



Article

Evaluating AI-Generated Tonsillectomy Guidance Against Clinical Practice Standards

Hetal Lad ^{1,*} , Shrey Shah ¹, Shreeya Bahethi ², Sudeepti Vedula ¹ and Brian Manzi ¹

¹ Department of Otolaryngology—Head and Neck Surgery, Rutgers New Jersey Medical School, Newark, NJ 07103, USA; sbs201@njms.rutgers.edu (S.S.); sv406@njms.rutgers.edu (S.V.); brian.manzi@rutgers.edu (B.M.)

² Department of Otolaryngology—Head and Neck Surgery, Hackensack Meridian School of Medicine, Nutley, NJ 07103, USA; shreeya.bahethi@hmsom.edu

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

Abstract

Background/Objectives: Artificial intelligence (AI) tools are used in clinical education and patient communication. Concordance with American Academy of Otolaryngology—Head and Neck Surgery Foundation clinical practice guidelines (CPG) and overall readability remain poorly defined. **Methods:** Over 2 months, ChatGPT, Google Gemini, and Google Search AI were queried using a single LF prompt and a layered set of sub-questions. Responses were scored with a 12-point rubric from the CPG plain-language tonsillectomy guideline. Readability was assessed using Flesch Reading Ease (FRE) and Flesch-Kincaid Grade Level (FKGL). Temporal trends and between-group differences were analyzed using regression and repeated-measures methods. **Results:** Layered prompts demonstrated greater alignment with guideline subtopics than LF (89.5% vs. 71.4%, $p < 0.001$). For LF queries, mean concordance was 74.5% (95% CI 72.0–77.0) for Google AI, 71.3% (68.9–73.7) for Gemini, and 68.4% (66.0–70.9) for ChatGPT. Overall, Google AI outperformed Gemini and ChatGPT on pairwise comparisons ($p = 0.001$). Google AI demonstrated improvement in concordance (slope = 0.03, $p = 0.003$), while Gemini showed no change ($p > 0.05$) and ChatGPT worsened over time (slope = -0.03 , $p < 0.001$). Readability differed by strategy and model. Layered prompts produced more accessible text (FKGL 44.6, FRE 10.3) compared with LF outputs (FKGL 35.2, FRE 11.3) and the CPG (FKGL 10.6, FRE 41.7). IRR was high (ICC = 0.96). **Conclusions:** Model type and prompt structure shaped the quality of AI-generated tonsillectomy guidance. Google AI consistently outperformed ChatGPT, with Gemini intermediate, and uniquely showed temporal improvement. Layered prompting produced more CPG-aligned and readable responses across all models. Readability was consistently above a 7th grade level.

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



Academic Editor: Yu Sun

Received: 14 August 2026

Revised: 21 September 2026

Accepted: 24 September 2026

Published: 30 September 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 Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Digital technologies have become integral to modern healthcare delivery, supporting physicians in clinical decision-making, documentation, and patient communication, while simultaneously expanding access to medical information for patients and caregivers. As these technologies become embedded in clinical practice, they are increasingly adopted by patients as well, reflecting a growing perception that medical information once limited to clinicians is now accessible to the public. While broader access to medical information is

beneficial, understanding the accuracy of these resources is essential to guide physicians in counseling patients and recommending reliable sources.

Large language models (LLMs) are trained on expansive datasets to generate human-like text responses to user prompts by referencing large bodies of text, often derived from the internet or other large databases [1]. These responses are generated as probabilistic sequences of words informed by training data and guided by the structure of the input prompt [2]. As a result of this probabilistic framework, LLM outputs may contain internal inconsistencies or present inaccurate information with apparent confidence [3,4].

Since the advent of ChatGPT in November 2022, LLMs have rapidly expanded into education, business, and healthcare settings, with claims of improved efficiency and accessibility [5]. These models have demonstrated notable performance benchmarks, including achieving passing scores on the United States Medical Licensing Examination [6] and performing at or near human levels on sample otolaryngology board examinations administered by the Royal College of Physicians and Surgeons of Canada [7,8]. While such findings offer reassurance regarding technical competency, the accuracy and appropriateness of LLM-generated information for patient-facing use remains inconsistent across studies [9].

Despite growing public reliance on LLMs, limited research has examined their role in patient decision-making within otolaryngology. One prior study evaluated postoperative instructions generated by ChatGPT and Google Search for common pediatric otolaryngology procedures, including tympanostomy tube placement, tonsillectomy and adenoidectomy, inferior turbinate reduction, tympanoplasty, cochlear implantation, neck mass resection, microdirect laryngoscopy and bronchoscopy, and tongue-tie release [10]. However, this analysis preceded the launch of Google Gemini in March 2023 and the integration of AI-generated responses into Google Search.

Tonsillectomies are among the most frequently performed pediatric otolaryngologic procedures, with approximately 289,000 annual cases in children under 15 years of age [11]. Caregivers commonly seek to understand why tonsillectomy is recommended, how the procedure may affect their child, and the logistical and emotional burden surgery may place on families. While these discussions ideally occur with an otolaryngologist, caregivers increasingly turn to online resources during their personal time for supplemental information. Although tonsillectomies offer well-established benefits, potential complications include throat pain, postoperative nausea and vomiting, delayed oral intake, postoperative bleeding, and, rarely, death are important to understand [12,13]. The gold-standard source for patient-facing information on tonsillectomies is the Clinical Practice Guideline: Tonsillectomy in Children, published in 2019 by the American Academy of Otolaryngology—Head and Neck Surgery [14]. Particularly, the plain-language summary is written for patients and caregivers to understand [14]. However, systematic comparisons of contemporary AI-generated responses against this guideline have not yet been performed.

