

Comparative evaluation of four large language models for postoperative patient education after septorhinoplasty: A bilingual cross-sectional study

Zülküf Küçüktağ¹, Aykut Özdoğan¹, Gökçe Saygı Uysal¹, Yağmur Uluhan¹

¹Department of Otorhinolaryngology-Head and Neck Surgery, Etlik City Hospital, Ankara, Türkiye

Cite as: Küçüktağ Z, Özdoğan A, Saygı Uysal G, Uluhan Y. Comparative evaluation of four large language models for postoperative patient education after septorhinoplasty: A bilingual cross-sectional study. Northwestern Med J. 2026;6(4):317-329.

Abstract

Aim: To compare the quality, readability, and emergency referral performance of four major language models in answering post-septorhinoplasty patient questions in Turkish and English.

Materials and Methods: A cross-sectional evaluation study comprised 31 questions: 20 routine postoperative and 11 emergency/red-flag scenarios. Each question was submitted once to ChatGPT, Google Gemini, DeepSeek, and Claude in Turkish and English within independent conversation windows. Two otolaryngologists independently evaluated the responses using the Quality Analysis of Medical Artificial Intelligence tool. English responses were additionally assessed using Flesch Reading Ease, Flesch-Kincaid Grade Level, and Simple Measure of Gobbledygook indices. Emergency referral accuracy was evaluated separately.

Results: Total quality scores differed significantly among models for Turkish routine postoperative questions ($p<0.001$, Kendall's $W=0.775$) and for Turkish emergency/red-flag questions ($p=0.009$, $W=0.350$), whereas the corresponding overall differences for English questions were not statistically significant. Within-model language comparisons showed significant Turkish-English differences for routine questions across all four models, whereas no language-related differences were observed for emergency/red-flag questions. ChatGPT, DeepSeek, and Claude provided appropriate emergency referrals in all 22 responses, whereas Gemini did so in 16 of 22 responses (72.7%). All models exceeded recommended readability thresholds for patient education materials.

Conclusion: The performance of large language models in post-septorhinoplasty patient education varied by language context and question category. Although emergency referral performance was generally high, model-specific omissions occurred, and all models exceeded recommended readability thresholds. These findings support language-specific evaluation, simplification to improve readability, and expert review before clinical use.

Keywords: artificial intelligence, large language models, patient education, rhinoplasty, septorhinoplasty, postoperative care

Corresponding author: Zülküf Küçüktağ **E-mail:** zkucuk79@yahoo.com

Received: 11.06.2026 **Accepted:** 23.09.2026 **Published:** 25.10.2026

Copyright © 2026 The Author(s). This is an open-access article published by Bolu İzzet Baysal Training and Research Hospital under the terms of the [Creative Commons Attribution License \(CC BY\)](#) which permits unrestricted use, distribution, and reproduction in any medium or format, provided the original work is properly cited.

INTRODUCTION

Artificial intelligence (AI)-based large language models (LLMs) show increasing potential to support patient communication and education in healthcare. Postoperative information provision is one area where this potential is particularly evident; in recent years, studies evaluating the responses provided by LLMs to postoperative patient questions have begun to be published, primarily in the field of otolaryngology but also across various other surgical specialties (1,2). These studies address various aspects, including the alignment of AI tools with clinical guidelines, their potential contributions to patient accessibility, and the challenges in integrating them into healthcare systems (3,4).

In response to this need, studies investigating the use of LLM models in postoperative patient counseling for rhinoplasty and related facial plastic surgery are on the rise (5,6). These studies examine the capacity of AI models to provide reliable and patient-centered information on common concerns related to rhinoplasty and postoperative recovery (7,8). The postoperative course following septorhinoplasty includes a range of expected symptoms, such as swelling, bruising, nasal obstruction, and discomfort that patients may find difficult to distinguish from findings requiring urgent medical assessment. Patients may therefore seek rapid guidance from easily accessible AI-based tools when deciding whether a symptom represents a normal part of recovery or a potential complication. In this context, inaccurate reassurance, incomplete advice, or failure to recommend timely medical evaluation could have direct patient-safety implications. Evaluating not only the overall quality of information generated by LLMs but also their ability to distinguish routine postoperative concerns from emergency and red-flag scenarios is therefore clinically relevant.

The performance of LLMs in different languages and their reliability for patients constitute another important aspect of research in this field. Multilingual comparative studies have shown that LLMs can generate competitive responses in languages other than English (9); however, there are concerns regarding the readability and comprehensibility of AI responses for patients (10). From the patient's perspective,

trust in and acceptance of AI tools have been shown to vary according to demographic characteristics and technology use; however, whether patient trust aligns with clinical reliability remains a subject of debate (11).

However, the vast majority of existing studies evaluate a single model, are conducted in English only, and do not systematically address the distinction between routine postoperative questions and emergency clinical situations. Furthermore, the performance of AI tools in different language groups, such as Turkish-speaking patient populations, remains largely unknown. Although recent studies have compared multiple LLMs in rhinoplasty-related patient education, including the evaluation of postoperative red-flag scenarios (12,13), evidence regarding their comparative performance across different languages remains limited.

This study aimed to evaluate and compare responses from ChatGPT, Gemini, DeepSeek, and Claude to post-septorhinoplasty patient questions in Turkish and English using the Quality Analysis of Medical Artificial Intelligence (QAMAI) tool; assess total QAMAI scores and effect sizes; determine English response readability via Flesch Reading Ease (FRE), Flesch-Kincaid Grade Level (FKGL), and Simple Measure of Gobbledygook (SMOG) indices; and compare emergency referral accuracy across languages. To our knowledge, this is the first study to evaluate questions from septorhinoplasty patients using a bilingual (Turkish-English) design and a systematic distinction between routine and emergency cases. The primary research question was whether the quality of LLM-generated responses to questions from post-septorhinoplasty patients differs across AI models, languages (Turkish vs English), and question types (routine postoperative vs emergency/red-flag). We hypothesized that response quality and emergency referral performance would vary across models and language contexts.

MATERIALS AND METHODS

Study design and ethics

This study is a cross-sectional evaluation comparing the responses of four AI models to patient questions posed following septorhinoplasty. Only text generated by AI systems was analyzed; no patient data were used.

