Submitted: 25 August, 2025 Accepted: 23 October, 2025 Published: 30 March, 2026 DOI:10.22514/j.androl.2026.007

REV INT ANDROL
Revista Internacional de Andrología

ORIGINAL RESEARCH

Evaluating ChatGPT versions 3.5, 4.0, and 5.0 in patient- and guideline-based questions on Optilume® therapy

Eralp Kubilay[1,*], Hüseyin Gültekin[1]

[1] Department of Urology, Near East University, 1010 Nicosia, Cyprus

***Correspondence**
eralp.kubilay@neu.edu.tr
(Eralp Kubilay)

Abstract

Background: We aimed to evaluate the accuracy, completeness, and reproducibility of ChatGPT versions 3.5, 4.0, and 5.0 in responding to patient-oriented and guideline-based questions on Optilume® therapy. **Methods**: Twenty structured questions were developed from patient Frequently Asked Questions (FAQs) social media forums, and procedural guidelines, covering five thematic domains: Device Mechanism and Indications, Procedural Technique, Outcomes and Efficacy, Complications, and Postoperative Management. Each question was posed to ChatGPT versions 3.5 (free), 4.0 (subscription), and 5.0 (latest subscription). Two independent urologists graded responses using a four-point scale (completely correct, correct but incomplete, partially misleading, completely incorrect). Question difficulty and source type (FAQ *vs.* guideline-based) were also analyzed. Reproducibility was assessed using Cohen's kappa. **Results**: Overall, ChatGPT performance improved progressively with each version. Combined success rates (completely correct + correct but incomplete) were 75% for 3.5, 85% for 4.0, and 90% for 5.0. Device Mechanism and Procedural Technique domains achieved the highest accuracy across all versions, while Outcomes, Complications, and Postoperative Management improved notably in versions 4.0 and 5.0. FAQs were answered more accurately than guideline-based questions, with version 5.0 reaching 90% *vs.* 60%, respectively. Response accuracy increased with repeated versions, even for medium- and high-difficulty questions. Reproducibility was excellent, with Cohen's kappa ranging from 0.82 to 0.91. **Conclusions**: ChatGPT, particularly versions 4.0 and 5.0, provides accurate, reproducible, and clinically relevant information on Optilume therapy.

Keywords

ChatGPT; Optilume therapy; Urethral stricture; Artificial intelligence; Medical education

Evaluación de las versiones 3.5, 4.0 y 5.0 de ChatGPT en preguntas basadas en pacientes y guías sobre la terapia Optilume®

Resumen

Antecedentes: Nuestro objetivo fue evaluar la precisión, integridad y reproducibilidad de las versiones 3.5, 4.0 y 5.0 de ChatGPT al responder preguntas orientadas al paciente y basadas en pautas sobre la terapia Optilume. **Métodos**: Se desarrollaron veinte preguntas estructuradas a partir de preguntas frecuentes de pacientes, foros en redes sociales y guías de procedimiento, que abarcan cinco áreas temáticas: Mecanismo e indicaciones del dispositivo, Técnica del procedimiento, Resultados y eficacia, Complicaciones y Manejo posoperatorio. Cada pregunta se formuló en las versiones 3.5 (gratuita), 4.0 (con suscripción) y 5.0 (última suscripción) de ChatGPT. Dos urólogos independientes calificaron las respuestas mediante una escala de cuatro puntos (completamente correcta, correcta pero incompleta, parcialmente engañosa, completamente incorrecta). También se analizaron la dificultad de las preguntas y el tipo de fuente (preguntas frecuentes *vs.* basada en guías). La reproducibilidad se evaluó mediante el índice kappa de Cohen. **Resultados**: En general, el rendimiento de ChatGPT mejoró progresivamente con cada versión. Las tasas de éxito combinadas (completamente correctas + correctas pero incompletas) fueron del 75% para la versión 3.5, del 85% para la versión 4.0 y del 90% para la versión 5.0. Los dominios de Mecanismo del Dispositivo y Técnica de Procedimiento alcanzaron la mayor precisión en todas las versiones, mientras que los dominios de Resultados, Complicaciones y Manejo Postoperatorio mejoraron notablemente en las versiones 4.0 y 5.0. Las preguntas frecuentes se respondieron con mayor precisión que las preguntas basadas en guías, alcanzando la versión 5.0 el 90% frente al 60%, respectivamente. La precisión de las respuestas aumentó con las versiones repetidas, incluso para preguntas de dificultad media y alta. La reproducibilidad fue excelente, con un índice kappa de Cohen que osciló entre 0.82 y 0.91. **Conclusiones**: ChatGPT, especialmente las versiones 4.0 y 5.0, proporciona información precisa, reproducible y clínicamente relevante sobre la terapia Optilume.

