



Article

Benchmarking Large Language Model Responses Against Surgical Clinical Practice Guidelines for Chronic Rhinosinusitis: The Importance of User Prompts

Hetal Lad , Emily Kwon , Ayushi Chadha , Sean Z. Haimowitz, Brandon S. Gold, Rachel Kaye and Wayne D. Hsueh

Department of Otolaryngology—Head and Neck Surgery, Rutgers New Jersey Medical School, Newark, NJ 07103, USA; esk104@njms.rutgers.edu (E.K.); ac2489@scarletmail.rutgers.edu (A.C.); sean.haimowitz@rutgers.edu (S.Z.H.); bsg101@rutgers.edu (B.S.G.); kayera@njms.rutgers.edu (R.K.); wayne.hsueh@rutgers.edu (W.D.H.)

* Correspondence: hl1006@njms.rutgers.edu; Tel.: +1-973-972-2548

Abstract

Background/Objectives: As patients increasingly rely on large language models (LLMs) for Chronic Rhinosinusitis (CRS) diagnosis, surgical candidacy, and perioperative care, evaluating the accuracy of LLM-generated information against established clinical practice guidelines for surgical management of CRS is essential. **Methods:** ChatGPT, Google AI, Google Gemini, and Grok were queried using a 21-question guideline-mapped prompt set (long) and a single patient-focused prompt (short). Two physician reviewers independently scored responses using a 3-point rubric across 21 fields. Primary outcomes were guideline-concordant scores; secondary outcomes included readability measured with the Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL). Inter-rater reliability (IRR) was assessed using the intraclass correlation coefficient (ICC). Analyses were performed in SPSSv31. **Results:** Guideline concordance ranged from 55.36% to 77.98% ($p > 0.05$), highest for Grok (77.98%, 95% CI 63.67–92.28), followed by Google Gemini (66.67%, 95% CI 35.42–97.91), ChatGPT (55.95%, 95% CI 29.25–82.65), and Google AI (55.36%, 95% CI 29.97–80.75), with Grok significantly outperforming both ChatGPT and Google AI. Prompt structure significantly affected scores. Long-form prompting resulted in higher guideline concordance scores than short-form prompting (+26.19, $p < 0.001$). The CPG demonstrated a more readable structure, with a higher FRE (44.1), exceeding scores generated by Grok (31.7), ChatGPT (39.9), Gemini (39.2), and Google AI (32.5). In contrast, the CPG was a higher reading grade level (FKGL score of 11.7) than Grok (11.4), Gemini (10.5), and ChatGPT (10.3), but was lower in reading grade compared to Google AI, which produced the highest FKGL score (12.5). IRR was high (ICC = 0.961). **Conclusions:** LLMs demonstrated similar guideline concordance, suggesting patients can expect comparable accuracy across platforms. While LLMs generally improved FKGL scores compared to the AAO-HNS CPG, they demonstrated lower FRE scores, indicating mixed results on overall readability. However, longer prompt structure meaningfully influenced output quality, highlighting how a user's ability to frame precise prompts is critical to obtaining accurate information.

Keywords: chronic rhinosinusitis; large language models; artificial intelligence; patient education; prompt engineering; clinical practice guidelines



Academic Editor: Agnieszka Szczepiek

Received: 29 June 2026

Revised: 28 July 2026

Accepted: 28 July 2026

Published: 1 August 2026

Copyright: © 2026 by the authors.

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

Chronic rhinosinusitis (CRS) is a common inflammatory disorder affecting more than 10% of adults and remains one of the leading indications for endoscopic sinus surgery [1]. Although most patients initially receive medical therapy, surgery is recommended for those with persistent symptoms despite appropriate medical management. Advances in the understanding of CRS pathophysiology have refined surgical decision-making by incorporating disease endotypes, inflammatory patterns, and extent of sinonasal involvement into treatment planning. These developments have shifted surgical management from a standardized approach toward more individualized procedures, with the goal of improving symptom control, reducing disease recurrence, and optimizing outcomes, particularly in patients with severe disease or nasal polyposis [1]. As surgical management of CRS has become increasingly individualized, patients are also turning to online resources to better understand their diagnosis, treatment options, and expectations for surgery. Consequently, ensuring the accuracy of readily accessible medical information has become an important component of patient education.

The advent of large language models (LLMs) has transformed medicine, with applications ranging from passing radiology fellowship board exams [2] and interpreting chest X-rays [3] to generating patient communication and clinical documentation [4]. Despite these advances, concerns regarding the accuracy and reliability of LLMs remain. A 2025 study evaluating ChatGPT's management of thyroid nodules found that while responses were generally consistent with American Thyroid Association guidelines, physicians remained hesitant to rely on it given its variable accuracy [5]. Performance can also vary across platforms: Sami et al. reported that ChatGPT answered pediatric radiology questions with 83.5% accuracy compared to Gemini's 68.4% [6] and Draelos et al. reported that ChatGPT, Claude, Gemini, and Llama each provided unsafe responses to patient concerns across numerous fields [7]. Nevertheless, patients continue to engage with LLMs daily. According to a recent OpenAI report, more than 5% of all ChatGPT messages globally relate to healthcare, averaging billions of queries weekly [8]. This growing reliance emphasizes the need to rigorously evaluate how these platforms respond to patient-oriented questions. The usefulness of AI-generated information, however, also depends on whether patients can understand it. Health literacy is limited for a large share of the population, with more than half of American adults reading below a sixth-grade level [9]. Online materials frequently exceed this level, and whether AI-generated content is any more accessible remains largely uncharacterized in rhinology.

The American Academy of Otolaryngology—Head and Neck Surgery (AAO-HNS) 2025 guideline on Surgical Management of Chronic Rhinosinusitis (CRS) offers this benchmark for evaluation [10]. As patients increasingly rely on LLMs for information on CRS diagnosis, surgical candidacy, and perioperative care, assessing the concordance between AI-generated information and expert consensus is essential. This study aimed to evaluate the concordance of commonly used AI platforms against the AAO-HNS guidelines and to examine how prompt length influences response accuracy.

