Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Assessment of ChatGPT-5 and Gemini 2.5 Pro Responses to Frequently Asked Patient Questions in Oral Surgery: A Study on Validity, Readability, and Temporal Consistency

Aylin Ekmekcioglu

aekmekcioglu@mehmetakif.edu.tr

Mehmet Akif Ersoy University

Burcu Yeliz Kollayan

Mehmet Akif Ersoy University

Research Article

Keywords: Artificial intelligence, ChatGPT-5, Gemini 2.5 Pro, dental implant surgery, impacted tooth extraction, patient education, readability, temporal consistency

Posted Date: May 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9542161/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background:

Artificial intelligence (AI) chatbots are increasingly used by patients as easily accessible sources of health-related information. In dentistry, particularly in oral and maxillofacial surgery and dentomaxillofacial radiology, the accuracy, readability, and stability of chatbot-generated responses are of critical importance because misleading or overly complex information may negatively affect patient understanding and decision-making. This study aimed to evaluate the validity, reliability, readability, and temporal consistency of responses generated by ChatGPT-5 and Gemini 2.5 Pro to frequently asked patient-centered questions related to dental implant surgery and impacted tooth extraction.

Methods:

This comparative observational benchmarking study included 40 frequently asked patient-centered questions, equally distributed between implant surgery and impacted tooth extraction. Questions were developed and refined by an oral and maxillofacial surgeon and a dentomaxillofacial radiologist. Each question was submitted to ChatGPT-5 and Gemini 2.5 Pro across six independent rounds at one-week intervals using official APIs under standardized zero-shot conditions. Responses were independently evaluated by two experts using a 5-point Likert scale for clinical accuracy and appropriateness, followed by consensus scoring. Readability was assessed using the Ateşman readability formula, which is specifically adapted for Turkish texts. Inter-rater reliability was analyzed using intraclass correlation coefficient (ICC), and internal consistency was assessed with Cronbach's alpha. Differences between models were analyzed using non-parametric tests, with $p < 0.05$ considered statistically significant.

Results:

A total of 480 chatbot responses were evaluated, including 240 responses generated by ChatGPT-5 and 240 responses generated by Gemini 2.5 Pro 2.5 Pro. Both models showed generally high clinical performance, with median Likert scores of 4 (IQR: 4–5). Mean scores were 4.18 ± 0.71 for ChatGPT-5 and 4.17 ± 0.78 for Gemini 2.5 Pro 2.5 Pro, with no statistically significant difference between the models ($p = 0.801$). Temporal consistency analysis showed no significant variation across six rounds for either ChatGPT-5 ($p = 0.106$) or Gemini 2.5 Pro 2.5 Pro ($p = 0.128$), with low Kendall's W values indicating high short-term stability. Readability analysis demonstrated that ChatGPT-5 produced significantly more readable responses than Gemini 2.5 Pro 2.5 Pro, with higher Ateşman readability scores (83.43 vs. 71.86, $p < 0.001$). Subgroup analysis revealed no significant differences between implant-related and impacted tooth-related questions for either chatbot. Inter-rater reliability was excellent (ICC = 0.92), and internal consistency was high (Cronbach's alpha = 0.94).

Conclusions:

ChatGPT-5 and Gemini 2.5 Pro both generated generally accurate, clinically appropriate, and temporally stable responses to common patient questions regarding implant surgery and impacted tooth

extraction. While overall clinical validity was comparable between the models, ChatGPT-5 demonstrated superior readability, suggesting a potential advantage in patient-facing communication. AI chatbots may serve as useful adjunctive tools for dental patient education; however, their outputs should remain under professional supervision and should not replace individualized clinical consultation.

Introduction

Artificial intelligence (AI) technologies have driven a rapid transformation in the healthcare sector in recent years, fundamentally reshaping patient–physician interactions. AI-based systems offer significant potential to enhance diagnostic accuracy, optimize clinical workflows, and support patient-centered care (1, 2). In particular, AI-driven chatbots have emerged as accessible digital tools that facilitate patients' access to health-related information by providing instant responses regarding symptoms, treatment options, and preventive care, thereby redefining traditional models of health communication (1).

In dentistry, AI applications are increasingly utilized across a wide range of subspecialties, including prosthodontics, orthodontics, periodontology, and oral and maxillofacial surgery (3). Previous studies have demonstrated that AI-based algorithms can achieve high accuracy rates, particularly in diagnostic tasks relying on optical and radiographic imaging (3, 4). Deep learning–based convolutional neural network models, for example, have shown strong performance in the detection and diagnosis of dental caries using radiographic data, underscoring the growing role of AI in dental diagnostics (4).

Alongside diagnostic applications, AI-based chatbots have gained attention as tools for patient education and public access to dental information(1). Recent studies evaluating chatbot performance in implant dentistry have reported variable levels of clinical accuracy, completeness, and reliability in responses to frequently asked patient questions (5–7). Comparative analyses involving multiple AI platforms have further highlighted inconsistencies in the quality and safety of information provided regarding oral implant surgery, emphasizing the need for systematic evaluation of these systems before their widespread adoption in patient communication (5–7).

