NORTHWESTERN MEDICAL JOURNAL
Northwestern Med J. 2025;5(4):209-221
DOI: 10.54307/NWMJ.2025.220

RESEARCH ARTICLE

Evaluation of ChatGPT-4.5 and DeepSeek-V3-R1 in answering patient-centered questions about orthognathic surgery: a comparative study across two languages

İpek Necla Güldiken[1], **Emrah Dilaver**[1]

[1]Department of Oral and Maxillofacial Surgery, Faculty of Dentistry, İstinye University, İstanbul, Türkiye

Cite as: Güldiken İN, Dilaver E. Evaluation of ChatGPT-4.5 and DeepSeek-V3-R1 in answering patient-centered questions about orthognathic surgery: a comparative study across two languages. Northwestern Med J. 2025;5(4):209-221.

ABSTRACT

Aim: Patients undergoing orthognathic surgery frequently seek online resources to better understand the procedure, risks, and outcomes. As generative artificial intelligence (AI) models are increasingly integrated into healthcare communication, it is essential to evaluate their ability to deliver accurate, comprehensive, and readable patient information.

Methods: This study conducted a comparative assessment of two large language models (LLMs)—ChatGPT-4.5 and DeepSeek-V3-R1—in answering frequently asked orthognathic patient questions, analyzing accuracy, completeness, readability, and quality across English (EN) and Turkish (TR). Twenty-five patient-centered questions categorized into five clinical domains yielded 200 AI-generated responses, independently evaluated by two oral and maxillofacial surgeons (OMFSs) using a multidimensional framework. Statistical analyses included non-parametric tests and inter-rater reliability assessments (Intraclass Correlation Coefficient (ICC), and Cohen's Kappa).

Results: Significant differences emerged across clinical categories in difficulty and accuracy scores ($p < 0.05$). Questions in the "Postoperative Complications & Rehabilitation" category were least difficult, while those in "Diagnosis & Indication" category were rated most difficult but achieved the highest accuracy and quality ratings. English (EN) responses significantly outperformed Turkish (TR) responses in readability, word count, and accuracy ($p < 0.05$), though completeness and quality did not differ significantly by language. No significant performance differences were found between the two chatbots. Inter-observer agreement was generally high, except for completeness ($p = 0.001$), where Observer-I assigned higher scores.

Conclusion: Both LLMs effectively generated clinically relevant responses, demonstrating substantial potential as supplemental tools for patient education, although the superior performance of EN responses emphasizes the need for further multilingual optimization.

Keywords: ChatGPT, DeepSeek, large language models, orthognathic surgery, patient education

Corresponding author: İpek Necla Güldiken **E-mail:** ipek.guldiken@istinye.edu.tr
Received: 18.08.2025 **Accepted:** 22.09.2025 **Published:** 26.10.2025

INTRODUCTION

The use of the internet for accessing health-related information has markedly increased in past three years (1). Orthognathic surgery, essential for managing dentofacial deformities, prompts many patients to seek information online due to limited access to specialists and the complexity of the treatment process. Patients typically investigate the surgical procedure, recovery timeline, associated risks, costs, and expected outcomes (2). Research shows that this information-seeking behavior intensifies during the decision-making period and preoperative phase, primarily via medical websites, blogs, and forums (3). Postoperative information needs also persist and are often addressed through professional communication and support networks (4). This trend highlights both the procedure's growing prevalence and the critical need for reliable, accessible information.

Findings over the past two years highlight AI's growing effectiveness and reliability in medical contexts, underscoring its expanding role in patient counselling and information dissemination. Recent studies have assessed how LLMs—including ChatGPT, DeepSeek—respond to patient queries in healthcare (5-9). ChatGPT-4.5 (developed by OpenAI, USA) and DeepSeek-V3-R1 (developed by DeepSeek-AI, China) are increasingly used in healthcare for responding to patient queries, owing to their advanced natural language processing capabilities. ChatGPT-4.5 is the latest model in OpenAI's GPT series, while DeepSeek-V3-R1 is the latest interactive model trained on large-scale datasets in DeepSeek-AI models. Both models stand out for their ability to generate fast and comprehensive responses to complex medical questions (10).

This study aimed to compare ChatGPT-4.5 and DeepSeek-V3-R1 in terms of response accuracy, comprehensiveness, readability, and overall quality for frequently asked patient questions on orthognathic surgery. It also evaluated the effects of language (EN vs. TR) and question difficulty on model performance. As the first study to provide a cross-linguistic comparison of LLM outputs in this context, it underscores the potential of these models to enhance access to trustworthy health information and reduce language-based communication barriers. Supporting standardized, multilingual communication in global healthcare was a key motivation for this research.

MATERIAL AND METHODS

Study design

Since this research did not involve human subjects or personal health data, formal ethical approval was not required. Nonetheless, all testing was conducted in a neutral setting to uphold the integrity of the study. To enhance transparency and ensure methodological consistency, the METRICS framework (Model, Evaluation, Timing, Range/Randomization, Individual factors, Count, and Specificity of prompts and language) was adopted (11). This structured approach also contributes to standardizing AI evaluations in healthcare and minimizing potential sources of bias. The methodological workflow of the study is summarized in Figure 1.