2. Materials and Methods

2.1. Data Collection

The 2025 AAO-HNS guidelines were chosen as the gold standard for this study, as it was developed by a task force selected for their expertise on CRS surgical management. The guidelines were accessed in October 2025. The “Plain Language Summary: Surgical Management of Chronic Rhinosinusitis” provides information to adult patients on when to consider moving forward with surgical interventions for those who are evaluated and diagnosed with CRS. 21 questions were derived from the frequently asked questions.

2.2. AI Response Generation

ChatGPT (v5.1), Google AI (v2.5), Google Gemini (v3), and Grok (v4.1) platforms, reflecting those most updated and accessible to patients at the time of collection, were queried using a 21-question guideline-mapped prompt set (long) and a single patient-focused prompt (short) (Table A1). In November 2025, both long and short prompts were queried to the most recent version of ChatGPT-5 (ChatGPT-5, released 7 August 2025), Google AI (released 14 May 2024), Google Gemini-3 (released 18 November 2025), and Grok 4.1 (released 17 November 2025) with a “new chat” to avoid bias and influence on subsequent questions within the same chat. Responses were copied and pasted into a separate blinded Word document for each platform. All text formatting was removed from these sources so as not to distinguish responses based on appearance.

2.3. Physician Assessment

For each prompt on the Word document, a guideline-based scoring rubric grading out of a 42-point score (2 points for each question listed) was utilized by graders. 0 generally referred to an incorrect answer or lack of explanation. 1 referred to some accuracy and inclusion of material. 2 provided a comprehensive explanation (Table A2). AAO-HNS guidelines received a 42/42 by default. Responses were graded by 2 otolaryngology physicians (BG and SZH). Each physician had their own version of the original Word document so as not to be influenced by each other’s responses. All physicians were blinded to which document correlated with in regard to AI models. After each physician completed their gradings, results were input into a Microsoft Excel (Version 16.78.3, Microsoft Corporation, Redmond, Washington) file for analysis.

2.4. Readability Assessment

Responses from all AI models were separately graded by the author (HL) via Flesch Reading Ease (FRE) score and Flesch Kincaid Grade Level (FKGL) [11]. The FRE score is measured on a scale of 0–100 that measures the reading difficulty of a text based on word count and syllables, with higher scores corresponding to easier readability. A FRE score of above 80 correlates with a 6th-grade reading level. FKGL is also measured based on word count and syllables but is scored on a 0–18 scale. Each FKGL score correlates with a specific grade level, with numbers 0–12 correlated with their respective grade levels and numbers above 12 with college and graduate levels. The American Medical Association (AMA) recommends that all patient education materials be written at a sixth-grade reading level [12].

2.5. Statistical Analysis

Inter-rater reliability (IRR) was assessed using the intraclass correlation coefficient (ICC). Descriptive statistics are presented as mean scores for ease of interpretation with one-way ANOVA for comparison of means. Readability scores are presented as Tukey-adjusted pairwise comparisons for FRE and FKGL across AI platforms and AAO-HNS guidelines. Analysis was performed using R. This non-human subject research was exempted by the Institutional Review Board (Pro2026000677).

3. Results

This study was conducted prospectively across four LLMs using two prompt structures. A total of 8 prompts were imputed by 2 study members, from which 8 responses were evaluated by 2 separate study members. The reviewers demonstrated a high intraclass correlation coefficient of 0.961.

Guideline concordance ranged from 55.36% to 77.98% ($p > 0.05$), highest for Grok (77.98%, 95% CI 63.67–92.28), followed by Google Gemini (66.67%, 95% CI 35.42–97.91), ChatGPT (55.95%, 95% CI 29.25–82.65), and Google AI (55.36%, 95% CI 29.97–80.75). Pairwise comparisons demonstrated that Grok achieved significantly higher concordance scores than ChatGPT ($t = 5.51$, $p = 0.0035$), although this difference remained significant only after Bonferroni correction (adjusted $p = 0.021$). Similarly, Grok significantly outperformed Google AI ($t = 5.88$, $p = 0.0024$; adjusted $p = 0.015$). No statistically significant differences were observed between Grok and Gemini ($t = 2.49$, $p = 0.064$; adjusted $p = 0.385$), ChatGPT and Gemini ($t = -1.98$, $p = 0.097$; adjusted $p = 0.581$), ChatGPT and Google AI ($t = 0.12$, $p = 0.907$; adjusted $p = 1.000$), or Gemini and Google AI ($t = 2.13$, $p = 0.079$; adjusted $p = 0.476$) (Figure 1). Responses assigned a score of 0 reflected omission of important guideline-recommended information rather than hallucinated or harmful recommendations. While missing safety counseling was observed, no model generated responses containing clearly dangerous medical advice.