However, the majority of existing studies rely on single time-point (static) assessments, providing limited insight into the stability of chatbot-generated responses over time (5, 6). AI models are continuously updated and are known to carry the risk of generating incorrect or fabricated information, commonly referred to as "hallucinations"(1, 2). As a result, the temporal consistency of chatbot responses remains insufficiently explored, representing a significant gap in the current literature (7).

Surgical procedures requiring detailed radiological assessment, such as dental implant placement and impacted tooth extraction, are associated with a higher risk of intraoperative and postoperative complications. Accurate and reliable patient information regarding these procedures is therefore critical (5–7). Inaccurate or incomplete information in these areas may compromise patient safety, making them particularly suitable domains for evaluating the reliability of AI-based chatbots (5–7).

Recently developed next-generation AI models, such as GPT-5 and Gemini 2.5 Pro, offer enhanced reasoning capabilities and improved contextual coherence compared with earlier versions. Evaluating the performance of these advanced models is essential for establishing future standards in patient communication and education within dental practice (5–7).

The null hypothesis (H0) of this study is that there is no significant difference between the two AI chatbots in terms of clinical accuracy and temporal consistency. The alternative hypothesis (H1) proposes that differences exist between the chatbots with respect to these parameters.

Accordingly, the aim of this study was to evaluate the validity, reliability, and temporal consistency of responses generated by ChatGPT-5 and Gemini 2.5 Pro to frequently asked questions posed by patients considering dental implant treatment or scheduled for impacted tooth extraction. These two platforms were selected because they represent widely used and readily accessible AI-based chatbots with increasing relevance for patient education and health information seeking in contemporary clinical and public settings.

Materials and Methods

Study Design

This study was designed as a comparative observational benchmarking study with repeated measurements over time. The primary aim was to evaluate the validity, reliability, readability, and temporal consistency of responses generated by artificial intelligence (AI) chatbots to patient-centered questions related to oral and maxillofacial surgery and dentomaxillofacial radiology.

Data Collection

A total of 40 frequently asked patient-centered questions related to implant surgery and impacted tooth extraction were included in the study. The questions covered key clinical domains, including surgical procedures, radiographic evaluation, prognosis, potential complications, adverse effects, and pre-, intra-, and postoperative considerations.

Of the 40 questions, 20 pertained to implant surgery and 20 to impacted tooth extraction. For each topic, 10 questions were independently developed by an oral and maxillofacial surgeon and an oral and maxillofacial radiologist based on commonly encountered patient inquiries in routine clinical practice. The remaining 10 questions per topic were initially generated using an AI chatbot. Although some initial questions were generated with AI assistance, all items were independently reviewed, modified, and finalized by two domain experts. The AI-generated wording was not used directly; instead, the questions were clinically restructured to reflect real-world patient concerns. The complete list of questions is presented in Fig. 1 and flow chart of the study design is presented in Fig. 2.

AI Chatbot Interaction Protocol

AI Chatbot Interaction Protocol
Each question was submitted to two large language models across six independent rounds: ChatGPT-5, accessed via the OpenAI API using the gpt-5 model endpoint, and Gemini 2.5 Pro 2.5 Pro, accessed via the Google Gemini 2.5 Pro API using the Gemini 2.5 Pro-2.5-pro model endpoint. All interactions were conducted between January 9, 2026, and February 13, 2026. To improve methodological reproducibility and avoid variability associated with web-interface settings, all queries were performed through official API endpoints. No system-level instructions, predefined roles, contextual prompts, or additional explanatory information were provided. Each interaction consisted solely of the patient-centered question entered verbatim under zero-shot conditions. A new independent session was used for each query to prevent contextual carryover effects. No external tools, web browsing, retrieval-augmented generation, or file-upload functions were enabled. The default generation settings of the respective API platforms were used, and no additional parameter adjustments were applied by the researchers.

To evaluate the stability and reproducibility of model responses over time, each question was repeated across independent rounds separated by time intervals rather than being administered on consecutive days. This design was intended to minimize short-term response similarity and to better capture temporal variability in model outputs, while also reflecting real-world patterns of repeated patient inquiries occurring at different time points.

Ethical Considerations

Ethics committee approval was not required, as the study did not involve human participants, patient data, or identifiable personal information. According to national and institutional regulations, studies based exclusively on publicly available AI-generated content are exempt from ethical approval.

Scoring and Evaluation

Each chatbot response was independently evaluated by an oral and maxillofacial surgeon and an oral and maxillofacial radiologist using a 5-point Likert scale:

- **1 (Strongly disagree)**: Completely inaccurate or irrelevant
- **2 (Disagree)**: Largely incorrect with limited correct information
- **3 (Neutral)**: Partially correct but containing notable errors, omissions, or irrelevant content
- **4 (Agree)**: Mostly accurate with minor inaccuracies or missing details
- **5 (Strongly agree)**: Fully accurate, comprehensive, and clinically appropriate(25)

Initially, all responses were scored independently, resulting in 480 individual ratings per evaluator (40 questions × 2 chatbots × 6 rounds). Discrepancies in scoring between the evaluators were resolved

through structured consensus discussions, with reference to the AAOMS clinical guideline on the management of impacted third molars for impacted tooth–related assessments, and the EAO and ITI consensus reports for implant-related assessments(22, 23). Consensus scores were used for the final statistical analysis.