The purpose of this study is thus to evaluate how contemporary LLMs portray the balance between risks and benefits of tonsillectomy in patient-facing responses. Additionally, we aim to assess the readability of these responses relative to established plain-language guidelines, recognizing that effective patient communication requires not only accuracy but also accessibility that clinicians are trained to do but AI might not be.

2. Materials and Methods

2.1. Study Design

A prospective comparative study was conducted to evaluate the performance of 3 free and easily accessible artificial intelligence (AI) tools in the context of patient-oriented questions about tonsillectomies: ChatGPT (v5.1, released 7 August 2025), Google Gemini

(v3, released 14 May 2024), and Google AI (v2.5, released 18 November 2025). This non-human-subjects research was determined to be exempt by the Institutional Review Board (Pro2026000677). The 2019 AAO-HNS guidelines were chosen as the gold standard for accuracy of outputs as they were developed by a task force selected for their expertise on tonsillectomy management. The guidelines were accessed in September 2025. The “Plain Language Summary: Tonsillectomy in Children” provides information about pediatric tonsillectomy and supporting discussions among patients, caregivers, and clinicians regarding indications for surgery, available management options, and perioperative care. Questions were derived from the subsections of these guidelines to form a long-form prompt (LF) and layered-form prompt (YF) for entry into the AI tools. This study was conducted over 60 days, daily for the LF and weekly for the YF. All three tools were accessed through their publicly available web interfaces rather than application programming interfaces (APIs). Therefore, generation hyperparameters such as temperature, top_p, and max_tokens were not user configurable or available for reporting. A new chat session in a private browser window was initiated for each prompt administration to avoid prior conversational carry-over. Publicly available interfaces may undergo unannounced model or system updates, and their reported version labels and study dates describe the specific interfaces available during the study. They do not guarantee exact reproducibility at later dates.

2.2. Prompt Construction

Each AI platform was queried using two distinct prompt strategies designed to reflect common real-world information-seeking behaviors among parents and caregivers. The prompts were developed by clinical providers based on questions commonly raised by caregivers during pediatric tonsillectomy consultations. The long-form (LF) prompt consisted of a single comprehensive free-text query incorporating common caregiver concerns: “Why is my ENT recommending tonsillectomy? When and why should it be done? What technique should they use? What are the pros and cons of each technique? What else should I know about this procedure?”.

The YF was a structured series of eight sequential sub-questions, each targeting a specific subtopic within tonsillectomy counseling: Why is my ENT recommending tonsillectomy for my child?; When is the best age to do the surgery, and why?; What could happen if we delay or don’t have the surgery?; What techniques are available for tonsillectomy, and what are the pros and cons of each?; Which technique do you recommend for my child?; What risks or complications should I be aware of?; How long will my child need to stay in the hospital, and what should I expect at discharge?; What else should I know before deciding?

Prompt wording was standardized across models and remained unchanged throughout the study period to minimize variability attributable to prompt engineering rather than model behavior.

2.3. Physician Scoring Rubric

All AI-generated responses were independently evaluated by two physician raters using a standardized 12-point scoring rubric derived from the 2019 American Academy of Otolaryngology—Head and Neck Surgery Foundation Clinical Practice Guideline for Tonsillectomy in Children. Responses were blinded to AI platform identity prior to evaluation. Each physician independently assigned concordance scores according to the predefined rubric criteria. Inter-rater reliability for total guideline concordance scores was assessed using the intraclass correlation coefficient.

Model responses were evaluated against a 12-point scoring rubric derived from the American Academy of Otolaryngology—Head and Neck Surgery Foundation plain-

language tonsillectomy guideline, which served as the reference standard (Table A1). The rubric evaluated key guideline subtopics relevant to patient education using domain-specific scoring ranges (Table 1). Most domains were scored from 0 to 2, with 2 indicating the closest alignment with guideline recommendations, 1 indicating incomplete alignment, and 0 indicating absent or discordant information. Four more narrowly defined domains—length of stay and discharge criteria, postoperative care guidance, practical implementation considerations, and transparent and reliable resources—were scored from 0 to 1 according to whether the specified information was adequately addressed. These binary domains were intended to assess the presence of defined informational elements rather than distinguish multiple degrees of completeness. Postoperative care guidance was evaluated on a binary domain because omissions of critical safety information and potentially dangerous recommendations were assessed in an additional safety table.

Table 1. Scoring Rubric against Tonsillectomy CPG.

Domain	Max Points	Scoring Criteria
Comprehensive Diagnostic Assessment (Diagnostic criteria)	2	0 = Content lacks evidence of a diagnostic patient assessment.
		1 = Partial or incomplete evidence of diagnostic assessment.
		2 = Comprehensive, well-documented diagnostic patient assessment consistent with guideline criteria including details on both infection and sleep apnea.
Comprehensive Preoperative Assessment (Risks, Benefits, Complications)	2	0 = No evidence of a thorough preoperative assessment.
		1 = Partial evidence of a preoperative assessment including some risks, benefits, and complications.
		2 = Comprehensive, well-documented preoperative patient assessment consistently discussing risks, benefits, and complications as pros and cons of a tonsillectomy.
Clarity on Surgical Procedure	2	0 = Lacks clarity or omits essential procedural details.
		1 = Partial clarity on the surgical procedure identifying only one technique.
		2 = Clear and detailed information on the surgical procedure discussing both intracapsular and extracapsular tonsillectomy.
Effective Risk Mitigation Strategies	2	0 = No or insufficient coverage of complication-mitigation strategies.
		1 = Partial coverage of risk mitigation strategies.
		2 = Comprehensive coverage of evidence-based strategies to mitigate complications including details about diet and hydration.
Length of Stay and Discharge Criteria	1	0 = Inadequate or absent information on length of stay or discharge criteria.
		1 = Clear and specific information on length of stay and discharge criteria.