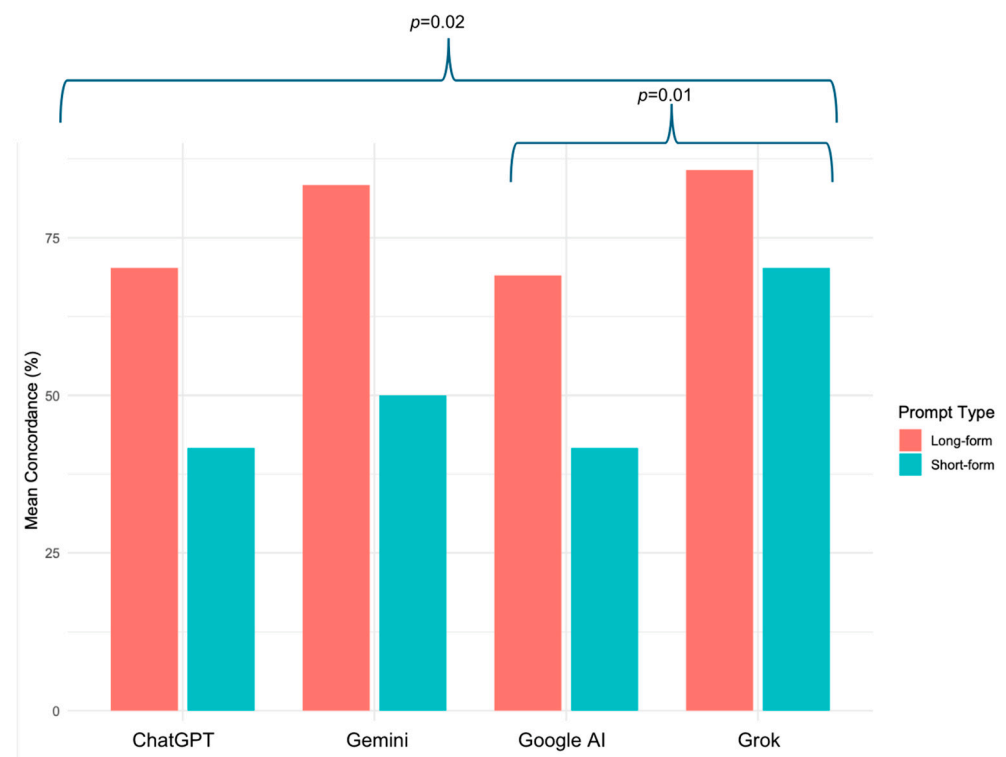


Figure 1. Mean concordance to clinical practice guidelines separated by prompt type.

Across all inputs, prompt structure significantly affected scores, with long-form guideline-mapped prompts outperforming short patient-focused prompts (+26.19, $p < 0.001$). Across all four models, long-form prompting resulted in higher guideline concordance scores than short-form prompting. Concordance increased from 70.2% to 85.7% for Grok, 41.7% to 70.2% for ChatGPT, 50.0% to 83.3% for Google Gemini, and 41.7% to 69.0% for Google AI.

Improvements in readability were less consistent. ChatGPT and Grok produced lower reading grade levels with long-form prompting, whereas Google Gemini and Google AI generated responses with higher grade levels. FRE scores varied between models, as Grok generated responses with significantly higher FRE scores than both ChatGPT (adjusted $p = 0.02$) and Google AI (adjusted $p = 0.01$) (Figure 2). No significant differences were observed between Grok and Gemini or among the remaining model comparisons after

correction for multiple testing. Overall, Grok produced the most readable responses, whereas ChatGPT and Google AI generated responses with lower reading ease scores.

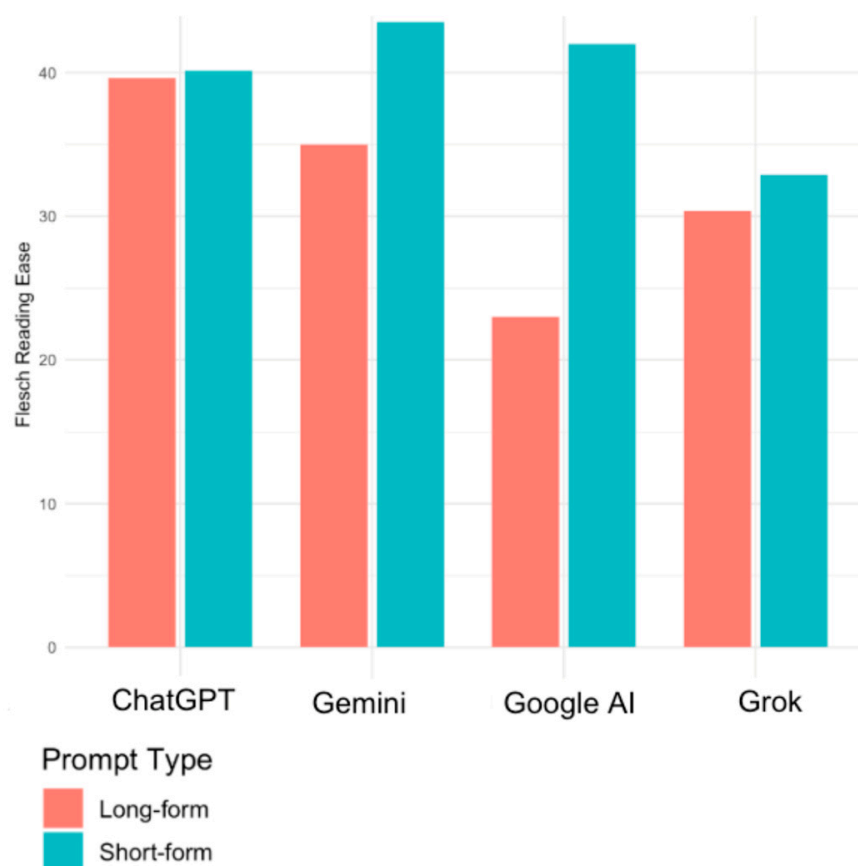


Figure 2. Flesch reading ease scores of large language model outputs. Higher scores indicate easier reading levels.

FRE scores were largely unchanged by prompt structure for ChatGPT but decreased for the remaining models. FRE scores were consistently higher with long-form prompting in comparison with short-form prompting, indicating improved readability. Across models, the greatest improvements were observed for ChatGPT and Google AI, while Grok and Gemini also demonstrated increases in reading ease. Grok generated significantly more readable responses than both ChatGPT (Bonferroni-adjusted $p = 0.02$) and Google AI (Bonferroni-adjusted $p = 0.01$). No other pairwise differences reached statistical significance after adjustment for multiple comparisons. FRE was highest and FKGL was lowest for Google Gemini and ChatGPT (FRE 39.25, 39.85 and FKGL 10.20, 10.50 respectively) ($p > 0.05$).

FKGL scores were the highest with Google AI at a mean score of 12.5, indicating a reading level above the twelfth grade. Grok produced responses with a mean FKGL of 11.4, followed by Google Gemini at 10.5 and ChatGPT at 10.3. The prompting structure had a variable effect on FKGL across LLMs. Long-form prompting reduced the reading grade level of responses generated by Grok (11.9 vs. 10.8) and ChatGPT (11.2 vs. 9.3), indicating improved readability. In contrast, long-form prompting increased the FKGL of responses produced by Google Gemini (9.9 vs. 11.1) and Google AI (10.5 vs. 13.6), reflecting more complex text. Among all models, ChatGPT generated the lowest FKGL score with long-form prompting, while Google AI produced the highest (Figure 3).