Readability Analysis

In addition to expert-based evaluation, readability was objectively assessed to evaluate the quality of patient-centered communication. Given that the target patient population of the present study consisted of Turkish-speaking individuals, all patient-centered questions were presented to the chatbots in Turkish to better reflect the intended clinical communication context. Therefore, the readability of the generated Turkish responses was assessed using the Ateşman readability formula, which was developed specifically for Turkish texts. This formula evaluates text difficulty based on sentence length and word length, with higher scores indicating easier readability. The use of a Turkish-specific readability index was preferred to ensure linguistic appropriateness and methodological validity. Readability analyses were conducted using automated text analysis tools, and the results were interpreted according to established readability levels for Turkish. (24).

Hallucination and Reference Accuracy Assessment

All chatbot responses were further examined for hallucinations, defined as factually incorrect statements, unsupported clinical claims, or fabricated scientific references. The presence of references to scientific literature, clinical guidelines, or statistical data was recorded.

When references were included, their authenticity and relevance were verified using established databases and guideline documents. Fabricated or unverifiable references were documented and considered during the qualitative interpretation of chatbot performance.(25)

Statistical Analysis

Expert ratings based on a 5-point Likert scale were considered the primary outcome measure. Given the ordinal nature of the data and repeated-measures design, non-parametric statistical methods were used. Inter-rater reliability between the two evaluators was assessed using the intraclass correlation coefficient (ICC) based on a two-way random-effects model with absolute agreement (ICC [2,1]). Internal consistency of the rating scale was evaluated using Cronbach's alpha. Descriptive statistics were presented as median and interquartile range (IQR), as well as mean ± standard deviation (SD). Differences between chatbot models were analyzed using the Wilcoxon signed-rank test. Effect size (r) was calculated for pairwise comparisons. Temporal consistency across six repeated measurements was evaluated using the Friedman test. Kendall's W was reported as a measure of effect size for repeated measures. Subgroup comparisons (implant-related vs. impacted tooth-related questions) were analyzed using the Wilcoxon signed-rank test. Readability scores were compared between models using the

Wilcoxon signed-rank test, and effect size (r) was calculated. A p-value < 0.05 was considered statistically significant. All analyses were performed using SPSS software (version 26.0; IBM Corp., Armonk, NY, USA).

Results

A total of 480 chatbot responses were evaluated across six independent rounds. Of these, 240 responses were generated by ChatGPT-5 and 240 responses were generated by Gemini 2.5 Pro 2.5 Pro. The median Likert score was 4 (IQR: 4–5) for both models, indicating generally accurate and clinically appropriate responses. The mean scores were 4.18 ± 0.71 for ChatGPT-5 and 4.17 ± 0.78 for Gemini 2.5 Pro 2.5 Pro. No statistically significant difference was observed between the two models (p = 0.801), and the effect size was negligible (r = 0.02) (Table 1).

Temporal consistency analysis across six rounds revealed no significant variation in performance for either model (ChatGPT: p = 0.106; Gemini 2.5 Pro: p = 0.128). Kendall's W values were low (ChatGPT: 0.045; Gemini 2.5 Pro: 0.043), indicating minimal variability and high stability of responses over time (Table 2).

Readability analysis demonstrated that ChatGPT-5 responses had significantly higher Ateşman readability scores compared with Gemini 2.5 Pro 2.5 Pro responses (83.43 vs. 71.86, p < 0.001), with a large effect size (r = 0.75), indicating substantially better patient-level readability (Table 3).

Subgroup analysis showed no statistically significant differences between implant-related and impacted tooth-related questions for ChatGPT (p = 0.312) or Gemini 2.5 Pro (p = 0.338), suggesting consistent performance across clinical domains (Table 4).

Reliability analysis showed excellent agreement between evaluators, with an intraclass correlation coefficient (ICC [2,1]) of 0.92. Additionally, the internal consistency of the rating scale was high, with a Cronbach's alpha of 0.94 (Table 5).

No fabricated references, unsupported statistical claims, or clinically relevant hallucinations were detected in either model.

Table 1
Overall Model Performance

Parameter	Model	Median [IQR]	Mean ± SD	p-value	Effect size (r)
Likert Score	ChatGPT	4 [4–5]	4.18 ± 0.71	0.801	0.02
	Gemini 2.5 Pro	4 [4–5]	4.17 ± 0.78		

IQR: Interquartile range; SD: Standard deviation.

[*] *Effect size (r) was calculated for Wilcoxon signed-rank test comparisons.*

Table 2
Temporal Consistency Across Six Rounds

Model	Friedman p-value	Kendall's W
ChatGPT	0.106	0.045
Gemini 2.5 Pro	0.128	0.043

* *Kendall's W represents the effect size for the Friedman test, indicating the degree of agreement across repeated measurements.*

Table 3
Readability Analysis

Metric	ChatGPT	Gemini 2.5 Pro	p-value	Effect size(r)
Ateşman readability score	83.43	71.86	**< 0.001***	0.75

* *Statistically significant at p < 0.05.*

Ateşman readability scores indicate readability, with higher values reflecting easier text comprehension.