To simulate an average user experience and reduce potential bias, both models were accessed via a newly created Google account using the "Continue with Google" option. Prior to testing, all browsing history, cookies, and cache were deleted by selecting the "Clear Browsing Data" option with the time range set to "All Time".

Question development, categorization and sampling

The sample of patient-centered questions was derived using a purposive sampling strategy, specifically combining expert clinical input with a structured review of existing patient education materials. The authors created a preliminary question pool by adapting content from published literature (2,12-14) and reviewing patient guidelines issued by professional bodies, such as the American Association of Oral and Maxillofacial Surgeons (AAOMS) and the British Association of Oral and Maxillofacial Surgeons (BAOMS). Additional questions were identified through a targeted Google search using keywords such as "orthognathic surgery", "patient FAQs (frequently

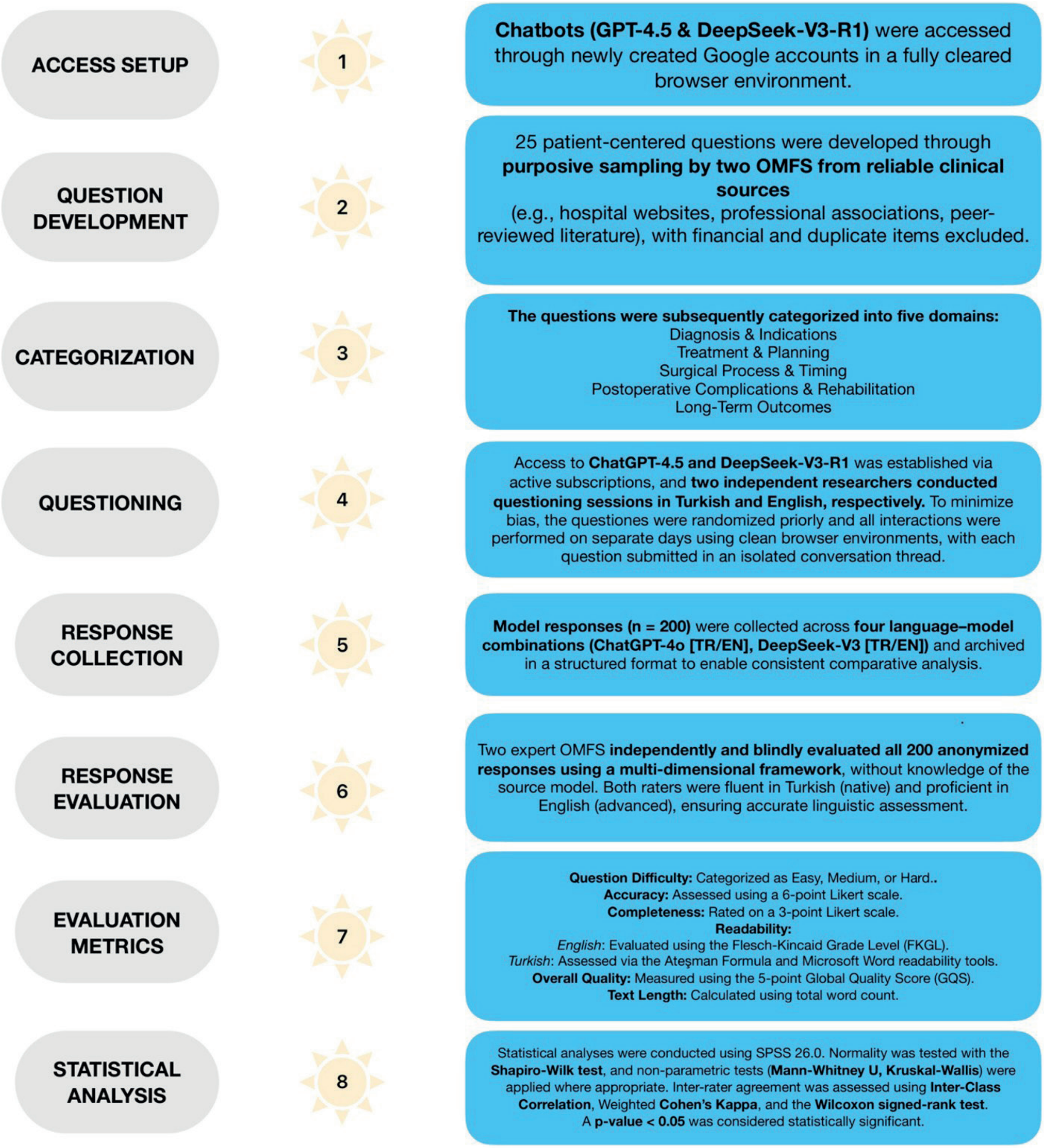

Figure 1. Flowchart of the study.

asked questions)", and "patient guide", focusing on the top-ranking results from reputable hospital and clinical websites, including large academic medical centers and national surgical institutes.

