

Article

Assessing ChatGPT Accuracy Across Versions for Patient and Guideline Queries in Sacral Neuromodulation

Kirolos Eskandar 

Faculty of Medicine and Surgery, Helwan University, Cairo 11211, Egypt; kirolos.eskandar@gmail.com or kirolos210575@med.helwan.edu.eg; Tel.: +49-1775965567

Abstract

Background/Objectives: Sacral neuromodulation (SNM) is an established therapy for refractory overactive bladder and non-obstructive urinary retention. With the rapid adoption of large language models (LLMs) such as ChatGPT, their accuracy in procedure-specific domains requires evaluation. The aim of this study was to compare the accuracy, completeness, and reproducibility of ChatGPT versions 3.5, 4.0, and 5.0 in answering patient- and guideline-based questions on SNM. **Methods:** Twenty questions were developed from international guidelines, device information, and common patient inquiries, covering five domains (mechanism, technique, outcomes, complications, postoperative management), two source types (frequently asked question [FAQs] vs. guideline), and three difficulty levels. These thematic domains were derived from core clinical counseling areas routinely addressed in SNM evaluation and follow-up. Each was submitted to ChatGPT versions 3.5, 4.0, and 5.0. Responses were rated independently by two urologists on a four-point accuracy scale. Combined success (Grades 1–2) and accuracy trends were compared across versions. Chi-square tests were used to assess differences across versions, Cramer's V to measure effect size, and Cohen's kappa to evaluate reproducibility. **Results:** Accuracy improved progressively across versions. Combined success rates rose from 70% in version 3.5 to 85% in 4.0 and 90% in 5.0 ($p = 0.031$, Cramer's V = 0.29). Highest accuracy was observed in mechanism and procedural technique, while complication- and guideline-based questions showed lower performance. FAQ and straightforward questions were answered more reliably than guideline-based or complex ones. Reproducibility was excellent across all versions ($\kappa = 0.81$ – 0.91). **Conclusions:** ChatGPT 4.0 and 5.0 show strong potential as adjunctive tools for patient education in SNM, particularly for FAQs and procedural explanations. However, because persistent limitations were observed in guideline interpretation and complication management, clinician oversight remains essential, and these models should not be regarded as substitutes for professional clinical judgment.

Keywords: patient education; clinical decision support; artificial intelligence in urology; health communication; medical guidelines



Academic Editor: Henry Woo

Received: 21 September 2025

Revised: 2 December 2025

Accepted: 5 January 2026

Published: 12 February 2026

Copyright: © 2026 by the author.

Published by MDPI on behalf of the

Société Internationale d'Urologie.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Overactive bladder (OAB), urgency urinary incontinence, and chronic non-obstructive urinary retention are prevalent and burdensome conditions, often refractory to first- and second-line therapies. For appropriately selected patients, sacral neuromodulation (SNM, e.g., InterStim®) is recommended as an effective minimally invasive option [1]. Major urological guidelines describe SNM as a well-established therapy with durable symptom improvement, with responder rates of approximately 70–80% in contemporary studies.

Although device-related adverse events are generally modest, proper patient selection, counseling, and adherence to guideline-based practice remain critical [2,3].

Large language models (LLMs), particularly OpenAI's ChatGPT series, have been rapidly adopted since 2022 by patients, educators, and clinicians. This trend is supported by recent literature showing increasing integration of LLMs into clinical workflows, patient education activities, and medical training environments [2,4]. Potential applications include generating patient education materials, answering frequently asked questions (FAQs), supporting medical training, and assisting with documentation [5]. Comparative analyses demonstrate that newer versions, especially GPT-4, outperform GPT-3.5 on standardized medical knowledge tests, although limitations such as factual errors, inconsistent reasoning, and reliance on outdated training data remain important concerns [6]. Specialty-focused evaluations have suggested that ChatGPT can produce useful patient-directed content in fields such as oncology and orthopedics, but variability in readability, accuracy, and guideline concordance underscores the need for careful validation before clinical integration [7]. Moreover, evolving digital media platforms are increasingly used to disseminate procedural medical knowledge, with studies identifying substantial variability in accuracy and reliability of content, particularly on publicly accessible video platforms [8]. This highlights the broader need to critically evaluate artificial intelligence (AI)-generated procedural information.

Despite the growing body of literature on ChatGPT in medicine, few studies have explored its role in procedural urology. Existing work has assessed chatbot performance in answering questions related to therapies such as onabotulinum toxin and SNM, but comparisons across multiple ChatGPT versions and across patient- versus guideline-based question formats are lacking [9,10]. Notably, while Hacıbey and Halis (2025) [10] evaluated chatbot responses to questions on both Botox and SNM, their analysis did not investigate performance differences across ChatGPT versions nor assess reproducibility through repeated querying. Thus, the extent to which version progression influences accuracy and consistency remains unclear. In parallel, emerging studies comparing other large language models—such as Claude, Gemini, and Med-PaLM—suggest variability in urological knowledge and adherence to guideline-based responses [11,12], further motivating performance benchmarking of specific model families.

The thematic domains used in the present study—mechanism, technique, outcomes, complications, and postoperative management—reflect the major content areas that typically guide SNM counseling, device education, and postoperative follow-up in clinical practice. These domains represent the core clusters of information routinely addressed during patient assessment and shared decision-making.