Therefore, approval from an ethics committee was not required. The study was reported with reference to the Chatbot Assessment Reporting Tool (CHART) (14), a reporting framework specifically developed for studies evaluating chatbot-generated health advice. Given the cross-sectional comparative design of the study, relevant principles of the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) (15) statement were also considered.

Artificial intelligence models

Four AI models were evaluated: ChatGPT (GPT-5.3 Instant, OpenAI, San Francisco, USA), Google Gemini 2.0 Flash (Google LLC, Mountain View, USA), DeepSeek V3 (DeepSeek AI, Hangzhou, China), and Claude Sonnet 4.6 (Anthropic, San Francisco, USA). These models were selected to represent widely used, general-purpose LLMs developed by different organizations that are capable of generating multilingual responses relevant to patient-facing health information. Their inclusion also enabled comparison across multiple contemporary LLM platforms, rather than limiting the evaluation to a single model or developer. ChatGPT was accessed through a ChatGPT Plus subscription (GPT-5.3 Instant), while Claude was accessed through a paid subscription. Gemini and DeepSeek were accessed through their web-based interfaces; their subscription status was not recorded during data collection. All queries were performed through the models' web-based user interfaces without API access. Data were collected between April 1 and April 30, 2026, using the following LLMs: ChatGPT (GPT-5.3 Instant), Google Gemini 2.0 Flash, DeepSeek V3, and Claude Sonnet 4.6. To implement a blind evaluation procedure, the models were coded as M1 (ChatGPT), M2 (Gemini 2.0 Flash), M3 (DeepSeek V3), and M4 (Claude). Each Turkish and English question was submitted once to each model in a new and independent conversation window to minimize contextual bias. A standard prompt structure was used for all queries: "You are answering a question from a patient who has undergone septorhinoplasty surgery. Provide an answer that is accurate, clear, and patient-friendly according to current medical standards." The methodological framework was designed to align with the multi-model AI comparison study in the field of otolaryngology published by Alsudays et al. (1).

Question groups

The question set was developed by the research team based on common postoperative patient concerns and clinically relevant warning signs encountered after septorhinoplasty, and was not derived from a single formal clinical guideline. The final set of 31 questions was selected to provide broad coverage of predefined routine and emergency/red-flag domains. Because this study evaluated AI-generated responses using a predefined question set rather than a patient- or population-based sample, no formal a priori sample size or power calculation was performed. The questions were divided into two main categories. Group A — Routine Postoperative Questions (n=20): questions covering expected postoperative findings, activity restrictions, and common patient concerns. These span a wide range, from managing swelling and bruising to interpreting asymmetrical swelling. Group B — Emergency and Red Flag Questions (n=11): This group includes clinical scenarios that require urgent medical evaluation, such as active nasal bleeding, signs of surgical site infection, possible vascular compromise, and orbital complications. All questions were prepared in both Turkish and English, and their translation accuracy and clinical content validity were verified by a bilingual ENT specialist. Questions were reviewed by two senior ENT specialists for clinical relevance and comprehensiveness prior to data collection, providing face validity for the question set. Each question was submitted separately in Turkish and English within independent conversation windows to prevent cross-language contextual bias.

Assessment tool

The responses were analyzed using the QAMAI tool, which was developed and validated by Vaira et al. (16) to assess the quality of health information provided by AI platforms. QAMAI measures five core criteria on a Likert scale ranging from 1 (inadequate) to 5 (excellent): Accuracy, Clarity, Relevance, Completeness, and Usefulness. The maximum total QAMAI score is 25; the total QAMAI score is defined as the sum of these five criteria (range: 5–25). Two independent ENT specialists (R1 and R2) with ≥10 years of clinical experience evaluated all responses independently, with the models' identities blinded. The study comprised 31 questions × 4 models × 2

languages, yielding 248 model-generated responses. Each response was evaluated by two raters across five QAMAI criteria, corresponding to 2,480 individual criterion-level ratings (248 responses $\times$ 5 criteria $\times$ 2 raters). For inter-rater agreement analyses, ratings from the two evaluators were paired for each criterion, yielding 1,240 expected criterion-level rating pairs (248 responses $\times$ 5 criteria). Because one criterion-level rating was missing, 1,239 complete criterion-level rating pairs were available for analysis.

Readability analysis

Three readability metrics were applied to the English responses: FRE, FKGL, and the SMOG index (17,18). The American Medical Association (AMA) and the U.S. National Institutes of Health (NIH) recommend the SMOG index as the primary measure of readability for patient education materials; they set the target level at SMOG ≤ 8 and FKGL ≤ 6 (6th-grade reading level). Since no psychometrically validated readability tool is available for Turkish, responses in Turkish were not included in this analysis.

Accuracy of urgent referral recommendations

For Group B responses, both raters independently assessed whether the AI model (a) correctly identified the clinical urgency of the scenario, and (b) included a clear recommendation for the patient to visit the emergency department or to contact their physician immediately. Responses were classified into two categories: correct (appropriate and sufficient emergency referrals) and incorrect (inadequate communication of urgency or a lack of referral).

Statistical Analysis

The distribution of continuous variables was formally assessed using the Shapiro–Wilk test. Response length was quantified as the number of whitespace-delimited words in each model-generated response, excluding the question text and model labels. Because the same questions were answered by all four models and the response-length distributions did not consistently satisfy parametric assumptions, between-model differences in response length were analyzed using the Friedman test, with the question as the matched unit. Because the same LLM responses were

independently evaluated by both raters, the two sets of QAMAI ratings were treated as paired rather than independent observations. For the primary QAMAI analyses, scores assigned by the two raters were averaged at the individual response level, yielding a single combined score for each question–model combination. Total QAMAI scores were calculated only when all five criterion scores were available from both raters; missing values were not imputed. Because QAMAI scores are bounded ordinal ratings and several distributions showed ceiling effects and departures from normality, non-parametric methods were used. Model comparisons within each language and question group were performed using the Friedman test, with question as the matched unit. Significant omnibus tests were followed by pairwise Wilcoxon signed-rank tests with Holm correction. Kendall’s W was reported as the effect-size measure for Friedman tests. For the mean total QAMAI scores, 95% confidence intervals were estimated using percentile bootstrap resampling of response-level scores (20,000 resamples).