The emphasis was placed on thematic representativeness and content saturation, rather than statistical generalizability, aligning with the principles of purposive sampling as outlined by Etikan et al.

(15). This approach ensured the inclusion of clinically relevant, high-frequency, and informative questions for comparative evaluation across AI models.

The final set of 25 patient-centered questions was selected after excluding those with financial content and removing duplicates (Table 1). To enhance analytical consistency, the questions were ontologically categorized according to the medical process-based framework proposed by Chong et al. (16), which classifies patient inquiries into five distinct domains: (1) Diagnosis & Indication, (2) Treatment & Planning, (3) Surgical Process & Timing, (4) Postoperative Complications & Rehabilitation, and (5) Long-Term Outcomes. This classification facilitated a structured analysis of AI model performance across different stages of the orthognathic surgery care continuum (Figure 2).

Table 1. Patient questions and the categories

ID	Question	Category
1	What exactly is orthognathic surgery?	
2	Who needs orthognathic surgery?	
3	Which conditions can orthognathic surgery address?	
4	Can this surgery be done purely for cosmetic reasons, even if there's no functional issue?	Diagnosis & Indication
5	At what age is orthognathic surgery performed?	
6	How do I know if I need orthognathic surgery vs. just orthodontic treatment?	
7	How do I find or choose a qualified surgeon and orthodontist for my treatment?	
8	Do I need braces before orthognathic surgery?	Treatment & Planning
9	How should I prepare for jaw surgery and how?	
10	How is orthognathic surgery performed?	
11	How long does orthognathic surgery take?	Operation Process & Time
12	What type of anesthesia is used?	
13	Will I need to have my jaws wired shut, and for how long?	
14	What is the typical recovery time after orthognathic surgery?	
15	Is the procedure painful, and what can I expect in terms of post-operative discomfort?	Postoperative Complications & Rehabilitation
16	How do I manage swelling and other potential side effects after surgery?	
17	Is numbness or loss of sensation normal after surgery, and will it go away over time?	
18	What are the potential risks and complications of orthognathic surgery?	
19	How soon can I return to work or school after surgery?	
20	Does having this surgery improve facial appearance as well as jaw function?	
21	Will orthognathic surgery affect speech or eating in the long run?	
22	Are there any long-term lifestyle changes required after orthognathic surgery?	
23	Can I have symmetry disorder in my face after surgery?	Long-Term Results
24	Does this surgery correct breathing or snoring problems?	
25	After orthognathic surgery, may I need to have another operation in the future?	

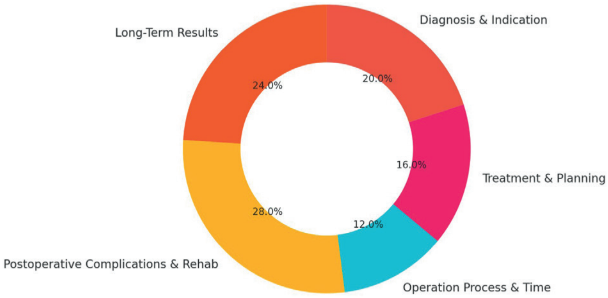

Figure 2. Question distrubution according to the categories.

Questioning

Access to ChatGPT (OpenAI) and DeepSeek AI was obtained through active subscriptions to ChatGPT-4.5 and DeepSeek-V3-R1, which offer improved processing power, priority usage, and increased accuracy through the latest updates. To systematically evaluate the models, two independent researchers conducted the questioning sessions. One researcher interacted with the models in Turkish (TR), while the other did so in English (EN).

All interactions were carried out using newly created user accounts in a clean browsing environment (with cleared cookies, cache, and history) to minimize contextual bias. Each session was conducted on a different day to reduce potential carry-over effects, and every question was submitted in a distinct conversation thread to ensure isolated contextual modeling using newly created accounts and a clear browsing environment.

Response collection

Model outputs were systematically collected across four distinct language–model pairings: ChatGPT (TR), ChatGPT (EN), DeepSeek (TR), and DeepSeek (EN). All responses were preserved in their entirety and archived in a structured format to ensure consistency and enable comparative evaluation.

Response evaluation

Model-generated answers were assessed using a multi-dimensional framework incorporating both quantitative and qualitative indicators (2,14). The evaluation was conducted by two board-certified OMFSs, who independently and blindly scored all 200 responses. The responses—randomized across both models—were anonymized such that evaluators were unaware of the originating model. Both raters possessed advanced proficiency in English and native-level fluency in Turkish, ensuring reliable linguistic judgment across both language sets.

