

Are AI Chatbots Ready for HoLEP Patient Education? A Comparative Evaluation of ChatGPT, Gemini, and Claude

 Engin Denizhan Demirkıran¹

¹ Department of Urology, Zonguldak Bülent Ecevit University, School of Medicine, Zonguldak, Türkiye

*Correspondence: Engin Denizhan Demirkıran; eddemirkiran@gmail.com

Cite as: Demirkıran ED. Are AI Chatbots Ready for HoLEP Patient Education? A Comparative Evaluation of ChatGPT, Gemini, and Claude. J Cukurova Anesth Surg. 2026;9(3):840-850. doi:10.36516/jocass.1974413

Received: 19.06.2026, Accepted: 20.09.2026, Published: 30.09.2026, CC BY-NC 4.0

Abstract

Objective: To compare the quality, guideline concordance, and readability of HoLEP-related patient education responses generated by ChatGPT, Gemini, and Claude. **Materials and Methods:** Thirty standardized patient questions on holmium laser enucleation of the prostate (HoLEP) were submitted independently to three conversational AI platforms, generating 90 responses. Response quality was evaluated using the 16-item DISCERN instrument, and clinical accuracy was assessed with question-specific four-item guideline-concordance checklists. Readability was measured using seven validated formulas. Platform differences were analyzed using the Friedman test and Bonferroni-corrected Wilcoxon signed-rank tests. **Results:** Significant differences were found among platforms for DISCERN scores, checklist performance, and all readability metrics. Claude achieved the highest mean DISCERN and checklist scores, followed by ChatGPT, whereas Gemini scored significantly lower. Claude also produced the most readable responses across all formulas. However, all three platforms exceeded the recommended sixth to eighth-grade reading level for patient education materials. **Conclusion:** Claude and ChatGPT provided higher-quality and more guideline-concordant HoLEP-related information than Gemini, while Claude showed the best readability. Nevertheless, all platforms remained above recommended patient readability levels and produced only poor-to-moderate DISCERN scores. Conversational AI may support HoLEP patient education, but clinician review, simplification, and guideline-based verification remain necessary.

Keywords: artificial intelligence; ChatGPT; Claude; Gemini; HoLEP; patient education; DISCERN; readability

1. Introduction

Benign prostatic hyperplasia (BPH) is among the most common urological conditions affecting aging men, and a substantial proportion of affected men ultimately require surgical intervention when symptoms prove refractory to medical therapy.¹ Holmium laser enucleation of the prostate (HoLEP) has emerged as a surgical option effective across the full range of prostate sizes, including glands exceeding 100 g, with five-year randomized trial data demonstrating outcomes comparable to open simple prostatectomy alongside a more favorable perioperative morbidity profile, and the procedure is included among the

options endorsed in the American Urological Association's guideline on the surgical management of LUTS attributed to BPH.^{2,3} Despite this growing endorsement, HoLEP carries technical and recovery-related nuances that require careful, accurate communication with patients during the surgical decision-making process, including a recognized surgeon learning curve estimated at approximately 25 to 50 cases, a comparatively elevated rate of transient postoperative stress incontinence, and a high likelihood of retrograde ejaculation.^{1,4}

Comprehensive preoperative counseling is therefore central to informed consent and realistic expectation-setting for HoLEP, particularly given that its perioperative course differs in clinically meaningful ways from more familiar alternatives such as transurethral resection of the prostate. Recent urology-specific literature has highlighted a growing role for decision aids and emerging digital tools, including conversational AI platforms such as ChatGPT, in supplementing traditional consent and counseling processes, and a 2026 systematic review specifically examined how artificial intelligence tools are being used to support patient communication and shared decision-making across urological practice.^{5,6}

Conversational artificial intelligence (AI) tools, including OpenAI's ChatGPT, Google's Gemini, and Anthropic's Claude, have rapidly become a common point of contact for patients seeking health information, often consulted before, after, or even instead of a clinical visit.⁶ Because these platforms now operate with live web-search capability and conversational fluency, patients can pose detailed, personalized questions about a planned surgical procedure and receive an immediate, seemingly authoritative answer.

Prior evaluations of large language models in surgical and urological patient education have largely compared successive versions of a single platform, such as ChatGPT 3.5 versus ChatGPT 4.0 in the context of spine surgery patient education, or a general-purpose model against a bespoke, guideline-fine-tuned system developed for a specific orthopedic condition, rather than the major commercially available conversational AI platforms against one another. As patients now move fluidly between these platforms, often without awareness that response quality may differ substantially by provider, a direct multi-platform comparison specific to a technically nuanced surgical procedure such as HoLEP is needed.