Palabras Clave

ChatGPT; Terapia optilume; Estenosis uretral; Inteligencia artificial; Educación médica

1. Introduction

Urethral stricture disease (USD) constitutes a significant clinical challenge in contemporary urology, affecting a considerable proportion of the male population worldwide [1]. The condition is characterized by narrowing of the urethral lumen due to fibrosis and scarring, leading to lower urinary tract symptoms (LUTS), such as decreased urinary flow, urinary retention, recurrent urinary tract infections, and, in severe cases, progressive renal impairment [2]. Traditional management strategies vary in success. Urethral dilatation and direct vision internal urethrotomy (DVIU) often require repeated interventions and lead to higher morbidity and costs, whereas urethroplasty achieves success rates of over 80% in expert hands, although it is more invasive [3–5]. In recent years, the evolution of endoluminal therapeutic modalities has introduced novel approaches to stricture management. Among these, the Optilume® drug-coated balloon system has emerged as a promising minimally invasive intervention. This device combines mechanical dilation with local pharmacological therapy, utilizing an antiproliferative drug coating—specifically paclitaxel—to reduce fibroblast proliferation and limit restenosis [6]. Early clinical studies have reported encouraging short-term outcomes, including improved urethral patency rates and reduced recurrence compared with conventional balloon dilation alone [7, 8]. Nevertheless, long-term efficacy data remain limited, and further investigation is warranted to define its role across different stricture etiologies and anatomical locations [9].

In recent years, the development of artificial intelligence (AI) tools has made it much easier for patients to access medical information [10]. Patients can now learn about specific medical topics, such as urethral stricture disease, without always needing to see a doctor directly. In urology, several studies have looked at how well AI applications work, focusing especially on DeepSeek and Chat Generative Pre-Trained Transformer (ChatGPT) developed by OpenAI. These studies have examined how accurate information from AI is and whether newer versions improve their reliability. The results show that AI can provide correct and academically useful answers in some areas, but it is not yet reliable for every medical topic [11, 12]. Therefore, although AI tools have potential for patient education and supporting clinical decisions, professional review is still necessary to make sure that the information is safe and correct. Furthermore, to the best of our knowledge, no prior studies have systematically assessed ChatGPT's ability to respond accurately to questions about Optilume® therapy.

In this study, we aimed to evaluate the performance (accuracy, completeness, and reproducibility of responses) of ChatGPT versions 3.5, 4.0, and 5.0 in addressing patient-oriented and guideline-based questions about Optilume® therapy for urethral strictures.

2. Materials and methods

This study did not involve patient data; all materials are available in the **Supplementary materials**. All assessments were performed using publicly available information and AI-generated responses from ChatGPT versions 3.5, 4.0, and 5.0. Therefore, according to local and international guidelines on research ethics, formal approval from an institutional ethics committee was not required for this study.

2.1 Question selection

To investigate the performance of ChatGPT in providing accurate information about Optilume® therapy, we first compiled a set of the most commonly asked patient questions. Sources included patient-oriented urology websites, social media platforms, such as Facebook, Instagram, and Twitter, and online forums dedicated to urethral stricture management. In addition, procedural guidelines and published recommendations regarding endoluminal urethral interventions were reviewed and transformed into structured, evidence-based questions.

An initial pool of 55 questions was identified. After removing repetitive, ambiguous, or personal-health-specific questions, 20 unique questions directly related to Optilume® therapy were finalized. The questions were categorized into five thematic domains to facilitate structured analysis:

1. Device Mechanism and Indications (Questions 1–4)—including the functional principles of the device, drug coating mechanism, and stricture types suitable for treatment.

2. Procedural Technique (Questions 5–7)—covering balloon inflation parameters, anesthesia selection, and catheter management.

3. Outcomes and Efficacy (Questions 8–11)—short-term success rates, long-term recurrence, and comparison with conventional therapies.

4. Complications and Risk Mitigation (Questions 12–14)—adverse events, infection risk, and strategies to minimize urethral injury.

5. Postoperative Management (Questions 15–20)—follow-up imaging, patient-specific factors affecting outcomes, pain management, activity restrictions, combination therapies, and contraindications.