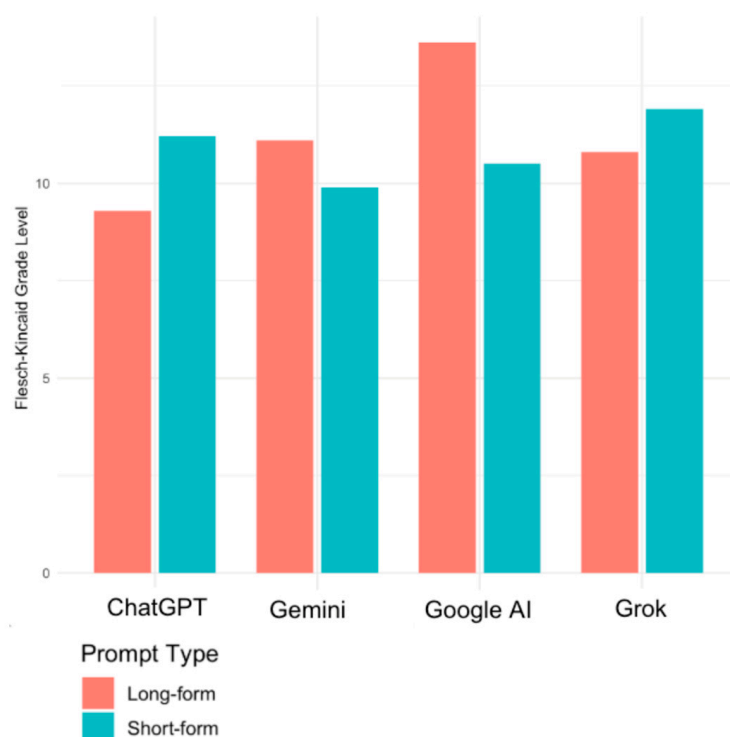


Figure 3. Flesch-Kincaid grade level readability scores of large language model outputs. Lower scores indicate easier reading levels.

To contextualize our findings within current clinical practice, comparisons against the AAO-HNSF 2025 CRS Guidelines were made. The CPG demonstrated a FRE score of 44.1 and a FKGL score of 11.7. The CPG thus demonstrated a more readable structure, with a higher FRE, exceeding scores generated by Grok (31.7), ChatGPT (39.9), Gemini (39.2), and Google AI (32.5). In contrast, the CPG required a higher reading grade level than Grok (11.4), Gemini (10.5), and ChatGPT (10.3), but was slightly easier to read than Google AI, which produced the highest FKGL score (12.5).

4. Discussion

The rapid adoption of LLMs has reshaped how patients access and interpret medical information, with studies demonstrating that many users prefer LLM-generated responses over traditional search engines due to their perceived clarity and conversational accessibility [13]. As such, this study is important to help direct patients to utilize the most accurate LLMs and optimize their search queries for more comprehensive results. The need for further patient guidance on the effective use of LLMs is emphasized by recent evidence showing that although new LLMs accurately identified clinical conditions in nearly 95% of simulated scenarios, users interacting with the same models were able to identify the correct condition in fewer than 34.5% of cases [14]. While LLMs are being evaluated against medically complex queries, there remains limited evidence evaluating how LLMs respond regarding surgical decision-making. Increased utilization of LLMs by patients as a primary source of medical information outside of traditional clinical encounters justifies the need for LLM evaluation by clinicians.

To address this gap, we conducted a structured evaluation of multiple widely used LLM platforms in CRS management. Independent physician grading of these models demonstrated highly consistent agreement between reviewers. The use of blinded physician reviewers and a standardized scoring framework provided structured benchmarking methods for evaluation of AI-generated medical information.

LLMs demonstrated moderate alignment with the 2025 AAO-HNS guideline on surgical management of CRS. Overall concordance ranged from 55.36% to 77.98%, indicating that models inconsistently reproduced the complete framework outlined in AAO-HNS recommendations. The variable concordance observed across models is consistent with prior literature evaluating LLM performance in patient counseling, with reported accuracy of responses in simple clinical scenarios, including patient FAQs, symptom categorization, and clinical recommendations, ranging from 20% to 95% across studies [15]. The observed moderate concordance in this study also mirrors prior findings from American Academy of Orthopaedic Surgeons guideline-based evaluations of rotator cuff injury management, in which ChatGPT and Gemini achieved 81% and 75% concordance with recommended treatments but still generated discordant recommendations in 19% and 25% of queried interventions [16].

While overall performance varied between platforms, Grok demonstrated the highest concordance, proving to be more clinically reliable than Google Gemini, ChatGPT, and Google AI. Notably, Grok significantly outperformed both ChatGPT and Google AI, suggesting that measurable differences in guideline concordance may exist between platforms. Prior work has similarly found that in a lung cancer clinical decision-making analysis, Grok-3 achieved the highest overall performance across accuracy, completeness, and practicality when compared to DeepSeek-R1 and GPT-4.5, supporting the clinical validity of Grok in responding to patient queries [17]. On the other hand, ChatGPT and Google Gemini have not demonstrated superiority in accuracy of medical recommendations. Alarifi et al. found no significant differences between ChatGPT, Gemini, and Claude in thyroid nodule risk-assessment recommendations [18], while Aliyeva et al. demonstrated 100% accuracy and relevance of ChatGPT responses regarding post-cochlear implantation care when graded by otolaryngologists [19]. The present findings reflect that consistent alignment with guideline-directed clinical decision-making remains an ongoing limitation across platforms.

An important finding of this study was that low concordance scores were primarily attributable to omission of guideline-recommended information rather than fabrication of overtly harmful medical recommendations. However, prior literature has demonstrated that hallucination remains an important limitation of LLMs in medical applications. In a comparative analysis of five models evaluating spine surgery topics, general-purpose LLMs produced substantial and variable citation errors, with fabrication rates as high as 26.2% for Gemini [20]. Importantly, model reliability appears to be improving over time, as GPT-4 demonstrated lower hallucination rates than its predecessor GPT-3.5 when tasked with reproducing systematic reviews of rotator cuff pathology (28.6% vs. 39.6%) [21], suggesting progressive improvements across successive model generations. Despite these advances, the persistence of fabricated or incomplete information across clinical domains highlights the need for continued patient caution when using LLMs for medical information.