This study aimed to compare ChatGPT, Gemini, and Claude in their responses to a standardized set of patient questions about HoLEP, spanning indication/diagnosis, procedural mechanics, risks/complications, and postoperative/functional outcomes. Response quality was assessed using the DISCERN instrument and a guideline-concordance checklist developed from current BPH/HoLEP clinical guidelines, while readability was evaluated using seven validated formulas. We hypothesized that the three platforms would differ in both accuracy and readability, with category-specific weaknesses most likely to emerge in technically dense or psychosocially sensitive domains such as perioperative risk communication.

2. Materials and Methods

2.1. Study design and overview

This study was designed as a cross-sectional comparative content analysis evaluating the accuracy, guideline concordance, and readability of patient education responses generated by three publicly available, general-purpose conversational AI platforms: ChatGPT (OpenAI), Gemini (Google), and Claude (Anthropic). All three platforms were accessed in their most recent publicly available consumer configuration, with live web-search functionality enabled, to reflect the default experience of a patient consulting these tools outside a research setting. The exact model name, version identifier, and release date were recorded for each platform at the time of data collection and are reported in Table 1.

2.2. Question development

Thirty standardized patient questions regarding holmium laser enucleation of the prostate (HoLEP) were compiled from outpatient urology consultations and common patient-facing queries on HoLEP. Questions were screened for direct relevance to HoLEP, suitability for general patient education, and non-redundancy, then organized into four categories reflecting the stages of a typical patient's decision-making and recovery pathway: indication/diagnosis (n=8), procedural mechanics (n=8), risks/complications (n=8), and postoperative/functional outcomes (n=6). All questions were written in English at approximately an eighth-grade reading level. The complete question set, together with a four-item guideline-concordance checklist developed for each question, is provided in Supplementary Material.

2.3. Data collection procedure

Data were collected between May 30 and June 10, 2026. The 30 questions were submitted to each platform in the same predefined order as independent, single-turn prompts. Each platform was configured as follows: ChatGPT (OpenAI GPT-5.5, reasoning effort set to "high"), Gemini (Google Gemini 3.5 Flash, extended thinking enabled), and Claude (Anthropic Claude Opus 4.8, effort set to "max"). All three platforms were accessed through subscription-based consumer versions with web-search functionality enabled and identical prompt wording, with no conversational history retained between questions. Each response was exported verbatim into a structured spreadsheet, and inline citations, hyperlinks, or source markers generated by the platforms' web-search features were removed prior to readability analysis to avoid distortion of sentence- and word-count-based metrics.

Table 1. AI platform specifications, access period, and access type at the time of data collection.

Platform	Model	Configuration	Release date	Access details	Web search
ChatGPT	GPT-5.5 (Thinking)	Reasoning effort: high	April 23, 2026	May 30–June 10, 2026; subscription-based	Enabled
Gemini	Gemini 3.5 Flash	Extended thinking	May 19, 2026	May 30–June 10, 2026; subscription-based	Enabled
Claude	Claude Opus 4.8	Effort: max	May 28, 2026	May 30–June 10, 2026; subscription-based	Enabled

2.4. Evaluation instruments

Response quality and reliability were assessed using two complementary instruments. First, the DISCERN instrument, a validated 16-item tool for judging the quality of written consumer health information on treatment choices, was applied to each response (score range 16–80).⁹ DISCERN scores were interpreted using the following six-category classification: very poor (<40), poor (40–49), moderate (50–59), good (60–69), very good (70–79), and excellent (80), consistent with prior AI-generated patient education studies.⁸ Second, a guideline-concordance checklist was developed specifically for this study, comprising four essential content elements per question derived from current evidence-based guidelines on the surgical management of LUTS/BPH, including the AUA Guideline on Surgical Management of LUTS Attributed to BPH and the 2026 EAU Guidelines on the Management of Non-neurogenic Male LUTS, supplemented by the HoLEP-specific literature cited in the Introduction.^{2,10} Each guideline-concordance checklist item was scored dichotomously as 0 = not met or 1 = met, yielding a total checklist score ranging from 0 to 4 for each response. Readability was assessed using seven validated English-language formulas: Flesch-Kincaid Grade Level,¹¹ Flesch Reading Ease,¹² SMOG Index,¹³ Gunning

Fog Index,¹⁴ Fry Readability Graph¹⁵, Raygor Estimate,¹⁶ and the FORCAST formula.¹⁷ All seven formulas were computed programmatically using Python 3.12 (scipy 1.13); in-line hyperlinks and reference markers were stripped from each response prior to calculation to prevent distortion of word- and syllable-count-based metrics. Because Fry, Raygor, and FORCAST formulas were originally developed for longer text samples, values obtained from short FAQ-style responses were interpreted as algorithmic readability estimates rather than strict grade-level assignments.