The aim of this study is to evaluate and compare ChatGPT versions 3.5, 4.0, and 5.0 in answering patient- and guideline-based questions on SNM therapy, with a focus on domain-specific accuracy, question difficulty, and reproducibility.

2. Methods

2.1. Study Design

This was a comparative evaluation study designed to assess the accuracy, completeness, and reproducibility of ChatGPT responses regarding SNM therapy for refractory overactive bladder and non-obstructive urinary retention. The study compared three versions of ChatGPT—3.5, 4.0, and 5.0—using a structured set of patient- and guideline-based questions. No human or animal participants were involved, and no identifiable patient data were used. All information was derived from publicly available sources and AI-generated outputs; therefore, according to local and international guidelines, formal approval from an institutional review board or ethics committee was not required. The four-point accuracy

scale used in this study was adapted from prior ChatGPT medical benchmarking frameworks [13–15] but has not undergone formal psychometric validation; this limitation is acknowledged both here and in the Supplementary Materials.

2.2. Question Development

A total of 20 questions was developed to cover key aspects of SNM therapy.

2.2.1. Sources

Patient FAQs were collected from online support groups, forums, and patient education websites (e.g., Inspire, Reddit OAB community).

Guideline-based questions were derived from international recommendations, including the American Urological Association (AUA), European Association of Urology (EAU), and UK National Institute for Health and Care Excellence (NICE) guidelines.

Device manufacturer resources (Medtronic InterStim®, Minneapolis, MN, USA, Axonics®, Irvine, CA, USA) were reviewed for technical details commonly sought by patients.

2.2.2. Identification of Patient FAQ Domains

FAQ themes were extracted by reviewing recurrent patient questions across multiple platforms and clustering them into conceptual categories (mechanism, procedure, outcomes, risks, and postoperative management). These categories reflect the core counseling domains routinely addressed during SNM evaluation and follow-up in clinical practice.

2.2.3. FAQ Prioritization

From an initial pool of 50 FAQ and guideline-derived questions, items were prioritized based on (a) frequency of appearance across platforms, (b) clinical relevance as determined by two urologists, and (c) representation of the five major thematic domains. Overly general, redundant, or patient-specific questions were excluded.

2.2.4. Selection of Guideline-Based Questions

Guideline questions were selected by reviewing the 2024 AUA/Society of Urodynamic, Female Pelvic Medicine & Urogenital Reconstruction (SUFU), 2024 EAU, and 2015 NICE SNM guidance [3,5] documents and identifying recommendations that required interpretation, synthesis, or risk stratification. Questions were chosen by one investigator and independently confirmed by two urologists using predefined criteria: (a) direct linkage to published guideline statements, (b) coverage of indications, contraindications, risks, and follow-up, and (c) representation across all stages of SNM therapy.

Final Selection: From an initial pool of 50 candidate questions, duplicates, overly general items, and patient-specific queries were excluded, leaving 20 unique questions.

Validation: The final 20 questions were independently reviewed and validated by two board-certified urologists prior to testing to ensure clinical accuracy, guideline relevance, and appropriate difficulty stratification.

Thematic Domains: Questions were organized into five categories for domain-specific analysis:

Mechanism & Indications (Q1–4)

Procedural Technique (Q5–7)

Outcomes & Efficacy (Q8–11)

Complications & Risk Mitigation (Q12–14)

Postoperative Management (Q15–20)

Table 1 provides the full list of questions.

Table 1. Patient-oriented and guideline-based questions regarding Sacral Neuromodulation (InterStim®) therapy for refractory overactive bladder and urinary retention.

No.	Question	Category
Mechanism & Indications		
1	How does sacral neuromodulation (SNM) work to improve bladder function?	Frequently asked questions (FAQ)
2	Who is eligible for SNM therapy?	Guideline-based
3	Can SNM be used for urinary retention as well as overactive bladder?	Guideline-based
4	Is SNM indicated for patients with neurological conditions (e.g., multiple sclerosis, spinal cord injury)?	Guideline-based
Procedural Technique		
5	What is the difference between the trial (test phase) and the permanent implant in SNM?	FAQ
6	Is the SNM procedure done under local or general anesthesia?	FAQ
7	How is the lead positioned during the SNM procedure?	Guideline-based
Outcomes & Efficacy		
8	What are the success rates of SNM for overactive bladder?	FAQ
9	How long does the benefit of SNM usually last?	FAQ
10	How does SNM compare to Botox injections for overactive bladder?	Guideline-based
11	Can SNM therapy be repeated or adjusted if symptoms return?	FAQ
Complications & Risk Mitigation		
12	What are the most common complications of SNM (e.g., lead migration, infection)?	FAQ
13	How often do patients need revision surgery after SNM?	Guideline-based
14	Is SNM safe for patients who may require magnetic resonance imaging (MRI) scans in the future?	Guideline-based
Postoperative Management		
15	How long is the recovery period after SNM implantation?	FAQ
16	How long does the SNM battery last, and how is it replaced?	FAQ
17	What activity restrictions should patients follow after SNM implantation?	FAQ
18	Can SNM be used together with medications for overactive bladder?	Guideline-based
19	What follow-up care and monitoring are recommended after SNM?	Guideline-based
20	Are there contraindications for SNM, such as certain medical conditions or prior surgeries?	Guideline-based

2.3. Classification of Questions

By Source Type: Questions were classified as FAQ-based ($n = 10$) or guideline-based ($n = 10$).