Prompt structure emerged as a major driver of performance, with long-form prompting improving guideline concordance and readability across all models. Long-form guideline-mapped prompts achieved an average concordance of 77.1%, significantly higher than the 50.8% concordance observed with short-form patient-style prompts. This difference may be explained by prior work demonstrating that the sequential and chain-of-thought prompting seen in the long-form prompts can improve medical reasoning [22]. These findings are particularly important because they suggest that intra-model variability may be partially mitigated through prompt design alone. They also reflect that user ability and familiarity with medical terminology may substantially influence the quality of AI-generated counseling received.

Readability analysis further demonstrated that all evaluated outputs, including the 2025 AAO-HNS guideline, exceeded recommended readability thresholds for patient education materials. In this study, all FKGL scores exceeded the recommended reading level. FKGL scores improved with LLM use compared to the AAO-HNS CPG for all platforms except Google AI. FRE scores similarly reflected relatively difficult reading material, suggesting that AI use without consideration of existing disparities in health literacy may therefore further exacerbate communication barriers rather than improve patient understanding.

Short-form outputs generally demonstrated higher FRE scores, supporting greater readability relative to long-form responses. However, this relationship varied by model, with Gemini and Google AI demonstrating lower FKGL scores in short-form outputs, while ChatGPT and Grok demonstrated lower reading levels in long-form outputs. Notably, improved readability in shorter responses frequently occurred alongside reductions in guideline concordance, indicating current tradeoffs between completeness and comprehension. This pattern is consistent with evaluations of older ChatGPT and Grok models, where higher-quality outputs were less readable and more readable outputs tended to be less complete [23].

This study found that both model selection and prompt structure substantially influence LLM performance. Long-form prompting consistently improved guideline adherence and readability across platforms even compared to CPG. Despite improvements, readability remains above ideal patient levels, limiting direct clinical use. This study contributes to a growing body of evidence supporting the importance of patient guidance on what LLM to use and prompt engineering in optimizing AI outputs for clinical use, as past research has demonstrated that structured prompt engineering can improve the clarity of LLM-generated patient education materials [24]. For example, a mixed-methods study evaluating 123 structured patient-style prompts across cardiovascular disease topics found that prompts generated by users with higher health literacy consistently produced higher-quality responses, highlighting that prompt formulation and user understanding significantly influence the clinical appropriateness of LLM-generated medical advice [25].

Existing research has primarily focused on how prompt formulation influences LLM performance, with considerably less attention given to the effects of prompt length and the amount of contextual information provided. Instead, prior studies have largely examined whether longer AI-generated responses are associated with greater accuracy. However, response length alone does not appear to predict content quality. Zhou et al. reported that although DeepSeek generated substantially more verbose responses than comparator models, the additional detail did not consistently improve factual accuracy, clinical reasoning, or overall response quality [26]. Similarly, Marcaccini et al. [27] found that excessively lengthy AI-generated outputs may reduce readability, increasing the cognitive burden for patients interpreting complex medical information. Collectively, these findings suggest that the relationship between prompt detail, response length, and response quality is multifactorial. In contrast, the influence of prompt length on adherence to clinical practice guidelines remains largely unexplored. This study proposes that providing greater clinical context through more detailed prompts may enable LLMs to better interpret user intent and generate more clinically relevant responses, although this additional context must be balanced against the need for concise, understandable patient education. Ultimately, AI tools should serve as adjuncts to clinician counseling rather than replacements.

While this study assesses how well LLMs perform in generating CRS education materials, several limitations of the present study should be acknowledged. First, the clinician-designed queries may not accurately reflect the way real patients would phrase their prompts, potentially limiting the broad application of these findings. The long-form

prompts were derived directly from the guideline question framework to ensure comprehensive assessment of its recommendations, while three team members (BG, SH, and WH) who regularly interact with patients contributed insight into short-form patient phrasing. Second, LLM outputs are inherently stochastic, so responses may vary across repeated queries or alternative wording; this study therefore does not fully assess reproducibility or capture the full range of outputs each model may generate. Caution should be exercised when generalizing to all patient interactions. Third, the evaluation was limited to four models and, because AI systems are continuously updated, the performance characteristics observed here may not hold for future iterations or reflect the broader generative AI landscape. Future work should evaluate a larger and more diverse set of patient-generated prompts with repeated queries per prompt to better characterize response variability, include additional commercial and open-source models such as Claude and DeepSeek, and extend this framework to clinical practice guidelines across other medical and surgical specialties to better define the generalizability of LLMs in patient education.

Although the CRS guidelines evaluated in this study were published in 2025, the use of LLMs for obtaining medical information has continued to increase since their release. As such, this study provides a current assessment of how current AI platforms respond to postoperative patient-oriented queries in comparison with established clinical practice guidelines.

5. Conclusions

Accurate patient education in CRS remains challenging because management often involves prolonged medical therapy and individualized surgical decision-making. In this study, LLMs demonstrated moderate overall concordance with the 2025 AAO-HNS guideline on surgical management of CRS, with Grok significantly outperforming both ChatGPT and Google AI. However, overall, LLMs demonstrated similar guideline concordance, suggesting patients can expect comparable accuracy across platforms. Prompt structure significantly influenced response quality, with long-form guideline-mapped prompts producing significantly higher concordance scores than short patient-focused prompts, highlighting how users' ability to frame precise prompts is critical to obtaining accurate information. While LLMs generally improved FKGL scores compared to the AAO-HNS CPG, they demonstrated lower FRE scores, indicating a mixed result on overall readability. Overall, readability remained at a high school to college level, indicating persistent difficulty for typical patient comprehension. This study demonstrates that both model type and prompt design significantly influence AI-generated clinical content, though this may be limited by the need for higher user reading levels.

Author Contributions: Conceptualization, H.L. and W.D.H.; methodology, H.L., E.K. and W.D.H.; formal analysis, H.L. and E.K.; investigation, H.L., E.K., S.Z.H. and B.S.G.; data curation, H.L. and E.K.; writing—original draft preparation, H.L., E.K. and A.C.; writing—review and editing, H.L., E.K., A.C., S.Z.H., B.S.G., R.K. and W.D.H.; supervision, W.D.H.; project administration, W.D.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study by the Rutgers University Institutional Review Board as it conducts non-human subjects research using publicly available resources (New Brunswick, NJ, USA).