Table 1. *Cont.*

Domain	Max Points	Scoring Criteria
e	1	0 = Lacks or provides insufficient postoperative care guidance.
		1 = Clear and concise information on postoperative care including recovery time and what to expect.
Practical Implementation Considerations	1	0 = Recommendations are impractical or infeasible in real-world settings.
		1 = Recommendations are practical and feasible in clinical practice.
Transparency of Sources and Rationale	1	0 = Lack of transparency in source attribution or rationale.
		1 = Transparent documentation of sources and rationale for recommendations.
e	12	Higher scores indicate greater concordance with guideline-based, patient-centered counseling.

This safety evaluation includes omission of clinically indicated warnings regarding postoperative hemorrhage, respiratory compromise, and severe dehydration; recommendations involving contraindicated or potentially dangerous medications; inappropriate postoperative monitoring recommendations for high-risk patients; and fabricated contraindications or other potentially harmful advice. Critical safety omissions were defined as failures to provide essential safety information when clinically relevant to the prompt and necessary to guide appropriate caregiver action (Table A1).

Responses were scored based on the presence and accuracy of guideline-concordant information, with higher scores reflecting greater alignment with the reference standard. Total concordance scores were calculated as percentages to facilitate comparison across prompt strategies and models.

2.4. Readability Assessment

Responses were then compared to different readability scales: Flesch Reading Ease (FRE) and the Flesch–Kincaid Grade Level (FKGL). FRE functions on a scale where higher scores indicate easier readability. On the other hand, higher scores on the FKGL indicate higher educational reading levels.

Readability scores were computed for each response and compared across prompt strategies and models. For contextual benchmarking, readability scores were also compared with those of the AAO-HNSF plain-language guideline.

2.5. Statistical Analysis

Descriptive statistics were used to summarize guideline concordance and readability outcomes. LF and YF responses were analyzed separately because of their differing administration frequencies. LF prompts were administered daily over the 60-day study period, whereas YF prompts were administered approximately weekly (Table A2). For each assessment, the eight individual question scores were combined into a single overall guideline concordance percentage for each AI platform.

Separate repeated-measures analyses of variance (ANOVAs) were conducted to compare guideline concordance and readability outcomes across AI platforms within each prompt condition. Assessment dates served as matched measurement occasions across the three platforms. Pairwise comparisons between AI platforms were conducted following the omnibus analyses, with adjustments for multiple comparisons.

To directly compare prompting strategies, a matched-date analysis was conducted using assessment dates on which both LF and YF responses were available. For each AI platform and matched assessment date, the difference in overall guideline concordance was calculated by subtracting the LF score from the YF score. These differences were averaged across the three AI platforms within each assessment date, yielding one mean difference per assessment occasion. The mean difference across assessment dates was evaluated using a one-sample *t*-test against zero. The mean paired difference and corresponding 95% confidence interval were reported. LF responses collected on dates without corresponding YF assessments were retained in the separate LF analyses but excluded from the matched-date comparison.

Temporal trends in guideline concordance were assessed separately for each AI platform using linear regression, with concordance score as the dependent variable and study day as the independent variable. Statistical significance was defined as a two-sided *p*-value < 0.05. All analyses were performed using R version 4.4.2.

3. Results

A total of 168 prompts were evaluated across three LLMs using two distinct prompt structures. Inter-rater reliability for total guideline concordance scores was high at an ICC of 0.96 (95% CI: 0.95–0.99). Across prompt structures, Google AI consistently demonstrated the highest average concordance, with Gemini demonstrating intermediate performance and ChatGPT demonstrating the lowest concordance (Figure 1). However, outputs generated using YF demonstrated significantly higher alignment with guideline subtopics compared with LF, with mean concordance values of 89.5% and 71.4%, respectively ($p < 0.001$). This difference in concordance between prompt strategies was consistently observed across all three evaluated models. The relative separation between YF and LF remained stable throughout the study period, indicating that the observed effect of prompt structure was persistent and not driven by temporal fluctuations.

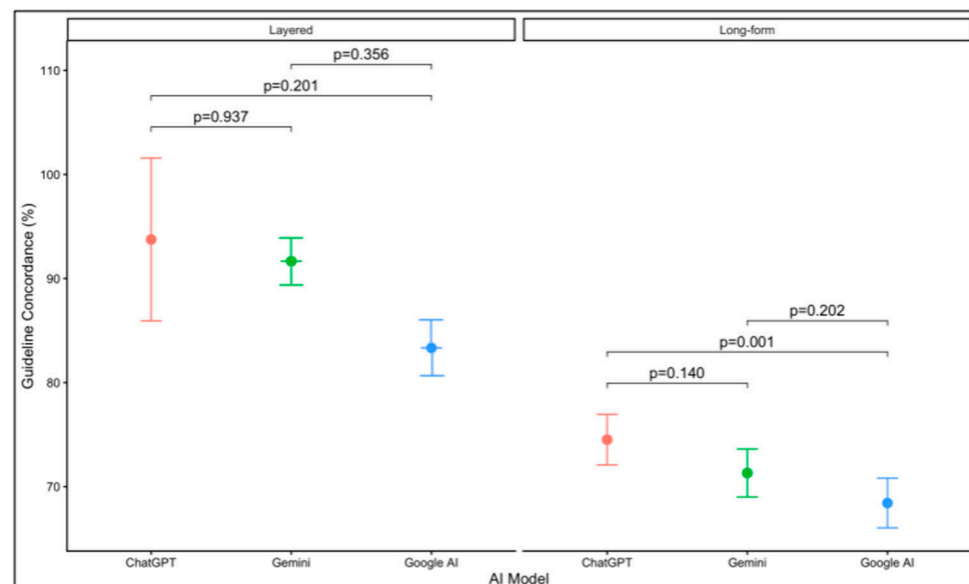


Figure 1. Mean Concordance to Clinical Practice Guidelines Separated by Prompt Type. Mean CPG concordance (%) is shown for layered and long-form prompts across ChatGPT, Gemini, and Google AI, with error bars representing 95% CIs.

