ORIGINAL PAPER

UROLOGICAL ONCOLOGY

Accuracy of large language models in answering urological questions on lower urinary tract symptoms: comparison with the EAU 2025 guidelines

Mohamed Eldaneen[1,2], Ahmed Eissa[2], Magdy Sabaa[2], Panagiotis Nikolinakos[1,3], Mohammed Saber-Khalaf[1,4], Karl H. Pang[1,5]

[1]Department of Urology, Chelsea and Westminster Hospital NHS Foundation Trust, London, United Kingdom
[2]Department of Urology, Faculty of Medicine, Tanta University, Egypt
[3]Department of Pediatric Surgery, National and Kapodistrian University of Athens, Greece
[4]Department of Urology, Sohag University Hospital, Sohag University, Sohag, Egypt
[5]Division of Surgery and Interventional Science, University College London, United Kingdom

Citation: Eldaneen M, Eissa A, Sabaa M, et al. Accuracy of large language models in answering urological questions on lower urinary tract symptoms: comparison with the EAU 2025 guidelines. Cent European J Urol 2026; 79: 152-160.

Article history
Submitted: Jan. 9, 2026
Accepted: Feb. 22, 2026
Published online: Mar. 16, 2026

Corresponding author
Karl H. Pang
Chelsea and Westminster Hospital NHS Foundation Trust,
Division of Surgery and Interventional Science,
University College London,
London, United Kingdom
karlpang@doctors.org.uk

Introduction The aim of the study was to evaluate the accuracy of responses from three common artificial intelligence (AI) tools – ChatGPT, Claude, and DeepSeek – to patient enquiries regarding lower urinary tract symptoms (LUTS), comparing these responses to European Association of Urology (EAU) 2025 guidelines. As patients increasingly turn to large language models (LLMs), it is crucial to assess their reliability.

Material and methods Prospective face-to-face and telephone general urology clinics were conducted between April and May 2025, during which consented patients were asked to state the questions they would submit to an AI tool if they were to enquire about their LUTS. These questions were submitted to ChatGPT (GPT-4), Claude (Sonnet 4.0), and DeepSeek (V3), and the responses were summarised and compared against the EAU guidelines. Each response was categorised as correct, missing key elements from the guidelines, or incorrect/misleading.

Results Sixteen patients participated in the study, and following removal of duplicate questions, a total of 13 were included for analysis. These questions covered symptom causation, diagnostic workup and management of LUTS. All models provided correct information in 92.3% of their responses when compared to the EAU guidelines. However, 92.3–100% of answers omitted key elements from the guidelines, and 30.8–92.3% contained incorrect or misleading content.

Conclusions Current LLMs provide readily accessible guidance on LUTS; however, their unsupervised use in clinical decision-making remains a concern and may be considered premature.

Key Words: ChatGPT ‹› DeepSeek ‹› Claude ‹› artificial intelligence ‹› lower urinary tract symptoms

INTRODUCTION

Lower urinary tract symptoms (LUTS) affect both men and women, with a global prevalence of 63.2% [1]. Despite this, approximately two-thirds of individuals with LUTS do not discuss their symptoms with a healthcare professional, often due to embarrassment or the perception that such symptoms are a normal part of ageing [2, 3].

Digital technologies, particularly artificial intelligence (AI), are enhancing medical practice by improving patient communication and diagnostic accuracy [4]. One of the most prominent applications of AI in medicine involves large language models

(LLMs), which generate human-like text in response to user input.

Transformers such as ChatGPT, which can generate contextually relevant medical information, have proven helpful in explaining various urological topics. Other emerging models, including Claude and DeepSeek, aim to provide safe and informative responses [5]. This represents a potentially transformative opportunity for patients who face barriers to traditional pathways for seeking medical advice [6].

A recent health survey reported that 48% of consumers actively use generative AI for health-related enquiries [7]. Another survey of 607 participants found that 78.4% were willing to use ChatGPT for self-diagnosis [8]. However, several challenges remain, including potential declines in patient trust, concerns about AI accuracy, and risks related to data breaches [9, 10].

Therefore, it is crucial to assess how LLMs interpret patient questions related to LUTS. This study aimed to compare responses from GPT-4, Claude (Sonnet 4.0), and DeepSeek (V3) against the current European Association of Urology (EAU) guidelines, evaluating their reliability in addressing patient-style LUTS queries in accordance with evidence-based standards.