Informed Consent Statement: Patient consent was waived as there were no human subjects involved and no patient information was accessed, collected, or reported. Thus no patients were contacted for consent.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: During the preparation of this manuscript, the author(s) used ChatGPT (v5.1), Google AI, Google Gemini (v3), and Grok (v4.1) for the purposes of querying models using long and short prompts to analyze each model's responses for completeness and accuracy. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CRS	Chronic Rhinosinusitis
LLMs	Large Language Models
FRE	Flesch Reading Ease
FKGL	Flesch-Kincaid Grade Level
IRR	Inter-rater Reliability
ICC	Intraclass Correlation Coefficient
AAO-HNS	American Academy of Otolaryngology—Head and Neck Surgery
AAO-HNSF	American Academy of Otolaryngology–Head and Neck Surgery Foundation
AMA	American Medical Association

Appendix A

Table A1. Prompt Structure.

Single Patient-Focused Prompt	
1	I've been dealing with sinus problems for a long time, and my doctor mentioned surgery might help. Can you explain what chronic rhinosinusitis is, how it's diagnosed, when and what surgery might be needed, the possible benefits and risks, what recovery is like, and how I can take care of myself afterward?
Guideline Mapped Prompt	
1	What is chronic rhinosinusitis (CRS)?
2	How is the diagnosis of CRS confirmed?
3	What is sinus surgery?
4	How do I know if I need sinus surgery?
5	What are the benefits of verifying my diagnosis before surgery?
6	What happens during the assessment for surgery?
7	What are the risks of verifying my diagnosis or assessing my candidacy for surgery?
8	What makes me a good surgical candidate?
9	Can I play a role in deciding if surgery is right for me?
10	Are antibiotics helpful or not helpful for my chronic sinus disease?
11	Will surgery replace my need for most sinonasal medications?
12	Are there alternative treatments to surgery?
13	What is the difference between sinus dilation surgery and sinus surgery creating wide sinus openings?
14	How does creating wide sinus openings help my chronic sinus disease?
15	What should I expect after surgery? Should I expect pain?
16	What type of postoperative care and/or medications will I need, and for how long?
17	How many postoperative visits will I have, and what is the timing for these?
18	What will be covered or done at these postoperative visits?
19	What are the limitations after surgery?
20	How does the extent of surgery impact my healing after surgery?
21	Will my physician provide any resources about sinus surgery for me?

Table A2. Clinical Guideline Concordance Scoring Rubric.

Questions	Rubric Scores
What is chronic rhinosinusitis (CRS?)	Signs and Symptoms (2 points): 0: No mention or incorrect explanation of chronic rhinosinusitis (CRS) symptoms or diagnostic timeframe. 1: Partial or vague description (e.g., mentions sinus congestion or drainage but omits duration, inflammation, or multiple symptom criteria). 2: Comprehensive and accurate explanation consistent with CPG—defines CRS as ≥ 12 weeks of ≥ 2 symptoms (nasal obstruction/congestion, discolored drainage, facial pain/pressure, or reduced sense of smell)
How is the diagnosis of CRS confirmed?	Diagnosis Confirmation 0: Missing or inaccurate discussion of how CRS is confirmed. 1: Mentions evaluation or imaging without specificity. 2: Clearly describes confirmation by objective evidence of sinonasal inflammation (nasal endoscopy or CT findings) and distinguishes CRS from allergy, acute sinusitis, or ear-related symptoms.
What is sinus surgery?	0 points: No or incorrect explanation of sinus surgery. 1 point: Accurately explains endoscopic sinus surgery (ESS) as a telescope-guided, minimally invasive procedure performed through the nostrils to open blocked sinus pathways, clear infection or inflammation, and improve medication delivery. 2 points: Provides a comprehensive explanation including purpose (to improve breathing and reduce infections), benefits (better sinus drainage, reduced symptoms), and sets realistic expectations about recovery and outcomes.
How do I know if I need sinus surgery?	Need for Sinus Surgery (2 points): 0: No or inappropriate explanation of indications for surgery. 1: Mentions failure of medication or need for imaging without details. 2: Explains that surgery is considered when medical therapy (e.g., saline irrigation, intranasal/oral steroids, antibiotics in select cases) fails; identifies subtypes benefiting most (CRSwNP, AFRS, EMRS, fungal disease, bony obstruction); and emphasizes individualized timing and absence of a one-size-fits-all prerequisite regimen.
What are the benefits of verifying my diagnosis before surgery?	Benefits of Verifying Diagnosis Before Surgery (2 points): 0: No discussion of benefits. 1: Mentions excluding other diagnoses (migraines, asthma, etc) to guide treatment plan 2: Describes surgical planning benefit and assessing the success of surgery with a combination of symptoms and objective evidence
What happens during the assessment for surgery?	How to Verify Diagnosis Before Surgery (2 points): 0: No pre-operative assessment guidance. 1: Mentions CT or general evaluation only. 2: Details comprehensive assessment: symptom severity, quality-of-life impact, endoscopic findings, radiographic imaging, subtype identification (CRSwNP, EMRS, AFRS), and prior medical response.
What are the risks of verifying my diagnosis or assessing my candidacy for surgery?	Risks of Verifying Diagnosis Before Surgery (2 points): 0: No discussion of need to check, review, or repeat certain tests 1: Explains need to check, review, or repeat certain tests without mention of specifics. 2: Explains additional costs and minor risks of nasal endoscopy or imaging
What makes me a good surgical candidate?	Good Surgical Candidate Criteria (2 points): 0: No mention or incorrect explanation of who qualifies for sinus surgery. 1: Partial or vague criteria (e.g., mentions “failed medical therapy” without specifying what that means or which subtypes benefit most). 2: Comprehensive and guideline-consistent description—identifies candidates as adults with CRS who meet diagnostic criteria, have persistent symptoms or impaired quality of life despite appropriate medical therapy, and whose subtype or disease severity suggests surgery would provide greater benefit than continued medical management.
Can I play a role in deciding if surgery is right for me?	0: No mention of patient input 1: Partial mention of patient input 2: Mentions input of patient is vital, asked about CRS impact on patient life, talks about doctor sharing process of decision making
Are antibiotics helpful or not helpful for my chronic sinus disease?	0: No mention of abx use 1: Mentions antibiotic use but not specifically when indicated 2: Refers to pathophysiology of CRS and why abx may not be indicated, discusses that abx for bacterial infection and CRS is not necessarily triggered by bacteria. Refers to CRS time course, relates side effects of antibiotics and dangers of overuse, reviews signs of active infection and clinician performing exam to determine

Table A2. Cont.