Inter-model differences were most pronounced for LF, whereas under YF conditions, concordance values for all three models converged toward higher overall alignment ($p > 0.05$ for all YF comparisons). When analyses were restricted to LF prompts, mean

guideline concordance differed significantly by model. Google AI demonstrated the highest mean concordance at 74.5% (95% CI, 72.0–77.0), followed by Gemini at 71.3% (95% CI, 68.9–73.7), and ChatGPT at 68.4% (95% CI, 66.0–70.9). Mean CPG concordance for Google AI responses was significantly higher than ChatGPT concordance ($p = 0.001$). Gemini demonstrated a 2.6-percentage-point higher mean concordance than ChatGPT ($p = 0.202$), while Google AI exceeded Gemini by 2.8 percentage points ($p = 0.140$).

Temporal analyses demonstrated model-specific trends in guideline concordance. Google AI demonstrated a statistically significant increase in concordance over time, with a positive linear slope of 0.03 ($p = 0.003$), corresponding to incremental improvements in guideline alignment across successive evaluations (Figure 2A). In contrast, Gemini demonstrated no statistically significant temporal change in concordance over time ($p > 0.05$) (Figure 2B), while ChatGPT significantly decreased in concordance over time ($p < 0.001$) (Figure 2C).

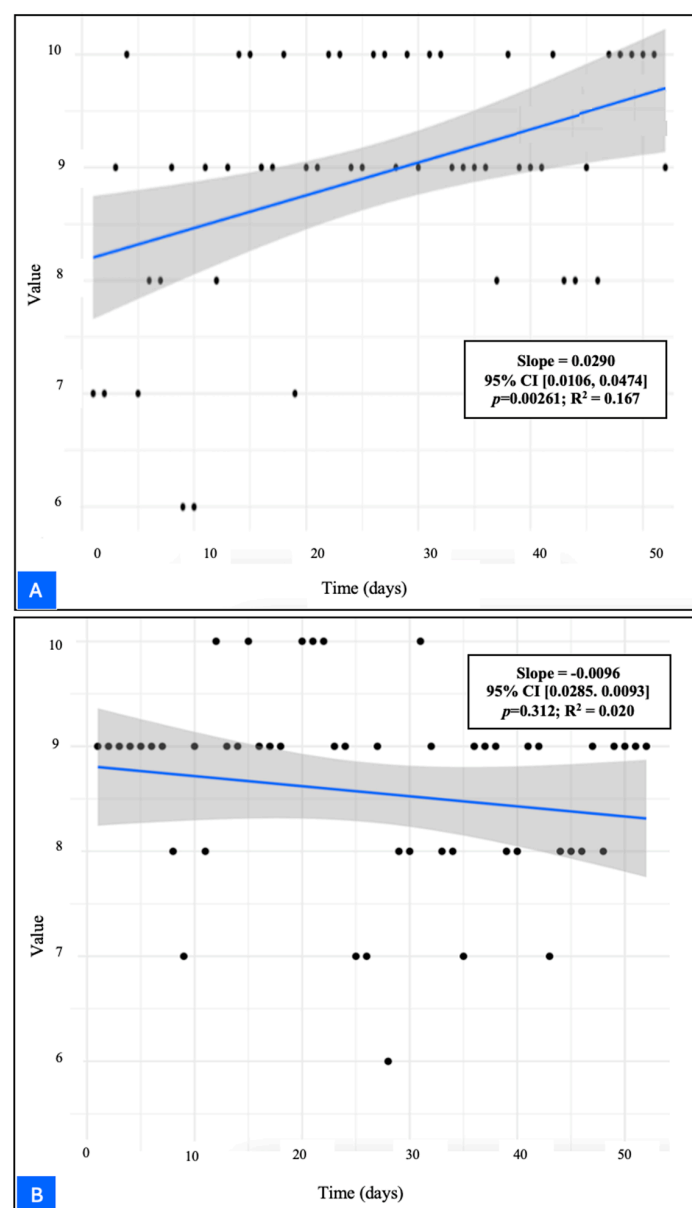


Figure 2. Cont.

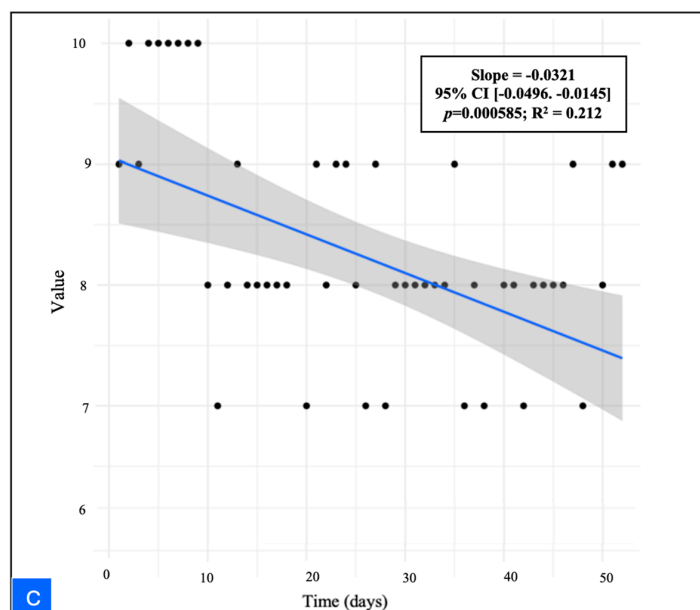


Figure 2. CPG Concordance Averages Over Time. Scatter plots with regression lines show CPG concordance over time for Google AI (A), Gemini (B), and ChatGPT (C). Shaded areas indicate 95% CIs.