MATERIAL AND METHODS

This study was registered with the local Quality and Clinical Governance Department (PCD1237). From 8 April 2025, patients with LUTS attending general urology clinics – either face-to-face or via telephone – were invited to participate in a study in which they would be asked what questions they might submit to an AI tool such as ChatGPT regarding their condition. Inclusion criteria comprised new or follow-up adult patients (aged >18 years), male or female, who presented with LUTS and were able to provide consent.

Patients who gave verbal consent were invited by a single clinician (ME), conducting the clinics, to pose questions related to their condition. These questions were recorded in an Excel spreadsheet, and duplicate entries were removed. The study concluded once 3–6 unique questions had been collected in each of the following thematic categories: 1) symptom causation; 2) diagnostic approaches (both non-invasive and invasive); and 3) medical and surgical treatments, including the safety of interventions and side-effect profiles. The study ended on 8 May 2025.

The final set of questions was submitted to three widely available LLMs:

- Chat Generative Pre-trained Transformer (ChatGPT-4, OpenAI, April 2025 version);
- Claude (Sonnet 4.0, Anthropic, 2025 access);
- DeepSeek (2025 release).

The responses generated by each model were independently reviewed by senior urologists (KHP, AE). Decisions on agreement are summarised in Suppl. Material (suppl. File 1, suppl. Tables 1, 2), and any discrepancies were resolved through discussion. Evaluation was based on compliance with the EAU 2025 guidelines (publicly available online). Depending on the questions posed, the two relevant EAU guidelines reviewed were:

- non-neurogenic female LUTS [11];
- management of non-neurogenic male LUTS [12].

Each answer was assessed using three predefined criteria:

- Whether the response included accurate information consistent with EAU guidance.
- Whether it omitted key content highlighted in the guidelines.
- Whether it contained incorrect or misleading statements.

Analysis

A structured assessment table was used to record binary outcomes (Yes/No) for each criterion, per model, and per question. Any disagreements between reviewers were resolved through consensus discussion. The grading criteria and examples of grading are demonstrated in Suppl. Material (Suppl. File 1, suppl. Tables 1, 2).

The primary outcome was the proportion of responses per model that fully aligned with the EAU 2025 guidelines. Secondary outcomes included the frequency and nature of missed information and incorrect content. Figures are reported as whole numbers, accompanied by percentages and ranges. For continuous data, the mean and range were used to present averages.

Tables and the Figure were created using Microsoft Excel (Microsoft Corporation, 2025, version 16).

RESULTS

Between 8 Apr 2025 and 8 May 2025, five face-face and one telephone clinic were conducted (by ME), attended by 72 patients (45 [62.5%] males, 27 [37.5%] females). Sixteen patients participated in the study (9 [56.25%] males, 7 [43.75%] females). The mean (range) age in years was 61.85 (22–88) overall; male 54.43 (22–76); female 60.5 (41–88). Following removal of duplicate questions, a total of 13 were included for analysis

(Table 1). These questions covered symptom causation, diagnostic workup and management of LUTS. All models provided correct information in 92.3% (n = 12) of their responses when compared to the EAU guidelines. However, 92.3–100% (n = 12–13) of answers omitted key elements from the guidelines. Regarding incorrect information, 30.8% (n = 4) of ChatGPT responses were incorrect, compared to 61.5% (n = 8) for Claude and 92.3% (n = 12) for DeepSeek. The comparative performance of the three AI tools is illustrated in Figure 1. ChatGPT delivered structured and mostly accurate responses that closely adhered to the EAU guidelines. Claude provided structured and clinically sensible answers but frequently omitted key details or included inaccuracies relative to EAU standards. DeepSeek produced the most comprehensive responses in terms of format, but often included outdated, tangential, or guideline-inconsistent information.

Analysis of responses

The summary of AI responses to each patient question, alongside the corresponding EAU 2025 guidelines, is presented in Table 2.

Question 1: Causes of nocturia

All AI models provided good overviews but overlooked nocturnal and 24-hour polyuria, which are key points in the EAU guidelines. DeepSeek was the most comprehensive, Claude included tangential factors such as pregnancy, and ChatGPT offered a well-structured overview.

Question 2: Causes of LUTS

While all models listed the correct causes, they missed critical conditions, such as nocturnal polyuria and chronic pelvic pain syndrome, both of which are highlighted in EAU guidelines.