Question difficulty was categorized as easy, medium, or hard based on Goodman et al. (17). Accuracy was rated on a 6-point Likert scale, and completeness on a 3-point scale, with intermediate scores reflecting partial correctness. Readability was assessed using language-specific tools: the Flesch-Kincaid Grade Level (FKGL) for English (7,18) and the Ateşman Formula for Turkish (19-21). Overall answer quality was evaluated using the 5-point Global Quality Score (GQS), reflecting scientific accuracy, clarity, and informativeness (22). Text length was calculated as total word count using Microsoft Word (Microsoft Corporation, Redmond, WA, USA) (see Table 2 for detailed scoring criteria).

The FKGL score is derived from average sentence length and syllables per word, with lower scores indicating simpler, more accessible language (18). For TR responses, readability was evaluated using the Ateşman Formula, a well-established metric adapted from the Flesch Reading Ease Index (FREI) for the TR language. The formula incorporates average word and sentence lengths to generate a numerical readability score, where higher values indicate easier comprehension (19,23). In addition, Microsoft Word's built-in readability analysis was used as a supplementary tool to validate the scoring consistency.

Statistics

All statistical analyses were performed using SPSS version 26.0 (IBM Corp., Armonk, NY, USA). Descriptive statistics were used to summarize participant and response characteristics. The normality of continuous variables was assessed using the Shapiro-Wilk test ($p < 0.001$). For variables that did not follow a normal distribution, non-parametric tests were employed, including the Mann-Whitney U test and Kruskal-Wallis test for group comparisons. The inter-rater agreement between the two evaluators was analyzed using the

Table 2. Evaluation framework for model responses

Evaluation Dimension	Metric/Tool	Scale/Range	Description
Question difficulty	Cathegorical	Easy Medium Hard	Classification based on clinical and linguistic complexity
Accuracy	6-Point Likert Scale	1–6	1 = Completely incorrect 2 = Mostly incorrect 3 = Contains some factual errors 4 = Partially correct 5 = Mostly accurate 6 = Completely accurate
Completeness	3-Point Likert Scale	1–3	1 = Inadequate 2 = Moderately complete 3 = Comprehensive
Readability (English)	Flesch-Kincaid Grade Level (FKGL)	Grade level (e.g., 8.0)	1–5 = Very easy to read 6–8 = Easy to read 9–12 = Fairly difficult / Standard 13–16 = Difficult to read 17+ = Very difficult / Academic or technical material
Readability (Turkish)	Ateşman Formula / Microsoft Word	Score (e.g., 0–100)	90–100 = Very easy to read 70–89 = Easy to read 50–69 = Fairly difficult / Standard 30–49 = Difficult to read 0–29 = Very difficult / Academic or technical material
Overall quality	Global Quality Score (GQS)	1–5	1 = Very poor 2 = Poor 3 = Moderate 4 = Good 5 = Excellent quality
Text length	Microsoft Word (Microsoft Corp., USA)	Numeric word count	Total number of words in the model response

Intraclass Correlation Coefficient (ICC) and Weighted Cohen's Kappa. Furthermore, the inter-observer consistency was appraised by means of the Wilcoxon signed-rank test. All results were interpreted within a 95% confidence interval, and a p-value <0.05 was considered statistically significant.

RESULTS

The majority of the questions were categorized under the headings "Postoperative Complications & Rehabilitation" and "Long-Term Outcomes," with a balanced distribution across language groups, chatbots, and observers. Descriptive statistics for question difficulty, word count, accuracy, completeness, readability, and overall quality are presented in Table 3.

Statistical comparisons revealed significant differences in difficulty scores based on question category (p <0.001), with “Diagnosis & Indication” questions rated as significantly more difficult than those in “Surgical Process & Timing” and “Postoperative Complications & Rehabilitation”. Word count did not vary significantly by category. Despite the higher difficulty scores, accuracy scores were also significantly higher in the “Diagnosis & Indication” compared to “Postoperative Complications & Rehabilitation” category (p = 0.013) (Table 4).

Table 3. Descriptive statistics about questionnaire		
Variables	**Mean. ± S.D.**	**Median (Min.Max.)**
Difficulty	1.9 ± 0.7	2 (1-3)
Word count	289.8 ± 108.3	279 (81-596)
Accuracy	4.9 ± 0.9	5 (3-6)
Completeness	2.7 ± 0.5	3 (2-3)
Readability	3.9 ± 1.1	4 (2-6)
Quality	3.9 ± 0.9	4 (2-5)
Category	**n**	**%**
Diagnosis & Indication	40	20
Treatment & Planning	32	16
Operation process & Time	24	12
Postoperative Complications & Rehabilitation	56	28
Long term results	48	24
Language	**n**	**%**
TR	100	50
ENG	100	50
Chatbot	**n**	**%**
ChatGPT	100	50
DeepSeek	100	50
Observer	**n**	**%**
Observer-I	100	50
Observer-II	100	50

Language comparisons showed that EN responses had significantly higher word counts ($p<0.001$), higher accuracy ($p = 0.047$), and higher readability scores ($p <0.001$) compared to TR responses (Table 5). No statistically significant differences were observed between ChatGPT and DeepSeek models in terms of difficulty, word count, accuracy, completeness, readability, or quality (Table 4). Similarly, completeness and quality scores did not significantly differ by language (Table 5).

