0.48) (p = 0.008).
- Management Strategies: Perplexity (4.74 ± 0.22) slightly outperformed ChatGPT and Gemini, whereas Copilot remained significantly lower (3.70 ± 0.42) (p = 0.021).
- Special Populations & Research: ChatGPT, Gemini, and Perplexity each scored similarly high (~4.56–4.58), while Copilot was significantly lower (3.66 ± 0.43) (p = 0.025).
- Gemini's performance was more variable—comparable to

Table 1. Comparative performance of four AI models across thematic categories related to nocturia and nocturnal polyuria

Topic	ChatGPT	Gemini	Copilot	Perplexity	p-value
FAQs (n=25)	4.61 ± 0.32[a]	4.35 ± 0.28[a]	3.63 ± 0.45[b]	4.52 ± 0.30[a]	0.032
General Understanding	4.86 ± 0.21[a]	3.62 ± 0.38[b]	3.58 ± 0.36[b]	4.52 ± 0.29[a]	0.018
Etiology & Pathophysiology	4.28 ± 0.30[a]	4.44 ± 0.38[a]	3.82 ± 0.41[b]	4.30 ± 0.31[a]	0.047
Diagnostic Work-Up	4.80 ± 0.27[a]	4.64 ± 0.24[a]	3.42 ± 0.48[b]	4.50 ± 0.29[a]	0.008
Management Strategies	4.54 ± 0.33[a]	4.50 ± 0.25[a]	3.70 ± 0.42[b]	4.74 ± 0.22[a]	0.021
Special Populations & Research	4.58 ± 0.29[a]	4.56 ± 0.27[a]	3.66 ± 0.43[b]	4.56 ± 0.26[a]	0.025

Superscript lower-case letters are used to identify statistically significant differences between groups. The same letters (e.g., a-a) indicate no significant difference, while different letters (e.g., a-b) indicate a significant difference (p<0.05).

FAQs: frequently asked questions.

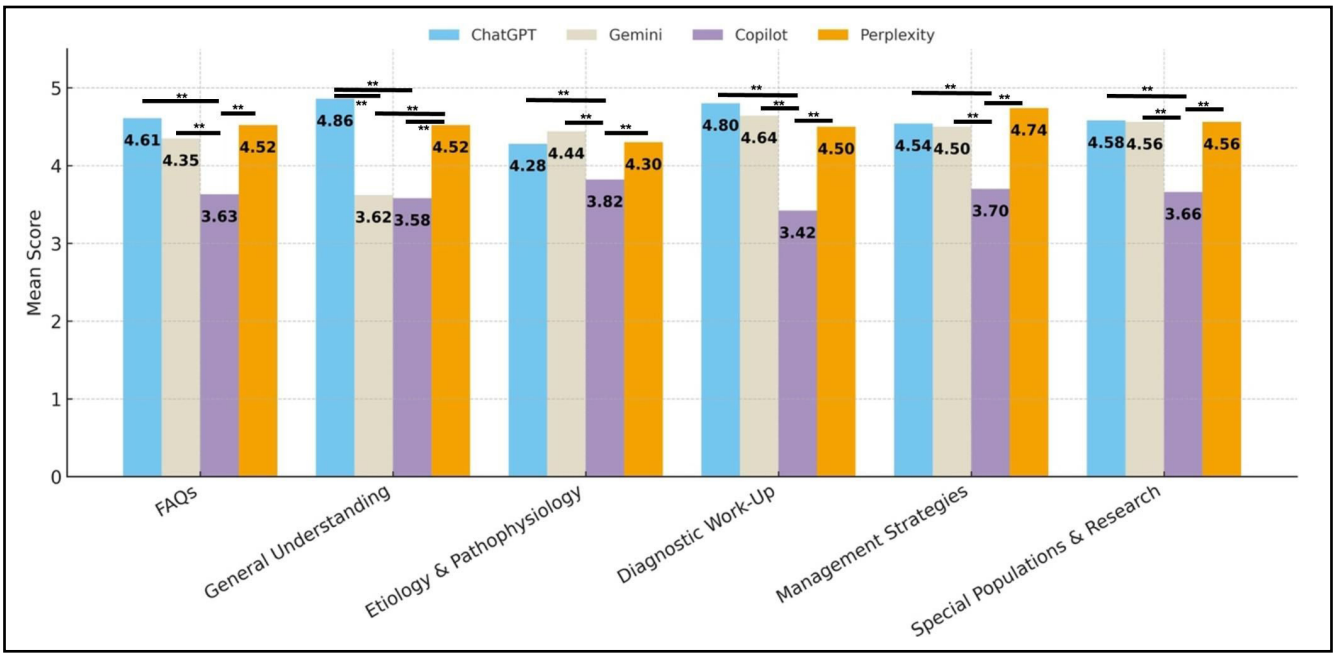

Figure 1. Mean performance scores of four artificial intelligence models across five thematic domains related to nocturia and nocturnal polyuria. FAQs, frequently asked questions. Asterisks indicate statistically significant differences between models ($p < 0.05$).