2.5. Scoring protocol and reliability

All 90 responses (30 questions × 3 platforms) were anonymized, assigned a random identifier, and presented in a fully randomized order for scoring; therefore, the evaluator was blinded to platform identity during DISCERN and checklist assessment. DISCERN and checklist scores were assigned in a single sitting (T0) by the corresponding author, a board-certified urologist with subspecialty experience in endourological surgery. To assess intra-rater reliability, the complete response set was re-anonymized, re-randomized, and re-scored for DISCERN alone, blind to T0 scores, approximately two weeks after the initial rating session (T1). Intra-rater reliability was quantified using the intraclass correlation coefficient ICC(3,1) (two-way mixed model, consistency definition). ICC values were interpreted as: poor (<0.50), moderate (0.50–0.75), good (0.75–0.90), and excellent (>0.90).¹⁸

2.6. Statistical analysis

Continuous variables are presented as mean ± standard deviation. Because three platforms were compared across 30 paired questions, between-group differences in DISCERN score, checklist score, and each readability metric were assessed using the Friedman test. Pairwise post-hoc comparisons were performed using the Wilcoxon signed-rank test with Bonferroni correction for three comparisons (corrected $\alpha = 0.017$). Effect sizes were expressed as $r = Z/\sqrt{N}$, where Z is the standardized test statistic and N is the number of paired observations ($N = 30$). The same approach was applied to subgroup analyses by question category. Friedman and Wilcoxon analyses were performed using jamovi (Version 2.7.31; The jamovi project, Sydney, Australia); bootstrap ICC estimation was performed using Python 3.12 (scipy 1.13). Statistical significance was set at $p < 0.05$.

3. Results

3.1. Intra-rater reliability

Intra-rater reliability for DISCERN scoring was assessed by re-evaluating all 90 anonymized and re-randomized responses approximately two weeks after the initial rating session (T1), using ICC(3,1) (two-way mixed, consistency model) as appropriate for single-rater repeated measurement designs. Overall intra-rater reliability was good (ICC(3,1) = 0.876, 95% bootstrap CI [0.825, 0.913]). Agreement was good across all three platforms: ChatGPT (ICC = 0.775, 95% CI [0.535, 0.881]), Claude (ICC = 0.841, 95% CI [0.693, 0.927]), and Gemini (ICC = 0.856, 95% CI [0.758, 0.915]).

3.2. DISCERN quality scores

Significant differences in DISCERN scores were identified across the three platforms (Friedman $\chi^2(2) = 32.123$, $p < 0.001$) (Table 2). Claude produced the highest mean DISCERN score (47.13 ± 4.31), followed closely by ChatGPT (46.67 ± 4.70), with Gemini scoring substantially lower (40.37 ± 4.80). According to established DISCERN classification thresholds, Claude responses were classified as poor in 70%, moderate in 27%, and very poor in 3% of cases, while ChatGPT responses were classified as poor in 70%, moderate in 23%, and very poor in 7% of cases. Gemini responses were distributed between poor (53%) and very poor (47%) categories, with none reaching the moderate threshold. Post-

hoc pairwise comparisons with Bonferroni correction ($\alpha = 0.017$) revealed no significant difference between Claude and ChatGPT ($W = 128$, $p = 0.351$, $r = 0.17$), while both significantly outperformed Gemini (Claude vs. Gemini: $W = 4$, $p < 0.001$, $r = 0.844$; ChatGPT vs. Gemini: $W = 21$, $p < 0.001$, $r = 0.795$) (Table 3).

Table 2. Descriptive statistics and Friedman test results for DISCERN scores, guideline-concordance checklist scores, and readability metrics across the three AI platforms (n = 30 questions per platform).

Metric	Ideal standard	ChatGPT (mean $\pm$ SD)	Gemini (mean $\pm$ SD)	Claude (mean $\pm$ SD)	Friedman $\chi^2(2)$	p value
Quality assessment						
DISCERN score (16–80)	–	46.67 $\pm$ 4.70	40.37 $\pm$ 4.80	47.13 $\pm$ 4.31	32.123	<0.001
Checklist items met (0–4)	–	3.07 $\pm$ 0.87	2.30 $\pm$ 1.21	3.40 $\pm$ 0.77	21.390	<0.001
Readability						
FKGL	≤ 8	14.56 $\pm$ 4.41	15.09 $\pm$ 3.62	12.21 $\pm$ 2.86	13.400	0.001
FRES	≥ 60	26.19 $\pm$ 20.77	21.03 $\pm$ 19.44	45.90 $\pm$ 12.03	36.600	<0.001
SMOG	≤ 8	15.45 $\pm$ 3.31	15.89 $\pm$ 3.15	13.30 $\pm$ 2.26	10.867	0.004
Gunning Fog	≤ 8	17.86 $\pm$ 4.65	18.60 $\pm$ 4.46	14.69 $\pm$ 3.12	16.467	<0.001
Raygor	≤ 8	13.80 $\pm$ 1.47	14.19 $\pm$ 1.16	12.27 $\pm$ 1.38	29.193	<0.001
Fry	≤ 8	12.63 $\pm$ 1.61	12.98 $\pm$ 1.64	10.53 $\pm$ 1.25	34.067	<0.001
FORCAST	≤ 8	12.79 $\pm$ 1.40	13.15 $\pm$ 1.21	11.42 $\pm$ 0.91	32.067	<0.001