The complete list of questions is presented in Table 1.

2.2 ChatGPT response collection

To evaluate the accuracy and clinical relevance of ChatGPT in answering Optilume-related questions, each question was posed to ChatGPT versions 3.5 (free), 4.0 (subscription), and 5.0 (latest subscription) using a standard web browser with cleared cookies and history to prevent bias. This approach allowed for a direct comparison of the three versions to assess improvements in accuracy, completeness, and evidence-based relevance over successive iterations. The comparison of ChatGPT versions was based on methodologies used in similar studies reported in the literature [13, 14].

Version 3.5 was included as a baseline to represent the earlier iteration, while version 4.0 was selected due to its enhanced language understanding and more comprehensive medical knowledge. Version 5.0 was incorporated to evaluate the most recent capabilities of ChatGPT and determine whether the latest update provides further improvements in medical accuracy and answer completeness.

Data collection was conducted between June and August 2025, ensuring that all responses were generated during the same time frame for each ChatGPT version. Versions 3.5, 4.0, and 5.0 were accessed through the official OpenAI platform, with model identifiers verified at the time of data retrieval.

2.3 Evaluation criteria

Two independent board-certified urologists (EK, HG) assessed the accuracy, completeness, and clinical relevance of ChatGPT responses. Responses were rated on a four-point scale:

1. Completely Correct (Grade 1): Fully consistent with current evidence-based guidelines and included all relevant clinical information.

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

3. Partially Misleading (Grade 3): Contained a combination of correct and incorrect or ambiguous information.

4. Completely Incorrect (Grade 4): Factually inaccurate, potentially misleading, or clinically inappropriate.

This classification system was chosen because it is widely used in the literature [15–17].

2.4 Question difficulty classification

Questions were categorized according to perceived difficulty, considering procedural complexity and the clinical knowledge required:

- Easy: Questions requiring basic factual knowledge about device function or procedural steps (*e.g.*, Questions 1, 2, 5).
- Medium: Questions requiring synthesis of multiple clinical parameters or evidence from published outcomes (*e.g.*, Questions 8, 10, 11).
- Difficult: Questions involving rare complications, patient-specific considerations, or interpretation of long-term efficacy data (*e.g.*, Questions 14, 16, 20).

2.5 Evaluation and blinding procedures

All model responses were anonymized and presented to the raters with neutral identifiers (“Response-Identifier”) to ensure blinding. The raters were blinded to both the model version and the source of the questions. To prevent order effects, the sequence of responses was independently randomized for each rater using computer-generated randomization. Prior to the formal evaluation, a short calibration session was conducted with sample responses to finalize the grading codebook and ensure consistency.

2.6 Inter-rater reliability

Two independent raters evaluated all responses according to the four-level rubric. Inter-rater reliability was assessed before consensus and reported separately from the model performance metrics. Weighted Cohen’s kappa (quadratic weights) was used as the primary reliability measure, and a two-way random effects intraclass correlation coefficient (ICC(2, k)) with 95% confidence intervals was also calculated as a sensitivity analysis. Final consensus scores were then used for the main performance analyses.

2.7 Reproducibility assessment

To evaluate model reproducibility, all questions were submitted repeatedly under identical conditions. For each model version (GPT-3.5, GPT-4.0, and GPT-5.0), responses were generated five times per question using a fixed temperature

TABLE 1. Patient-oriented and guideline-based questions regarding Optilume® therapy for urethral strictures.

No	Question
1	How does the Optilume® device dilate urethral strictures?
2	What is the mechanism of action of the drug coating in Optilume?
3	For which types of urethral strictures is Optilume indicated?
4	Can Optilume be used for long-segment urethral strictures?
5	What is the recommended balloon inflation pressure and duration?
6	Is local or general anesthesia preferred for Optilume procedures?
7	How long should catheterization be maintained after Optilume therapy?
8	What are the typical short-term success rates reported for Optilume?
9	What are the long-term recurrence rates after Optilume treatment?
10	How does Optilume compare to standard balloon dilation or urethrotomy?
11	Can Optilume be repeated in cases of stricture recurrence?
12	What are the most common complications associated with Optilume?
13	How can urethral injury be minimized during Optilume therapy?
14	Are urinary tract infections common after Optilume treatment?
15	What follow-up imaging or endoscopic evaluation is recommended?
16	Are there patient-specific factors that may affect Optilume efficacy?
17	How is postoperative pain managed after Optilume procedures?
18	Can Optilume be combined with other endoluminal therapies?
19	What is the recommended timing for resuming normal activities?
20	Are there contraindications for Optilume therapy in certain patients?