Table 2. Comparative quality domain scores of AI models in answering clinical questions on nocturia and nocturnal polyuria

Topic	ChatGPT	Gemini	Copilot	Perplexity	p-value
Relevance	4.80 ± 0.26[a]	4.64 ± 0.31[a]	4.08 ± 0.42[b]	4.78 ± 0.24[a]	0.015
Clarity	4.64 ± 0.23[a]	4.38 ± 0.30[a]	3.64 ± 0.39[b]	4.60 ± 0.27[a]	0.012
Structure	4.66 ± 0.25[a]	4.40 ± 0.28[a]	3.60 ± 0.43[b]	4.44 ± 0.26[a]	0.010
Utility	4.48 ± 0.30[a]	4.20 ± 0.29[a]	3.32 ± 0.40[b]	4.36 ± 0.25[a]	0.005
Factual Accuracy	4.48 ± 0.27[a]	4.14 ± 0.32[b]	3.54 ± 0.38[c]	4.44 ± 0.23[a]	0.001

Superscript lower-case letters in the tables (e.g., *a, b, c*) denote statistically distinct groups; values sharing the same letter are not significantly different ($p < 0.05$).

ChatGPT and Perplexity in some categories (e.g., *Etiology & Pathophysiology*), yet significantly lower in others (e.g., *General Understanding* and *Special Populations*). This variability suggests that while Gemini can produce high-quality responses in certain contexts, its consistency remains limited.

Performance Across Quality Domains

Evaluation across the five quality domains revealed consistent patterns of performance superiority by ChatGPT and Perplexity Pro (Table 2, Fig. 2):

- Relevance: ChatGPT (4.80 ± 0.26), Perplexity (4.78 ± 0.24), and Gemini (4.64 ± 0.31) all scored significantly higher than Copilot (4.08 ± 0.42) ($p = 0.015$).
- Clarity: ChatGPT (4.64 ± 0.23) and Perplexity (4.60 ± 0.27) demonstrated excellent clarity, outperforming Gemini and Copilot, the latter scoring significantly lower (3.64 ± 0.39) ($p = 0.012$).
- Structure: Similar trends were observed, with ChatGPT and Perplexity again leading, while Copilot had the lowest structure score (3.60 ± 0.43) ($p = 0.010$).
- Utility: ChatGPT (4.48 ± 0.30) and Perplexity (4.36 ± 0.25) offered the most clinically useful responses, whereas Copilot was substantially weaker (3.32 ± 0.40) ($p = 0.005$).

- Factual Accuracy: The most notable disparities were observed in the factual accuracy domain, where Copilot scored the lowest (3.54 ± 0.38) and each model was assigned a different statistical grouping (a, b, c in Table 2), indicating highly significant differences between all AI models (p = 0.001). Similarly, Copilot's clarity and structure scores were significantly lower, reflecting limitations in presenting responses in a logically organized and easy-to-understand manner.

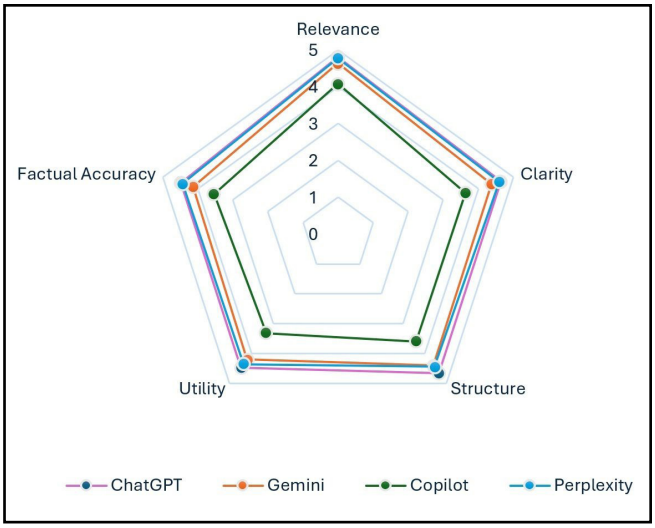

Figure 2. Radar chart illustrating the comparative quality performance of four AI models in answering clinical questions on nocturia and nocturnal polyuria. Higher values reflect better domain-specific performance on a 5-point Likert scale.