Turkish–English comparisons of total QAMAI scores within each model were performed using paired Wilcoxon signed-rank tests, with the same question serving as the matching unit. Holm correction was applied within each question group for the four prespecified model-wise language comparisons, and matched-pairs rank-biserial correlation (r_{rb}) was reported as the effect-size measure.

Inter-rater agreement analyses were performed using available complete rating pairs. Linearly weighted Cohen’s kappa was calculated for individual QAMAI criteria, and intraclass correlation coefficients [ICC (2,1) and ICC (2,k)] were calculated for total QAMAI scores. For total QAMAI ICC estimates, 95% confidence intervals were obtained using nonparametric bootstrap resampling of complete response-level pairs (20,000 resamples).

Emergency referral accuracy was compared across the four models within each language using Cochran’s Q test, accounting for the matched structure of the 11 emergency/red-flag questions. When the omnibus test was significant, pairwise comparisons were performed using exact McNemar tests with Holm correction. Turkish–English differences in emergency referral

accuracy within each model were assessed using exact McNemar tests for paired binary data. For emergency-referral accuracy proportions, 95% confidence intervals were calculated using Wilson score intervals. Overall, Turkish–English emergency referral rates were reported only descriptively.

Because the Turkish and English datasets comprised paired versions of the same questions, inferential analyses were performed separately for each language; no inferential analysis was conducted on pooled Turkish–English data. All analyses were performed using Python/SciPy v1.11. Statistical significance was set at $p < 0.05$.

RESULTS

Inter-rater agreement

The inter-rater reliability analysis yielded an overall linear-weighted κ of 0.284 at the criterion level. For total QAMAI scores, the overall ICC (2,1) was 0.364 (95% CI, 0.260–0.457) and ICC (2,k) was 0.534 (95% CI, 0.412–0.627) (Table 1). Across language–question group datasets, total-score ICC (2,1) values ranged from 0.415 to 0.553 in Turkish and from 0.179 to 0.208 in English. Overall, 1,218 of 1,239 complete paired criterion-level ratings (98.3%) differed by no more than 1 scale point, and no discrepancy exceeded 2 points. One criterion-level rating was inadvertently

Table 1. Inter-rater agreement across all response categories.

Dataset	Simple κ	Weighted κ (linear)	Interpretation	Exact agreement	± 1 agreement	ICC (2,1) [95% CI]	ICC (2,k) [95% CI]
Criterion-level scores (5 criteria $\times$ all responses)							
Turkish — Group A (n=400)	0.252	0.338	Fair	255/400 (63.7%)	393/400 (98.2%)	0.432	0.603
Turkish — Group B (n=220)	0.323	0.363	Fair	149/220 (67.7%)	216/220 (98.2%)	0.409	0.581
English — Group A (n=399)	0.188	0.190	Slight	261/399 (65.4%)	392/399 (98.2%)	0.161	0.277
English — Group B (n=220)	0.219	0.234	Fair	145/220 (65.9%)	217/220 (98.6%)	0.253	0.404
Overall (n=1,239)	0.239	0.284	Fair	810/1,239 (65.4%)	1,218/1,239 (98.3%)	0.332	0.499
Total QAMAI score (range 5–25)							
Turkish — Group A (n=80)	—	—	—	—	—	0.415 [0.248–0.555]	0.587 [0.397–0.713]
Turkish — Group B (n=44)	—	—	—	—	—	0.553 [0.321–0.718]	0.713 [0.486–0.836]
English — Group A (n=79)	—	—	—	—	—	0.208 [–0.004–0.394]	0.344 [–0.008–0.565]
English — Group B (n=44)	—	—	—	—	—	0.179 [–0.120–0.432]	0.304 [–0.273–0.603]
Overall (n=247)	—	—	—	—	—	0.364 [0.260–0.457]	0.534 [0.412–0.627]

Weighted κ (linear weights) = primary metric for ordinal Likert data. ICC(2,1) = single-rater; ICC(2,k) = average-measures intraclass correlation coefficient (two-way random effects, absolute agreement). Values in brackets represent 95% confidence intervals. The 95% confidence intervals for total QAMAI ICC estimates were obtained using nonparametric bootstrap resampling of complete response-level pairs (20,000 resamples). κ interpretation (Landis & Koch, 1977): 0.00–0.20 = Slight; 0.21–0.40 = Fair; 0.41–0.60 = Moderate. Overall, 98.3% of complete paired criterion-level ratings differed by no more than 1 scale point. One criterion-level rating (Evaluator 2, Accuracy, English Group A) was missing, resulting in 399/400 complete criterion-level rating pairs and 79/80 complete total QAMAI rating pairs; no imputation was performed. QAMAI: Quality Analysis of Medical Artificial Intelligence.

left blank by Evaluator 2 in the English routine dataset and was therefore excluded from the corresponding complete-pair analyses; no imputation was performed.

Total QAMAI score

As shown in Table 2, total QAMAI scores differed significantly across models for Turkish routine postoperative questions ($\chi^2=46.52$, $p<0.001$, Kendall's $W=0.775$). Pairwise Wilcoxon signed-rank tests with Holm correction showed significant differences between ChatGPT and Gemini (adjusted $p=0.001$), between ChatGPT and DeepSeek (adjusted $p<0.001$), between ChatGPT and Claude (adjusted $p=0.002$), between Gemini and DeepSeek (adjusted $p<0.001$), and between DeepSeek and Claude (adjusted $p<0.001$); however, the difference between Gemini and Claude was not significant (adjusted $p=0.207$). No significant overall model differences were observed for English routine postoperative questions ($\chi^2=7.60$, $p=0.055$, Kendall's $W=0.133$). For emergency/red-flag questions, a significant overall model difference was observed in Turkish ($\chi^2=11.56$, $p=0.009$, Kendall's $W=0.350$), although, after the Holm correction, no pairwise comparison remained significant. The

corresponding overall model difference for English was not statistically significant ($\chi^2=7.63$, $p=0.054$, Kendall's $W=0.231$).