Table 4
Subgroup Analysis

Model	Group	Mean ± SD	p-value
ChatGPT	Implant	4.22 ± 0.68	0.312
	Impacted teeth	4.14 ± 0.73	
Gemini 2.5 Pro	Implant	4.20 ± 0.75	0.338
	Impacted teeth	4.13 ± 0.80	

* *SD: Standard deviation.*

Group comparisons were performed using the Wilcoxon signed-rank test.

Table 5
Reliability Analysis

Metric	Value
Cronbach's Alpha	0.94
ICC (2, 1)	0.92

* *Cronbach's alpha values > 0.90 indicate excellent internal consistency.*

* *ICC: Intraclass correlation coefficient; values > 0.90 indicate excellent agreement.*

Discussion

Artificial intelligence–based chatbots have become increasingly prominent in healthcare, particularly in the context of patient education and digital health communication. Patients' growing demand for rapid and accessible information regarding treatment options, possible complications, preoperative preparation, and recovery processes has led to the consideration of these systems as complementary tools in healthcare delivery (1, 2). Similarly, in dentistry, chatbots have been reported to offer potential benefits in patient education, answering frequently asked questions related to treatment, and facilitating access to online health information. Nevertheless, the accuracy, clinical appropriateness, readability, and temporal consistency of the information provided by these systems remain important areas of debate (3, 5–7). This issue is particularly relevant in oral and maxillofacial surgery, where patient information generated by artificial intelligence must be evaluated carefully because clinical decision-making often involves radiological assessment, individualized risk evaluation, and procedures associated with a higher potential for complications (12).

In this context, the present study evaluated the validity, reliability, readability, and temporal consistency of responses generated by ChatGPT-5 and Gemini 2.5 Pro to patient-centered questions related to dental implant surgery and impacted tooth extraction. The findings demonstrated that both chatbots generally produced responses with a high level of clinical appropriateness and comparable accuracy according to expert evaluation. In addition, the absence of a significant change in performance across repeated measurements over a six-week period suggests that the responses of both models remained relatively stable in the short term. However, the readability analysis revealed that ChatGPT-5 performed significantly better than Gemini 2.5 Pro, indicating that clinical accuracy and patient comprehensibility do not always progress in parallel.

The findings of the present study are broadly consistent with the current literature suggesting that artificial intelligence–based chatbots may be useful for patient-oriented information delivery in dentistry. Previous studies in implant dentistry have shown that AI-generated responses are often clinically acceptable, informative, and potentially useful for patient education. Tuzlalı et al. reported that AI chatbots could demonstrate notable performance in responding to frequently asked questions about dental implant treatment, although marked differences were observed between platforms (5). Similarly, Binaljadm et al. emphasized that chatbots could provide generally reliable information in the context of implant dentistry, but that their level of validity may be limited when stricter evaluation criteria are applied (7). In the field of endodontics, Mohammad-Rahimi et al. likewise showed that chatbots may generate responses that initially appear useful and acceptable, while reliability and validity may vary depending on the model (8). Taken together, these findings support our results, indicating that AI-assisted patient education is promising, yet the quality of responses may vary according to the model and clinical context.

One of the noteworthy findings of this study was the absence of a statistically significant difference between ChatGPT-5 and Gemini 2.5 Pro in terms of overall expert evaluation scores. This result suggests

that, in the context of routine patient-centered questions in oral and maxillofacial surgery, both models have reached a comparable level of basic clinical accuracy. This finding is partly in line with recent comparative studies in implantology and oral surgery. For example, Taymour et al. reported that ChatGPT-3.5, ChatGPT-4, and Gemini 2.5 Pro demonstrated acceptable performance in terms of overall usefulness and reliability when responding to dental implantology questions (11). Similarly, Aziz et al. showed that ChatGPT and Gemini 2.5 Pro were generally capable of producing reliable responses to patient questions related to orthognathic surgery (15). However, some studies in the literature have reported more pronounced performance differences between chatbots, particularly when reference accuracy, comprehensiveness, higher-level validity criteria, and topic-specific clinical details were assessed (7, 8). Therefore, the lack of a significant difference in the present study may be attributable to the fact that both models represent more recent-generation systems and were tested under the same standardized, context-independent, zero-shot conditions. This finding may suggest that the performance gap observed more prominently in earlier-generation chatbot studies is becoming narrower among newer-generation models.

One of the most distinctive aspects of our study is that temporal consistency was directly assessed through a repeated-measures design. In the existing dental literature, the performance of AI chatbots has mostly been evaluated at a single time point; therefore, the extent to which their responses change over time has remained largely unclear. However, because AI models are continuously updated, the possibility that the same question may elicit different responses at different time points is particularly important in healthcare from the perspective of clinical reliability. In the present study, neither ChatGPT-5 nor Gemini 2.5 Pro demonstrated a significant change in performance over the six-week evaluation period, suggesting that both models were capable of producing relatively stable responses in the short term. The low Kendall's W values further support this interpretation and indicate limited variability across repeated measurements. Rather than contradicting previous studies, this finding complements them, as the current literature has generally focused on chatbot accuracy at a specific time point, whereas our study provides an additional methodological contribution by examining short-term reproducibility. In this respect, our study offers a more comprehensive framework by evaluating not only the ability of chatbots to produce correct responses but also whether such responses are maintained over time (1, 2, 7–9, 11, 13).