When readability metrics were averaged across models, Google AI, Gemini, and ChatGPT demonstrated similar performance. Mean FKGL scores across models clustered between approximately 10 and 12 depending on prompt structure, while mean FRE scores clustered between approximately 30 and 50 (Table 2). No statistically significant differences in readability metrics were observed between models (Table 2).

Table 2. Mean Readability Scores across Model and Prompt Type After Bonferroni Correction.

Prompt Type	Model	Mean Score	Significance vs. ChatGPT	Significance vs. Gemini	Significance vs. Google AI
FRE					
LF	ChatGPT	41.2	—	$p = 0.079$	$p < 0.001$
	Gemini	42.1	$p = 0.079$	—	$p < 0.001$
	Google AI	22.2	$p < 0.001$	$p < 0.001$	—
YF	ChatGPT	44.8	—	$p = 0.023$	$p = 0.019$
	Gemini	40.4	$p = 0.023$	—	$p = 0.721$
	Google AI	48.6	$p = 0.019$	$p = 0.721$	—
FKGL					
LF	ChatGPT	9.8	—	$p < 0.001$	$p < 0.001$
	Gemini	10.6	$p < 0.001$	—	$p < 0.001$
	Google AI	13.6	$p < 0.001$	$p < 0.001$	—
YF	ChatGPT	9.9	—	$p = 0.002$	$p = 0.053$
	Gemini	11.1	$p = 0.002$	—	$p = 0.053$
	Google AI	10	$p = 0.053$	$p = 0.053$	—

Readability metrics differed significantly by prompting strategy and model. When analyzing FRE scores, where higher scores indicate increased readability, Google AI responses to LF performed the worst. Across LF, ChatGPT (41.2) and Gemini (42.1) demonstrated comparable FRE scores, both of which were significantly higher than Google AI (22.2; $p < 0.001$) (Table 2). In contrast, for YF, Google AI achieved the highest FRE score (48.6), outperforming ChatGPT (44.8; $p = 0.019$), while ChatGPT remained significantly more readable than Gemini (40.4; $p = 0.023$). Similar trends were observed with FKGL, where

lower scores indicate easier readability. For LF, ChatGPT (9.8) produced significantly lower grade levels than both Gemini (10.6; $p < 0.001$) and Google AI (13.6; $p < 0.001$). For YF, ChatGPT (9.9) remained significantly lower than Gemini (11.1; $p = 0.002$). Overall, ChatGPT demonstrated the most consistent readability across both prompt types, whereas Google AI showed marked improvement with YF prompting.

These readability differences changed significantly over time. In Google AI, FKGL decreased (-0.09 , $p < 0.001$) while FRE increased (0.55 , $p < 0.001$), indicating improved readability over time. In contrast, ChatGPT demonstrated a modest increase in FRE (0.02 , $p = 0.04$) with minimal change in FKGL. As a result, Google AI achieved the lowest FKGL and highest FRE at later time points, whereas no significant differences were observed between Gemini and ChatGPT (Figure 3).

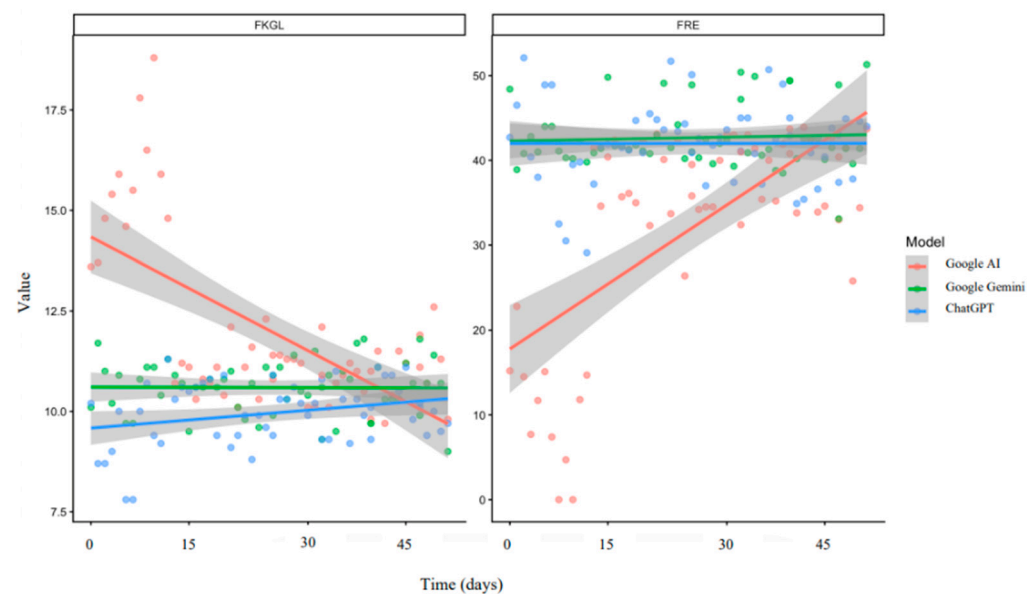


Figure 3. Readability Trends Over Time by Model. Scatter plots with regression lines show FKGL and FRE over time for Google AI, Gemini, and ChatGPT. Shaded areas indicate 95% CIs.