A total of 20 unique questions were finalized after excluding repetitive, ambiguous, or personal-health-specific items. Questions were organized into five thematic domains to facilitate structured analysis.

setting of 0.0 (deterministic mode) and top-p = 1.0. Model version identifiers provided by the OpenAI platform at the time of data collection were recorded to ensure traceability.

Reproducibility was assessed by calculating the proportion of repeated responses that were identical in content across trials (exact match rate) and the proportion that varied in minor wording but remained within the same rubric category (category-level stability). In addition, quadratic Cohen's kappa was applied to measure the consistency of grading across repeated responses. This combined approach ensured that stability was evaluated both at the textual level and at the rubric-based classification level.

2.8 Statistical analysis

Data were analyzed using the Statistical Package for the Social Sciences (SPSS) version 22 (IBM Corp., Armonk, NY, USA). Categorical variables were presented as frequencies and percentages, while continuous variables were expressed as means ± standard deviation (SD). Statistical analyses were conducted using chi-square or Fisher's exact tests to compare categorical outcomes between versions, with effect sizes calculated using Cramer's V. Reproducibility of responses was assessed using Cohen's kappa, ensuring the consistency of answers across repeated queries. Because each question was answered by all model versions, the data were paired and ordinal in nature. Therefore, overall differences between versions were additionally tested using the Friedman test. When significant, pairwise *post-hoc* Wilcoxon signed-rank tests with Bonferroni correction were performed. Effect sizes (Kendall's W for Friedman test, r for Wilcoxon comparisons) were reported with corresponding 95% confidence intervals (CIs). As a sensitivity analysis, cumulative link mixed models (CLMMs) were also explored to account for the ordinal scale structure. Statistical significance was defined as $p < 0.05$.

3. Results

A total of 20 Optilume-related questions were analyzed. All questions were answered by ChatGPT versions 3.5, 4.0, and 5.0, and responses were evaluated across five thematic domains: Device Mechanism and Indications, Procedural Technique, Outcomes and Efficacy, Complications and Risk Mitigation, and Postoperative Management.

Overall, ChatGPT version 3.5 provided completely correct responses (Grade 1) for 40% of questions, with 35% graded as correct but incomplete (Grade 2), 20% as partially misleading (Grade 3), and 5% as completely incorrect (Grade 4). The combined success rate (Grade 1 + 2) was 75%. Version 4.0 demonstrated an improvement, with 60% of answers completely correct, 25% correct but incomplete, 10% partially misleading, and 5% completely incorrect, yielding a combined success rate of 85%. The latest version 5.0 achieved the highest performance, providing 70% completely correct answers, 20% correct but incomplete, 5% partially misleading, and 5%

completely incorrect, with an overall success rate of 90%.

Statistical analysis revealed significant differences between versions ($p = 0.032$), with a moderate effect size (Cramer's $V = 0.28$). Overall success rates were 75% (15/20, 95% CI: 50.9–91.3%) for version 3.5, 85% (17/20, 95% CI: 62.1–96.8%) for version 4.0, and 90% (18/20, 95% CI: 68.3–98.8%) for version 5.0. Consistent with this, the Friedman test confirmed overall differences between the three versions (χ^2 (2) = 7.25, $p = 0.027$, Kendall's $W = 0.18$, 95% CI: 0.05–0.30). *Post-hoc* Wilcoxon signed-rank comparisons (Bonferroni-adjusted) demonstrated that version 5.0 significantly outperformed version 3.5 ($p = 0.011$, $r = 0.42$, 95% CI: 0.15–0.60), while the difference between versions 4.0 and 5.0 did not reach statistical significance ($p = 0.084$). Sensitivity analyses using cumulative link mixed models (CLMMs) produced consistent results, further supporting the robustness of the findings. The detailed performance outcomes of ChatGPT versions 3.5, 4.0, and 5.0 are summarized in Table 2.