Another important finding of this study was that ChatGPT-5 demonstrated significantly better readability than Gemini 2.5 Pro. This result highlights that health information generated by artificial intelligence should be evaluated not only in terms of clinical accuracy but also in terms of patient comprehensibility. In the context of patient education, medically accurate information alone is not sufficient; it must also be communicated in a manner appropriate to the patient's educational level, health literacy, and cognitive load. Recent studies across different fields of dentistry support this view. In particular, studies in oral and maxillofacial radiology, dental trauma, and oral health education have reported that the informational quality and readability of chatbot responses do not always align. Some systems may provide accurate information but fail to communicate it in a patient-friendly manner. From this perspective, our finding suggests that although the two models demonstrated similar clinical performance, they did not provide

the same level of communication quality from the patient's standpoint. Therefore, readability should not be considered merely a secondary outcome in the evaluation of AI-based patient information tools, but rather an independent and clinically meaningful outcome variable (9, 16, 19, 21).

The subgroup analysis also yielded noteworthy findings. In the present study, neither ChatGPT-5 nor Gemini 2.5 Pro demonstrated a significant performance difference between implant-related questions and impacted tooth extraction–related questions. This suggests that both models exhibited a similar response profile across two clinically distinct and important areas of oral and maxillofacial surgery. However, these two topic areas differ substantially in their informational content. Implant-related questions generally focus on osseointegration, bone support, healing, imaging, and long-term success, whereas questions related to impacted tooth extraction more commonly involve surgical risks, anatomical proximity, the possibility of nerve injury, postoperative complications, and patient anxiety. Therefore, the observation of similar performance may indicate that advanced AI models can generate balanced responses not only in a narrow clinical topic but also across different subdomains of patient-centered oral surgery. Since many studies in the literature focus on a single topic area, our findings are also important in that they provide a comparative evaluation across different clinical categories. Nevertheless, the finding of similar performance across topics does not necessarily mean that the models produced equally detailed and error-free responses in all subdomains. Rather, it indicates similarity at the general score level and underscores the need for more detailed topic-based analyses in future studies (5, 8, 10, 15, 16).

The high inter-rater agreement and internal consistency observed in this study also represent important methodological strengths that support the reliability of the findings. The high intraclass correlation coefficient indicates that the oral and maxillofacial surgeon and the dentomaxillofacial radiologist evaluated the chatbot responses in a largely similar manner. Likewise, the high Cronbach's alpha value demonstrates strong internal consistency of the scoring approach used in the study. Since subjective elements based on expert interpretation cannot be completely eliminated when evaluating AI-generated texts, high agreement coefficients are important in showing that the findings were not primarily driven by random differences between evaluators. Although previous studies have also reported acceptable levels of agreement in expert-based assessments, the high reliability coefficients in our study allow both the similarities and the differences between models to be interpreted more robustly (5, 13).

Nevertheless, the present findings should not be interpreted as indicating that AI-based chatbots are ready to replace clinician-led patient counseling. Although both models generally produced accurate and clinically acceptable responses, previous studies have shown that important limitations may emerge when stricter evaluation criteria are applied, particularly with respect to validity, reference accuracy, comprehensiveness, empathy, and contextual appropriateness. Studies conducted in implant dentistry, endodontics, oral cancer, trauma, and oral surgery have reported that AI chatbots may provide useful supportive information in some situations, yet may also generate incomplete, misleading, or insufficient guidance in complex clinical scenarios. Furthermore, patient communication requires not only informational accuracy but also consideration of the individual clinical situation, personalized

explanation of risks, and communication tailored to the psychological needs of the patient. In this regard, AI chatbots should, at best, be regarded as supportive educational tools and should not be considered independent systems capable of diagnosis, treatment planning, or providing counseling equivalent to informed consent (7, 8, 14, 17, 18, 20).

The relatively favorable results obtained in this study may be explained by several methodological and technological factors. First, both platforms evaluated in this study represent current and advanced large language models that were widely used during the study period. This may have contributed to the generation of more fluent, contextually coherent, and structured responses compared with earlier-generation models. Second, although the questions used in the study were patient-centered, they were reviewed and refined by an oral and maxillofacial surgeon and a dentomaxillofacial radiologist to reflect clinical reality. This approach may have provided a more realistic and balanced basis for evaluation compared with ambiguous or artificially constructed question sets. In addition, all questions were submitted through official APIs under standardized zero-shot conditions and in new sessions each time. This minimized contextual carryover effects and contributed to obtaining results in a more controlled and reproducible manner. Therefore, the high clinical appropriateness and short-term stability observed in our study should be interpreted not only as a result of model capacity but also as a consequence of a carefully designed methodological framework (5, 8, 9, 11).