FKGL = Flesch-Kincaid Grade Level; FRES = Flesch Reading Ease Score; SMOG = Simple Measure of Gobbledygook; FORCAST = Ford, Caylor, Sticht formula. Ideal standards reflect the 6th–8th grade reading level recommended by the AMA/NIH for patient education materials; for FRES, higher values indicate easier readability; for all other readability formulas, lower values indicate easier readability. Bold values indicate statistical significance

Category-level subgroup analysis demonstrated significant between-platform differences in DISCERN scores in the indication/diagnosis ($\chi^2(2) = 7.55$, $p = 0.023$, $n = 8$), procedure mechanics ($\chi^2(2) = 10.40$, $p = 0.006$, $n = 8$), and risks/complications ($\chi^2(2) = 9.48$, $p = 0.009$, $n = 8$) categories. Differences in the postoperative/functional outcomes category did not reach statistical significance ($\chi^2(2) = 5.55$, $p = 0.063$, $n = 6$), likely reflecting the reduced statistical power associated with the smaller number of questions in this category (Table 4).

3.3. Guideline-concordance checklist

Significant differences in checklist scores were observed across platforms (Friedman $\chi^2(2) = 21.390$, $p < 0.001$) (Table 2). Claude achieved the highest mean number of guideline-concordant content elements per response (3.40 ± 0.77 ; 85.0% of items met), followed by ChatGPT (3.07 ± 0.87 ; 76.7%), with Gemini scoring substantially lower (2.30 ± 1.21 ; 57.5%). Post-hoc analysis confirmed a significant advantage of Claude over Gemini ($W = 5$, $p < 0.001$, $r = 0.717$) and of ChatGPT over Gemini ($W = 48$, $p = 0.009$, $r = 0.474$); the difference between Claude and ChatGPT did not survive Bonferroni correction ($W = 27$, $p = 0.046$, $r = 0.364$) (Table 3).

Category-level analysis revealed significant between-platform differences in guideline concordance for procedure mechanics ($\chi^2(2) = 6.74$, $p = 0.034$) and risks/complications ($\chi^2(2) = 6.46$, $p = 0.040$), while differences in indication/diagnosis ($\chi^2(2) = 4.73$, $p = 0.094$) and postoperative/functional outcomes ($\chi^2(2) = 4.93$, $p = 0.085$) did not reach significance (Table 4). Across all categories, Gemini's checklist performance was consistently the lowest, with mean item concordance ranging from 2.00 (postoperative/functional outcomes) to 2.50 (indication/diagnosis), compared with 3.17–3.62 for Claude and 2.88–3.38 for ChatGPT.

3.4. Readability

Significant between-platform differences were detected across all seven readability metrics (all Friedman $p \leq 0.005$; Table 2). Claude produced the most readable responses by every measure (FKGL 12.21 ± 2.86 ; FRES 45.90 ± 12.03 ; SMOG 13.30 ± 2.26 ; Gunning Fog 14.69 ± 3.12 ; Raygor 12.27 ± 1.38 ; Fry 10.53 ± 1.25 ; FORCAST 11.42 ± 0.91), followed by ChatGPT (FKGL 14.56 ± 4.41 ; FRES 26.19 ± 20.77 ; SMOG 15.45 ± 3.31 ; Gunning Fog 17.86 ± 4.65), with Gemini producing the least readable responses across all formulas (FKGL 15.09 ± 3.62 ; FRES 21.03 ± 19.44 ; SMOG 15.89 ± 3.15 ; Gunning Fog 18.60 ± 4.46).

ChatGPT and Gemini showed marked within-platform variability in FRES scores, with certain individual responses yielding negative values (ChatGPT minimum -19.0 ; Gemini minimum -14.8), reflecting isolated responses consisting of a single long, polysyllabic-dense sentence; no Claude response produced a negative FRES value (minimum $+6.9$). Despite Claude's consistent advantage, all three platforms substantially exceeded the 6th–8th grade reading level recommended by the American Medical Association and National Institutes of Health for patient education materials across all seven metrics.

Post-hoc comparisons consistently confirmed Claude's superiority over both ChatGPT and Gemini on every readability metric (all $p < 0.001$ after Bonferroni correction, with large effect sizes: $r = 0.591$ – 1.000 for Claude vs. ChatGPT; $r = 0.657$ – 0.972 for Claude vs. Gemini). In contrast, ChatGPT and Gemini did not significantly differ from each other on any readability metric (all $p \geq 0.096$) (Table 3).