QAMAI criteria scores-group A: Routine postoperative questions

As shown in Table 3, significant differences among models were observed for the Turkish routine questions in Accuracy ($\chi^2=30.86$, $p<0.001$, Kendall's $W=0.514$), Clarity ($\chi^2=25.90$, $p<0.001$, Kendall's $W=0.432$), Completeness ($\chi^2=50.83$, $p<0.001$, Kendall's $W=0.847$), and Usefulness ($\chi^2=37.79$, $p<0.001$, Kendall's $W=0.630$); Relevance did not differ significantly ($p=0.392$). Holm-adjusted pairwise comparisons showed significant differences between DeepSeek and each of the other three models for Accuracy. For clarity, significant differences were observed between Gemini and each of ChatGPT, DeepSeek, and Claude. All six pairwise comparisons were significant for both Completeness and Usefulness. For English routine questions, significant model differences were observed for Clarity ($\chi^2=22.96$, $p<0.001$, $W=0.383$) and Completeness ($\chi^2=9.12$, $p=0.028$, $W=0.152$), whereas Accuracy, Relevance, and Usefulness did not

Table 2. Total QAMAI score by model, language, and question group.

Group	Language	ChatGPT Mean±SD [95% CI]	Gemini Mean±SD [95% CI]	DeepSeek Mean±SD [95% CI]	Claude Mean±SD [95% CI]	χ^2	p	Kendall's W
Routine Postoperative Questions	TR	23.20±0.70 [22.88–23.45]	24.08±0.80 [23.73–24.40]	20.88±1.18 [20.38–21.38]	24.35±0.69 [24.02–24.60]	46.52	<0.001*	0.775
Routine Postoperative Questions†	EN	23.85±0.92 [23.45–24.23]	23.00±1.30 [22.40–23.50]	23.68±0.77 [23.35–24.00]	23.66±1.25 [23.11–24.18]	7.60	0.055	0.133
Emergency and Red Flag Questions	TR	23.82±1.10 [23.14–24.36]	22.45±1.59 [21.55–23.32]	22.41±1.28 [21.64–23.09]	24.00±1.95 [22.77–24.82]	11.56	0.009*	0.350
Emergency and Red Flag Questions	EN	23.95±0.69 [23.55–24.32]	22.68±1.12 [22.05–23.32]	23.00±1.18 [22.32–23.64]	23.91±1.04 [23.32–24.45]	7.63	0.054	0.231

Values are presented as mean ± SD [95% CI] based on response-level scores obtained by averaging the two raters. The 95% confidence intervals for mean total QAMAI scores were estimated using percentile bootstrap resampling (20,000 resamples). Model comparisons within each language and question group were performed using the Friedman test, with question as the matched unit. Kendall's W is reported as the effect-size measure. When the omnibus Friedman test was significant, pairwise Wilcoxon signed-rank tests with Holm correction were performed. †For English routine postoperative questions, one Claude response with an incomplete rater assessment was excluded; Claude $n=19$, while the other model-specific descriptive estimates used $n=20$. The complete-case Friedman analysis used $n=19$ matched questions. TR = Turkish; EN = English. Total QAMAI score range: 5–25. $p < 0.05$.

differ significantly. Pairwise comparisons for Clarity were significant for ChatGPT versus Gemini, ChatGPT versus Claude, Gemini versus DeepSeek, and DeepSeek versus Claude after Holm correction. Although the omnibus test for Completeness was significant, no pairwise comparison remained significant after Holm correction.

QAMAI criteria scores-group B: Emergency and red flag questions

As shown in Table 3, significant model differences were observed for Turkish emergency/red-flag questions in Clarity ($\chi^2=16.50$, $p<0.001$, Kendall's $W=0.500$) and Completeness ($\chi^2=16.42$, $p<0.001$, Kendall's $W=0.498$); whereas Accuracy, Relevance, and

Table 3. QAMAI criterion scores by question group and language.

Group	Lang.	Criterion	n	ChatGPT Mean±SD	Gemini Mean±SD	DeepSeek Mean±SD	Claude Mean±SD	χ^2	p	Kendall's W
Routine	TR	Accuracy	20	4.90±0.26	4.90±0.21	4.38±0.46	4.95±0.22	30.86	<0.001*	0.514
Routine	TR	Clarity	20	4.47±0.11	4.90±0.21	4.40±0.42	4.58±0.24	25.90	<0.001*	0.432
Routine	TR	Relevance	20	4.97±0.11	5.00±0.00	5.00±0.00	5.00±0.00	3.00	0.392	0.050
Routine	TR	Completeness	20	4.40±0.21	4.65±0.33	3.17±0.44	4.92±0.18	50.83	<0.001*	0.847
Routine	TR	Usefulness	20	4.45±0.15	4.62±0.32	3.92±0.44	4.90±0.26	37.79	<0.001*	0.630
Routine	EN	Accuracy	19	4.82±0.34	4.71±0.51	4.84±0.24	4.76±0.31	1.19	0.754	0.021
Routine	EN	Clarity	20	4.97±0.11	4.53±0.38	4.85±0.24	4.50±0.43	22.96	<0.001*	0.383
Routine	EN	Relevance	20	5.00±0.00	5.00±0.00	5.00±0.00	5.00±0.00	0.00	1.000	0.000
Routine	EN	Completeness	20	4.45±0.32	4.40±0.31	4.47±0.38	4.72±0.41	9.12	0.028*	0.152
Routine	EN	Usefulness	20	4.60±0.31	4.38±0.36	4.50±0.28	4.65±0.43	5.82	0.121	0.097
Emergency/ Red flag	TR	Accuracy	11	4.77±0.34	4.55±0.47	4.64±0.45	4.77±0.61	4.62	0.202	0.140
Emergency/ Red flag	TR	Clarity	11	4.86±0.23	4.45±0.15	4.41±0.20	4.73±0.34	16.50	<0.001*	0.500
Emergency/ Red flag	TR	Relevance	11	5.00±0.00	5.00±0.00	5.00±0.00	5.00±0.00	0.00	1.000	0.000
Emergency/ Red flag	TR	Completeness	11	4.55±0.35	4.32±0.51	3.95±0.27	4.82±0.46	16.42	<0.001*	0.498
Emergency/ Red flag	TR	Usefulness	11	4.64±0.45	4.14±0.60	4.41±0.58	4.68±0.64	6.22	0.101	0.188
Emergency/ Red flag	EN	Accuracy	11	4.91±0.20	4.73±0.34	4.73±0.26	4.73±0.26	3.69	0.297	0.112
Emergency/ Red flag	EN	Clarity	11	4.91±0.20	4.45±0.42	4.91±0.20	4.64±0.45	10.66	0.014*	0.323
Emergency/ Red flag	EN	Relevance	11	5.00±0.00	5.00±0.00	4.91±0.20	4.82±0.25	8.25	0.041*	0.250
Emergency/ Red flag	EN	Completeness	11	4.55±0.27	4.27±0.61	4.18±0.40	5.00±0.00	17.07	<0.001*	0.517
Emergency/ Red flag	EN	Usefulness	11	4.59±0.30	4.23±0.41	4.27±0.41	4.73±0.26	11.04	0.012*	0.335