Several limitations of this study should be acknowledged. First, the study included only two AI chatbots and was limited to dental implant surgery and impacted tooth extraction; therefore, the findings cannot be directly generalized to all dental disciplines or to all subfields of oral and maxillofacial surgery. Second, the six-week repeated-measures design allowed assessment of short-term temporal consistency; however, it may not reflect long-term model behavior, particularly after major model updates, retraining processes, or interface-level modifications. Third, readability assessment was based on established text analysis metrics and did not reflect more direct clinical outcomes such as actual patient comprehension, reduction in anxiety, or confidence in decision-making. Fourth, although hallucination and reference verification were included in the evaluation framework, these variables were not presented as separate quantitative outcome categories, which limits the depth of interpretation regarding source reliability and misinformation generation. Finally, although high inter-rater agreement was achieved in the expert-based scoring approach, it is not possible to completely eliminate a certain degree of subjective interpretation when evaluating AI-generated texts. Therefore, the present results should be interpreted cautiously and supported by future studies with larger samples, multilingual designs, and patient-based outcome measures.

Conclusion

In conclusion, this study demonstrated that ChatGPT-5 and Gemini 2.5 Pro were generally capable of producing accurate, clinically appropriate, and short-term stable responses to frequently asked patient questions related to dental implant surgery and impacted tooth extraction. However, similar levels of clinical validity do not necessarily indicate similar communication quality; notably, ChatGPT-5

demonstrated superior readability performance. Our findings suggest that AI-based chatbots may serve as useful supportive tools for patient education, while also clearly indicating that these systems cannot replace professional clinical evaluation and individualized patient counseling. Therefore, the future use of AI-assisted patient education tools should involve a combined assessment of accuracy, readability, consistency, and clinical supervision.

Future studies may expand this line of research in several important directions. First, larger-scale and multilingual study designs would allow a more comprehensive evaluation of AI chatbot performance in patient communication across different linguistic and cultural contexts. In particular, in languages other than English, not only clinical accuracy but also communication-related dimensions such as readability, empathy, and cultural appropriateness should be assessed separately. In addition, studies with longer follow-up periods are needed to determine whether temporal stability is maintained after major model updates. Future research may also compare standardized zero-shot queries with patient-friendly, empathy-oriented, or simplified prompting strategies. Most importantly, in addition to expert ratings, the evaluation of direct patient-centered outcomes such as patient comprehension, anxiety reduction, trust, satisfaction, and confidence in decision-making would provide a more accurate understanding of the real clinical utility of these systems.

Abbreviations

AAOMS
American Association of Oral and Maxillofacial Surgeons
AI
Artificial Intelligence
API
Application Programming Interface
EAO
European Association for Osseointegration
ICC
Intraclass Correlation Coefficient
IQR
Interquartile Range
ITI
International Team for Implantology
LLM
Large Language Model
OMFS
Oral and Maxillofacial Surgery
SD
Standard Deviation
SPSS

Statistical Package for the Social Sciences

Declarations

Ethics approval and consent to participate

Ethics committee approval was not required because this study did not involve human participants, patient records, biological samples, animal subjects, or identifiable personal information. The study was based exclusively on patient-centered questions and responses generated by artificial intelligence chatbots under standardized experimental conditions. Therefore, informed consent was not applicable.

Clinical trial number

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was conducted without any external funding or financial support from public, commercial, or not-for-profit funding agencies.

Author contributions

AE and BYK contributed to the conception and design of the study. Question development, refinement, chatbot querying, and data collection were performed by AE and BYK. Expert evaluation of chatbot responses was performed independently by AE and BYK, and discrepancies were resolved through consensus discussions. Statistical analysis, data interpretation, and manuscript preparation were carried out by AE and BYK. Both authors critically reviewed previous versions of the manuscript, read, and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 2019;6(2):94–8.
2. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2(4):230–43.
3. Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. J Dent Res. 2020;99(7):769–74.
4. Lee JH, Kim DH, Jeong SN, Choi SH. Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. J Dent. 2018;77:106–11.
5. Tuzlalı M, Baki N, Aral K, Aral CA, Bahçe E. Evaluating the performance of AI chatbots in responding to dental implant FAQs: A comparative study. BMC Oral Health. 2025;25:1548.
6. Özden Yüce M, Adalı E, Yaman İ, Alpaguter B, Karadeniz B. Exploring AI-Driven Insights on Oral Implant Surgery: A Study of Four Different AI Applications. Int Arch Dent Sci. 2025;46(1):19–25.
7. Binaljadm TM, Alqutaibi AY, Zafar MS, Saker S, Halboub E. Artificial Intelligence Chatbots as Sources of Implant Dentistry Information for the Public: Validity and Reliability Assessment. [journal metadata PubMed/Thieme listing]. 2025. 10.1055/s-0045-1809155.
8. Mohammad-Rahimi H, Ourang SA, Pourhoseingholi MA, Dianat O, Dummer PMH, Nosrat A. Validity and reliability of artificial intelligence chatbots as public sources of information on endodontics. Int Endod J. 2024;57(3):305–14. 10.1111/iej.14014.
9. Helvacioglu-Yigit D, Demirturk H, Ali K, Tamimi D, Koenig L, Almashraqi A. Evaluating artificial intelligence chatbots for patient education in oral and maxillofacial radiology. Oral Surg Oral Med Oral Pathol Oral Radiol. 2025;139(6):750–9. 10.1016/j.oooo.2025.01.001.
10. Yagcı F, Eraslan R, Albayrak H, İpekten F. Accuracy and Reliability of Artificial Intelligence Chatbots as Public Information Sources in Implant Dentistry. Int J Oral Maxillofac Implants. 2026;41(2):254–62. 10.11607/jomi.11280.
11. Taymour N, Fouda SM, Abdelrahaman HH, Hassan MG. Performance of the ChatGPT-3.5, ChatGPT-4, and Google Gemini 2.5 Pro large language models in responding to dental implantology inquiries. J Prosthet Dent. 2025;134(6):2427–34. 10.1016/j.prosdent.2024.12.016.
12. Vázquez-Sebrango G, Anitua E, Macía I, Arganda-Carreras I. The role of artificial intelligence in implant dentistry: a systematic review. Int J Oral Maxillofac Surg. 2025;54(11):1098–122. 10.1016/j.ijom.2025.04.005.
13. Liu Z, Nalley A, Hao J, et al. The performance of large language models in dentomaxillofacial radiology: a systematic review. Dentomaxillofac Radiol. 2025;54(8). 10.1093/dmfr/twaf060.