Question 3: Causes of overactive bladder

All models included relevant factors but failed to exclude conditions such as urinary tract infection and bladder outlet obstruction, which are essential

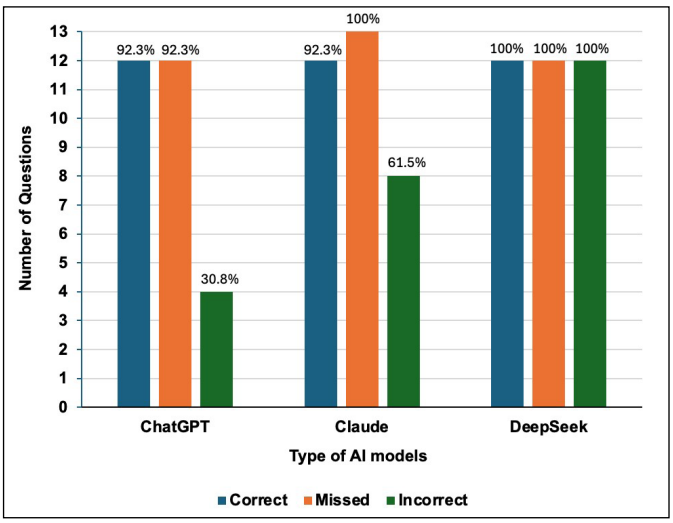

Figure 1. *Performance comparison of AI models (ChatGPT, Claude, and DeepSeek) in responding to LUTS-related patient questions.*

Responses were categorised as correct, missing key contents or incorrect/misleading responses, benchmarked against the 2025 EAU guidelines.
EAU – European Association of Urology; LUTS – lower urinary tract symptoms

Table 1. *Questions patients would submit to an AI tool regarding their symptoms*

Question number	Question category	Patient's question
1	Symptom causes	Why do I wake up at night to urinate (nocturia)?
2	Symptom causes	Why am I experiencing urinary symptoms such as frequent urination, urgency, and a weak stream?
3	Symptom causes	Why do I have an overactive bladder?
4	Diagnostic approach (invasive)	Can a flexible cystoscopy help determine the cause of my urinary symptoms?
5	Diagnostic approach (non-invasive)	What kind of scans or imaging can identify the cause of my urinary symptoms?
6	Diagnostic approach (non-invasive)	How can doctors check my urinary symptoms without doing any invasive tests?
7	Diagnostic approach (invasive)	What are the more invasive tests to investigate my urinary symptoms?
8	Medical treatment	What medicines can help treat my urinary symptoms?
9	Medical treatment	What are the side effects of solifenacin, and how does it work?
10	Conservative treatment	Are there any lifestyle changes or non-medical treatments I can try to manage my urinary symptoms?
11	Surgical treatment	How safe is prostate surgery, such as transurethral resection?
12	Medical treatment	What are the side effects of taking medicines for my urinary symptoms?
13	Surgical treatment	What types of surgery are available for treating urinary or prostate problems?

considerations according to the EAU guidelines. In addition, they lacked a clear definition of overactive bladder.

Question 4: Role of flexible cystoscopy

ChatGPT incorrectly stated that flexible cystoscopy is routinely used for the initial evaluation of LUTS. Claude omitted bladder neck dynamics, while DeepSeek provided the most detailed anatomical insights.

Question 5: Imaging for LUTS

All models correctly identified ultrasound (US) as the initial imaging method but also recommend-

Table 2. *Summary of AI responses to patient questions and comparison with EAU recommendations*