Questions	Rubric Scores
Will surgery replace my need for most sinonasal medications?	0: Does not answer question 1: Mentions that it is part of treatment 2: Mentions that it is not a replacement for medications and part of treatment and allows topicals to work better
Are there alternative treatments to surgery?	0: Describes no alternatives 1: Describes some alternative treatments 2: Describes alternatives and when surgery is considered
What is the difference between sinus dilation surgery and sinus surgery creating wide sinus openings?	0: Does not define the differences between surgeries 1: Defines only one of the surgeries 2: Defines both surgeries
How does creating wide sinus openings help my chronic sinus disease?	0: Refers to only 1 way 1: Refers to only 2–3 out of the 4 ways 2: Describes all 4 benefits
What should I expect after surgery? Should I expect pain?	0: No postoperative expectations 1: Gives only some expectations and some suggestions for dealing with pain 2: Gives full expectations and describes level of pain, recommends possible pain meds and when narcotics indicated, describes some symptoms that may occur and when to resume nasal irrigations and sprays
What type of postoperative care and/or medications will I need, and for how long?	0: Doesn't describe any care or meds 1: Describes only meds or describes only timeline 2: Gives nasal care and timeline
How many postoperative visits will I have, and what is the timing for these?	0: Doesn't answer how many visits 1: Does not give time course of how regular the visits will be 2: Conveys patients will be seen regularly and at specific intervals
What will be covered or done at these postoperative visits?	0: Does not review what will happen at postop visits 1: Mentions that visits may involve nasal endoscopy and debridement 2: Mentions that visits may involve nasal endoscopy and debridement and defines, talks about taking OTC pain med before appointments and assessing symptoms and surgical recovery, modify meds, perform physical exam
What are the limitations after surgery?	0: No specifics on limitations after surgery 1: Give some examples of how long maybe be out of work or when to stop taking stronger pain med 2: Also avoids heavy physical activity, and avoiding sneezing, and not going underwater
How does the extent of surgery impact my healing after surgery?	0: Does not explain extent of surgery impact 1: Goes into some detail of process of healing 2: Describes process of healing and what may take longer vs. shorter
Will my physician provide any resources about sinus surgery for me?	0: Provides no resources 1: Provides some resources 2: Provides educational materials in paper/electronic format, including restrictions, important meds, signs and symptoms for urgent eval during post op

References

1. Fokkens, W.J.; Lund, V.J.; Hopkins, C.; Hellings, P.W.; Kern, R.; Reitsma, S.; Toppila-Salmi, S.; Bernal-Sprekelsen, M.; Mullol, J.; Alobid, I.; et al. European Position Paper on Rhinosinusitis and Nasal Polyps 2020. *Rhinology* **2020**, *58*, 1–464. [[CrossRef](#)] [[PubMed](#)]
2. Ariyaratne, S.; Jenko, N.; Mark Davies, A.; Iyengar, K.P.; Botchu, R. Could ChatGPT Pass the UK Radiology Fellowship Examinations? *Acad. Radiol.* **2024**, *31*, 2178–2182. [[CrossRef](#)] [[PubMed](#)]
3. Siam, M.K.; Faruk, M.J.H.; Cheng, J.Q.; Gu, H. Fusion-Augmented Large Language Models: Boosting Diagnostic Trustworthiness via Model Consensus. In *Proceedings of the 2025 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*; IEEE: New York, NY, USA, 2025; pp. 1–7. [[CrossRef](#)]
4. Hirani, H.; Saboo, B.; Modi, A.; Modi, P.; Samajdar, S.S.; Saboo, B.; Chawla, M.; Maheshwari, A.; Gupta, A.; Parikh, R.; et al. Clinical Assessment of Large Language Models: A Comprehensive Multi-domain Performance Study for Healthcare Applications. *Int. J. Diabetes Technol.* **2025**, *4*, 159–165. [[CrossRef](#)] [[PubMed](#)]