When compared with plain-language guideline benchmarks, YF outputs increased target readability values. YF produced a mean FKGL of 10.3, slightly below the CPG value of 10.6 and a mean FRE of 44.6, exceeding the CPG value of 41.7, indicating an overall more readable output. In contrast, LF consistently exceeded the CPG's FKGL at 11.3 and demonstrated substantially lower FRE, 35.2, indicating reduced accessibility relative to guideline standards (Table 2).

4. Discussion

Prior studies evaluating large language models (LLMs) in clinical settings have demonstrated variable accuracy, incomplete guideline adherence, and inconsistent performance across models and prompt types [6,15–17]. However, few studies have specifically examined LLM performance in the context of tonsillectomy-related clinical guidance, where patient-facing questions often span indications, perioperative expectations, and postoperative care. Our findings build on this literature by demonstrating that prompt structure is a large determinant of guideline concordance, with layered prompting producing consistently higher concordance across all models. While earlier work has largely focused on comparing models head-to-head or assessing overall accuracy [6,18], fewer studies have systematically evaluated how prompt design and temporal model training influence coverage of discrete guideline subtopics relevant to common otolaryngologic procedures such as tonsillectomy.

Across all analyses, statistically significant differences were observed in CPG concordance by prompt structure. Under both types of queries, Google AI demonstrated higher guideline concordance than both Gemini and ChatGPT, while Gemini consistently outperformed ChatGPT but did not reach the performance of Google AI. Model-specific differences in concordance were most pronounced under LF conditions, whereas under YF, concordance values converged across all models with no statistically significant differences observed. Together, these findings suggest that differences in model architecture may influence performance when organization of input is required. Notably, these inter-model differences were attenuated when layered prompting was used, consistent with prior work demonstrating that structured prompts can narrow performance gaps between models [19,20]. The observed advantage of YF aligns with prior research in prompt engineering and task decomposition, which has shown that breaking complex queries into smaller, sequential components improves factual completeness and reduces omission errors in LLM outputs [19,21,22]. Importantly, this effect was observed consistently across Google AI, Gemini, and ChatGPT in the present study, suggesting that structured prompting may mitigate underlying architectural differences between models. CPGs are typically organized in modular formats, with distinct sections addressing diagnosis, staging, management options, and follow-up. Prompting strategies that reflect this structure may therefore facilitate more consistent retrieval and synthesis of guideline-relevant information, as has been suggested in prior evaluations of chain-of-thought and stepwise prompting approaches [20,21].

Temporal trends were also observed for LLM training. Improvement in guideline concordance was observed for Google AI, while Gemini demonstrated no significant change, and ChatGPT demonstrated a significant decline in concordance over time. Notably, for Google AI, FKGL decreased and FRE increased, whereas ChatGPT had an increase in FRE, and Gemini scores remained stable, further highlighting model-specific temporal dynamics. Prior longitudinal analyses of LLM performance have reported heterogeneous trends over time, with some models demonstrating incremental gains following updates and others showing stable performance across evaluation periods [23,24]. The present findings extend this literature by demonstrating that temporal change may differ even among patient-facing systems, underscoring the importance of reporting model identity and evaluation time frame in clinical LLM studies. No platform updates were reported during the 60-day study period; however, publicly accessible AI systems may undergo changes without notification to users. Consequently, we believe the observed temporal trends may reflect undisclosed platform updates and changes in the responses generated by publicly accessible platforms during the study period.

Readability outcomes have been less frequently evaluated in studies of LLM-generated clinical content, despite their relevance for patient-facing applications. Although FRE and FKGL provide standardized, quantitative measures of linguistic readability, they have limitations when applied to medical information. Multisyllabic medical terminology may increase estimated reading difficulty even when appropriately explained, and these measures do not fully capture semantic accessibility or caregiver comprehension. To complement these readability assessments, we employed a customized scoring rubric derived from the AAO-HNSF's patient-facing tonsillectomy guideline materials to independently evaluate the clinical content and guideline concordance of AI-generated responses. Together, these approaches provided complementary assessments of linguistic readability and clinical accuracy. Existing work has shown that LLM outputs often exceed recommended reading levels for patient education materials [25–27]. In our study, readability did exceed the recommended 7th grade level but differed significantly by both prompting strategy and model. While overall averages across models appeared similar, stratified analyses revealed that Google AI performed worst under LF prompting but achieved the highest readability

under YF prompting. Overall, layered prompting was associated with lower reading grade levels and higher reading ease scores across all three systems, bringing outputs closer to plain-language guideline targets. While Google AI, Gemini, and ChatGPT differ in baseline performance and temporal stability, prompt structure exerts a unifying influence on output quality across models.

Collectively, these results reinforce emerging recommendations that evaluations of LLMs in healthcare should consider not only model identity but also prompt design and temporal stability [17,28]. Prompt engineering represents a modifiable factor that may improve guideline adherence without requiring access to proprietary model updates or specialized fine-tuning. This has implications for clinicians, educators, and institutions to accurately guide patients using LLMs for clinical clarification after a visit.