Values are mean±SD based on criterion-level scores averaged across the two raters. Model comparisons were performed using the Friedman test, with Kendall's W as the effect size; significant omnibus tests were followed by pairwise Wilcoxon signed-rank tests with Holm correction. †For English routine Accuracy, one incomplete rater assessment resulted in n=19; all other routine analyses included n=20 and emergency/red-flag analyses n=11. TR=Turkish; EN=English. $p<0.05$.

Usefulness did not differ significantly. Holm-adjusted pairwise comparisons showed significant differences in Clarity between ChatGPT and Gemini and between ChatGPT and DeepSeek; for Completeness, only the comparison between DeepSeek and Claude remained significant. For English emergency/red-flag questions, significant overall differences were observed for Clarity ($\chi^2=10.66$, $p=0.014$, $W=0.323$), Relevance ($\chi^2=8.25$, $p=0.041$, $W=0.250$), Completeness ($\chi^2=17.07$, $p<0.001$, $W=0.517$), and Usefulness ($\chi^2=11.04$, $p=0.012$, $W=0.335$); whereas Accuracy did not differ significantly. After the Holm correction, no individual pairwise comparisons remained significant for Clarity, Relevance, or Usefulness. For completeness, significant differences remained between Claude and each of ChatGPT, Gemini, and DeepSeek.

Readability Analysis and Response Length

As shown in Table 4, the mean readability scores of all models exceeded the recommended AMA/NIH targets for patient education materials (SMOG ≤ 8 and FKGL ≤ 6). SMOG scores ranged from 11.0 ± 0.2 (DeepSeek, Group A) to 12.4 ± 0.2 (Claude, Group B), corresponding to a reading level of grades 11–12, while FRE scores ranged from 36.2 ± 0.5 to 48.6 ± 0.7 across models and question groups. Regarding response length, DeepSeek generated the shortest responses in Turkish Group A (41.8 ± 8.4 words), whereas Claude generated the longest responses in English Group B (153.0 ± 43.2 words). Between-model differences in response length were statistically significant in all four language–question group combinations (Turkish routine: $p<0.001$; English routine: $p=0.034$; Turkish emergency/red-flag: $p=0.003$; English emergency/red-flag: $p<0.001$).

Table 4. Response characteristics and readability metrics.

Metric	ChatGPT Mean $\pm$ SD	Gemini Mean $\pm$ SD	DeepSeek Mean $\pm$ SD	Claude Mean $\pm$ SD	p
Group A: Routine Postoperative Questions (n=20)					
FRE score†	39.6 $\pm$ 0.7	44.5 $\pm$ 0.6	48.6 $\pm$ 0.6	36.7 $\pm$ 0.5	<0.001*
FKGL score†	11.0 $\pm$ 0.3	10.6 $\pm$ 0.3	9.8 $\pm$ 0.2	11.5 $\pm$ 0.2	<0.001*
SMOG index†	11.9 $\pm$ 0.2	11.7 $\pm$ 0.2	11.0 $\pm$ 0.2	12.2 $\pm$ 0.2	<0.001*
Word count (TR)	76.1 $\pm$ 22.7	96.6 $\pm$ 32.5	41.8 $\pm$ 8.4	86.3 $\pm$ 27.5	<0.001*
Word count (EN)	110.7 $\pm$ 40.8	116.5 $\pm$ 45.0	87.3 $\pm$ 21.7	115.8 $\pm$ 44.8	0.034*
Group B: Emergency and Red Flag Questions (n=11)					
FRE score†	39.3 $\pm$ 0.8	45.6 $\pm$ 0.5	48.6 $\pm$ 0.7	36.2 $\pm$ 0.5	<0.001*
FKGL score†	11.1 $\pm$ 0.4	10.1 $\pm$ 0.3	9.8 $\pm$ 0.3	11.7 $\pm$ 0.2	<0.001*
SMOG index†	12.1 $\pm$ 0.2	11.3 $\pm$ 0.2	11.1 $\pm$ 0.2	12.4 $\pm$ 0.2	<0.001*
Word count (TR)	90.2 $\pm$ 30.8	106.9 $\pm$ 57.5	50.0 $\pm$ 29.6	96.5 $\pm$ 45.5	0.003*
Word count (EN)	138.9 $\pm$ 20.3	148.9 $\pm$ 54.3	91.4 $\pm$ 22.5	153.0 $\pm$ 43.2	<0.001*
AMA/NIH Recommended Threshold for Patient Education Materials					
Target reading level	FRE ≥ 61	FKGL ≤ 6	SMOG ≤ 8	—	All models exceeded recommended threshold

Mean $\pm$ SD. †Applied to English responses only (no validated readability instruments for Turkish). FRE=Flesch Reading Ease (higher=easier to read). FKGL=Flesch-Kincaid Grade Level. SMOG=Simple Measure of Gobbledygook (preferred for medical patient education per AMA/NIH). AMA/NIH patient education target: SMOG ≤ 8 , FKGL ≤ 6 (sixth-grade reading level). Word-count comparisons across models were performed using the Friedman test, with question as the matched unit. $p<0.05$.