Question	ChatGPT	Claude	DeepSeek	EAU
1. Causes of nocturia	Urological, systemic, behavioural lifestyle factors, and ageing	Medical and lifestyle causes, including stress and pregnancy	Medical, neurological causes	Bladder storage problems; 24-hour polyuria; nocturnal polyuria; sleep disorders; mixed aetiology [12]
2. Causes of LUTS	BPH, OAB, infections, stones, cancer, and neurological conditions. Classified by symptom	Includes medications, BOO, ageing; suggests urinalysis and imaging	Adds neurogenic bladder and diabetes; outlines next steps (PSA, US)	Emphasises missed causes such as chronic pelvic pain syndrome and detrusor underactivity; symptom classification is essential [12]
3. Causes of OAB	Neurological diseases, ageing, BOO, diabetes, and idiopathic conditions	Hormonal and dietary causes; stresses thorough evaluation	Details on detrusor activity, obesity, and diagnostics	Highlights the need to exclude infection and obvious pathology, as defined by the EAU [11]
4. Role of flexible cystoscopy	Not first-line; used for complex or unresolved LUTS	Second-line for haematuria, cancer suspicion, or failed treatments	Detects structural issues; used after initial tests	Not routine; consider for patients with red flag symptoms [12]
5. Imaging for LUTS	US, CT urogram, MRI, cystoscopy	Prioritises US, CT, or MRI for red flags	Prefers US, CT, or MRI as needed	Recommends US first: PVR, prostate size, and hydronephrosis [11, 12]
6. Non-invasive tests	International Prostate Symptom Score, urinalysis, PVR volume, uroflowmetry, and US	Transrectal US, bladder diary, pad weight test, stress test	Medical history, neurological screening, and imaging	Urinalysis, PVR volume, and uroflowmetry should be the initial steps; US is recommended for treatment selection or evaluation of hydronephrosis. PSA and creatinine are optional, based on clinical suspicion or patient age [12]
7. Invasive tests	Cystoscopy, urodynamic studies, biopsy, contrast studies	Includes intravenous pyelogram (outdated)	Adds transrectal US and biopsy, urethral pressure profile	Cystoscopy before minimally invasive or surgical treatment if findings could alter approach; pressure-flow studies when indicated (failed treatment, age extremes, low flow, high PVR volume); retrograde urethrogram or voiding cystourethrogram only if urethral stricture suspected [11, 12]
8. Medical treatment	Alpha-blockers, 5ARi, PDE5i, overactive bladder medications, and oestrogen	Mentions bladder sling, artificial urinary sphincter, botulinum toxin, minimally invasive and surgical options	Lists transrectal US biopsy treatment, emerging therapies (prostatic artery embolisation, Rezum, UroLift, botulinum toxin, sacral neuromodulation)	Watchful waiting for men with mild, non-bothersome symptoms; alpha-blockers for moderate to severe LUTS; 5ARi for men with prostate size >40 ml or high risk of progression; combination therapy for larger prostates with significant symptoms also EAU recommends anticholinergics for storage LUTS if PVR <150 ml and offer combination therapy with mirabegron if monotherapy not effective [12]
9. Solifenacin safety profile	Muscarinic M3 receptor blocker; side effects include dry mouth, urinary retention, visual changes, and cognitive effects	Adds information on drug metabolism and caution in elderly patients	Lists more contraindications; focuses on central nervous system effects	Emphasises cognitive risks, caution in BPH, and the need for careful use of antimuscarinic agents in elderly patients [11, 12]
10. Conservative measures	Fluid management, pelvic floor muscle training, and bladder scheduling	Diet modification and smoking cessation	Broad and accurate interventions include posture, medications, and herbal remedies	Watchful waiting for mild symptoms; encourages lifestyle modifications, bladder training, pelvic floor muscle training; allows combination of conservative measures; percutaneous tibial nerve stimulation can be considered [11, 12]
11. TURP safety	Success rate of 80–90%; risks include bleeding and retrograde ejaculation	Adds mortality and safety statistics; mentions bipolar TURP	Compares laser procedures and risk reduction; retrograde ejaculation "up to 90%"	Indicated for prostates 30–80 ml; retrograde ejaculation (~65–75%), erectile dysfunction (~5–10%), urethral stricture (~2–10%), urinary incontinence (<1%); bipolar TURP reduces transurethral resection syndrome; HoLEP recommended for larger prostates; bladder neck incision for small [12]

Table 2. *Continued*

Question	ChatGPT	Claude	DeepSeek	EAU
12. Medical therapy side effects	Alpha-blockers: dizziness, intraoperative floppy iris syndrome; 5ARi: erectile dysfunction; antimuscarinic agents: cognitive effects; beta-3 adrenergic agonists: hypertension, headache; PDE5i: headache, flushing, hypotension	Notes mood changes with 5ARi and increased burden with combination therapy	Notes mood changes and visual disturbances with PDE5i	Silodosin has fewer hypotensive effects; tamsulosin is associated with floppy iris syndrome. Caution is advised with antimuscarinic agents in older people due to cognitive risks. PDE5i are contraindicated in combination with nitrates [11, 12]
13. Surgical options	TURP as gold standard; HoLEP for large prostates (>80 ml); minimally invasive options include UroLift, Rezum, prostatic artery embolisation; includes radical prostatectomy	Covers traditional (TURP, open surgery) and minimally invasive surgical therapies (UroLift, Rezum, temporary implantable nitinol device, iTind); addresses ejaculatory preservation; mentions advanced techniques such as aquablation	Mentions transurethral needle ablation and transurethral microwave therapy (both obsolete per EAU guidelines)	Classifies treatments by mechanism: TURP, vaporisation, enucleation, ablative, non-ablative, prostatic artery embolisation; treatment choice based on prostate size, bleeding risk, ejaculatory function preservation, patient preference, and comorbidities; does not recommend radical prostatectomy [12]

5ARi – 5-α reductase inhibitor; BPH – benign prostatic hyperplasia; BMI – body mass index; BOO – bladder outlet obstruction; CT – computed tomography; EAU – European Association of Urology; HoLEP – holmium laser enucleation of the prostate; iTind – temporary implantable nitinol device; LUTS – lower urinary tract symptoms; MRI – magnetic resonance imaging; OAB – overactive bladder; PDE5i – phosphodiesterase-5 inhibitors; PVR – postvoid residual; PSA – prostate-specific antigen; Rezum – water vapour thermal therapy; TURP – transurethral resection of the prostate; US – ultrasound; UroLift – prostatic urethral lift

ed non-recommended modalities such as computed tomography (CT) and magnetic resonance imaging (MRI). None of the models mentioned the specific EAU guidelines regarding the use of US to assess prostate volume or postvoid residual (PVR) volume.