This study is limited by the inherent nuance of prompt design. CPG concordance was assessed and may not capture clinical nuance in individualized patient scenarios, such as questions regarding specific foods or activities. However, this limitation has been noted in prior studies, and we therefore focused on generalizable prompts to maintain consistency [6,18]. Additionally, although FRE and FKGL provide standardized measures of linguistic readability, they offer indirect estimates of accessibility and do not directly assess comprehension or usability among patients and caregivers [27]. To complement these measures, we employed a customized scoring rubric derived from the AAO-HNSF clinical practice guideline and its patient-facing materials to evaluate clinical accuracy and guideline concordance. Future studies may incorporate complementary instruments, such as the Suitability Assessment of Materials (SAM) and Patient Education Materials Assessment Tool (PEMAT), alongside direct caregiver assessments to evaluate broader aspects of understandability and actionability. Finally, the evaluated models and performance metrics reflect a specific time period, and ongoing model updates may alter relative performance, as demonstrated in prior longitudinal analyses [24,29].

5. Conclusions

Both model type and prompt structure significantly influenced the quality of AI-generated tonsillectomy guidance. Among the evaluated systems, Google AI demonstrated the highest overall guideline concordance, with Gemini showed intermediate performance and ChatGPT demonstrated lower concordance, while Google AI was also the only model to exhibit measurable improvement over time. Across all platforms, layered prompting consistently yielded responses that were more closely aligned with established clinical guidelines and demonstrated improved readability compared with LF prompting. However, overall readability remained above recommended patient-education thresholds, highlighting the persistent gap between AI-generated clinical content and materials suitable for direct patient use. Collectively, these findings emphasize that prompt design is a critical determinant of output quality, and that model selection alone is insufficient to ensure guideline-concordant, patient-appropriate guidance. Until further advances are made, AI-generated content related to tonsillectomy should be used as an adjunct to, rather than a replacement for, clinician judgment and oversight.

Author Contributions: Conceptualization, H.L., S.V. and B.M.; methodology, H.L., S.V. and B.M.; formal analysis, H.L.; investigation, H.L., S.S., S.B. and S.V.; data curation, H.L., S.S., S.B. and S.V.; writing—original draft preparation, H.L. and S.V.; writing—review and editing, H.L., S.S., S.B. and S.V.; supervision, B.M.; project administration, B.M. 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 due to exemption by the Institutional Review Board (Pro2026000677).

Informed Consent Statement: Not applicable.

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 authors used ChatGPT (v5.1), Google AI (v2.5), and Google Gemini (v3), for the purposes of querying models using a long and short prompt to evaluate each models' 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:

LLMs	Large Language Models
FRE	Flesch Reading Ease
FKGL	Flesch–Kincaid Grade Level
AAO-HNS	American Academy of Otolaryngology—Head and Neck Surgery
LF	Long-form
YF	Layered-form

Appendix A

Table A1. Critical Safety Omissions and Potentially Harmful Recommendations in AI-Generated Pediatric Tonsillectomy Guidance.

Safety Category	Critical Omission or Potentially Harmful Recommendation
Postoperative hemorrhage	Failure to recognize significant primary or secondary bleeding or recommend urgent evaluation
Respiratory compromise	Failure to recognize airway obstruction, apnea, cyanosis, or severe breathing difficulty
Severe dehydration	Failure to recognize clinically significant dehydration or inability to maintain oral intake
Medication-related harm	Recommendation of contraindicated codeine or tramadol, dangerous medication dosing, or failure to recognize serious opioid-related respiratory depression
High-risk postoperative monitoring	Inappropriate recommendation for routine discharge without indicated monitoring in children younger than 3 years or with severe obstructive sleep apnea
Fabricated contraindications or harmful advice	Invented contraindications, inappropriate recommendations to delay emergency care, or other potentially life-threatening advice

Table A2. Distribution of AI-Generated Responses by Platform and Prompt Type.

AI Platform	LF	YF	Total
ChatGPT	60	8	68
Gemini	60	8	68
Google AI	60	8	68
Total	180	24	204