DISCUSSION

As generative AI becomes increasingly embedded in clinical informatics, evaluating its reliability in domain-specific contexts such as urology is essential. This study provides a systematic evaluation of four widely used LLMs—ChatGPT-4.0, Gemini 1.5 Pro, Copilot (GPT-4-based), and Perplexity Pro—in the context of nocturia and nocturnal polyuria, two highly prevalent and distressing lower urinary tract conditions frequently encountered in urological practice. While all four models successfully produced responses to expert-formulated clinical questions, their overall performance varied substantially across thematic domains and quality criteria. To the best of our knowledge, this is the first study to systematically evaluate the performance of LLMs in addressing clinical content specifically related to nocturia and nocturnal polyuria.

Consistent with prior research evaluating LLMs in urology-related topics such as urolithiasis management (18), our findings revealed that ChatGPT-4.0 and Perplexity Pro consistently outperformed Gemini and Copilot in key areas such as diagnostic clarity, clinical accuracy, and procedural explanation. In particular, ChatGPT achieved the highest average score across all five evaluation domains—relevance, clarity, structure, utility, and factual accuracy—while Copilot scored the lowest, often failing to provide guideline-based or adequately detailed responses. Gemini performed comparably to ChatGPT and Perplexity Pro in all thematic domains except 'General Understanding', where it scored significantly lower. This suggests that while Gemini's content accuracy is largely consistent, its introductory clarity or foundational summarization may require improvement. This domain-specific inconsistency is critical, given that nocturnal polyuria and nocturia often require nuanced diagnostic differentiation and personalized treatment planning.

These findings reinforce earlier reports in the literature demonstrating ChatGPT's high accuracy in specialty-specific medical contexts. For example, Zhu et al. compared five large language models by posing 22 questions on prostate cancer, and ChatGPT achieved the highest accuracy rate among them (19). Similarly, Caglar et al. found that ChatGPT maintained a guideline adherence rate exceeding 90% in pediatric urology, highlighting its potential in medical education and patient counseling (20). Hacıbey and Halis further supported these results by showing that ChatGPT outperformed other LLMs in addressing clinically relevant questions regarding onabotulinum toxin and sacral neuromodulation (SNM) in the treatment of overactive bladder (15). Consistent with these studies, our evaluation showed that ChatGPT achieved near-perfect scores in the "General Understanding" and "Diagnostic Work-Up" domains.

Interestingly, Gemini exhibited high scores in the "Etiology and Pathophysiology" category, suggesting a potential strength in conceptual reasoning. However, both Gemini and Copilot showed limitations in domains requiring the synthesis of clinical guidelines and nuanced patient-centered reasoning. Copilot consistently scored the lowest across all evaluated domains, with particularly poor performance in factual accuracy and utility. While some of these shortcomings may stem from inherent architectural limitations or reliance

on a general-purpose training corpus, other contributing factors likely include insufficient exposure to domain-specific medical content, lack of clinical fine-tuning, and potential dataset bias. These deficits are particularly critical in clinical communication contexts, where precision, guideline adherence, and applicability are essential. The findings underscore the necessity for future LLMs to be trained on structured, peer-reviewed clinical corpora and to undergo post-hoc validation aligned with specialty-specific standards. Supporting this, a recent evaluation of the Me-LLaMA model demonstrated that LLMs with access to curated clinical datasets significantly outperformed those trained primarily on unfiltered web-based content (21).

From a clinical utility standpoint, these findings carry significant implications. Nocturia and nocturnal polyuria are associated with sleep disturbances, falls, cardiovascular morbidity, and reduced quality of life—especially in older adults (2,3). Providing patients and clinicians with accurate, easily digestible information is essential for safe and effective management.