Question 6: Non-invasive tests

ChatGPT aligned well with the EAU guidelines but, like the others, failed to stress the appropriate timing of uroflowmetry and US. Claude included unnecessary incontinence-specific tests, while DeepSeek incorrectly categorised cystoscopy and urodynamics as non-invasive.

Question 7: Invasive tests

ChatGPT included appropriate tools such as cystoscopy and urodynamic studies but incorrectly listed prostate biopsy, which is not a standard part of LUTS evaluation. Claude provided the most accurate information aligned with the guidelines. DeepSeek, however, confused LUTS workup with cancer evaluation, mentioning tests such as transrectal US-guided biopsy and the rarely used Whitaker test. None of the models addressed specific EAU guidelines concerning urodynamic studies or the use of retrograde urethrograms and voiding cystourethrograms for suspected urethral stricture. They also overlooked the need for flexible cystoscopy prior to minimally invasive or surgical treatments, an omission that may impact treatment decision-making.

Question 8: Medical treatment

ChatGPT provided mostly accurate information on medication but missed key details such as watchful waiting for men with mild symptoms and the prostate size (≥40 ml) required to initiate 5ARi. Claude summarised treatment options well but included non-medical interventions such as bladder slings, which are not typical for LUTS, and slightly overstated the use of Botox. DeepSeek provided thorough responses regarding age and symptom severity but also overlooked the prostate size threshold and the role of watchful waiting. It incorrectly included prostate cancer tests within LUTS treatment and mentioned newer procedures without clarifying their experimental status. None of the models fully aligned with the EAU guidelines concerning risk assessment or follow-up recommendations.

Question 9: Solifenacin profile

ChatGPT provided a brief overview but lacked detail on contraindications and considerations for elderly patients. Claude offered a more balanced pharmacological profile but still missed some essential information. DeepSeek delivered the most complete and clinically detailed summary.

Question 10: Conservative measures
All three AI models provided adequate information on the conservative management of LUTS. DeepSeek offered practical details but included herbal therapies that are not endorsed by the EAU. However, none of the models mentioned percutaneous tibial nerve stimulation or emphasised the combined conservative approach recommended by the EAU for selected patients. They also omitted the role of watchful waiting, which is an essential first-line strategy for men with mild symptoms.

Question 11: Transurethral resection of the prostate (TURP) safety
ChatGPT accurately reported risks, Claude offered a well-summarised explanation, and DeepSeek provided detailed comparisons but slightly overstated sexual side effects. However, none of the models mentioned key EAU details such as typical prostate size (30–80 ml) for TURP or transurethral incision of the prostate for smaller prostates. They also omitted rates for significant side effects, including retrograde ejaculation (~65–75%) and erectile dysfunction (~5–10%).

Question 12: Medical therapy side effects
ChatGPT provided an overview but missed key safety details, such as silodosin's lower risk of hypotension and tamsulosin's association with floppy iris syndrome. Claude was well structured but overlooked specific risks related to tamsulosin and the nitrate warning associated with phosphodiesterase-5 inhibitors. DeepSeek included various details but neglected silodosin's hypotension risk and tadalafil's contraindications. None of the models fully covered the safety warnings outlined in the EAU guidelines.

Question 13: Surgical options
ChatGPT correctly listed procedures such as TURP, holmium laser enucleation of the prostate (HoLEP), and minimally invasive surgical therapies (MIST) but mistakenly included radical prostatectomy, which is not for LUTS, and overlooked individual treatment considerations. Claude provided a practical summary on functional goals and newer treatments but lacked structured guidance on prostate size and placed undue emphasis on experimental options such as the temporary implantable nitinol device. DeepSeek included outdated methods such as transurethral needle ablation (TUNA) and transurethral microwave thermotherapy (TUMT), did not align with the EAU's surgical framework, and failed to mention that prostate artery embolisation is investigational and not considered first-line therapy. None of the models fully adopted the EAU's approach of personalising procedures to prostate size, bleeding risk, and ejaculation preservation.