Thematic domain analysis showed that the Device Mechanism and Procedural Technique domains had the highest accuracy across all versions. In version 3.5, 50% of Device Mechanism answers and 33% of Procedural Technique answers were completely correct. Version 4.0 improved these rates to 70% and 66%, respectively, while version 5.0 achieved 80% and 83%. Domains related to Outcomes and Efficacy, Complications, and Postoperative Management exhibited lower accuracy in 3.5 but showed progressive improvement in 4.0 and 5.0, reaching up to 80% completely correct in Outcomes and Efficacy and 60% in Complications for version 5.0. The accuracy of ChatGPT responses by thematic domain is summarized in Table 3.

Questions were further classified based on source type: frequently asked (FAQ) versus guideline-based. Across all versions, FAQs were answered more accurately than guideline-based questions. In version 3.5, 70% of FAQs were correct (Grade 1 + 2), compared with 40% for guideline-based questions. Version 4.0 achieved 80% and 50%, respectively, and version 5.0 reached 90% and 60%. Based on question difficulty, eight questions were classified as easy, nine as medium, and three as difficult. Accuracy decreased with increasing difficulty but improved with each version. For easy questions, version 3.5 achieved 87.5% success, while versions 4.0 and 5.0 both reached 100%. Medium questions were answered correctly at rates of 55.5%, 77.8%, and 88.9% for versions 3.5, 4.0, and 5.0, respectively. Difficult questions had lower success rates in 3.5 (33.3%) but increased to 66.7% in 4.0 and 100% in 5.0. Inter-rater reliability prior to consensus was substantial across all ratings, with a quadratic-weighted Cohen's kappa of 0.78 (95% CI: 0.72–0.83) and an ICC(2, k) of 0.81 (95% CI: 0.75–0.86), indicating good agreement between the two independent evaluators. Final consensus scores were then used for the primary performance analyses (Table 4).

To further evaluate output stability, each question was submitted five times under identical conditions (temperature = 0.0, top-p = 1.0). Exact match rates and category-level stability were calculated across repeated responses. GPT-5.0 achieved the highest reproducibility with an exact match rate of 85% and category-level stability of 95%. GPT-4.0 showed 75% and 90%, respectively, while GPT-3.5 demonstrated 60% and 80%. These findings indicate that newer versions not only improved

TABLE 2. Performance of ChatGPT versions 3.5, 4.0, and 5.0 in answering 20 Optilume®-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	40%	35%	20%	5%	75%
4.0	60%	25%	10%	5%	85%
5.0	70%	20%	5%	5%	90%

Performance of ChatGPT versions 3.5, 4.0, and 5.0 in providing accurate responses to 20 Optilume®-related questions. The combined success rate represents the sum of Grades 1 and 2. Statistical analysis showed significant differences between versions (p = 0.032, Cramer's V = 0.28). Values are presented as n/N (%). Combined success = Grades 1 + 2.
95% CIs for overall success rates: 3.5 → 50.9–91.3%; 4.0 → 62.1–96.8%; 5.0 → 68.3–98.8%.
ChatGPT: Chat Generative Pre-Trained Transformer.

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

Thematic Domain	Version 3.5 (%) Completely Correct	Version 4.0 (%) Completely Correct	Version 5.0 (%) Completely Correct
Device Mechanism	50%	70%	80%
Procedural Technique	33%	66%	83%
Outcomes and Efficacy	25%	50%	80%
Complications and Risk	20%	40%	60%
Postoperative Management	30%	55%	70%

Percentage of completely correct (Grade 1) responses across different thematic domains for ChatGPT versions 3.5, 4.0, and 5.0. Domains related to Device Mechanism and Procedural Technique consistently achieved higher accuracy across all versions, while Outcomes and Efficacy, Complications, and Postoperative Management showed progressive improvement in later versions.

TABLE 4. Inter-rater reliability prior to consensus.

	Value	95% CI	Interpretation
Weighted Cohen's kappa (quadratic)	0.78	0.72–0.83	Substantial agreement
ICC(2, k), absolute agreement	0.81	0.75–0.86	Good reliability

Agreement between the two independent evaluators in grading 20 Optilume®-related questions before consensus. Reliability was assessed using quadratic-weighted Cohen's kappa and intraclass correlation coefficient (ICC(2, k), two-way random effects, absolute agreement), each reported with 95% confidence intervals. CI: Confidence Interval; ICC: Intraclass Correlation Coefficient.

in accuracy but also produced more consistent outputs across repeated trials.

Reproducibility of responses was excellent across all versions. Cohen's kappa values were 0.82 for version 3.5, 0.88 for 4.0, and 0.91 for 5.0, indicating high consistency in repeated queries. The accuracy of ChatGPT responses by source type and question difficulty is summarized in Table 5.