By Difficulty: Two urologists independently categorized questions into three levels:

Easy: factual knowledge (e.g., device mechanism, anesthesia type).

Medium: questions requiring synthesis of multiple data points or interpretation of comparative evidence (e.g., success rates, SNM vs. Botox).

Difficult: questions requiring nuanced interpretation of guidelines, contraindications, safety protocols, long-term outcomes, or risk stratification.

Disagreements were resolved by consensus, and classification criteria were applied consistently across all items.

2.4. ChatGPT Response Collection

Each of the 20 questions was posed to ChatGPT versions 3.5 (free), 4.0 (subscription), and 5.0 (latest subscription) between 1 July and 1 September 2025.

Standardization: Queries were submitted using a web browser with cleared cookies and history. Each question was entered exactly as written in Table 1.

Response Randomization and Blinding: All responses were anonymized, randomized, and stripped of metadata (including timestamps and model identifiers) before evaluation to ensure that raters were fully blinded to the ChatGPT version and response time.

Response Recording: The first complete answer generated was copied verbatim into a study database. No follow-up prompts or regenerations were allowed.

Health-Literacy Neutral Prompting: No attempts were made to adjust, simplify, or enhance health literacy; all questions were asked as written, without instructing the model to generate patient-friendly explanations.

Reproducibility Check: Each question was re-asked in triplicate for each version at different time points, and response consistency was measured statistically.

2.5. Evaluation of Responses

Two board-certified urologists with expertise in functional urology and neuromodulation independently evaluated responses using a four-point scale adapted from prior ChatGPT benchmarking studies [13–15]:

Completely Correct (Grade 1): Fully accurate, guideline-concordant, and comprehensive.

Correct but Incomplete (Grade 2): Accurate but lacking essential details.

Partially Misleading (Grade 3): Containing both accurate and inaccurate elements, or ambiguity with potential to mislead.

Completely Incorrect (Grade 4): Factually wrong, unsafe, or contradictory to guidelines.

Validation Status: This scale has not undergone formal psychometric validation and is used here as an adaptation of previously published LLM evaluation frameworks.

Clinical Importance of Errors: Raters also identified whether any incorrect or incomplete elements were considered clinically important (e.g., omissions related to magnetic resonance imaging (MRI) safety, infection risk, or postoperative restrictions). These determinations informed the interpretation of model performance in the Results and Discussion sections.

Discrepancy Resolution: Disagreements were resolved by discussion; if consensus was not achieved, the more conservative (lower) grade was assigned. The arithmetic mean of reviewer scores was used for statistical analysis.

2.6. Statistical Analysis

Primary Outcome: Combined success rate (percentage of responses graded as completely correct or correct but incomplete).

Secondary Outcomes: Accuracy by domain, source type, and difficulty level; reproducibility across repeated queries.

Tests Used: Chi-square or Fisher's exact test compared categorical outcomes across versions. Effect sizes were measured with Cramer's V.

Two Levels of Agreement Measurement:

Inter-rater Agreement: Cohen's weighted κ was used to assess agreement between urologist evaluators.

Model Repeatability: A second κ metric was applied to assess reproducibility of ChatGPT responses across repeated runs of identical queries.

Definition of Clinical Meaningfulness: Differences between model versions were considered "clinically meaningful" when improvements occurred in domains with direct implications for patient safety or counseling accuracy—particularly complication management, MRI safety, postoperative restrictions, and guidance-dependent decision making.

Software: Analyses were conducted using Jeffreys's Amazing Statistics Program (JASP) (Version 0.95.1).

Significance Threshold: $p < 0.05$.

Supplementary Material Statement: The Supplementary Materials include exact ChatGPT build/version identifiers (Table S1), query timestamps (Tables S2 and S3), and all response-generation parameters (Tables S4 and S5).

3. Results

A total of 20 structured SNM-related questions were analyzed across ChatGPT versions 3.5, 4.0, and 5.0 (Table 1). All questions were answered by each model and evaluated across five domains: Mechanism & Indications, Procedural Technique, Outcomes & Efficacy, Complications & Risk Mitigation, and Postoperative Management.

3.1. Overall Accuracy Across Versions

Accuracy improved progressively with newer models (Table 2).

Table 2. Performance of ChatGPT versions 3.5, 4.0, and 5.0 in answering 20 sacral neuromodulation (SNM)-related questions.

ChatGPT Version	Completely Correct (Grade 1)	Correct but Incomplete (Grade 2)	Partially Misleading (Grade 3)	Completely Incorrect (Grade 4)	Combined Success Rate (Grade 1 + 2)
3.5	7/20 (35%)	7/20 (35%)	4/20 (20%)	2/20 (10%)	14/20 (70%)
4.0	11/20 (55%)	6/20 (30%)	2/20 (10%)	1/20 (5%)	17/20 (85%)
5.0	13/20 (65%)	5/20 (25%)	1/20 (5%)	1/20 (5%)	18/20 (90%)

Version 3.5 achieved a combined success rate of 14/20 (70.0%), increasing to 17/20 (85.0%) in version 4.0 and 18/20 (90.0%) in version 5.0. Correspondingly, the proportion of misleading or incorrect responses declined from 6/20 (30.0%) in version 3.5 to 2/20 (10.0%) in version 5.0.