DISCUSSION

In this study, we examined the types of questions patients might pose to an AI tool regarding their LUTS and compared the responses from three LLMs – ChatGPT, Claude, and DeepSeek – with the EAU guidelines. While all three tools provided substantial information, they occasionally overlooked key elements of the guidelines or included inaccurate details. This analysis is the first of its kind to evaluate model performance in LUTS interpretation, diagnostic evaluation, and treatment, with a focus on accuracy, adherence to guidelines, and clinical relevance. Although this was not a definitive assessment of these AI tools, the results provide a snapshot of their performance.
The AI tools provide information for patients; however, during clinical consultations, patients must be aware that the information they obtain is for reference only and must not dictate their decision-making. Any decisions on management must be shared between the patient and clinician.
Regarding LUTS aetiology, DeepSeek provided the most detailed responses but occasionally included irrelevant explanations, echoing prior findings on ChatGPT's non-specific pathophysiological reasoning in benign prostate hyperplasia (BPH) cases.[13]
In the diagnostic workup, ChatGPT generally adhered to EAU-recommended non-invasive tools; however, it inaccurately included prostate biopsy, reflecting earlier observations where unnecessary investigations were suggested in other contexts. [14] Claude omitted critical thresholds (e.g., PVR volume >150 ml) that typically guide decisions regarding the need for urodynamic pressure studies.
In treatment recommendations, ChatGPT did not mention conservative management. All models referenced common pharmacological options, but none applied EAU-specific criteria, such as a prostate volume of ≥40 ml for initiating 5-α reductase inhibitors (5ARi) or a PVR of <150 ml to commence anti-cholinergic therapy. These gaps reflect earlier concerns from Caglar et al. [13] about failure to appropriately escalate treatment within BPH management.
Regarding medications, Claude and DeepSeek correctly identified typical side effects but overlooked key distinctions – for example, silodosin's lower risk of orthostatic hypotension in elderly patients.
In surgical management, all models listed standard

procedures, such as TURP, HoLEP, and MIST, but failed to adhere to the EAU's volume-based criteria (e.g., TURP for prostates of 30–80 ml or HoLEP for >80 ml). ChatGPT inaccurately recommended radical prostatectomy for LUTS, reflecting a broader issue of recommending high-risk treatments.

Overall, ChatGPT produced the most structured and generally safe outputs but lacked specificity. Claude was readable and cautious but often superficial. DeepSeek offered the most comprehensive content but was prone to including outdated, tangential, or non-guideline-compliant information. Despite their varying strengths, none of the models demonstrated consistently, guideline-aligned responses across all clinical domains.

Recent studies have evaluated ChatGPT's performance in urology with similar results to our study. Talyshinskii et al. [15] mentioned that GPT-4 handles patient urolithiasis questions well but struggles with specialist-level queries. Its diagnostic accuracy is good, though surgical planning and EAU guideline adherence are weaker. Cikar et al. [14] found that while GPT-4 addressed urolithiasis appropriately, it often lacked guideline specificity. Caglar et al. [13] assessed its performance in BPH and prostate cancer, noting sound reasoning but inappropriate content, such as unnecessary recommendations for radical prostatectomy. A similar study in paediatric urology highlighted strong language fluency, but there were gaps in protocol adherence [16]. While these studies provide insights into ChatGPT's performance in specific domains, they do not compare multiple models or evaluate guideline alignment across LUTS topics.

These findings highlight a key theme: although LLMs can communicate medical reasoning in accessible terms, their adherence to specific guidelines remains inconsistent. As Sallam (2023) noted, the lack of explainability and self-validation against authoritative sources renders current AI tools unreliable for unsupervised clinical use [17]. Unlike validated clinical tools, LLMs lack source citations and transparent reasoning, making verification of their accuracy challenging [18].

Given these limitations, the development of specialty-specific AI tools is essential. The EAU has recently piloted digital engagement tools, including an AI chatbot and interactive guideline summaries. However, our evaluation of the EAU chatbot revealed that responses were often brief and redirected users to full documents rather than providing context-aware guidance [19].

This study has several limitations. It analysed a selection of representative questions derived from patients, providing only a snapshot comparison of accuracy against the EAU guidelines. The number of questions determined was arbitrary, but we ensured that these represented the full spectrum of LUTS-related domains in the EAU guidelines. In addition, since there are limited studies published on this topic, the number of questions chosen was based on the study published by Talyshinskii et al. on urolithiasis (11 questions) [15]. Moreover, the limited sample may restrict generalisability and statistical robustness.