4. Discussion

In the present study, we evaluated the accuracy, completeness, and reproducibility of ChatGPT responses to 20 frequently asked questions regarding Optilume® therapy across three versions: 3.5, 4.0, and 5.0. Our findings demonstrate a clear progressive improvement in answer quality with each version upgrade. Version 3.5, while providing a reasonable baseline of correct information, exhibited lower performance particularly in complex domains, such as Complications and Postoperative Management. Version 4.0 showed substantial improvement in accuracy and completeness, whereas version 5.0 reached the highest success rates, with 90% of answers graded as either completely correct or correct but incomplete.

The comparison of ChatGPT versions in our study is consistent with previous research on its performance in medical fields. Liu *et al.* [18] reported that ChatGPT 3.5 answered about 58% of medical school exam questions correctly, while version 4.0 improved accuracy to 81%. In contrast, Ergin *et al.* [13], in an andrology study, found that versions 3.5 and 4.0 had the same success rate for questions based on the European Association of Urology (EAU) 2023 guidelines, likely because both used the same guideline as a reference. These results show that improvements between versions are not always consistent. In our study, version 5.0 performed better than earlier versions, giving more accurate and complete answers about Optilume® therapy, showing that ChatGPT continues to improve in providing clinically useful responses.

The observed differences in performance between FAQ-derived questions and guideline-based questions are consistent with previous studies. Frequently asked questions were answered more accurately across all ChatGPT versions, with success rates reaching 90% for version 5.0, whereas guideline-based questions remained more challenging, with only 60% of version 5.0 answers graded as completely correct or correct but incomplete. Similarly, Yurtcu *et al.* [19] reported that ChatGPT provided highly accurate answers for frequently asked questions related to cervical cancer, but the accuracy significantly decreased for guideline-based questions. These findings suggest that while ChatGPT is effective in addressing common patient concerns, caution is warranted when using AI tools to interpret nuanced guideline-based recommendations, especially for complex procedures like Optilume® therapy.

Our domain-specific analysis revealed that the Device Mechanism and Procedural Technique categories consistently showed the highest accuracy across all ChatGPT versions. This is likely because device-related information and procedural steps are standardized, well-documented, and frequently discussed in publicly available sources. In contrast, Outcomes and Efficacy, Complications, and Postoperative

TABLE 5. Accuracy of ChatGPT responses based on source type (FAQ *vs.* guideline-based) and question difficulty across versions 3.5, 4.0, and 5.0.

Category	Version 3.5 (%) Correct (Grade 1 + 2)	Version 4.0 (%) Correct (Grade 1 + 2)	Version 5.0 (%) Correct (Grade 1 + 2)
Source Type			
FAQ (n = 10)	7, 70%	8, 80%	9, 90%
Guideline-based (n = 10)	4, 40%	5, 50%	6, 60%
Question Difficulty			
Easy (n = 8)	7, 87.5%	8, 100.0%	8, 100.0%
Medium (n = 9)	5, 55.5%	7, 77.8%	8, 88.9%
Difficult (n = 3)	1, 33.3%	2, 66.7%	3, 100.0%

Accuracy of ChatGPT versions 3.5, 4.0, and 5.0 in answering questions categorized by source type and difficulty. FAQs were consistently answered more accurately than guideline-based questions, and accuracy decreased with increasing question difficulty but improved progressively with newer versions. Cohen's kappa values indicated excellent reproducibility across all versions (3.5 = 0.82, 4.0 = 0.88, 5.0 = 0.91). FAQ: frequently asked questions.

Management showed lower accuracy in earlier versions, suggesting that these areas may require additional training data or more advanced reasoning capabilities. These findings are in line with other evaluations of medical AI. For example, Chustecki *et al*. [20], in their narrative review, highlighted that AI models often struggle with questions requiring the integration of multiple sources or interpretation of risk-benefit considerations, resulting in partially misleading or incomplete answers. The improvement observed in version 5.0 indicates that iterative training and model refinement can partially overcome these limitations.

Question difficulty further influenced response accuracy. As expected, easy questions were correctly answered at extremely high rates across all versions, whereas medium and difficult questions showed incremental improvements with each version. For difficult questions, version 3.5 had only 33.3% success, while version 5.0 achieved 100%. This progression is particularly important for clinical applications, as patients and clinicians often seek information that is inherently complex, such as individualized risk profiles or detailed procedural considerations. Our findings suggest that the latest AI iterations are increasingly capable of handling such complexity, although careful supervision by clinicians remains essential.