Accuracy of Emergency Referral

As shown in Table 5, ChatGPT, DeepSeek, and Claude provided appropriate emergency referrals in all 22 responses (100%). Gemini provided an appropriate emergency referral in 16 of 22 responses (72.7%). When emergency referral accuracy was compared across the four models using Cochran's Q test, a significant overall difference was observed for the Turkish responses ($Q=12.00$, $p=0.007$), whereas no significant difference was found for the English responses ($Q=6.00$, $p=0.112$). Pairwise exact McNemar tests with Holm correction showed no statistically significant pairwise differences between models in Turkish (all adjusted $p \geq 0.750$). In paired Turkish-English comparisons within each model, no significant language-related difference in

emergency referral accuracy was observed (Gemini: exact McNemar $p=0.500$; ChatGPT, DeepSeek, and Claude: $p=1.000$). In Turkish, Gemini failed to provide an appropriate emergency referral for Q22 (fever and foul-smelling discharge), Q23 (post-traumatic deformity), Q25 (marked respiratory distress), and Q28 (post-tamponade bleeding).

Comparison of Response Quality Between Turkish and English

For routine postoperative questions, paired Turkish-English comparisons showed significant differences in total QAMAI scores for all four models after Holm correction (Table 6). English scores were higher than Turkish scores for ChatGPT ($p=0.004$, $r_{rb}=-0.813$) and

Table 5. Emergency referral accuracy-group B: emergency and red flag questions.

Language	n	ChatGPT	Gemini	DeepSeek	Claude	Cochran's Q	p
Turkish	11	11/11 (100.0%)	7/11 (63.6%)	11/11 (100.0%)	11/11 (100.0%)	12.00	0.007*
		[95% CI 74.1–100.0%]	[95% CI 35.4–84.8%]	[95% CI 74.1–100.0%]	[95% CI 74.1–100.0%]		
English	11	11/11 (100.0%)	9/11 (81.8%)	11/11 (100.0%)	11/11 (100.0%)	6.00	0.112
		[95% CI 74.1–100.0%]	[95% CI 52.3–94.9%]	[95% CI 74.1–100.0%]	[95% CI 74.1–100.0%]		
Overall†	22	22/22 (100.0%)	16/22 (72.7%)	22/22 (100.0%)	22/22 (100.0%)	—	—
		[95% CI 85.1–100.0%]	[95% CI 51.8–86.8%]	[95% CI 85.1–100.0%]	[95% CI 85.1–100.0%]		

Values are presented as n/N (%) [95% CI]. The 95% confidence intervals for referral accuracy proportions were calculated using Wilson score intervals. Within-language differences among the four models were evaluated using Cochran's Q test because the same emergency questions were assessed across models. When the omnibus test was significant, pairwise exact McNemar tests with Holm correction were used. †Overall values are descriptive only and combine the Turkish and English responses; no pooled inferential test was performed. $p < 0.05$.

Table 6. Language comparison of total QAMAI scores (Turkish vs English).

Group	Model	n	Turkish Mean±SD	English Mean±SD	p	r _{rb}	Direction
Group A: Routine Postoperative Questions							
Routine	ChatGPT	20	23.20±0.70	23.85±0.92	0.004*	-0.813	EN>TR
Routine	Gemini	20	24.08±0.80	23.00±1.30	0.001*	+1.000	TR>EN
Routine	DeepSeek	20	20.88±1.18	23.68±0.77	<0.001*	-0.990	EN>TR
Routine	Claude	19†	24.47±0.42	23.66±1.25	0.014*	+0.691	TR>EN
Group B: Emergency and Red Flag Questions							
Emergency	ChatGPT	11	23.82±1.10	23.95±0.69	1.000	-0.028	ns
Emergency	Gemini	11	22.45±1.59	22.68±1.12	1.000	0.000	ns
Emergency	DeepSeek	11	22.41±1.28	23.00±1.18	1.000	-0.400	ns
Emergency	Claude	11	24.00±1.95	23.91±1.04	1.000	+0.356	ns

Values are mean±SD based on response-level total QAMAI scores averaged across the two raters. Turkish-English comparisons were performed using paired Wilcoxon signed-rank tests with Holm correction; r_{rb} indicates matched-pairs rank-biserial correlation (positive=TR>EN; negative=EN>TR). †Claude routine analysis included $n=19$ because of one incomplete English assessment; all other routine analyses included $n=20$ and emergency/red-flag analyses $n=11$. TR=Turkish; EN=English; ns=not significant. $p < 0.05$.

DeepSeek ($p < 0.001$, $r_{rb} = -0.990$), whereas Turkish scores were higher for Gemini ($p = 0.001$, $r_{rb} = +1.000$) and Claude ($p = 0.014$, $r_{rb} = +0.691$). For emergency or red-flag questions, no significant difference between Turkish and English was observed in any model after Holm correction (all adjusted $p = 1.000$).

DISCUSSION

This study provides a comparative evaluation of four major language models for post-septorhinoplasty patient education, using a bilingual Turkish-English design and systematically distinguishing between routine and emergency/red-flag questions. The findings of this study highlight model differences that have direct implications for patient safety, health equity, and the clinical integration of AI tools.

A clinically relevant finding was model-specific omissions in emergency referral recommendations, including infection, post-tamponade bleeding, and marked respiratory distress. These findings align with Alsudays et al. (1), who demonstrated inter-model differences in FESS patient education, with Gemini generally receiving lower ratings than ChatGPT-4 and DeepSeek. Related otolaryngology studies evaluating postoperative guidance and emergency management have also identified limitations in AI-generated medical advice, supporting the need for continued clinical oversight (19,20). Although Gemini showed lower descriptive emergency-referral accuracy in Turkish than in English, the paired-language comparison was not statistically significant, indicating that this apparent difference should be interpreted with caution. Our findings should also be considered in the context of recent comparative studies evaluating LLMs for rhinoplasty-related patient communication. Meyer et al. (12) compared ChatGPT, Gemini, and Claude in answering rhinoplasty consultation questions and identified differences in response quality among models, while also highlighting limitations in completeness and readability. Similarly, Palmer et al. (13) evaluated ChatGPT, Claude, Perplexity, and Gemini using patient-initiated questions following functional rhinoplasty, including red-flag scenarios, and reported generally favorable performance, particularly for

ChatGPT, although readability remained an important limitation. These findings are broadly consistent with our results, which demonstrate generally high response quality but also meaningful model-specific differences. The present study extends this literature by evaluating both Turkish and English responses within the same framework, thereby allowing potential language-related differences in response quality and emergency referral performance to be examined.