Assessment of the AI responses was performed by two urologists. The decisions made were subjective and hence subject to bias. However, any discrepancies were discussed among the assessors.

The use of patient-style questions does not fully replicate real-world consultations, and subjective judgment in assessing guideline adherence could introduce bias, despite evaluations being based on explicit EAU standards. Furthermore, the performance of LLMs may vary over time due to updates and retraining, affecting reproducibility in the absence of strict version control. Categorising AI responses as correct, missing, or incorrect when comparing to EAU may oversimplify performance, particularly in cases where answers are partially correct yet incomplete or unclearly phrased. These responses may still convey misleading reassurance or omit clinically important nuances, which could influence patient understanding or decision-making. Lastly, responses were evaluated solely in English, potentially overlooking disparities in multilingual performance [20].

The analysis represented a temporal snapshot based on AI model performance at the time of data collection. As LLMs are continuously updated and retrained, their accuracy and alignment with guidelines may change over time. Therefore, the findings should be interpreted as specific to the models' versions tested, highlighting the need for periodic benchmarking of AI outputs.

Although the questions originated from real patient enquiries, the study did not evaluate patients' perceptions, comprehension, or satisfaction with the AI-generated responses. Incorporating patient-centred metrics in future research will be essential to assess the practical utility and readability of AI-provided health information.

CONCLUSIONS

The snapshot results indicate that while LLMs such as ChatGPT, Claude, and DeepSeek provide accessible communication, they frequently fail to deliver consistent guideline-concordant advice for LUTS and its management. Their use as standalone tools

in urological care is premature, especially for nuanced clinical decisions.
Nevertheless, they may serve as valuable adjuncts for patient education with appropriate safeguards. These findings underscore the need for continuous benchmarking against specialty guidelines and the development of clinically validated AI support systems.

CONFLICTS OF INTEREST
The authors declare no conflict of interest.

FUNDING
This research received no external funding.

ETHICS APPROVAL STATEMENT
The ethical approval was not required.

SUPPLEMENTARY MATERIALS

Suppl. File 1. Evaluation criteria and examples

Definition of terms
Complete: The AI response included all relevant points covered by the 2025 EAU guideline for that topic.
Missing: The response omitted one or more essential elements described in the guideline.
Incorrect: The response included information that contradicted EAU guideline recommendations or introduced non-evidence-based or misleading statements.

These criteria were applied consistently across all responses by two independent reviewers with discrepancies resolved through discussion.

Examples

Suppl. Table 1. *Example 1 – diagnostic workup (non-invasive): How to investigate LUTS symptoms by non-invasive tests*

Source	Correct information	Incorrect information	Missed EAU guidance
ChatGPT	– Urinalysis, PVR, ultrasound, and uroflowmetry – PSA/creatinine listed as optional based on age/suspicion – Identifies invasive tests (e.g., cystoscopy, urodynamics) as second-line – Includes IPSS, bladder diary	NO	– Could state more clearly that USS and uroflowmetry are required before initiating treatment, as per EAU
Claude	– PSA, urinalysis, ultrasound, and appropriately includes TRUS – Lists history, IPSS, PVR, uroflowmetry – Symptom-specific functional tools (pad test, bladder diary)	– Pad weight test and stress test are less relevant unless evaluating incontinence specifically, which is not universal to all LUTS cases	– No mention of uroflowmetry being needed before medical or surgical therapy, as stated in EAU – Omits guidance on timing of USS use
DeepSeek	– Comprehensive: DRE, urinalysis, PVR, ultrasound, PSA, creatinine, medication review – IPSS, bladder diary	– Lists urodynamics and cystoscopy within general non-invasive section, though these are not non-invasive nor part of initial evaluation per EAU	– No clear emphasis on sequencing of tests before treatment – Does not flag USS as necessary before BPH surgical treatment (per EAU)

BPH – benign prostate hyperplasia; DRE – digital rectal examination; EAU – European Association of Urology; IPSS – International Prostate Symptom Score; LUTS – lower urinary tract symptoms; PSA – prostate specific antigen; PVR – postvoid residual volume; TRUS – transrectal ultrasound; USS – urethral stricture score