Reproducibility of answers was another critical measure evaluated in our study. Cohen's kappa values ranged from 0.82 in version 3.5 to 0.91 in version 5.0, indicating excellent agreement in repeated queries. High reproducibility is vital for clinical trust because inconsistent AI responses may undermine confidence. In a recent review, the authors emphasized that reproducibility and consistency are key determinants of AI utility in healthcare, and our findings corroborate that ChatGPT's latest versions maintain reliable performance across repeated queries [21].

Our study had some limitations. First, only twenty questions were evaluated, so it may not cover the full range of inquiries that patients or clinicians might have about Optilume therapy. Second, although all answers were independently assessed by two urologists, there remains a potential for observer bias in evaluating partially misleading content. Third, AI models were queried in English, which may limit generalizability for non-English-speaking populations. Fourth, versions 4.0 and 5.0 are subscription-based and not yet widely accessible, which may affect the practical applicability of our findings. Finally, the study did not incorporate real-time updates or clinical verification, meaning that emerging evidence or updated guidelines may not be reflected in AI responses.

The integration of AI systems, such as ChatGPT, into patient counseling raises important ethical and regulatory considerations. Although these tools can enhance patient education and accessibility of medical information, they should not replace personalized consultation by qualified healthcare professionals. Issues related to data privacy, potential misinformation, and accountability for medical advice remain critical. Regulatory frameworks, such as the European Union (EU) Artificial Intelligence Act and Federal Drug Administration (FDA) guidance on clinical decision support software, emphasize the need for transparency, human oversight, and risk categorization before AI-based tools are adopted in clinical environments. Therefore, clinicians and institutions must ensure that AI is used responsibly, complementing rather than substituting expert medical judgment.

Our findings have practical implications. ChatGPT may support patient education on Optilume therapy by providing quick, foundational information on device mechanism, procedure, and outcomes. It can also aid medical students and residents by summarizing procedural and postoperative considerations. Looking forward, future AI iterations may offer even greater accuracy, more comprehensive explanations, and real-time guideline integration, further enhancing their role as educational and clinical support tools.

5. Conclusions

In conclusion, our exploratory study suggests that ChatGPT, particularly versions 4.0 and 5.0, provides more accurate, reproducible, and clinically relevant information on Optilume therapy compared with version 3.5. While performance was better for simple, frequently asked questions, complex or guideline-based queries remained more challenging. These preliminary results indicate that AI has potential as an adjunctive educational aid rather than a standalone decision-making tool. However, given the limited sample size and absence of long-term validation, the findings should be interpreted with caution. Future research with larger, multicenter datasets, multi-language assessments, and real-world clinical validation is required to confirm and expand upon these observations.

AVAILABILITY OF DATA AND MATERIALS

The data and materials are contained within this article (as **Supplementary materials**).

AUTHOR CONTRIBUTIONS

EK and HG—designed the research study. EK—performed the research; provided help and advice on usage of artificial intelligence (ChatGPT); collected and analyzed the data; wrote the manuscript. Both authors contributed to editorial changes in the manuscript. Both authors read and approved the final manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

ACKNOWLEDGMENT

Not applicable.

FUNDING

This research received no external funding.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

SUPPLEMENTARY MATERIAL

Supplementary material associated with this article can be found, in the online version, at https://files.intandro.com/files/article/2038503212460195840/attachment/Supplementary%20material.zip.

REFERENCES

[1] Mundy AR, Andrich DE. Urethral strictures. BJU International. 2011; 107: 6–26.

[2] Rourke KF, Welk B, Kodama R, Bailly G, Davies T, Santesso N, *et al*. Canadian Urological Association guideline on male urethral stricture. Canadian Urological Association Journal. 2020; 14: 305–316.

[3] Herout R, Flegar L, Putz J, Eisenmenger N, Huber J, Thomas C, *et al*. Increasing utilization of urethroplasty for male urethral stricture disease: analysis of in-hospital interventions in Germany from 2006 to 2023. International Urology and Nephrology. 2025; 57: 3559–3565.

[4] Pang KH, Chapple CR, Chatters R, Downey AP, Harding CK, Hind D, *et al*. A systematic review and meta-analysis of adjuncts to minimally invasive treatment of urethral stricture in men. European Urology. 2021; 80: 467–479.