Table 3A. ChatGPT vs Claude. Post-hoc pairwise Wilcoxon signed-rank comparisons between platforms across all outcomes (Bonferroni-corrected $\alpha = 0.017$).

Metric	W	p	r
DISCERN	128	0.351	0.170
Checklist	27	0.046	0.364
FKGL	73	<0.001	0.626
FRES	2	<0.001	1.000
SMOG	75	0.001	0.591
Gunning Fog	46	<0.001	0.756
Raygor	37	<0.001	0.734
Fry	3	<0.001	0.862
FORCAST	19	<0.001	0.913

Table 3B. ChatGPT vs Gemini. Post-hoc pairwise Wilcoxon signed-rank comparisons between platforms across all outcomes (Bonferroni-corrected $\alpha = 0.017$).

Metric	W	p	r
DISCERN	21	<0.001	0.795
Checklist	48	0.009	0.474
FKGL	190	0.393	0.156
FRES	171	0.213	0.227
SMOG	187	0.360	0.167
Gunning Fog	194	0.428	0.145
Raygor	186	0.503	0.122
Fry	185	0.329	0.178
FORCAST	151	0.096	0.304

Table 3C. Claude vs Gemini. Post-hoc pairwise Wilcoxon signed-rank comparisons between platforms across all outcomes (Bonferroni-corrected $\alpha = 0.017$).

Metric	W	p	r
DISCERN	4	<0.001	0.844
Checklist	5	<0.001	0.717
FKGL	62	<0.001	0.677
FRES	11	<0.001	0.972
SMOG	58	<0.001	0.657
Gunning Fog	51	<0.001	0.731
Raygor	20	<0.001	0.798
Fry	12	<0.001	0.830
FORCAST	13	<0.001	0.956

W = Wilcoxon signed-rank statistic; r = effect size ($Z/\sqrt{N}$); ns = not significant ($p \geq 0.017$). Bold values indicate statistical significance after Bonferroni correction.

Table 4A. Category-level subgroup analysis: DISCERN score (mean)

Category	n	ChatGPT	Gemini	Claude	p
Indication / Diagnosis	8	48.5	42.8	49.6	0.023
Procedure Mechanics	8	42.4	37.5	43.2	0.006
Risks / Complications	8	48.6	41.4	48.4	0.009
Postoperative / Functional Outcomes	6	47.3	39.7	47.3	0.063

Table 4B. Category-level subgroup analysis: Checklist items met (mean)

Category	n	ChatGPT	Gemini	Claude	p
Indication / Diagnosis	8	2.88	2.50	3.50	0.094
Procedure Mechanics	8	2.88	2.38	3.25	0.034
Risks / Complications	8	3.38	2.25	3.62	0.040
Postoperative / Functional Outcomes	6	3.17	2.00	3.17	0.085

Friedman p-values reflect between-platform comparison within each question category. Postoperative/functional outcomes category contained 6 questions (n = 6); all other categories contained 8 questions (n = 8). Bold values indicate statistical significance.

4. Discussion

In this cross-sectional comparative analysis of three widely available conversational AI platforms, we found meaningful differences in the quality, guideline concordance, and readability of responses to patient-oriented HoLEP questions. Claude achieved the highest mean DISCERN and checklist scores, closely followed by ChatGPT, whereas Gemini consistently performed lower across both quality metrics. However, the difference between Claude and ChatGPT was not statistically significant for DISCERN, and the checklist advantage of Claude over ChatGPT did not remain significant after Bonferroni correction. These findings suggest that, although current large language models (LLMs) can generate broadly acceptable patient-facing information on HoLEP, performance is platform-dependent and remains insufficiently reliable for uncensored use in surgical counseling.

The observed DISCERN scores should be interpreted in the context of the instrument itself. DISCERN was developed to evaluate the quality of written consumer health information on treatment choices, with particular emphasis on aims, source transparency,

balance, uncertainty, alternatives, and support for shared decision-making.⁹ In the present study, Claude and ChatGPT generally achieved scores in the poor-to-moderate range, while Gemini clustered toward the lower end of this spectrum. This does not necessarily mean that all responses were clinically incorrect. Rather, it indicates that AI-generated answers often lacked features expected from high-quality patient education material, such as explicit citation of evidence, balanced discussion of alternative options, clarification of uncertainty, and explanation of what may happen without treatment. Similar concerns have been raised in urology-focused evaluations of AI patient education, where LLM outputs were often fluent and plausible but required expert oversight before being considered suitable for patient use.¹⁹⁻²²

The checklist findings provide a more procedure-specific perspective. Claude and ChatGPT included most of the guideline-concordant content elements, whereas Gemini omitted a larger proportion of key items. This gap was most evident in procedure mechanics and risk/complication domains, which are particularly important in HoLEP counseling. HoLEP is technically distinct from conventional TURP because it involves anatomical enucleation, morcellation, and treatment across a wide range of prostate sizes; therefore, incomplete or vague explanations may lead to misunderstanding of both the procedural mechanism and expected recovery.^{2-4,10} The category-level results reinforce this point: between-platform differences were significant for DISCERN in indication/diagnosis, procedure mechanics, and risks/complications, but not for postoperative/functional outcomes. One possible explanation is that postoperative issues such as catheter duration, return to activity, hematuria, urgency, incontinence, and retrograde ejaculation are commonly described in online HoLEP materials, whereas accurate explanations of indication thresholds, guideline-based work-up, surgical anatomy, and risk trade-offs require more specialized synthesis.