Suppl. Table 1. *Example 2 – reasons for LUTS*

Source	Correct information	Incorrect information	Missed EAU information
ChatGPT	– Identifies BPH, BOO, UTI, bladder stones, OAB, neurological causes, and bladder cancer – Categorizes symptoms: storage, voiding, post-micturition – Lists appropriate diagnostics: urine tests, ultrasound, uroflowmetry		– Omits nocturnal polyuria, CPPS, foreign bodies, and detrusor underactivity
Claude	– Includes BPH, BOO, UTI, OAB, bladder stones, prostatitis, neurological issues – Mentions medications, which can influence LUTS – Recommends urinalysis, imaging, and urodynamic studies		– Omits nocturnal polyuria, CPPS, foreign bodies, and detrusor underactivity
DeepSeek	– Most comprehensive: includes BPH, UTI, BOO, OAB, neurological causes, diabetes, meds, interstitial cystitis, tumours – Differentiates male/female factors – Describes evaluations well: PVR, DRE, PSA, urodynamics		– Miss detrusor underactivity, CPPS, foreign bodies, and nocturnal polyuria

BOO – bladder outlet obstruction; BPH – benign prostate hyperplasia; CPPS – chronic pelvic pain syndrome; DRE – digital rectal examination; EAU – European Association of Urology; LUTS – lower urinary tract symptoms; OAB – overactive bladder; PSA – prostate specific antigen; PVR – postvoid residual volume; UTI – urinary tract infection

References

1. Huang J, Chan CK, Yee S, et al. Global burden and temporal trends of lower urinary tract symptoms: a systematic review and meta-analysis. Prostate Cancer Prostatic Dis. 2023; 26: 421-428.

2. Coyne KS, Barsdorf AI, Thompson C, et al. Moving towards a comprehensive assessment of lower urinary tract symptoms (LUTS). Neurourol Urodyn. 2012; 31: 448-454.

3. Welch LC, Taubenberger S, Tennstedt SL. Patients' experiences of seeking health care for lower urinary tract symptoms. Res Nurs Health. 2011; 34: 496-507.

4. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med. 2022; 28: 31-38.

5. Kim SH, Tae JH, Chang IH, et al. Changes in patient perceptions regarding ChatGPT-written explanations on lifestyle modifications for preventing urolithiasis recurrence. Digit Health. 2023; 9: 20552076231203940.

6. Onyeaka HK, Romero P, Healy BC, Celano CM. Age Differences in the Use of Health Information Technology Among Adults in the United States: An Analysis of the Health Information National Trends Survey. J Aging Health. 2021; 33: 147-154.

7. Deloitte. Can GenAI Help Make Health Care Affordable? Consumers Think So. 2023 [cited 2025]. Available at: https://www.deloitte.com/us/en/Industries/life-sciences-health-care/blogs/health-care/can-gen-ai-help-make-health-care-affordable-consumers-think-so.html.

8. Shahsavar Y, Choudhury A. User Intentions to Use ChatGPT for Self-Diagnosis and Health-Related Purposes: Cross-sectional Survey Study. JMIR Hum Factors. 2023; 10: e47564.

9. Smith H. Clinical AI: opacity, accountability, responsibility and liability. Ai & Soc. 2021; 36: 535-545.

10. Moy S, Irannejad M, Manning SJ, et al. Patient Perspectives on the Use of Artificial Intelligence in Health Care: A Scoping Review. J Patient Cent Res Rev. 2024; 11: 51-62.

11. Urology, E.A.o. EAU Guidelines on Non-Neurogenic Female Lower Urinary Tract Symptoms. 2023 3 July 2025]; Available at: https://uroweb.org/guidelines/non-neurogenic-female-luts.

12. Urology, E.A.o. EAU Guidelines on Management of Non-Neurogenic Male LUTS. 2024 [cited 3 July 2025; Available at: https://uroweb.org/guidelines/management-of-non-neurogenic-male-luts.

13. Caglar U, Yildiz O, Meric A, et al. Evaluating the performance of ChatGPT in answering questions related to benign prostate hyperplasia and prostate cancer. Minerva Urol Nephrol. 2023; 75: 729-733.

14. Cakir H, Caglar U, Yildiz O, Meric A, Ayranci A, Ozgor F. Evaluating the performance of ChatGPT in answering questions related to urolithiasis. Int Urol Nephrol. 2024; 56: 17-21.

15. Talyshinskii A, Juliebø-Jones P, Zeeshan Hameed BM, et al. 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.

16. Caglar U, Yildiz O, Meric A, et al. Evaluating the performance of ChatGPT in answering questions related to pediatric urology. J Pediatr Urol. 2024; 20: 26.e1-26.e5.

17. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel). 2023; 11: 887.

18. Marey A, Arjmand P, Alerab ADS, et al. Explainability, transparency and black box challenges of AI in radiology: impact on patient care in cardiovascular radiology. Egypt J Radiol Nucl Med. 2024; 55: 183.

19. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023; 6: 1169595.

20. Naik N, Hameed BMZ, Shetty DK, et al. Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility? Front Surg. 2022; 9: 862322.