5. Moise, A.; Tatar, L.; Sela, N.; da Silva, S.D.; Kouz, J.; Tamilia, M.; Hier, M.P.; Forest, V.-I.; Payne, R.J. Thyroid Nodule Experts Evaluating ChatGPT's Assessment of Thyroid Nodules Classified by the Bethesda System for Reporting Thyroid Cytopathology. *J. Otolaryngol.-Head. Neck Surg.* **2025**, *54*, 19160216251387617. [CrossRef] [PubMed]
6. Abdul Sami, M.; Abdul Samad, M.; Parekh, K.; Suthar, P.P. Comparative Accuracy of ChatGPT 4.0 and Google Gemini in Answering Pediatric Radiology Text-Based Questions. *Cureus* **2024**, *16*, e70897. [CrossRef] [PubMed]
7. Draelos, R.L.; Afreen, S.; Blasko, B.; Brazile, T.L.; Chase, N.; Desai, D.P.; Evert, J.; Gardner, H.L.; Herrmann, L.; House, A.V.; et al. Large language models provide unsafe answers to patient-posed medical questions. *arXiv* **2025**, arXiv:2507.18905. [CrossRef]
8. US Patients are Turning to ChatGPT to Navigate Healthcare. Available online: <https://www.fiercehealthcare.com/ai-and-machine-learning/40m-people-use-chatgpt-answer-healthcare-questions-openai-says> (accessed on 17 February 2026).
9. Zhang, D.; Earp, B.E.; Kilgallen, E.E.; Blazar, P. Readability of Online Hand Surgery Patient Educational Materials: Evaluating the Trend Since 2008. *J. Hand Surg. Am.* **2022**, *47*, 186.e1–186.e8. [CrossRef] [PubMed]
10. Shin, J.J.; Wilson, M.; McKenna, M.; Rosenfeld, R.; Ammon, K.; Crosby, D.; Fuchs, J.M.; Hensler, J.B.; Illing, E.A.; Lam, K.; et al. Clinical Practice Guideline: Surgical Management of Chronic Rhinosinusitis. *Otolaryngol. Head Neck Surg.* **2025**, *172*, S1–S47. [CrossRef] [PubMed]
11. Wang, L.-W.; Miller, M.J.; Schmitt, M.R.; Wen, F.K. Assessing readability formula differences with written health information materials: Application, results, and recommendations. *Res. Soc. Adm. Pharm.* **2013**, *9*, 503–516. [CrossRef] [PubMed]
12. Eltorai, A.E.; Ghanian, S.; Adams, C.A., Jr.; Born, C.T.; Daniels, A.H. Readability of patient education materials on the american association for surgery of trauma website. *Arch. Trauma Res.* **2014**, *3*, e18161. [CrossRef] [PubMed] [PubMed Central]
13. Carl, N.; Haggenmüller, S.; Wies, C.; Nguyen, L.; Winterstein, J.T.; Hetz, M.J.; Mangold, M.H.; Hartung, F.O.; Grüne, B.; Holland-Letz, T.; et al. Evaluating interactions of patients with large language models for medical information. *BJU Int.* **2025**, *135*, 1010–1017. [CrossRef] [PubMed]
14. Bean, A.M.; Payne, R.E.; Parsons, G.; Kirk, H.R.; Ciro, J.; Mosquera-Gómez, R.; M, S.H.; Ekanayaka, A.S.; Tarassenko, L.; Rocher, L.; et al. Reliability of LLMs as medical assistants for the general public: A randomized preregistered study. *Nat. Med.* **2026**, *32*, 609–615. [CrossRef] [PubMed]
15. Geracitano, J.; Anderson, B.; Coffel, M.; Rosenzweig, M.; Dorn, S.D.; Khairat, S.; Conklin, J. The Accuracy of ChatGPT in Answering FAQs, Making Clinical Recommendations, and Categorizing Patient Symptoms: A Literature Review. *Adv. Health Inf. Sci. Pract.* **2025**, *1*, VXUL2925. [CrossRef] [PubMed]
16. Megafu, M.; Guerrero, O.; Yendluri, A.; Parsons, B.O.; Galatz, L.M.; Li, X.; Kelly, J.D.; Parisien, R.L. ChatGPT and Gemini Are Not Consistently Concordant with the 2020 American Academy of Orthopaedic Surgeons Clinical Practice Guidelines When Evaluating Rotator Cuff Injury. *Arthroscopy* **2025**, *41*, 2753–2757. [CrossRef] [PubMed]
17. Zhang, Y.; Yang, D.; Shi, Y.; Liu, Y. Performance of Large Language Models in Lung Cancer Clinical Decision-Making: A Comparative Analysis Based on DeepSeek, Grok, and GPT. *Cureus* **2025**, *17*, e99026. [CrossRef] [PubMed]
18. Alarifi, M. Appropriateness of Thyroid Nodule Cancer Risk Assessment and Management Recommendations Provided by Large Language Models. *J. Imaging Inform. Med.* **2025**, *38*, 4324–4335. [CrossRef] [PubMed]
19. Aliyeva, A.; Sari, E.; Alaskarov, E.; Nasirov, R. Enhancing Postoperative Cochlear Implant Care with ChatGPT-4: A Study on Artificial Intelligence (AI)-Assisted Patient Education and Support. *Cureus* **2024**, *16*, e53897. [CrossRef] [PubMed]
20. McLaughlin, N.D.; Srinivas, A.N.; Lowe, Z.F.; Botterbush, K.S.; Patel, M.S.; Avila, M.J. Large Language Model Hallucinations in Spine Surgery: A Comparative Analysis of Clinician vs Patient-Level Prompts. *Neurosurg. Pract.* **2026**, *7*, e000244. [CrossRef] [PubMed]
21. Chelli, M.; Descamps, J.; Lavoué, V.; Trojani, C.; Azar, M.; Deckert, M.; Raynier, J.-L.; 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] [PubMed]
22. Shah, H.A.; Househ, M. Chain of Thought Strategy for Smaller LLMs for Medical Reasoning. *Stud. Health Technol. Inform.* **2025**, *327*, 783–787. [CrossRef] [PubMed]
23. Chen, G.; Lin, C.; Kim, J.H.; Du, F.; Luo, Z.; Shin, Y.S.; Li, X. Readability and information quality of LLM-Generated HIV content: A methodological content evaluation. *BMC Infect. Dis.* **2025**, *25*, 1624. [CrossRef] [PubMed]
24. Mudrik, A.; Nadkarni, G.N.; Efros, O.; Soffer, S.; Klang, E. Prompt engineering in large language models for patient education: A systematic review. *medRxiv* **2025**. [CrossRef]
25. Lautrup, A.D.; Hyrup, T.; Schneider-Kamp, A.; Dahl, M.; Lindholt, J.S.; Schneider-Kamp, P. Heart-to-heart with ChatGPT: The impact of patients consulting AI for cardiovascular health advice. *Open Heart* **2023**, *10*, e002455. [CrossRef] [PubMed]

26. Zhou, M.; Pan, Y.; Zhang, Y.; Song, X.; Zhou, Y. Evaluating AI-generated patient education materials for spinal surgeries: Comparative analysis of readability and DISCERN quality across ChatGPT and DeepSeek models. *Int. J. Med. Inform.* **2025**, *198*, 105871. [[CrossRef](#)] [[PubMed](#)]
27. Marcaccini, G.; Seth, I.; Xie, Y.; Susini, P.; Pozzi, M.; Cuomo, R.; Rozen, W.M. Breaking bones, breaking barriers: ChatGPT, DeepSeek, and Gemini in hand fracture management. *J. Clin. Med.* **2025**, *14*, 1983. [[CrossRef](#)] [[PubMed](#)]

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.