Model performance varied across language contexts, with differences between models more evident in Turkish routine postoperative responses than in the corresponding English responses. This finding highlights the importance of evaluating LLMs in the language patients are expected to use them, rather than extrapolating performance from English-only assessments. Sallam et al. (9) reported no significant language-related differences in the performance of DeepSeek-R1 and ChatGPT-4o across Arabic and English ophthalmology queries, whereas in our study DeepSeek showed lower response quality in Turkish than in English for routine postoperative questions. Differences across clinical domains and languages may therefore affect the generalizability of multilingual LLM evaluations, supporting language- and context-specific assessment before clinical implementation.

Sari et al. (10) reported mean FRE scores of 28.77 for DeepSeek-V3 and 30.4 for ChatGPT-4, with responses from both models classified as very difficult to read; they further noted that such complex language may reduce accessibility for patients with limited health literacy. The readability findings in Table 4 show that all models exceeded the recommended thresholds for patient education materials (SMOG ≤ 8 , FKGL ≤ 6). This result is consistent with the study by Ayoub et al. (3), which reported lower understandability and actionability scores for ChatGPT-generated postoperative instructions compared with institutional materials. Our findings also suggest a need for a balance between content coverage and readability. Indeed, Claude, which generated the most comprehensive responses, also provided the responses with the lowest readability levels (Group B SMOG: 12.4 ± 0.2). This suggests that more detailed and comprehensive information is often presented using a more complex

language structure. Therefore, it is important that AI-generated patient education materials be reviewed for readability and be simplified as needed before clinical use.

Notably, Claude's Turkish response to the persistent bright-red nasal bleeding scenario appropriately recognized the urgency of the situation and recommended immediate contact with the surgeon or immediate evaluation in the emergency department. However, it advised against applying pressure to the soft part of the nose, which represents a contradictory first-aid instruction. Although this response was classified as correct in terms of emergency referral accuracy, it illustrates that appropriate recognition of an emergency and a referral recommendation do not necessarily ensure that all accompanying first-aid instructions provided by an LLM are clinically appropriate. These findings are consistent with Alghamdi et al. (5), who reported that ChatGPT-4 provided reliable responses to common post-rhinoplasty questions, with performance comparable to that of a senior consultant facial plastic surgeon, and with Ibas et al. (6), who found that ChatGPT-4o provided clinically accurate and safe responses to common postoperative rhinoplasty questions, with no critical errors identified by expert evaluators. Additionally, Capelleras et al. (7) highlighted the potential of ChatGPT to support patient education during rhinoplasty recovery by providing general information and reassurance. Taken together, these findings indicate that an appropriate emergency referral should not be interpreted as evidence that all accompanying clinical advice is accurate; they highlight the need for multidimensional assessment and expert oversight of LLM-generated postoperative guidance.

In Group B, no significant between-model differences in accuracy were observed for either Turkish ($p=0.202$) or English ($p=0.297$). However, significant between-model differences were observed in other quality dimensions, particularly Clarity and Completeness, suggesting that the evaluation of responses to emergency/red-flag questions should not rely solely on the accuracy criterion. These findings indicate that clinically significant differences may stem not only from the accuracy of medical information but also from whether this information is conveyed in a

complete and understandable manner. Therefore, when evaluating the clinical reliability of AI tools, it appears more appropriate to use multidimensional quality assessment tools rather than focusing solely on accuracy metrics.

Patient trust in AI tools does not necessarily reflect clinical reliability. Mayo-Yáñez et al. (11) demonstrated that patient trust in and acceptance of AI may vary according to demographic characteristics and technology use; however, the model-specific variation in emergency referral performance observed in our study suggests that such trust should not be assumed to indicate consistent performance in safety-critical scenarios. Aliyeva et al. (2) demonstrated the potential utility of ChatGPT-4 as a supplementary information resource for postoperative patients, particularly in settings with limited access to healthcare professionals. More broadly, Crowson and Nwosu (4) emphasized the importance of careful clinical integration, ethical considerations, and appropriate oversight when implementing AI tools in otolaryngology practice.

Inter-rater agreement was stronger in the Turkish datasets than in the English datasets, although agreement estimates were not uniformly high, supporting cautious interpretation of between-model differences, particularly when effect sizes were small.

Our findings have several clinical implications. First, AI models should be evaluated within the target patient population's language context, as English-only assessments may not reflect multilingual performance. Second, accuracy alone is insufficient for emergency scenarios; completeness and emergency referral accuracy must also be considered. Third, since all models exceeded recommended readability thresholds, AI-generated materials should be reviewed and simplified before clinical use. Finally, the observed variation in emergency referral performance across models highlights the importance of model- and language-specific safety evaluation before the use of LLMs in safety-critical patient education.

This study has several limitations. First, the question set was developed by the research team and reviewed for face validity by two senior ENT specialists; however, it was not derived directly from patient-reported

questions or subjected to formal psychometric validation. Second, each question was submitted only once to each model. Therefore, the potential variability of LLM-generated responses across repeated queries was not assessed. Third, the findings reflect the specific model versions evaluated during the study period, and subsequent model updates may limit the temporal generalizability of the results. Fourth, the analysis was limited to Turkish and English, and the findings cannot necessarily be generalized to other languages or linguistic contexts. Fifth, no validated Turkish readability instrument was available, precluding a direct comparison of readability between the Turkish and English responses. Sixth, evaluation by only two ENT specialists may have limited the precision of inter-rater reliability estimates and introduced some degree of subjectivity into QAMAI scoring. Finally, this study evaluated AI-generated responses under standardized experimental conditions rather than their use by actual patients; therefore, the effects of these responses on patient comprehension, decision-making, satisfaction, adherence, and clinical outcomes were not assessed. Prospective studies incorporating patient-generated questions, repeated model outputs, additional languages, larger expert panels, and patient-centered outcome measures are warranted.

CONCLUSION

This study compares four major language models for post-septorhinoplasty patient education across two languages and question categories. Significant between-model differences were observed for Turkish routine postoperative questions (Kendall's $W=0.775$), whereas the corresponding overall comparison in English did not reach statistical significance. Emergency referral accuracy was high for several models, although model-specific omissions were observed, and differences in individual QAMAI dimensions varied by language and question category. All models exceeded recommended readability thresholds, underscoring the need for simplification before clinical use. AI chatbots may support postoperative patient education but cannot replace professional medical evaluation in safety-critical scenarios.