A comparison of the evaluators' scores revealed significant discrepancies in the completeness ratings ($p <0.001$), with Observer I assigned higher ratings than Observer II. In a similar vein, Observer I provided substantially higher quality scores. Despite these differences, the inter-rater reliability was found to be strong to perfect across all evaluation criteria, with ICC and weighted Cohen's Kappa values ranging from 0.374 to 1.000 ($p <0.001$ for all metrics). These findings suggest that while subjective interpretation may influence certain dimensions, particularly completeness and quality, the overall scoring framework demonstrated a high level of consistency and reliability between evaluators (Table 5).

DISCUSSION

In this study, responses generated by two large language models—ChatGPT-4.5 and DeepSeek-V3-R1—to patient questions on orthognathic surgery were evaluated using a multidimensional framework. Analysis of 200 outputs showed that both models produced responses with high accuracy and quality, and similar word count and scope. However, EN responses significantly outperformed TR ones in terms of accuracy and readability, likely due to the uneven distribution of training data across languages (24). Notably, the "Diagnosis & Indication" category, despite its higher difficulty level, received the highest accuracy scores—contrary to previous findings (6). This suggests that structured knowledge domains, such as diagnostic content, may enhance model performance even on complex queries.

The methodological framework of this study aligns with prior research in the field (6,17,25,26). In designing the evaluation protocol, several established health content quality assessment guidelines were reviewed, including METRICS, CLEAR (Communication, Language, Evaluation, and Review), and MI-CLEAR-LLM (Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Healthcare) (11,27,28). The most applicable and pragmatic elements from these frameworks were selectively integrated into the METRICS-based assessment applied in this study.

Table 4. Comparative metrics by question type

Variables		N	Difficulty Mean. ± S.D.	Difficulty Median (Min.Max.)	Difficulty p	Word count Mean. ± S.D.	Word count Median (Min.Max.)	Word count p	Accuracy Mean. ± S.D.	Accuracy Median (Min.Max.)	Accuracy p
Category	Diagnosis& indication	40	2.3 ± 0.8	2.5 (1-3)	p<0.001	289.6 ± 110.5	281.5 (141-593)	0.117	5.2 ± 0.6	5 (4-6)	0.013*
	Treatment& planning	32	2.0 ± 0.7	2 (1-3)		330.0 ±116.5	324 (189-593)		4.8 ± 1.1	5 (3-6)	
	Surgical process& time	24	1.7 ± 0.8	1.5 (1-3)		294.1 ± 134.8	268.5 (149-593)		5.0 ± 0.8	5 (4-6)	
	Postoperative complication & rehabilitation	56	1.4 ± 0.5	1 (1-3)		262.0 ± 107.7	238 (81-489)		4.6 ± 0.9	5 (3-6)	
	Long-term results	48	2.1 ± 0.5	2 (1-3)		293.5 ± 77.6	296 (130-436)		5.0 ± 0.9	5 (3-6)	
Language	TR	100	1.8 ± 0.7	2 (1-3)	0.069	255.3 ± 101.9	230 (81-593)	p<0.001	4.8 ± 0.9	5 (3-6)	0.047*
	EN	100	2.0 ± 0.8	2 (1-3)		324.3 ± 103.8	333.5 (134-596)		5.0 ± 0.8	5 (3-6)	
Chatbot	ChatGPT	100	1.9 ± 0.7	2 (1-3)	0.727	278.1 ± 96.8	263.5 (134-593)	0.107	4.9 ± 0.8	5 (3-6)	0.276
	DeepSeek	100	1.8 ± 0.7	2 (1-3)		301.5 ± 117.9	292 (81-596)		5.0 ± 0.9	5 (3-6)	
Obserever	Observer-I	100	1.9 ± 0.7	2 (1-3)	1.000	289.8 ± 108.5	279 (81-596)	1.000	5.0 ± 1.0	5 (3-6)	0.068
	Observer-II	100	1.9 ± 0.7	2 (1-3)		289.8 ± 108.5	279		4.8 ± 0.8	5 (3-6)	

*$p<0.05$

Table 5. Comparisons of completeness, readability and quality on characteristic specialities of questionnaires