While LLMs generally demonstrated strong linguistic fluency, our results highlight that this does not always ensure clinical reliability. Copilot and, to a lesser extent, Gemini frequently produced responses lacking clinical precision, especially in diagnostic and management-related areas. Similar concerns have been echoed in recent literature, including studies evaluating AI in radiology (22), oncology (23), and urology (24), where model outputs sometimes conflicted with current standards of care.

Recent studies have demonstrated both the potential and the limitations of AI in clinical urology and broader healthcare. For example, Shah et al. reported that AI models have achieved promising results in the detection and grading of prostate cancer and the prediction of kidney stone composition. However, they cautioned that clinical integration requires large-scale validation and careful management of ethical concerns (25). Similarly, de Hond et al. reviewed the development and validation of AI-based prediction models, emphasizing that many published models lack sufficient external validation and are often built on data that do not fully represent real-world clinical diversity, thereby limiting their generalizability (26). Saraswat et al. further highlighted that the lack of explainability in "black-box" AI models creates barriers to clinical trust, citing specific cases where clinicians were reluctant to accept algorithmic recommendations without clear, interpretable reasoning (27). Our findings resonate with these prior observations: while advanced LLMs such as ChatGPT and Perplexity performed well on structured, guideline-based questions, they were less reliable in nuanced, case-based scenarios—underscoring the continued need for explainable, validated, and context-aware AI tools in clinical practice.

The implications of these findings are particularly relevant in the context of increasing reliance on generative AI for patient counseling, academic learning, and even clinical triage. Although advanced LLMs show promising performance and may serve as supportive tools in clinical education and communication, their use in diagnostic or therapeutic decision-making should be approached with caution (28). Importantly, none of the models evaluated in this study disclosed uncertainty levels or cited peer-reviewed sources—features that are essential for safe clinical integration. Based on these findings, several practical pathways exist for integrating LLMs into clinical and educational workflows in urology. Beyond educational and supportive roles, LLMs could be integrated into real-world urological practice through their deployment in clinical decision support systems, patient-facing triage tools, and automated guideline consultation platforms. For example, AI-powered chatbots could provide initial guidance for patients reporting nocturia symptoms, assist clinicians in reviewing complex cases, or streamline documentation by generating summaries and templated clinical notes. In training programs, LLMs may serve as interactive educational companions, simulating patient scenarios and reinforcing guideline-based reasoning. Successful integration will require rigorous validation, clear scope definition, and ongoing human oversight to ensure patient safety and high-quality care.

In the context of growing clinical reliance on AI, the ethical and regulatory landscape for LLMs remains underdeveloped. Notably, none of the evaluated models provided explicit uncertainty estimates or confidence scores alongside their responses. This lack of "uncertainty calibration" poses a significant risk: users may assume an AI-generated answer is fully reliable, even when the underlying model is uncertain

or operating outside its domain of expertise. Furthermore, the absence of source attribution—meaning the models do not cite peer-reviewed guidelines, original studies, or medical authorities—makes it difficult for clinicians and patients to verify the validity of the information provided. These limitations heighten the risk of misinformation, misinterpretation, and over-reliance on AI in clinical settings. For LLMs to be safely integrated into healthcare, robust frameworks for uncertainty communication, mandatory source citation, and continuous safety oversight by human experts will be essential. Developers and regulatory bodies must prioritize the inclusion of these features to ensure transparency, accountability, and the ethical use of generative AI in medicine.

This study has several strengths. The use of a standardized, thematically organized question set enabled structured comparisons across five clinically relevant domains. Scoring by two blinded expert evaluators ensured high inter-rater reliability (ICC = 0.91), and the multidimensional evaluation system provided a robust and nuanced performance profile for each AI model.

Future research should explore the integration of LLMs into real-time clinical scenarios, comparing AI-assisted versus physician-led decision-making. Additionally, incorporating patient perspectives and evaluating user trust will be essential to determining the acceptability of these technologies in clinical environments. Developers of LLMs should also prioritize embedding up-to-date clinical guidelines, integrating source attribution, and designing models that can flag uncertain or lower-confidence responses.