Future studies should extend comparisons to other otolaryngology procedures and prospectively

assess the impact on clinical outcomes and patient satisfaction.

Ethical approval

In this study, since no patient or experimental animal data were used, ethical committee approval was not required.

Author contribution

Concept: ZK; Design: ZK, GSU; Data Collection or Processing: GSU, AÖ, ZK, YU; Analysis or Interpretation: GSU, ZK, YU; Literature Search: ZK; Writing: ZK, AÖ. All authors reviewed the results and approved the final version of the article.

Source of funding

The authors declare the study received no funding.

Conflict of interest

The authors declare that there is no conflict of interest.

Use of artificial intelligence

During the preparation and revision of this manuscript between June and September 2026, ChatGPT (OpenAI, San Francisco, CA, USA) was used solely for English language editing to improve grammar, clarity, and readability. All scientific content, data interpretation, and conclusions were determined by the authors. The authors reviewed and approved the final manuscript and take full responsibility for its content.

REFERENCES

1. Alsudays AM, Almanea KA, Alhajlah AA, Alroqi A. Artificial intelligence in postoperative sinus care: a comparative study of ChatGPT-4, Google Gemini, and DeepSeek in patient education and support. *J Craniofac Surg.* 2026; 37(3-4): 490-3. [\[Crossref\]](#)
2. Aliyeva A, Alaskarov E, Sari E. Postoperative management of tympanoplasty with ChatGPT-4.0. *J Int Adv Otol.* 2025; 21(1): 1-6. [\[Crossref\]](#)
3. Ayoub NF, Lee YJ, Grimm D, Balakrishnan K. Comparison between ChatGPT and Google Search as sources of postoperative patient instructions. *JAMA Otolaryngol Head Neck Surg.* 2023; 149(6): 556-8. [\[Crossref\]](#)

4. Crowson MG, Nwosu OI. The integration and impact of artificial intelligence in otolaryngology-head and neck surgery: navigating the last mile. *Otolaryngol Clin North Am*. 2024; 57(5): 887-95. [\[Crossref\]](#)
5. Alghamdi AM, Alzahrani RS, Alghamdi FS, Albilasi TM, Doubi AO. The role of artificial intelligence in assisting patient education post-rhinoplasty. *International Journal of Medicine in Developing Countries*. 2024; 8(10): 2794-8. [\[Crossref\]](#)
6. Ibas M, Dursun S, Paksoy M, Ocal R, Karatas E. Accuracy and safety of ChatGPT-4o responses in rhinoplasty postoperative counseling: a panel-based study. *Acta Otolaryngol*. 2025; 145(9): 851-6. [\[Crossref\]](#)
7. Capelleras M, Soto-Galindo GA, Cruellas M, Apaydin F. ChatGPT and rhinoplasty recovery: an exploration of AI's role in postoperative guidance. *Facial Plast Surg*. 2024; 40(5): 628-31. [\[Crossref\]](#)
8. Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Ross R, Lee M. Aesthetic surgery advice and counseling from artificial intelligence: a rhinoplasty consultation with ChatGPT. *Aesthetic Plast Surg*. 2023; 47(5): 1985-93. [\[Crossref\]](#)
9. Sallam M, Alasfoor IM, Khalid SW, et al. Chinese generative AI models (DeepSeek and Qwen) rival ChatGPT-4 in ophthalmology queries with excellent performance in Arabic and English. *Narra J*. 2025; 5(1): e2371. [\[Crossref\]](#)
10. Sari F, Çelik Z, Mirza Y. ChatGPT-4 vs. DeepSeek-V3: a comparative study of response quality, reliability, usefulness, and readability for exercise and rehabilitation strategies in patients with ankylosing spondylitis. *Clin Rheumatol*. 2026; 45(1): 187-95. [\[Crossref\]](#)
11. Mayo-Yáñez M, Rameau A, Vaira LA, et al. Patient perceptions of artificial intelligence in otolaryngology-head and neck surgery: an international study. *Ear Nose Throat J*. 2025; undefined: 1455613251351774. [\[Crossref\]](#)
12. Meyer MKR, Kandathil CK, Davis SJ, et al. Evaluation of rhinoplasty information from ChatGPT, Gemini, and Claude for readability and accuracy. *Aesthetic Plast Surg*. 2025; 49(7): 1868-73. [\[Crossref\]](#)
13. Palmer WJ, Michlin D, Estephan L, et al. Utility and safety of artificial intelligence for patient-initiated contact after functional rhinoplasty. *OTO Open*. 2025; 9(4): e70170. [\[Crossref\]](#)
14. CHART Collaborative. Reporting guidelines for chatbot health advice studies: explanation and elaboration for the Chatbot Assessment Reporting Tool (CHART). *BMJ*. 2025; 390: e083305. [\[Crossref\]](#)
15. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet*. 2007; 370(9596): 1453-7. [\[Crossref\]](#)
16. Vaira LA, Lechien JR, Abbate V, et al. Validation of the Quality Analysis of Medical Artificial Intelligence (QAMAI) tool: a new tool to assess the quality of health information provided by AI platforms. *Eur Arch Otorhinolaryngol*. 2024; 281(11): 6123-31. [\[Crossref\]](#)
17. McLaughlin GH. SMOG grading - a new readability formula. *The Journal of Reading*. 1969; 12(8): 639-46.
18. Friedman DB, Hoffman-Goetz L. A systematic review of readability and comprehension instruments used for print and web-based cancer information. *Health Educ Behav*. 2006; 33(3): 352-73. [\[Crossref\]](#)
19. Dhar S, Kothari D, Vasquez M, et al. The utility and accuracy of ChatGPT in providing post-operative instructions following tonsillectomy: a pilot study. *Int J Pediatr Otorhinolaryngol*. 2024; 179: 111901. [\[Crossref\]](#)
20. Chen FJ, Nightingale J, You WS, Anderson D, Morrissey D. Assessment of ChatGPT vs. Bard vs. guidelines in the artificial intelligence (AI) preclinical management of otorhinolaryngological (ENT) emergencies. *Aust J Otolaryngol*. 2024; 7: 19. [\[Crossref\]](#)