			Completeness			Readability			Quality		
Variables		**N**	**Mean. ± S.D.**	**Median (Min.Max.)**	**p**	**Mean. ± S.D.**	**Median (Min.Max.)**	**p**	**Mean. ± S.D.**	**Median (Min.Max.)**	**p**
Catergory	Diagnosis & Indication	40	2.7 ±0.5	3 (2-3)	0.159	4.1 ± 1.1	4 (2-6)	0.001*	4.2 ± 0.7	4 (2-5)	0.026*
	Treatment & Planning	32	2.7 ±0.5	3 (2-3)		3.8 ± 0.9	4 (2-5)		3.8 ±1.0	4 (2-5)	
	Surgical process & Time	24	2.5 ± 0.5	3 (2-3)		3.8 ± 0.9	3.5 (3-5)		4.2 ±0.7	4 (3-5)	
	Postoperative complication & Rehabilitation	56	2.8 ± 0.4	3 (2-3)		3.5 ±1.2	3 (2-6)		3.7 ±0.9	4 (2-5)	
	Long-term results	48	2.5 ± 0.5	3 (2-3)		4.3 ± 0.9	4 (3-6)		3.8 ±0.8	4 (2-5)	
Language	TR	100	2.6 ± 0.5	3 (2-3)	0.237	3.2 ± 0.7	3 (2-5)	p<0.001	3.8 ± 0.9	4 (2-5)	0.236
	ENG	100	2.7 ± 0.5	3 (2-3)		4.6 ± 0.9	5 (3-6)		4.0 ± 0.8	4 (2-5)	
Chatbot	ChatGPT	100	2.7 ± 0.5	3 (2-3)	1.000	3.8 ± 1.0	3.5 (2-6)	0.370	3.8 ± 0.9	4 (2-5)	0.105
	DeePSeeK	100	2.7 ± 0.5	3 (2-3)		3.9 ± 1.1	4 (2-6)		4.0 ± 0.8	4 (2-5)	
Observer	Observer-I	100	2.8 ± 0.4	3 (2-3)	0.001*	3.9 ± 1.1	4 (2-6)	1.000	4.0 ±0.9	4 (2-5)	0.170
	Observer-II	100	2.5 ± 0.5	3 (2-3)		3.9 ± 1.1	4 (2-6)		3.8 ±0.8	4 (2-5)	

*p<0.05

Previous studies have shown that LLMs generally provide satisfactory responses to medical questions intended for patient education. Goodman et al. and Beheshti et al. reported that models such as ChatGPT are capable of generating accurate and useful medical content (17,29). Comparative evaluations of various LLMs—such as ChatGPT, Gemini, Bard, Claude, Copilot, and DeepSeek—indicate that ChatGPT is the most frequently studied model, with English being the dominant language of analysis (5-9,22). While direct comparisons between ChatGPT and DeepSeek remain limited, it has been showed that both models performed similarly when responding to complex, multi-domain queries—an observation that aligns with our own findings (10,23).

In the present study, open-ended, prompt-free questions were used to simulate a realistic conversational setting. The impact of prompt usage and question format on model performance has also been addressed in the literature, particularly in healthcare, where the absence of prompts in open-ended queries has been linked to accuracy issues (e.g., hallucination effects) (6,7,14,30-32). In their systematic review, Beheshti et al. highlighted that many studies failed to evaluate the influence of prompt design (29). Nevertheless, further research is warranted to clarify the role of prompt engineering in medical applications.

This study offers one of the first comparative analyses of two LLMs across both EN and TR, addressing the gap in the literature where evaluations are often limited to a single language (2,17,26). The dominance of EN in model training and the lack of standardized cross-linguistic evaluation metrics contribute to this limitation (24).

Readability was assessed using language-specific tools: the FKGL for EN and the Ateşman Formula for Turkish (a well-established metric adapted from the FREI) (18,21). To our knowledge, this is the first study to directly compare the readability of ChatGPT and DeepSeek responses in TR versus EN. While the American Medical Association (AMA) and National Institutes of Health (NIH) recommend health materials be written at a sixth-grade reading level (4), our findings, consistent with previous work (14,26), show that most responses exceed this threshold. Notably, TR responses exhibited lower (i.e., better) readability scores, suggesting improved accessibility for native speakers.

Prompt usage has been shown to enhance factual accuracy but not necessarily readability (7). Overall, the higher accuracy and readability of EN responses likely reflect the greater volume of training data in EN (24). These findings underscore the need for multilingual training and evaluation strategies, especially for low-resource languages like TR. Moreover, differences in text length and linguistic structure may also influence readability outcomes. By addressing the bilingual evaluation gap, this study contributes to promote equitable and comprehensible health communication (2,14,26).

Previous research indicates that online sources, including social media, provide predominantly low-quality orthognathic surgery information, characterized by subjective patient experiences (33). Bavbek et al. particularly highlighted the poor quality of Turkish-language online resources on orthognathic surgery (3). However, recent advances in AI-based chatbots, such as ChatGPT and Google Gemini, have significantly enhanced content reliability by incorporating academic literature and professional guidelines (2,14). Despite these improvements, issues regarding readability and accuracy persist, warranting cautious integration into clinical practice (4). Notably, no prior studies have evaluated DeepSeek's performance in orthognathic patient education, emphasizing our study's novel contribution in comparing model performance across languages, particularly in TR.

This study offers several strengths, including the evaluation of four model-language combinations, the ontological categorization of patient questions, blinded expert-based scoring, and a multidimensional assessment framework encompassing accuracy, completeness, readability, and quality. Furthermore, although there is currently no consensus on standardized criteria for evaluating chatbot performance in health literacy contexts (11,14,27,28), this study addressed that gap by adapting elements from existing guidelines—particularly the METRICS framework—to guide its evaluation process. However, certain limitations of this study should be acknowledged.