Study Limitations

This study has several limitations. First, the use of static, one-shot prompting does not reflect dynamic clinical questioning. Second, the models were evaluated without real-world patient interactions and without access to browsing-enabled features, which may limit the depth and currentness of responses. Third, in the context of increasing regulatory scrutiny over generative AI in healthcare (e.g., the EU AI Act), the absence of transparent traceability and confidence calibration mechanisms in LLM outputs remains a critical barrier to clinical adoption (29). In addition, none of the evaluated models provided explicit uncertainty estimates or cited peer-reviewed sources to support their answers. This lack of "uncertainty calibration" and "source attribution" may increase the risk of misinformation and over-reliance on AI-generated content. Until future LLMs can reliably communicate their confidence and directly attribute recommendations to established clinical guidelines, their use in unsupervised clinical decision-making should be approached with extreme caution and subject to ongoing human oversight. Fourth, the relatively limited sample size (25 questions) and the use of only two expert evaluators, although consistent with similar benchmarking studies, may restrict the generalizability of our results and reduce the ability to detect smaller differences between models. Future research involving larger and more diverse question sets, as well as additional expert reviewers, will be important to validate and extend these findings. Although mean scores and standard deviations were reported for ease of interpretation and comparison with previous studies, it should be acknowledged that Likert-type scale data are ordinal in nature. Therefore, medians and interquartile ranges may be more appropriate statistical measures for these data, as they better represent the central tendency and variability without assuming equal intervals between response categories. Future implementations in clinical decision support should include metadata layers that communicate uncertainty and cite sources to align with ethical standards of medical practice.

CONCLUSION

This study highlights that ChatGPT and Perplexity Pro currently represent the most reliable LLMs for generating clinically relevant information about nocturia and nocturnal polyuria. While they may assist in medical education and patient engagement, none of the evaluated models are ready for unsupervised clinical deployment. Their future integration must be supported by rigorous validation, expert oversight, and continuous alignment with updated medical guidelines.

Funding/Financial Disclosure: There was no institutional, commercial, or personal financial funding received for this research.

Presentation Information: The content of this manuscript has not been presented at any scientific meeting, nor has it been published in abstract or full-text form.

Conflict of Interest: The authors confirm that there is no financial or personal conflict of interest related to this study.

Ethical Approval: This study did not involve human participants, animal subjects, or patient data. Therefore, ethical approval was not required in accordance with institutional and national research committee standards. All AI models were accessed through publicly available platforms under their respective terms of use.

Author Contributions: Concept and Design: G.Ç., I.H.; Supervision: G.Ç., I.U., I.H.; Data Collection and/or Processing: G.Ç., I.U., I.H.; Analysis and/or Interpretation: G.Ç., I.U., I.H.; Literature Search: G.Ç., I.H.; Writing: G.Ç., I.U.; Critical Review: G.Ç., I.U., I.H.

REFERENCES

1. Tyagi S, Chancellor MB. Nocturnal polyuria and nocturia. Int Urol Nephrol 2023;55:1395-401. https://doi.org/10.1007/S11255-023-03582-5
2. Weiss JP, Everaert K. Management of Nocturia and Nocturnal Polyuria. Urology 2019;133:24-33. https://doi.org/10.1016/J.UROLOGY.2019.09.022
3. Lavadia AC, Kim JH, Yun SW, Noh T Il. Nocturia, Sleep Quality, and Mortality: A Systematic Review and Meta-Analysis. World J Mens Health 2025;43. https://doi.org/10.5534/WJMH.240237
4. Oelke M, De Wachter S, Drake MJ, Giannantoni A, Kirby M, Orme S, et al. A practical approach to the management of nocturia. Int J Clin Pract 2017;71:e13027. https://doi.org/10.5534/WJMH.240237
5. ChatGPT version 4.0 [Internet]. OpenAI [cited 2025 Apr 18]. Available from: https://chatgpt.com/
6. Gemini 1.5 Pro [Internet]. Google DeepMind [cited 2025 Apr 18]. Available from: https://gemini.google.com/
7. Perplexity Pro [Internet]. Perplexity AI [cited 2025 Apr 18]. Available from: https://www.perplexity.ai/
8. Copilot (GPT-4-based) [Internet]. GitHub [cited 2025 Apr 18]. Available from: https://copilot.microsoft.com/
9. Abd-Alrazaq A, AlSaad R, Alhuwail D, Ahmed A, Healy PM, Latifi S, et al. Large Language Models in Medical Education: Opportunities, Challenges, and Future Directions. JMIR Med Educ 2023;9:e48291. https://doi.org/10.2196/48291
10. Gupta R, Pedraza AM, Gorin MA, Tewari AK. Defining the Role of Large Language Models in Urologic Care and Research. Eur Urol Oncol 2024;7:1–13. https://doi.org/10.1016/j.euo.2023.07.017
11. Song H, Xia Y, Luo Z, Liu H, Song Y, Zeng X, et al. Evaluating the Performance of Different Large Language Models on Health Consultation and Patient Education in Urolithiasis. J Med Syst 2023;47:1–9. https://doi.org/10.1007/S10916-023-02021-3
12. Chervenak J, Lieman H, Blanco-Breindel M, Jindal S. The promise and peril of using a large language model to obtain clinical information: ChatGPT performs strongly as a fertility counseling tool with limitations. Fertil Steril 2023;120:575–83. https://doi.org/10.1016/J.FERTNSTERT.2023.05.151
13. Ferber D, Kather JN. Large Language Models in Uro-oncology. Eur Urol Oncol 2024;7:157–9. https://doi.org/10.1016/j.euo.2023.09.019
14. Şahin B, Genç YE, Doğan K, Şener TE, Şekerci ÇA, Tanıdır Y, et al. Evaluating the Performance of ChatGPT in Urology: A Comparative Study of Knowledge Interpretation and Patient Guidance. J Endourol 2024;38:799–808. https://doi.org/10.1089/END.2023.0413
15. Hacibey I, Halis A. Assessment of artificial intelligence performance in answering questions on onabotulinum toxin and sacral neuromodulation. Investig Clin Urol 2025;66:18893. https://doi.org/10.4111/ICU.20250040
16. Joshi A, Kale S, Chandel S, Pal D. Likert scale: Explored and explained. Br J Appl Sci Technol 2015;BJAST:157. https://doi.org/10.9734/BJAST/2015/14975.
17. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med 2016;15:155-63. https://doi.org/10.1016/J.JCM.2016.02.012
18. Altıntaş E, Ozkent MS, Gül M, Batur AF, Kaynar M, Kılıç Ö, et al. Comparative analysis of artificial