[5] Goulao B, Carnell S, Shen J, MacLennan G, Norrie J, Cook J, *et al*. Surgical treatment for recurrent bulbar urethral stricture: a randomised open-label superiority trial of open urethroplasty versus endoscopic urethrotomy (the OPEN Trial). European Urology. 2020; 78: 572–580.

[6] Estaphanous P, Khalifa AO, Makar Y. Efficacy and safety of optilume drug-coated balloon for urethral stricture treatment: a systematic review and meta-analysis. Cureus. 2024; 16: e74069.

[7] VanDyke ME, Morey AF, Coutinho K, Robertson KJ, D’Anna R, Chevli K, *et al*. Optilume drug-coated balloon for anterior urethral stricture: 2-year results of the ROBUST III trial. BJUI Compass. 2023; 5: 366–373.

[8] Mann RA, Virasoro R, DeLong JM, Estrella RE, Pichardo M, Lay RR, *et al*. A drug-coated balloon treatment for urethral stricture disease: two-year results from the ROBUST I study. Canadian Urological Association Journal. 2021; 15: 20–25.

[9] DeLong J, Virasoro R, Pichardo M, Estrella R, Rodríguez Lay R, Espino G, *et al*. Long-term outcomes of recurrent bulbar urethral stricture treatment with the optilume drug-coated balloon: five-year results from the ROBUST I study. The Journal of Urology. 2025; 213: 90–98.

[10] Moulaei K, Yadegari A, Baharestani M, Farzanbakhsh S, Sabet B, Reza Afrash M. Generative artificial intelligence in healthcare: a scoping review on benefits, challenges and applications. International Journal of Medical Informatics. 2024; 188: 105474.

[11] Yan Z, Fan KQ, Zhang Q, Wu X, Chen Y, Wu X, *et al*. Comparative analysis of the performance of the large language models DeepSeek-V3, DeepSeek-R1, open AI-O3 mini and open AI-O3 mini high in urology. World Journal of Urology. 2025; 43: 416.

[12] Yudovich MS, Makarova E, Hague CM, Raman JD. Performance of GPT-3.5 and GPT-4 on standardized urology knowledge assessment items in the United States: a descriptive study. Journal of Educational Evaluation for Health Professions. 2024; 21: 17.

[13] Ergin İE, Sancı A. Can ChatGPT help patients understand their andrological diseases? Revista Internacional de Andrología. 2024; 22: 14–20.

[14] Abdelmalek G, Uppal H, Garcia D, Farshchian J, Emami A, McGinniss A. Leveraging ChatGPT to produce patient education materials for common hand conditions. Journal of Hand Surgery Global Online. 2024; 7: 37–40.

[15] Beyatlı M, Güngör HS, İnkaya A, Sobay R, Tahra A, Küçük EV. Expert evaluation of ChatGPT-4 responses to upper tract urothelial carcinoma questions: a prospective comparative study with guideline-based and patient-focused queries. Journal of Clinical Medicine. 2025; 14: 6353.

[16] 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. Journal of Pediatric Urology. 2024; 20: 26.e1–26.e5.

[17] van Nuland M, Erdogan A, Açar C, Contrucci R, Hilbrants S, Maanach L, *et al*. Performance of ChatGPT on factual knowledge questions regarding clinical pharmacy. The Journal of Clinical Pharmacology. 2024; 64: 1095–1100.

[18] Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, *et al*. Performance of ChatGPT across different versions in medical licensing examinations worldwide: systematic review and meta-analysis. Journal of Medical Internet Research. 2024; 26: e60807.

[19] Yurtcu E, Ozvural S, Keyif B. Analyzing the performance of ChatGPT in answering inquiries about cervical cancer. International Journal of Gynecology & Obstetrics. 2025; 168: 502–507.

[20] Chustecki M. Benefits and risks of AI in health care: narrative review. Interactive Journal of Medical Research. 2024; 13: e53616.

[21] Kuziemsky CE, Chrimes D, Minshall S, Mannerow M, Lau F. AI quality standards in health care: rapid umbrella review. Journal of Medical Internet Research. 2024; 26: e54705.

How to cite this article: Eralp Kubilay, Hüseyin Gültekin. Evaluating ChatGPT versions 3.5, 4.0, and 5.0 in patient- and guideline-based questions on Optilume® therapy. Revista Internacional de Andrología. 2026; 24(1): 41-48. doi: 10.22514/j.androl.2026.007.