Statistical comparison showed significant differences across versions ($\chi^2 (2) = 6.96$, $p = 0.031$), with a moderate effect size (Cramer's V = 0.29, 95% CI: 0.10–0.47), and accuracy improvements were considered clinically meaningful because gains occurred primarily in domains with direct patient-safety implications—specifically complication management, MRI conditionality, and postoperative restrictions (Figure 1).

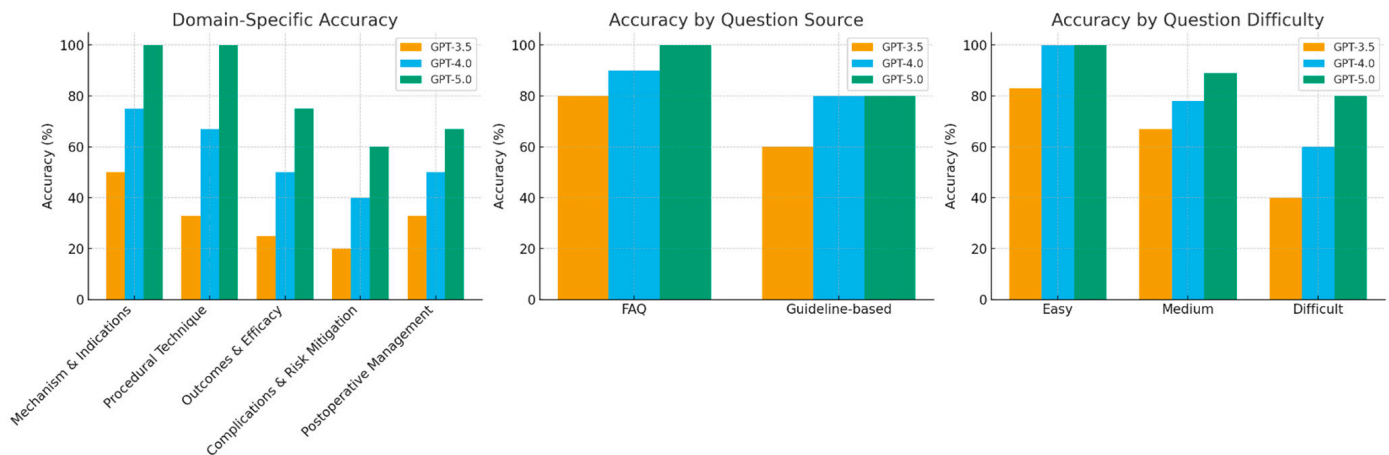


Figure 1. Accuracy of ChatGPT versions 3.5, 4.0, and 5.0 in answering sacral neuromodulation-related questions. FAQ: frequently asked questions.

3.2. Domain-Specific Accuracy

Accuracy varied across thematic domains (Table 3). Mechanism & Indications and Procedural Technique reached 4/4 and 3/3 completely correct answers, respectively, in version 5.0, whereas Complications & Risk Mitigation remained the most challenging domain, improving from 1/5 (20.0%) in version 3.5 to 3/5 (60.0%) in version 5.0.

Table 3. Accuracy of ChatGPT responses by thematic domain across versions 3.5, 4.0, and 5.0.

Thematic Domain	Version 3.5 (n/N, % Completely Correct)	Version 4.0 (n/N, % Completely Correct)	Version 5.0 (n/N, % Completely Correct)
Mechanism & Indications (Q1–4)	2/4 (50%)	3/4 (75%)	4/4 (100%)
Procedural Technique (Q5–7)	1/3 (33%)	2/3 (67%)	3/3 (100%)
Outcomes & Efficacy (Q8–11)	1/4 (25%)	2/4 (50%)	3/4 (75%)
Complications & Risk Mitigation (Q12–14)	1/5 (20%)	2/5 (40%)	3/5 (60%)
Postoperative Management (Q15–20)	2/6 (33%)	3/6 (50%)	4/6 (67%)

Postoperative Management also showed progressive gains from 2/6 (33.3%) in version 3.5 to 4/6 (66.7%) in version 5.0. Together, these results indicate that ChatGPT excelled in mechanistic and technical content while showing weaker performance in guideline-dependent and complication-focused questions. These areas represent clinically consequential domains, and therefore improvements in these categories further support the interpretation of clinical meaningfulness.

3.3. FAQ vs. Guideline-Based Questions

Performance also differed by question source (Table 4). FAQ items achieved full accuracy in version 5.0 (10/10, 100%), whereas guideline-based questions plateaued at 8/10 (80.0%). This pattern confirms greater reliability for straightforward patient education than for nuanced guideline interpretation.

Table 4. Accuracy of ChatGPT responses based on source type (frequently asked questions [FAQ] vs. guideline-based) and question difficulty across versions 3.5, 4.0, and 5.0.

Category	Version 3.5 (n/N, % Correct)	Version 4.0 (n/N, % Correct)	Version 5.0 (n/N, % Correct)
Source Type			
FAQ (<i>n</i> = 10)	8/10 (80%)	9/10 (90%)	10/10 (100%)
Guideline-based (<i>n</i> = 10)	6/10 (60%)	8/10 (80%)	8/10 (80%)
Question Difficulty			
Easy (<i>n</i> = 6)	5/6 (83%)	6/6 (100%)	6/6 (100%)
Medium (<i>n</i> = 9)	6/9 (67%)	7/9 (78%)	8/9 (89%)
Difficult (<i>n</i> = 5)	2/5 (40%)	3/5 (60%)	4/5 (80%)

3.4. Question Difficulty

When stratified by difficulty, easy items reached ceiling performance in both version 4.0 and 5.0 (6/6, 100%), whereas difficult items, despite improvement, lagged behind (version 3.5: 2/5, 40.0% → version 5.0: 4/5, 80.0%).