Our findings are consistent with emerging urological literature showing both promise and limitations of LLMs in patient communication. Paluszek and Loeb recently reviewed the use of AI chatbots for urological patient education and emphasized that these systems may improve access to information but should not replace clinician-validated counseling.²² Connors et al. compared ChatGPT and Google Bard with traditional endourology patient education resources and similarly highlighted the need to evaluate AI outputs against established educational standards rather than assuming accuracy from linguistic fluency alone.²¹ Shah et al. found that ChatGPT-generated men's health education materials could be competitive with traditional resources but still raised concerns regarding readability and clinical completeness.²⁰ More recently, Morrison et al. specifically evaluated whether ChatGPT responses to urologic inquiries were readable and supported by AUA guidelines, further underscoring the relevance of guideline concordance as an outcome measure in this field.¹⁹ Our study extends this literature by directly comparing three contemporary commercial platforms in a single technically nuanced endourological procedure.

Readability emerged as a major limitation. Although Claude generated significantly more readable responses than both ChatGPT and Gemini across all seven formulas, all three platforms exceeded the recommended sixth to eighth-grade level for patient education materials. The mean FKGL values of 12.21 for Claude, 14.56 for ChatGPT, and 15.09 for Gemini indicate that even the best-performing platform produced text closer to late high-school or college-level readability than to standard patient-facing recommendations. This is clinically important because high-quality information is not useful if patients cannot understand it. Prior studies across urology and other medical domains have reported similar discordance between AI-generated fluency and patient-level readability.^{19,20,23,24} The finding that Claude was both more readable and more guideline-concordant suggests that readability improvement does not necessarily require loss of informational quality.

However, even Claude's outputs remained too complex for general patient education use without further simplification.

The HoLEP-specific counseling context makes readability particularly important. Patients considering HoLEP must understand not only the benefits of durable obstruction relief but also the distinctive recovery profile of the procedure. Prior HoLEP studies have shown that patients may underestimate or misunderstand the expected duration of postoperative hematuria, dysuria, urgency, urinary incontinence, and ejaculatory changes despite preoperative counseling.²⁵ Transient urinary incontinence is a known early postoperative issue, particularly during the first weeks to months, although most cases improve over time.²⁶ Storage symptoms and urinary tract infection are also relevant early postoperative concerns that may be difficult for patients to distinguish without clear guidance.²⁷ Therefore, AI-generated HoLEP information should be judged not only by whether it mentions complications, but also by whether it explains their expected time course, warning signs, and implications in accessible language.

The lack of significant difference between Claude and ChatGPT in overall DISCERN score, despite Claude's superior readability, has practical implications. In clinical practice, HoLEP counseling must also be individualized according to prostate size, anticoagulation or antiplatelet status, baseline continence and storage symptoms, sexual priorities, comorbidities, and surgeon or center experience. For patient-facing HoLEP education, the optimal output is not simply the most complete or the most readable response, but one that balances accuracy, completeness, clarity, and risk communication. A highly detailed answer may perform well on content checklists but remain unsuitable for patients with limited health literacy. Conversely, an overly simplified answer may be readable but omit essential surgical risks or alternatives. This supports a clinician-in-the-loop model in which LLMs are used to draft or supplement educational material, but outputs are reviewed, simplified, and aligned with institutional practice before distribution. It also supports the use of structured prompting, guideline-grounded retrieval systems, or locally validated templates rather than unrestricted single-turn chatbot responses.

This study has several limitations. First, responses were generated from a single time window using specific platform configurations. Because LLMs are frequently updated, the findings represent a snapshot rather than a permanent ranking of platforms. Second, only English-language responses were assessed, limiting generalizability to non-English-speaking patient populations. Third, DISCERN and checklist scoring were performed by a single urologist, although intra-rater reliability was good. Fourth, while DISCERN is a validated tool for consumer health information, its application to short, single-question chatbot answers is imperfect; previous urological AI studies have similarly noted that DISCERN may not fully capture the quality of responses to specific surgical questions.²⁸ Fifth, the guideline-concordance checklist was developed for this study and was not independently validated. Finally, this analysis evaluated textual output quality, not patient comprehension, satisfaction, behavioral impact, or actual shared decision-making outcomes.