First, the limited sample size may not fully capture the range of responses these models can generate. Both question development and scoring were conducted by two OMFSs from the same institution, potentially introducing selection bias. Nonetheless, the inclusion of inter-rater reliability analysis partially mitigates this limitation.

The reproducibility of model outputs could not be assessed, as each question was asked only once and sessions were not repeated. Additionally, restricting the study to only two widely used LLMs limits the generalizability of the findings. The lack of transparency in source usage and inconsistent citation practices may also affect the perceived accuracy of content (14,29). Although both models exhibited citation behaviors, inconsistencies in source attribution prevented statistical analysis. Moreover, the assertive tone adopted by LLMs may create a false sense of confidence among users, which has been noted in previous studies (17,29).

Finally, while this study aimed to address the limitations of monolingual evaluations by including both EN and TR, structural differences between the two languages pose inherent challenges for direct comparison. Despite these limitations, this study remains one of the few bilingual and multidimensional evaluations of LLMs in the specialized field of orthognathic surgery, offering a foundation for future research.

CONCLUSIONS AND SUGGESTIONS

This study demonstrated that large language models, specifically ChatGPT-4.5 and DeepSeek-V3-R1, are capable of producing accurate and clinically relevant responses to patient-centered questions in orthognathic surgery. While both models performed similarly in overall quality, responses in EN showed higher readability and accuracy than those in TR. Differences across clinical categories and the presence of moderate inter-observer variability emphasize the need for standardized evaluation frameworks. With careful implementation and ongoing validation, LLMs may serve as valuable tools to support patient education and preoperative communication in oral and maxillofacial surgery settings.

Acknowledgments

The authors would like to thank the Department of Oral and Maxillofacial Surgery at İstinye University for supporting this research.

Ethical approval

The authors confirm that an ethical waiver was obtained due to the study design (retrospective evaluation of publicly accessible data without direct patient interaction), and that the study was conducted according to the Helsinki Declaration (2013).

Author contribution

Surgical and Medical Practices: ING, ED; Concept: ING, ED; Design: ING, ED; Data Collection or Processing: ING, ED; Analysis or Interpretation: ING, ED; Literature Search: ING, ED; Writing: ING. 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.

REFERENCES

1. Su H, Sun Y, Li R, et al. Large Language Models in Medical Diagnostics: Scoping Review With Bibliometric Analysis. J Med Internet Res. 2025; 27: e72062. **[Crossref]**
2. Aziz AAA, Abdelrahman HH, Hassan MG. The use of ChatGPT and Google Gemini in responding to orthognathic surgery-related questions: A comparative study. J World Fed Orthod. 2025; 14(1): 20-6. **[Crossref]**
3. Bavbek NC, Tuncer BB. Information on the Internet Regarding Orthognathic Surgery in Turkey: Is It an Adequate Guide for Potential Patients? Turk J Orthod. 2017; 30(3): 78-83. **[Crossref]**

4. Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? J Stomatol Oral Maxillofac Surg. 2023; 124(5): 101471. [Crossref]
5. Dursun D, Bilici Geçer R. Can artificial intelligence models serve as patient information consultants in orthodontics? BMC Med Inform Decis Mak. 2024; 24(1): 211. [Crossref]
6. Gumilar KE, Indraprasta BR, Faridzi AS, et al. Assessment of Large Language Models (LLMs) in decision-making support for gynecologic oncology. Comput Struct Biotechnol J. 2024; 23: 4019-26. [Crossref]
7. Gordon EB, Towbin AJ, Wingrove P, et al. Enhancing Patient Communication With Chat-GPT in Radiology: Evaluating the Efficacy and Readability of Answers to Common Imaging-Related Questions. J Am Coll Radiol. 2024; 21(2): 353-9. [Crossref]
8. Metin U, Goymen M. Information from digital and human sources: A comparison of chatbot and clinician responses to orthodontic questions. Am J Orthod Dentofacial Orthop. 2025; 168(3): 348-57. [Crossref]
9. Daraqel B, Wafaie K, Mohammed H, et al. The performance of artificial intelligence models in generating responses to general orthodontic questions: ChatGPT vs Google Bard. Am J Orthod Dentofacial Orthop. 2024; 165(6): 652-62. [Crossref]
10. Zhou M, Pan Y, Zhang Y, Song X, Zhou Y. Evaluating AI-generated patient education materials for spinal surgeries: Comparative analysis of readability and DISCERN quality across ChatGPT and deepseek models. Int J Med Inform. 2025; 198: 105871. [Crossref]
11. Sallam M, Barakat M, Sallam M. A Preliminary Checklist (METRICS) to Standardize the Design and Reporting of Studies on Generative Artificial Intelligence-Based Models in Health Care Education and Practice: Development Study Involving a Literature Review. Interact J Med Res. 2024; 13: e54704. [Crossref]
12. Modig M, Andersson L, Wårdh I. Patients' perception of improvement after orthognathic surgery: pilot study. Br J Oral Maxillofac Surg. 2006; 44(1): 24-7. [Crossref]
13. Lee S, McGrath C, Samman N. Impact of orthognathic surgery on quality of life. J Oral Maxillofac Surg. 2008; 66(6): 1194-9. [Crossref]
14. Yurdakurban E, Topsakal KG, Duran GS. A comparative analysis of AI-based chatbots: Assessing data quality in orthognathic surgery related patient information. J Stomatol Oral Maxillofac Surg. 2024; 125(5): 101757. [Crossref]
15. Etikan I, Musa SA, Alkassim RS. Comparison of convenience sampling and purposive sampling. American Journal of Theoretical and Applied Statistics. 2015; 5(1): 1-4. [Crossref]
16. Chong Q, Marwadi A, Supekar K, Lee Y. Ontology based metadata management in medical domains, Journal of Research and Practice in Information Technology. 2003; 35(2): 139-54.
17. Goodman RS, Patrinely JR, Stone CA, et al. Accuracy and Reliability of Chatbot Responses to Physician Questions. JAMA Netw Open. 2023; 6(10): e2336483. [Crossref]
18. Kincaid JP, Fishburne RP Jr, Rogers RL, Chissom BS. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel (Report No. ADA006655). 1975. [Crossref]
19. Ateşman E. Türkçede okunabilirliğin ölçülmesi. Dil Dergisi. 1997; 58: 71-4. Available at: https://web.archive.org/web/20201031111301/http://www.atesman.info/wp-content/uploads/2015/10/Atesman-okunabilirlik.pdf
20. Bezirci B, Yılmaz AE. A software library on measuring the readability of texts and a new readability criterion for Turkish. Dokuz Eylul University Faculty of Engineering Journal of Science and Engineering. 2010; 12(3): 49-62.
21. Er A, Ay IE, Horozoğlu Ceran T. Evaluating Turkish Readability and Quality of Strabismus-Related Websites. Cureus. 2024; 16(4): e58603. [Crossref]
22. Hancı V, Ergün B, Gül Ş, Uzun Ö, Erdemir İ, Hancı FB. Assessment of readability, reliability, and quality of ChatGPT®, BARD®, Gemini®, Copilot®, Perplexity® responses on palliative care. Medicine (Baltimore). 2024; 103(33): e39305. [Crossref]
23. Chan L, Xu X, Lv K. DeepSeek-R1 and GPT-4 are comparable in a complex diagnostic challenge: a historical control study. Int J Surg. 2025; 111(6): 4056-9. [Crossref]
24. Qin L, Chen Q, Zhou Y, et al. A survey of multilingual large language models. Patterns (N Y). 2025; 6(1): 101118. [Crossref]
25. Asfuroğlu ZM, Yağar H, Gümüşoğlu E. High accuracy but limited readability of large language model-generated responses to frequently asked questions about Kienböck's disease. BMC Musculoskelet Disord. 2024; 25(1): 879. [Crossref]
26. Hassona Y, Alqaisi D, Al-Haddad A, et al. How good is ChatGPT at answering patients' questions related to early detection of oral (mouth) cancer? Oral Surg Oral Med Oral Pathol Oral Radiol. 2024; 138(2): 269-78. [Crossref]
27. Park SH, Suh CH, Lee JH, Kahn CE, Moy L. Minimum Reporting Items for Clear Evaluation of Accuracy Reports of Large Language Models in Healthcare (MI-CLEAR-LLM). Korean J Radiol. 2024; 25(10): 865-8. [Crossref]
28. Kocak B, Baessler B, Bakas S, et al. CheckList for EvaluAtion of Radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. Insights Imaging. 2023; 14(1): 75. [Crossref]
29. Beheshti M, Toubal IE, Alaboud K, et al. Evaluating the Reliability of ChatGPT for Health-Related Questions: A Systematic Review. Informatics. 2025; 12(1): 9. [Crossref]
30. Huang AS, Hirabayashi K, Barna L, Parikh D, Pasquale LR. Assessment of a Large Language Model's Responses to Questions and Cases About Glaucoma and Retina Management. JAMA Ophthalmol. 2024; 142(4): 371-5. [Crossref]

31. Pandey S, Sharma S. A comparative study of retrieval-based and generative-based chatbots using Deep Learning and Machine Learning. Healthcare Analytics. 2023; 3: 100198. **[Crossref]**

32. Lehman E, Hernandez E, Mahajan D, et al. Do We Still Need Clinical Language Models? arXiv 2302.08091[Preprint]. 2023 Feb 16. **[Crossref]**

33. Ngo M, Jensen E, Meade M. The quality of orthognathic surgery information on social media: A scoping review. Int Orthod. 2025; 23(1): 100959. **[Crossref]**