These findings reinforce that ChatGPT accuracy decreases when questions require synthesis or interpretation rather than factual recall.

3.5. Reproducibility of Responses

Reproducibility remained excellent across all versions. Cohen's κ values were 0.81, 0.87, and 0.91 for versions 3.5, 4.0, and 5.0, respectively, with 95% confidence interval (CIs) of 0.68–0.93, 0.76–0.96, and 0.83–0.98. These findings confirm that observed performance differences were not driven by random variability in AI responses.

3.6. Weighted Sensitivity Analysis

To assess the effect of incomplete responses, a weighted analysis was performed (Grade-1 = 1.0, Grade-2 = 0.5, Grades-3/4 = 0.0). Weighted mean performance improved from 0.63 (3.5) → 0.76 (4.0) → 0.82 (5.0), confirming that models improved not only in correctness but also in quality and clinical usefulness of answers.

3.7. Examples of “Correct but Incomplete” (Grade-2) Responses

Common missing clinical details included:

- MRI safety lacking conditional rules (e.g., device must be switched to MRI mode).
- Infection risk omitted despite being a major cause of revision surgery.
- Ambiguous postoperative restrictions (e.g., not specifying no twisting, bending, or lifting >10 lbs for 4–6 weeks).

Short illustrative examples are provided in Supplementary Data (Tables S6–S9).

4. Discussion

In this study, a progressive improvement in ChatGPT's performance on SNM questions was observed, with combined success rates rising from 70% in version 3.5 to 85% in version 4.0 and 90% in version 5.0. Accuracy was highest for mechanism and procedural technique, while complication-related, postoperative, and guideline-derived queries remained more challenging. FAQs and straightforward factual questions were answered with greater precision than nuanced or guideline-based ones, and reproducibility across repeated runs was excellent, with kappa values exceeding 0.80. Because question domains were selected to reflect the main components of SNM counseling—mechanism, procedure, outcomes, complications, and postoperative management—these findings provide insight into how

LLMs perform across clinically relevant information categories. Importantly, version 3.5 represented a freely available model, whereas versions 4.0 and 5.0 require paid subscription access. The observed performance gap therefore also reflects meaningful differences between the model versions most likely to be used by patients (free-tier 3.5) and those used by clinicians or institutions (subscription models).

The version-dependent improvements identified here align with prior evidence that newer ChatGPT models outperform earlier releases across medical knowledge tasks. Meta-analyses have shown clear gains from GPT-3.5 to GPT-4 in licensing-style examinations and across multiple languages [16,17]. In oncology and decision-support settings, GPT-4 has produced higher-quality therapy recommendations and more reliable patient-facing explanations, although performance has varied with scenario complexity [14,18,19].

Within urology, evaluations have demonstrated moderate to strong alignment of ChatGPT responses with AUA and EAU guidelines, while highlighting persistent limitations in complication management and guideline-conditioned decisions [15,20–22]. Specialized or guideline-augmented LLMs can even surpass human benchmarks in board-style testing, indicating the potential of context-enriched models [23]. Only one published study has previously included SNM among overactive bladder interventions when assessing multiple chatbots, but SNM-specific accuracy and version-based comparisons were not analyzed [10]. By focusing exclusively on SNM, this study contributes novel insights while reinforcing the broader pattern of strong performance in procedural and mechanistic domains and weaker reliability for guideline-based or complication-related queries [14–16,20]. The relatively lower performance on guideline-based questions likely reflects several challenges: the static training cutoffs that prevent real-time incorporation of updated clinical guidelines, the difficulty of interpreting nuanced recommendations that often require contextual reasoning, and the tendency of LLMs to generalize from prior data rather than adapt to highly specific clinical standards.

ChatGPT has clear potential to serve as a supportive tool in both patient education and medical training. Patient-directed explanations and educational materials can be generated rapidly and at appropriate reading levels, improving health literacy and supporting preparation for consultations [4,24]. Prior studies have confirmed that ChatGPT can produce materials that meet recommended readability standards and provide useful summaries in oncology and other settings [4,7]. For clinicians, ChatGPT may function as an adjunct in counseling, generating handouts, simulating patient queries for trainees, and outlining procedural steps for resident teaching [25,26]. However, safe use requires oversight: AI-generated content should be reviewed by qualified clinicians before dissemination, and not used as an autonomous source of advice [24,27]. This study reinforces the necessity of clinician oversight by demonstrating that even the most advanced models struggle with high-complexity or guideline-dependent content. Safe integration into urology practice therefore requires workflows in which LLM-generated responses are used only as preliminary drafts that are verified, corrected, or expanded upon by clinicians. Importantly, its limitations include instances where raters identified certain errors or omissions as “clinically important,” such as incorrect or incomplete statements regarding MRI conditionality, infection risks, lead migration, or postoperative restrictions. These errors reinforce that ChatGPT can generate plausible but potentially unsafe outputs without expert review.