Despite these limitations, the study has several strengths. It used a standardized set of HoLEP-specific questions, evaluated three major AI platforms under comparable conditions, combined a validated general quality instrument with procedure-specific guideline-concordance scoring, and incorporated multiple readability formulas rather than relying on a single metric. The inclusion of intra-rater reliability further strengthens the credibility of the scoring process. These features provide a clinically relevant framework for evaluating AI-generated education in technically complex surgical domains.

In conclusion, contemporary conversational AI platforms can provide partially guideline-concordant information about HoLEP, but their outputs vary substantially by platform and remain too difficult to read for routine patient education. Claude and

ChatGPT outperformed Gemini in quality and guideline concordance, while Claude demonstrated the most favorable readability profile. Nevertheless, all platforms fell short of recommended patient readability standards, and none should be used as an unsupervised substitute for surgeon-led counseling. Future work should evaluate structured prompting, retrieval-augmented guideline-based systems, multilingual outputs, and patient comprehension outcomes to determine how LLMs can be safely integrated into HoLEP counseling and shared decision-making.

5. Conclusion

Claude and ChatGPT provided more guideline-concordant and higher-quality HoLEP-related patient education responses than Gemini, while Claude achieved the best readability scores. However, all three platforms exceeded the recommended sixth to eighth-grade reading level, and overall DISCERN scores remained poor to moderate. Current conversational AI tools may support the development of patient education materials, but their outputs require clinician review, simplification, and guideline-based verification before use in HoLEP counseling or informed consent.

Acknowledgements: With the aid of artificial intelligence (Claude Opus 4.8; Anthropic), the manuscript's language and structure were improved. The author has meticulously reviewed and confirmed the content, assuming complete responsibility for the scientific accuracy and integrity of the final document.

Funding: This research received no external funding.

Conflict of interest: The authors declare no competing interests.

Ethical considerations: This study did not involve human participants, identifiable patient data, or animal subjects; all data consisted of AI-generated text responses to standardized, non-patient-specific questions, so formal ethics committee approval was not required.

Data availability: Data can be provided upon request.

References

1. Lokeshwar SD, Harper BT, Webb E, Jordan A, Dykes TA, Neal DEJ, et al. Epidemiology and treatment modalities for the management of benign prostatic hyperplasia. *Transl Androl Urol*. 2019;8:529–39. doi:10.21037/tau.2019.10.01 PubMed PMID: 31807429.
2. Lerner LB, McVary KT, Barry MJ, Bixler BR, Dahm P, Das AK, et al. Management of Lower Urinary Tract Symptoms Attributed to Benign Prostatic Hyperplasia: AUA GUIDELINE PART II-Surgical Evaluation and Treatment. *J Urol*. 2021;206:818–26. doi:10.1097/JU.0000000000002184 PubMed PMID: 34384236.
3. Kuntz RM, Lehrich K, Ahyai SA. Holmium laser enucleation of the prostate versus open prostatectomy for prostates greater than 100 grams: 5-year follow-up results of a randomised clinical trial. *Eur Urol*. 2008;53:160–6. doi:10.1016/j.eururo.2007.08.036 PubMed PMID: 17869409.
4. Kampantais S, Dimopoulos P, Tasleem A, Acher P, Gordon K, Young A. Assessing the Learning Curve of Holmium Laser Enucleation of Prostate (HoLEP). A Systematic Review. *Urology*. 2018;120:9–22. doi:10.1016/j.urology.2018.06.012 PubMed PMID: 30403609.
5. Nedbal C, Naik N, Davis N, Madaan S, Tokas T, Cacciamani G, et al. Comprehensive consent in urology using decision aids, leaflets, videos and newer technologies: empowering patient choice and shared decision-making. *Ther Adv Urol*. 2024;16:17562872241301729. doi:10.1177/17562872241301729 PubMed PMID: 39583573.
6. Bhuiyan N, Bracey S, Pietropaolo A, Jones P, Tzelves L, Somani BK. The role of artificial intelligence in improving patient communication and shared decision-making in urology: A systematic review. *Scott Med J*. 2026;71:54–67. doi:10.1177/00369330261418599 PubMed PMID: 41642909.
7. Tabanlı A, Demirkiran ND. Comparing ChatGPT 3.5 and 4.0 in Low Back Pain Patient Education: Addressing Strengths, Limitations, and Psychosocial Challenges. *World Neurosurg*. 2025;196:123755. doi:10.1016/j.wneu.2025.123755 PubMed PMID: 39952398.
8. Öner SK, Demirkiran ND, Canlı EA, Bilir A. Gonarthrosis Advisor vs ChatGPT-5: quality and readability of artificial intelligence-generated patient education for knee osteoarthritis. *Acta Orthop Traumatol Turc*. 2026;60. doi:10.5152/j.aott.2026.25618

9. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *J Epidemiol Community Health* (1978). 1999;53:105–11. doi:10.1136/jech.53.2.105 PubMed PMID: 10396471.
10. Baboudjian M, Creta M, Pyrgidis N, Moris L, Duran MB, de Nunzio C, et al. Summary Paper on the 2026 European Association of Urology Guidelines on the Management of Non-neurogenic Male Lower Urinary Tract Symptoms. *Eur Urol*. 2026. doi:10.1016/j.eururo.2026.03.020 PubMed PMID: 41997777.
11. Kincaid JP, Fishburne RP Jr, Rogers RL, Chissom BS. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. 1975. doi:10.21236/ada006655
12. Flesch R. A new readability yardstick. *J Appl Psychol*. 1948;32:221–33. doi:10.1037/h0057532 PubMed PMID: 18867058.
13. McLaughlin GH. SMOG grading: a new readability formula. *Journal of Reading*. 1969;12:639–646.
14. Gunning R. The Technique of Clear Writing [Internet]. McGraw-Hill; 1952. Available from: <https://books.google.com.tr/books?id=ofl0AAAAMAAJ>
15. Fry EA. A readability formula that saves time. *Journal of Reading*. 1968;11:513-516,575-578.
16. Raygor AL. The Raygor readability estimate: A quick and easy way to determine difficulty. *Reading: Theory, research, and practice*. 1977;259:263.
17. Caylor JS, Sticht TG, Fox LC, Ford JP. Methodologies for Determining Reading Requirements Military Occupational Specialties. Human Resources Research Organization, Alexandria, VA. 1973;7:81.
18. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med*. 2016;15:155–63. doi:10.1016/j.jcm.2016.02.012 PubMed PMID: 27330520.
19. Morrison C, Vercnocke J, Moser AM, Cher ML, Lucas S, Cochrane J, et al. Are ChatGPT's Responses to Urologic Inquiries Readable and Supported by AUA Guidelines? *International Journal of Urological Nursing*. 2025;19. doi:10.1111/ijun.70023
20. Shah YB, Ghosh A, Hochberg AR, Rapoport E, Lallas CD, Shah MS, et al. Comparison of ChatGPT and Traditional Patient Education Materials for Men's Health. *Urol Pract*. 2024;11:87–94. doi:10.1097/UPJ.0000000000000490 PubMed PMID: 37914380.
21. Connors C, Gupta K, Khusid JA, Khargi R, Yaghoubian AJ, Levy M, et al. Evaluation of the Current Status of Artificial Intelligence for Endourology Patient Education: A Blind Comparison of ChatGPT and Google Bard Against Traditional Information Resources. *J Endourol*. 2024;38:843–51. doi:10.1089/end.2023.0696
22. Paluszek O, Loeb S. Artificial intelligence and patient education. *Curr Opin Urol*. 2025;35:219–23. doi:10.1097/MOU.0000000000001267 PubMed PMID: 39945126.
23. Rodler S, Cei F, Ganjavi C, Checucci E, De Backer P, Rivero Belenchon I, et al. GPT-4 generates accurate and readable patient education materials aligned with current oncological guidelines: A randomized assessment. *PLoS One*. 2025;20:e0324175. doi:10.1371/journal.pone.0324175 PubMed PMID: 40465696.
24. Kianian R, Sun D, Rojas-Carabali W, Agrawal R, Tsui E. Large Language Models May Help Patients Understand Peer-Reviewed Scientific Articles About Ophthalmology: Development and Usability Study. *J Med Internet Res*. 2024;26:e59843. doi:10.2196/59843 PubMed PMID: 39719077.
25. Assmus MA, Lee MS, Large T, Krambeck AE. Understanding holmium laser enucleation of the prostate (HoLEP) recovery: Assessing patient expectations and understanding. *Can Urol Assoc J*. 2022;16:E25–31. doi:10.5489/cuaj.7328 PubMed PMID: 34464254.
26. Doersch KM, Hines L, Campbell TD, Jain RK, Quarrier SO. Predictors of Postoperative Urinary Incontinence After Holmium Laser Enucleation of the Prostate (HoLEP) for Surgeons Early in Their Experience. *Low Urin Tract Symptoms*. 2024;16:e12533. doi:10.1111/luts.12533 PubMed PMID: 39267354.
27. Elsaqa M, Dowd K, El Mekresh A, Doersch KM, El Tayeb MM. Predictors of postoperative urinary tract infection following holmium laser enucleation of the prostate. *Can Urol Assoc J*. 2023;17:E364–8. doi:10.5489/cuaj.8269 PubMed PMID: 37549346.
28. Warren CJ, Edmonds VS, Payne NG, Voletti S, Wu SY, Colquitt J, et al. Prompt matters: evaluation of large language model chatbot responses related to Peyronie's disease. *Sex Med*. 2024;12:qfae055. doi:10.1093/sexmed/qfae055 PubMed PMID: 39257694.