14. Şişman AÇ, Acar AH. Artificial intelligence-based chatbot assistance in clinical decision-making for medically complex patients in oral surgery: a comparative study. BMC Oral Health. 2025;25(1):351. 10.1186/s12903-025-05732-w.
15. Aziz AAA, Abdelrahman HH, Hassan MG. The use of ChatGPT and Google Gemini 2.5 Pro in responding to orthognathic surgery-related questions: a comparative study. J World Fed Orthod. 2025. 10.1016/j.ejwf.2024.09.004.
16. Guven Y, Ozdemir OT, Kavan MY. Performance of Artificial Intelligence Chatbots in Responding to Patient Queries Related to Traumatic Dental Injuries: A Comparative Study. Dent Traumatol. 2024;41(6). 10.1111/edt.13020.
17. Johnson AJ, Singh TK, Gupta A, Sankar H, Gill I, Shalini M, Mohan N. Evaluation of validity and reliability of AI chatbots as public sources of information on dental trauma. Dent Traumatol. 2025.
18. Çege EE et al. Evaluation of the Performance of Artificial Intelligence-Based Chatbots in Providing First-Aid Information on Dental Trauma According to the ToothSOS Application. Dent Traumatol. 2025.
19. Özdemir ÖT et al. Evaluation of the readability, quality, and accuracy of AI chatbot responses to questions about deleterious oral habits. 2025. *BMC Oral Health* [or PubMed-indexed journal listing].
20. Topdağı B, Kavaz T. Assessment of information quality in contemporary artificial intelligence systems for digital smile design: A comparative analysis. J Prosthet Dent. 2025;134(4). 10.1016/j.prosdent.2025.06.030.
21. Rokhshad R, et al. Efficacy and empathy of AI chatbots in answering common oral cancer questions. [Elsevier oral/maxillofacial surgery journal listing in PubMed]; 2025.
22. American Association of Oral and Maxillofacial Surgeons. The Management of Impacted Third Molar Teeth. Rosemont (IL): AAOMS; 2024.
23. Donos N, Horvath A, Calciolari E, Mardas N. Impact of timing of dental implant placement and loading: Summary and consensus statements of group 1—The 6th EAO Consensus Conference 2021. Clin Oral Implants Res. 2021;32(Suppl 21).
24. Ateşman E. Measuring readability in Turkish. Dil Dergisi. 1997;58:71–4.
25. Roustan D et al. The Clinicians' Guide to Large Language Models. PMC / clinical review, 2025.

Figures

No.	Questions About Impacted Tooth Extraction	Questions About Dental Implant Surgery
1	Should every impacted tooth necessarily be extracted?	Can dental implants be placed in everyone, and am I a suitable candidate?
2	Is impacted tooth extraction different from routine tooth extraction?	How long does implant treatment take, and how many sessions are required?
3	Will I feel pain during impacted tooth extraction?	How long do pain and swelling last after implant surgery?
4	How long does the procedure take?	How long do dental implants last, and can I use them for a lifetime?
5	Will local anesthesia or general anesthesia be used?	Can my body reject the dental implant?
6	How long does pain last after extraction?	If there is insufficient bone, does this mean that implants cannot be placed?
7	When will swelling and bruising subside?	Is implant treatment expensive, and why does it cost more than other treatment options?
8	Are sutures placed after impacted tooth extraction?	Can I chew with an implant as I would with a natural tooth?
9	When are sutures removed, or do they dissolve on their own?	What should I pay attention to after implant surgery?
10	How long does bleeding last after extraction?	Can I undergo MRI or computed tomography after receiving an implant?
11	When can I eat after extraction?	Why are radiographs or computed tomography required before implant placement?
12	When can I smoke after impacted tooth extraction?	Can implant-related computed tomography (CBCT) harm me or expose me to radiation?
13	Will there be any permanent change in my face after extraction?	Is computed tomography necessary for every implant patient?
14	Is there a risk of infection after impacted tooth extraction?	What does it mean if computed tomography shows insufficient bone volume? Does it mean that implant placement is not possible?
15	When can I return to work or school after impacted tooth extraction?	Why is the location of the nerves in the mandible so important?
16	Why are radiographs or computed tomography taken before impacted tooth extraction?	Why are follow-up radiographs necessary after implant surgery?
17	Is computed tomography required for every impacted tooth?	Will I feel pain during implant surgery?
18	Can the radiographs or computed tomography harm me?	What are sinus lifting and bone grafting, and why might they be needed?
19	If the tooth appears close to the nerve on radiographs, does this mean that it cannot be extracted?	Does smoking affect implant success?
20	Why is the relationship with the sinus important in impacted teeth of the maxilla?	Is there an age limit for dental implant treatment?