Another important consideration is the absence of standardized SNM-specific question sets. Because no validated questionnaire or benchmark exists for evaluating AI performance in SNM counseling or decision-support, questions must be manually selected, which may introduce variability and limit cross-study comparability. Developing a standardized, expert-validated SNM question bank would allow more rigorous benchmarking and support the creation of safer AI-assisted educational tools.

Finally, although this study focused on ChatGPT, emerging evaluations of other LLM families—such as Claude, Gemini, and Med-PaLM—have revealed varying strengths in urological knowledge and procedural guidance [11,12]. These comparisons indicate that performance is model-specific rather than universal across LLMs, underscoring the need for direct, head-to-head evaluations to determine which systems are safest for patient-facing or clinician-support tasks.

The study confirmed persistent limitations in ChatGPT's performance, particularly in guideline-derived and high-difficulty queries, echoing prior evidence that LLMs struggle with nuanced interpretation and individualized risk assessment [13]. Risks of partially misleading or overconfident “hallucinations” have been documented and remain key safety concerns [28]. More generally, the “black-box” nature of LLMs undermines interpretability, motivating calls for explainable AI and model-augmentation strategies [29,30]. Static training cutoffs also limit the ability to reflect real-time guideline updates, and without clinician guidance, patient misinterpretation remains a concern [31]. Future integration of continuously updated guideline repositories or hybrid workflows where LLM-generated content is validated by clinicians could mitigate these risks and enhance reliability for guideline-dependent scenarios. Additionally, no explicit health-literacy prompting was used in this study, meaning responses reflected each model's default output style. Future studies may explore whether prompting models to generate patient-friendly wording improves reliability or reduces clinically important omissions. Finally, the grading scale used in this study was adapted from prior LLM evaluations but has not undergone formal psychometric validation; this limitation mirrors the clarification provided in the Supplementary data.

4.1. Study Limitations

- Only 20 questions were analyzed, which may not encompass the full spectrum of SNM-related queries. This relatively small sample constrains generalizability, as the performance trends observed may not fully capture the breadth of real-world patient or clinician inquiries.
- The evaluation was performed in English only, limiting generalizability to other languages.
- Responses were rated by two reviewers, leaving potential for residual subjectivity.
- Only ChatGPT was tested; other LLMs such as Claude, Gemini, and DeepSeek were not compared.
- SNM questions were selected manually because no standardized, validated SNM question bank exists, which may introduce selection bias and limit comparability across studies.
- The small question set ($n = 20$) limits the power of subgroup analyses, particularly for domain-level and difficulty-level comparisons.
- ChatGPT versions evolve continuously, and silent model updates during the study period may influence reproducibility, meaning results represent performance during a specific time window rather than a fixed model state.

Future studies should evaluate ChatGPT using larger, multilingual question sets to better reflect real-world use. Comparative studies with alternative LLMs (e.g., Claude, Gemini, DeepSeek, Med-PaLM) are warranted to contextualize performance. Benchmarking ChatGPT against these emerging systems will help determine whether observed limitations are model-specific or represent broader constraints of current LLMs. Integration with continuously updated guidelines could improve reliability, while patient-centered studies are needed to evaluate whether ChatGPT meaningfully enhances understanding, decision-making, and satisfaction in SNM care.

ChatGPT versions 4.0 and 5.0 demonstrated strong potential as adjunctive resources for patient education in SNM, particularly for FAQs and procedural knowledge. However, limitations in guideline interpretation and complication management necessitate clinician oversight. These models should be regarded as supportive aids that can enhance communication and patient engagement, but not as replacements for professional clinical judgment.

4.2. Recommendations for Safe Clinical Use of ChatGPT in SNM Counseling

Treat ChatGPT as a supplemental educational tool, not a standalone clinical advisor.

Review and verify all AI-generated explanations before sharing with patients.

Pay particular attention to complication management, MRI conditionality, and postoperative restrictions, which remain error-prone.

Use subscription-based models (e.g., GPT-4.0 or GPT-5.0) when possible, as free-tier models show notably lower accuracy.

Avoid relying on default responses for patient-facing communication; consider prompting for simplified or health-literate wording when appropriate.

Do not use ChatGPT for individualized recommendations, risk stratification, or complex guideline interpretation.

Integrate LLM-driven tools into supervised clinical workflows, with clinician oversight at every step.

5. Conclusions

This study demonstrated that ChatGPT versions 4.0 and 5.0 outperformed version 3.5 in addressing questions related to SNM therapy. Strongest performance was observed for mechanistic and procedural knowledge, while complication-related and guideline-derived queries remained more challenging. FAQs and factual items were answered with high reliability, highlighting utility for patient education and counseling. Nonetheless, limitations in guideline interpretation and risks of incomplete or misleading outputs emphasize the need for clinician oversight. Because version 3.5 represents the free, publicly accessible model—while versions 4.0 and 5.0 require subscription access—the performance gap has practical implications: most patients using free-tier versions may receive less accurate information.

Accuracy improvements were also considered clinically meaningful because gains occurred in safety-critical domains such as complication management, MRI restrictions, and postoperative care. Importantly, the persistently lower accuracy on guideline-based questions indicates that these models are not yet sufficient for independent clinical reliance. Direct comparison with other large language models (e.g., Claude, Gemini, Med-PaLM) will be essential to contextualize performance and determine whether these limitations are specific to ChatGPT or more broadly representative of current LLMs. ChatGPT shows promise as a supportive tool to improve patient understanding and facilitate shared decision-making in SNM, but it should serve as an adjunctive tool rather than an autonomous decision-making system, and future validation against other LLMs is needed.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/siuj7010011/s1>. **Supplement A:** ChatGPT Version Metadata, Query Parameters, and Reproducibility Protocol. **Supplement B:** Examples of Grade-2 (“Correct but Incomplete”) Clinical Responses and Error Typology.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study did not involve human participants, animal experiments, or identifiable personal data.