References

1. Esteva, A.; Robicquet, A.; Ramsundar, B.; Kuleshov, V.; DePristo, M.; Chou, K.; Cui, C.; Corrado, G.; Thrun, S.; Dean, J. A guide to deep learning in healthcare. *Nat. Med.* **2019**, *25*, 24–29. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Thirunavukarasu, A.J.; Ting, D.S.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S. Large language models in medicine. *Nat. Med.* **2023**, *29*, 1930–1940. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Dhar, S.; Kothari, D.; Vasquez, M.; Clarke, T.; Maroda, A.; McClain, W.G.; Sheyn, A.; Tuliszewski, R.M.; Tang, D.M.; Rangarajan, S.V. The utility and accuracy of ChatGPT in providing post-operative instructions following tonsillectomy: A pilot study. *Int. J. Pediatr. Otorhinolaryngol.* **2024**, *179*, 111901. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Langlie, J.; Kamrava, B.; Pasick, L.J.; Mei, C.; Hoffer, M.E. Artificial intelligence and ChatGPT: An otolaryngology patient's ally or foe? *Am. J. Otolaryngol.* **2024**, *45*, 104220. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Limna, P.; Kraiwanit, T.; Jangjarat, K.; Klayklung, P.; Chocksathaporn, P. The use of ChatGPT in the digital era: Perspectives on chatbot implementation. *J. Appl. Learn. Teach.* **2023**, *6*, 64–74. [\[CrossRef\]](#)
6. Kung, T.H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digit. Health* **2023**, *2*, e0000198. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Parikh, P.M.; Shah, D.M.; Parikh, K.P. Judge Juan Manuel Padilla Garcia, ChatGPT, and a controversial medicolegal milestone. *Indian J. Med. Sci.* **2023**, *75*, 3–8. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Mete, U. Evaluating the Performance of ChatGPT, Gemini, and Bing Compared with Resident Surgeons in the Otorhinolaryngology In-service Training Examination. *Turk. Arch. Otorhinolaryngol.* **2024**, *62*, 48–57. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Suresh, K.; Cohen, M.S.; Hartnick, C.J.; Bartholomew, R.A.; Lee, D.J.; Crowson, M.G. Making Use of Artificial Intelligence–Generated Synthetic Tympanic Membrane Images. *JAMA Otolaryngol. Head Neck Surg.* **2023**, *149*, 555–556. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Ayoub, N.F.; Lee, Y.-J.; Grimm, D.; Balakrishnan, K. Comparison Between ChatGPT and Google Search as Sources of Postoperative Patient Instructions. *JAMA Otolaryngol. Head Neck Surg.* **2023**, *149*, 556–558. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Hall, M.J.; Schwartzman, A.; Zhang, J.; Liu, X. *Ambulatory Surgery Data from Hospitals and Ambulatory Surgery Centers: United States 2010*; National Health Statistics Report; National Center for Health Statistics: Hyattsville, MD, USA, 2017; Volume 102, pp. 1–15.
12. Windfuhr, J.P.; Chen, Y.-S.; Remmert, S. Hemorrhage following tonsillectomy and adenoidectomy in 15,218 patients. *Otolaryngol. Head Neck Surg.* **2005**, *132*, 281–286. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Johnson, L.B.; Elluru, R.G.; Myer, C.M., III. Complications of Adenotonsillectomy. *Laryngoscope* **2002**, *112*, 35–36. [\[CrossRef\]](#)
14. Finestone, S.A.; Giordano, T.; Mitchell, R.B.; Walsh, S.A.; O'Connor, S.S.; Satterfield, L.M. Plain Language Summary for Patients: Tonsillectomy in Children. *Otolaryngol. Head Neck Surg.* **2019**, *160*, 206–212. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. Large language models encode clinical knowledge. *Nature* **2023**, *620*, 172–180. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Gilson, A.; Safranek, C.W.; Huang, T.; Socrates, V.; Chi, L.; Taylor, R.A.; Chartash, D. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med. Educ.* **2023**, *9*, e45312. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Nori, H.; King, N.; McKinney, S.M.; Carignan, D.; Horvitz, E. Capabilities of GPT-4 on Medical Challenge Problems. *arXiv* **2023**, arXiv:2303.13375.
18. Wang, X.; Ye, H.; Zhang, S.; Yang, M.; Wang, X. Evaluation of the Performance of Three Large Language Models in Clinical Decision Support: A Comparative Study Based on Actual Cases. *J. Med. Syst.* **2025**, *49*, 23. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv* **2023**, arXiv:2201.11903.
20. Liu, J.; Liu, F.; Wang, C.; Liu, S. Prompt Engineering in Clinical Practice: Tutorial for Clinicians. *J. Med. Internet Res.* **2025**, *27*, e72644. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Kojima, T.; Gu, S.S.; Reid, M.; Matsuo, Y.; Iwasawa, Y. Large Language Models are Zero-Shot Reasoners. *arXiv* **2022**, arXiv:2205.11916v4.
22. Zaghir, J.; Naguib, M.; Bjelogrić, M.; Névél, A.; Tannier, X.; Lovis, C. Prompt Engineering Paradigms for Medical Applications: Scoping Review. *J. Med. Internet Res.* **2024**, *26*, e60501. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Li, Z.; Xia, L.; Tang, J.; Xu, Y.; Shi, L.; Xia, L.; Yin, D.; Huang, C. UrbanGPT: Spatio-Temporal Large Language Models. In Proceedings of the KDD '24: The 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Barcelona, Spain, 25–29 August 2024; pp. 5351–5362.
24. Khraishi, R.; Zafar, I.; Myles, K.; Cowan, G.A. Evaluating Performance Drift From Model Switching in Multi-turn LLM Systems. *arXiv* **2026**, arXiv:2603.03111.

25. Friedman, C.; Rubin, J.; Brown, J.; Buntin, M.; Corn, M.; Etheredge, L.; Gunter, C.; Musen, M.; Platt, R.; Stead, W.; et al. Toward a science of learning systems: A research agenda for the high-functioning Learning Health System. *J. Am. Med. Inform. Assoc.* **2015**, *22*, 43–50. [[CrossRef](#)] [[PubMed](#)]
26. Eid, K.; Eid, A.; Wang, D.; Raiker, R.S.; Chen, S.B.; Nguyen, J. Optimizing Ophthalmology Patient Education via ChatBot-Generated Materials: Readability Analysis of AI-Generated Patient Education Materials and The American Society of Ophthalmic Plastic and Reconstructive Surgery Patient Brochures. *Ophthalmic Plast. Reconstr. Surg.* **2024**, *40*, 212–216. [[CrossRef](#)] [[PubMed](#)]
27. Aydin, S.; Karabacak, M.; Vlachos, V.; Margetis, K. Large language models in patient education: A scoping review of applications in medicine. *Front. Med.* **2024**, *11*, 1477898. [[CrossRef](#)] [[PubMed](#)]
28. Hopewell, S.; Chan, A.-W.; Collins, G.S.; Hróbjartsson, A.; Moher, D.; Schulz, K.F.; Tunn, R.; Aggarwal, R.; Berkwits, M.; Berlin, J.A.; et al. CONSORT 2025 statement: Updated guideline for reporting randomized trials. *Nat. Med.* **2025**, *31*, 1776–1783. [[CrossRef](#)] [[PubMed](#)]
29. Li, C.; Wang, Y.; Zhao, X.; Li, J.; Wang, H.; Ren, Y.; Sun, H.; Zhu, X.; Song, Q.; Wang, J. Comparative Analysis of Intestinal Inflammation and Microbiota Dysbiosis of LPS-Challenged Piglets between Different Breeds. *Animals* **2024**, *14*, 665. [[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.