intelligence chatbot recommendations for urolithiasis management: A study of EAU guideline compliance. French J Urol 2024;34:102666. https://doi.org/10.1016/J.FJUROL.2024.102666

19. Zhu L, Mou W, Chen R. Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? J Transl Med 2023;21:1-4. https://doi.org/10.1186/S12967-023-04123-5

20. 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. J Pediatr Urol 2024;20:26.e1-26.e5. https://doi.org/10.1016/J.JPUROL.2023.08.003

21. Xie Q, Chen Q, Chen A, Peng C, Hu Y, Lin F, et al. Medical foundation large language models for comprehensive text analysis and beyond. NPJ Digit Med 2025;8:11-0. https://doi.org/10.1038/S41746-025-01533-1

22. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD, et al. Evaluating ChatGPT as an Adjunct for Radiologic Decision-Making. MedRxiv 2023:2023.02.02.23285399. https://doi.org/10.1101/2023.02.02.23285399

23. Lombardo R, Gallo G, Stira J, Turchi B, Santoro G, Riolo S, et al. Quality of information and appropriateness of Open AI outputs for prostate cancer. Prostate Cancer Prostatic Dis 2024;28:229-31. https://doi.org/10.1038/S41391-024-00789-0

24. Şahin MF, Ateş H, Keleş A, Özcan R, Doğan Ç, Akgül M, et al. Responses of Five Different Artificial Intelligence Chatbots to the Top Searched Queries About Erectile Dysfunction: A Comparative Analysis. J Med Syst 2024;48:1-6. https://doi.org/10.1007/S10916-024-02056-0

25. Shah M, Naik N, Somani BK, Hameed BMZ. Artificial intelligence (AI) in urology-Current use and future directions: An iTRUE study. Turk J Urol 2020;46:S27. https://doi.org/10.5152/TUD.2020.20117

26. de Hond AAH, Leeuwenberg AM, Hooft L, Kant IMJ, Nijman SWJ, van Os HJA, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ Digit Med 2022;5:1–13. https://doi.org/10.1038/S41746-021-00549-7

27. Saraswat D, Bhattacharya P, Verma A, Prasad VK, Tanwar S, Sharma G, et al. Explainable AI for Healthcare 5.0: Opportunities and Challenges. IEEE Access 2022;10:84486–517. https://doi.org/10.1109/ACCESS.2022.3197671

28. 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;307:2023. https://doi.org/10.1148/RADIOL.230163

29. Almada M, Petit N. The EU AI Act: Between the rock of product safety and the hard place of fundamental rights. Common Market Law Review 2025;62:85-120. https://doi.org/10.54648/COLA2025004