Informed Consent Statement: Not applicable.

Data Availability Statement: Supplementary Materials accompanying this article include meta-data on ChatGPT version, query timestamps, response-generation parameters, and examples of Grade-2 responses.

Conflicts of Interest: The author declares no competing interests.

References

1. Amundsen, C.L.; Sutherland, S.E.; Kielb, S.J.; Dmochowski, R.R. Sacral and implantable tibial neuromodulation for the management of overactive bladder: A systematic review and meta-analysis. *Adv. Ther.* **2025**, *42*, 10–35. [CrossRef]
2. Busch, F.; Hoffmann, L.; Rueger, C.; Van Dijk, E.H.; Kader, R.; Ortiz-Prado, E.; Makowski, M.R.; Saba, L.; Hadamitzky, M.; Kather, J.N.; et al. Current applications and challenges in large language models for patient care: A systematic review. *Commun. Med.* **2025**, *5*, 26. [CrossRef]
3. Cameron, A.P.; Chung, D.E.; Dielubanza, E.J.; Enemchukwu, E.; Ginsberg, D.A.; Helfand, B.T.; Linder, B.J.; Reynolds, W.S.; Rovner, E.S.; Souter, L.; et al. The AUA/SUFU guideline on the diagnosis and treatment of idiopathic overactive bladder. *J. Urol.* **2024**, *212*, 11–20. [CrossRef]
4. Abdelmalek, G.; Uppal, H.; Garcia, D.; Farshchian, J.; Emami, A.; McGinniss, A. Leveraging ChatGPT to produce patient education materials for common hand conditions. *J. Hand Surg. Glob. Online* **2025**, *7*, 37–40. [CrossRef]
5. Sartori, A.M.; Kessler, T.M.; Castro-Díaz, D.M.; De Keijzer, P.; Del Popolo, G.; Ecclestone, H.; Frings, D.; Groen, J.; Hamid, R.; Karsenty, G.; et al. Summary of the 2024 update of the European Association of Urology guidelines on neuro-urology. *Eur. Urol.* **2024**, *85*, 543–555. [CrossRef] [PubMed]
6. Feloney, M.P.; Stauss, K.; Leslie, S.W. Sacral Neuromodulation. In *StatPearls*; StatPearls Publishing: Treasure Island, FL, USA, 2024. Available online: <https://www.ncbi.nlm.nih.gov/books/NBK567751/> (accessed on 15 August 2025).
7. Gibson, D.; Jackson, S.; Shanmugasundaram, R.; Seth, I.; Siu, A.; Ahmadi, N.; Kam, J.; Mehan, N.; Thanigasalam, R.; Jeffery, N.; et al. Evaluating the efficacy of ChatGPT as a patient education tool in prostate cancer: Multimetric assessment. *J. Med. Internet Res.* **2024**, *26*, e55939. [CrossRef] [PubMed]
8. Libretti, A.; Vitale, S.G.; Saponara, S.; Corsini, C.; Aquino, C.I.; Savasta, F.; Tizzoni, E.; Troia, L.; Surico, D.; Angioni, S.; et al. Hysteroscopy in the new media: Quality and reliability of hysteroscopy procedures on YouTube™. *Arch. Gynecol. Obstet.* **2023**, *308*, 1515–1524. [CrossRef]
9. National Institute for Health and Care Excellence. Sacral Nerve Stimulation for Idiopathic Chronic Non-Obstructive Urinary Retention. NICE Guideline IPG536. 2015. Available online: <https://www.nice.org.uk/guidance/ipg536> (accessed on 15 August 2025).
10. Hacibey, I.; Halis, A. Assessment of artificial intelligence performance in answering questions on onabotulinum toxin and sacral neuromodulation. *Investig. Clin. Urol.* **2025**, *66*, 188–193. [CrossRef]
11. Şahin, M.F.; Doğan, Ç.; Topkaç, E.C.; Şeramet, S.; Tuncer, F.B.; Yazıcı, C.M. Which current chatbot is more competent in urological theoretical knowledge? A comparative analysis by the European board of urology in-service assessment. *World J. Urol.* **2025**, *43*, 116. [CrossRef]
12. Malak, A.; Şahin, M.F. How Useful are Current Chatbots Regarding Urology Patient Information? Comparison of the Ten Most Popular Chatbots' Responses About Female Urinary Incontinence. *J. Med. Syst.* **2024**, *48*, 102. [CrossRef]
13. Chelli, M.; Descamps, J.; Lavoué, V.; Trojani, C.; Azar, M.; Deckert, M.; Raynier, J.; Clowez, G.; Boileau, P.; Ruetsch-Chelli, C. Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. *J. Med. Internet Res.* **2024**, *26*, e53164. [CrossRef]
14. Grilo, A.; Marques, C.; Corte-Real, M.; Carolino, E.; Caetano, M. Assessing the Quality and Reliability of ChatGPT's Responses to Radiotherapy-Related Patient Queries: Comparative Study With GPT-3.5 and GPT-4. *JMIR Cancer* **2025**, *11*, e63677. [CrossRef] [PubMed]
15. Tan, Y.Z.; Nah, S.A.; Saw, S.N.; Rajandram, R.; Ong, T.A. Evaluating the performance of artificial intelligence chatbots in answering urology questions derived from guidelines or board examinations: A systematic review. *Urol. Sci.* **2025**. [CrossRef]
16. Jin, H.K.; Lee, H.E.; Kim, E. Performance of ChatGPT-3.5 and GPT-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: a systematic review and meta-analysis. *BMC Med. Educ.* **2024**, *24*, 1013. [CrossRef]
17. Liu, M.; Okuhara, T.; Chang, X.; Shirabe, R.; Nishiie, Y.; Okada, H.; Kiuchi, T. Performance of ChatGPT across different versions in medical licensing examinations: Systematic review and meta-analysis. *J. Med. Internet Res.* **2024**, *26*, e60807. [CrossRef]
18. Stalp, J.L.; Denecke, A.; Jentschke, M.; Hillemanns, P.; Klapdor, R. Quality of ChatGPT-Generated Therapy Recommendations for Breast Cancer Treatment in Gynecology. *Curr. Oncol.* **2024**, *31*, 3845–3854. [CrossRef]
19. Chen, S.; Kann, B.H.; Foote, M.B.; Aerts, H.J.; Savova, G.K.; Mak, R.H.; Bitterman, D.S. The utility of ChatGPT for cancer treatment information. *medRxiv* **2023**. [CrossRef]