Figure 1

Patient-centered questions used to evaluate chatbot responses. The question set included 20 items related to impacted tooth extraction and 20 items related to dental implant surgery. Questions were designed to reflect common patient concerns regarding indications, procedural steps, postoperative care, radiological assessment, complications, and treatment prognosis

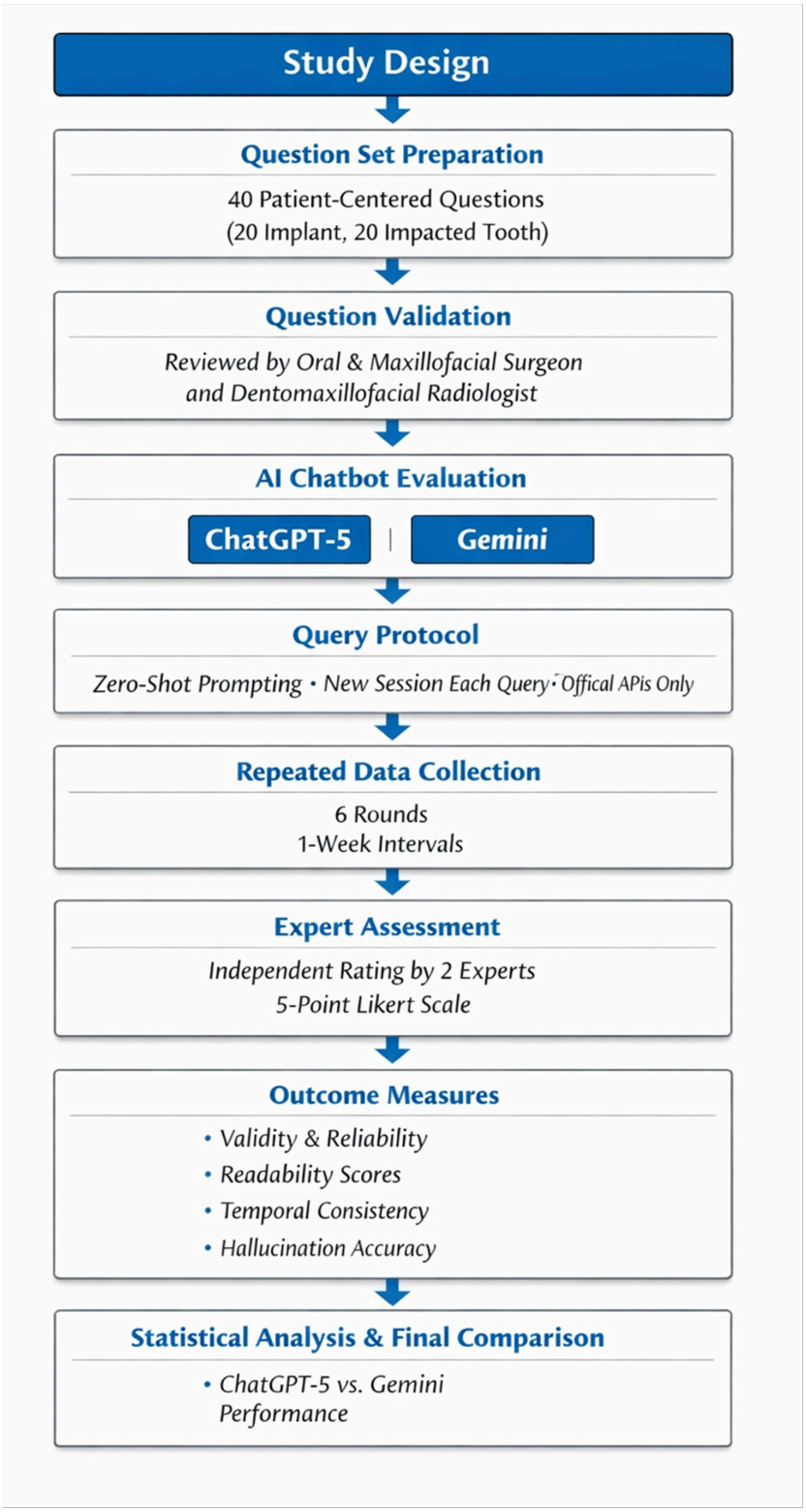


Figure 2

Flow chart of the study design, question development, chatbot interaction protocol, expert evaluation, and statistical analysis.