20. Scott, M.; Muncey, W.; Seranio, N.; Belladelli, F.; Del Giudice, F.; Li, S.; Ha, A.; Glover, F.; Zhang, C.A.; Eisenberg, M.L. Assessing Artificial Intelligence–Generated Responses to Urology Patient In-Basket Messages. *Urol. Pract.* **2024**, *11*, 793–798. [[CrossRef](#)]
21. Talyshinskii, A.; Juliebø-Jones, P.; Hameed, B.M.Z.; Naik, N.; Adhikari, K.; Zhanbyrbekuly, U.; Tzelves, L.; Somani, B.K. 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](#)]
22. Caglar, U.; Yildiz, O.; Meric, A.; Ayranci, A.; Gelmis, M.; Sarilar, O.; Ozgor, F. Evaluating the performance of ChatGPT in answering questions related to pediatric urology. *J. Pediatr. Urol.* **2024**, *20*, 26.e1–26.e5. [[CrossRef](#)] [[PubMed](#)]
23. Hetz, M.; Carl, N.; Hagggenmüller, S.; Wies, C.; Kather, J.; Michel, M.; Wessels, F.; Brinker, T. Superhuman performance on urology board questions using an explainable language model enhanced with European Association of Urology guidelines. *ESMO Real World Data Digit. Oncol.* **2024**, *6*, 100078. [[CrossRef](#)]
24. Sheikh, J.K.; Sohail, S.S.; Alam, S. Enhancing patient education with ChatGPT: Critical insights and future directions. *Indian J. Anaesth.* **2024**, *68*, 1112–1113. [[CrossRef](#)]
25. Armbruster, J.; Bussmann, F.; Rothhaas, C.; Titze, N.; Grütznier, P.A.; Freischmidt, H. “Doctor ChatGPT, Can You Help Me?” The Patient’s Perspective: Cross-Sectional Study. *J. Med. Internet Res.* **2024**, *26*, e58831. [[CrossRef](#)]
26. Tran, J.T.; Burghall, A.; Blydt-Hansen, T.; Cammer, A.; Goldberg, A.; Hamiwka, L.; Johnson, C.; Kehler, C.; Phan, V.; Rosaasen, N.; et al. Exploring the ability of ChatGPT to create quality patient education resources about kidney transplant. *Patient Educ. Couns.* **2024**, *129*, 108400. [[CrossRef](#)] [[PubMed](#)]
27. Haltaufderheide, J.; Ranisch, R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). *npj Digit. Med.* **2024**, *7*, 183. [[CrossRef](#)] [[PubMed](#)]
28. Asgari, E.; Montaña-Brown, N.; Dubois, M.; Khalil, S.; Balloch, J.; Yeung, J.A.; Pimenta, D. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digit. Med.* **2025**, *8*, 274. [[CrossRef](#)]
29. Sadeghi, Z.; Alizadehsani, R.; Cifci, M.A.; Kausar, S.; Rehman, R.; Mahanta, P.; Bora, P.K.; Almasri, A.; Alkhalwaldeh, R.S.; Hussain, S.; et al. A review of Explainable Artificial Intelligence in healthcare. *Comput. Electr. Eng.* **2024**, *118*, 109370. [[CrossRef](#)]
30. Muhammad, D.; Bendeche, M. Unveiling the black box: A systematic review of Explainable Artificial Intelligence in medical image analysis. *Comput. Struct. Biotechnol. J.* **2024**, *24*, 542–560. [[CrossRef](#)]
31. Iqbal, U.; Tanweer, A.; Rahmanti, A.R.; Greenfield, D.; Lee, L.T.; Li, Y.J. Impact of large language model (ChatGPT) in healthcare: an umbrella review and evidence synthesis. *J. Biomed. Sci.* **2025**, *32*, 45. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.