BMC Oral Health

RESEARCH

Open Access

Comparative performance of AI chatbots in dental implantology: insights and limitations

Sultan Merve Uçar[1*], Selin Gaş[2*] and Rafat Sasany[3]

Abstract

Objective This study critically evaluated the performance, accuracy, and clinical relevance of three large language models ChatGPT-4o, Claude 3.5, and Gemini 1.5 Pro when answering expert-generated questions on zygomatic implantology. The goal was to determine the extent to which such tools may function as educational or clinical decision supports in maxillofacial surgery.

Methods Thirty-eight standardized questions were developed by four oral and maxillofacial surgeons with advanced expertise in zygomatic implantology. Each model's responses were independently assessed by five calibrated clinical raters using validated metrics DISCERN, GQS, and a 5-point Accuracy Rubric to judge reliability, quality, and factual correctness. Non-parametric statistics (Kruskal–Wallis with Bonferroni post hoc correction; Spearman correlation) were used, and inter-rater reliability was quantified by ICC(2,1) = 0.86–0.91 ($p < 0.001$).

Results Gemini 1.5 Pro achieved slightly higher mean scores for response quality and accuracy, whereas Claude 3.5 and ChatGPT-4o performed comparably. However, absolute differences were modest (≤ 0.5 points on 5-point scales), indicating relative trends rather than decisive superiority. All models produced readable, clinically relevant content, though variability persisted in the depth and specificity of clinical guidance.

Conclusion Current AI language models exhibit moderate but inconsistent competency when addressing complex implantology scenarios. While Gemini 1.5 Pro scored marginally higher, these differences are unlikely to be of major practical consequence. Continuous validation, transparent reporting of model versions, and expert supervision remain essential before integrating such systems into routine dental education or clinical decision-making.

Keywords Artificial intelligence, Dental implantology, Dental education, Clinical decision-making

*Correspondence:
Sultan Merve Uçar
s.merveucr@gmail.com
Selin Gaş
selineren104@gmail.com
[1]Department of Prosthodontics, Faculty of Dentistry, Gelişim Üniversitesi, Istanbul, Turkey
[2]Department of Oral and Maxillofacial Surgery, Faculty of Dentistry, Gelişim Üniversitesi, Istanbul, Turkey
[3]Department of Prosthodontics, Faculty of Dentistry, Biruni University, Zeytinburnu, Turkey

Introduction

Rehabilitation of severely atrophic, fully edentulous maxilla remains challenging due to limited bone volume, complicating the placement of conventional implants [1]. Traditional approaches, such as bone augmentation, sinus lift, and Le Fort I osteotomy, achieve success rates of 60–90% but often require prolonged, multi-stage procedures that increase the risk of postoperative complications [2, 3]. Zygomatic implants offer an effective alternative by immediately anchoring in the zygomatic bone, bypassing the need for extensive grafting procedures and reducing treatment time, and have been used in conjunction with conventional-length dental implants in patients with severe maxillary resorption, demonstrating survival rates ranging from 96% to 100% [2, 4].

In recent years, the use of zygomatic implants has gained significant attention for the rehabilitation of atrophic edentulous maxillae. The 2023 ITI Consensus Workshop provides evidence-based guidance on the indications, surgical techniques, and long-term outcomes associated with these implants [5]. This consensus underscores the importance of careful case selection, particularly in severe maxillary resorption cases, and emphasizes advances in surgical protocols that have improved success rates. As our study's research questions and answers are aligned with the clinical recommendations and decisions outlined in this consensus, it serves as a foundational framework for evaluating the efficacy and safety of zygomatic implants in current practice.

Zygomatic implants are associated with several technical and prosthetic complications that can affect long-term success. One of the most common issues is screw or abutment loosening, which occurs frequently due to bending moments transmitted through the implant, especially in full-arch fixed reconstructions. These mechanical problems can lead to further prosthetic failures, such as fractures of the veneering material or loosening of components, often necessitating retightening or repairs [2, 6]. Placement challenges also contribute to complications; improper positioning or surgical technique can increase the risk of sinus perforation, oroantral communications, or implant malalignment, which may compromise osseointegration or lead to sinusitis. Moreover, the biomechanics of the implant support system can result in material fatigue, causing fractures of the prosthetic components. Effective surgical planning and patient education are essential to minimize these issues, and managing expectations is critical given the reported 54.1% complication rate over five years for screw-supported prostheses [1, 2, 7]. Addressing these complications promptly is vital for maintaining implant stability, prosthetic function, and overall treatment success.

Successful rehabilitation with zygomatic implants begins with thorough patient selection. Clinicians should carefully evaluate bone anatomy, sinus health, systemic conditions, and patient expectations to identify candidates who will most benefit from this graftless approach while minimizing risks [8]. Careful assessment and individualized planning are essential to optimize surgical outcomes and long-term implant success.

Artificial intelligence (AI) is a rapidly evolving technology designed to create intelligent systems capable of simulating human cognitive functions [9, 10]. Advances in big data, computational power, and algorithms over the past two decades have increasingly simplified and enhanced various aspects of daily life [11]. In healthcare, especially medicine and dentistry, AI's potential is growing, particularly in diagnostics, treatment planning, and education [12–14]. The integration of AI into routine clinical and academic tasks is transforming healthcare delivery and educational methods, offering significant benefits to both learners and practitioners [15, 16]. By processing vast data efficiently, AI models can offer objective and comprehensive clinical and histopathological analyses, improving treatment strategies and prognoses [17–19]. Recent studies have demonstrated that AI-driven approaches, particularly AI-enhanced cone beam computed tomography (CBCT) analysis, can significantly improve surgical outcome prediction and optimize prosthetic load distribution by processing complex anatomical data with greater precision. [20] These advances enable more accurate preoperative planning and personalized treatment strategies, ultimately leading to better patient prognoses and enhanced clinical outcomes.

Since its release on November 30,2022, ChatGPT by OpenAI has gained widespread attention for its ability to engage users in natural, language-based dialogues and generate diverse content such as essays, summaries, stories, and code. Trained on a vast collection of text from books, articles, and websites, it quickly attracted a massive user base, reaching 100 million users within two months [21]. As AI tools like ChatGPT become more integrated into education and clinical practice, assessing their accuracy and reliability is essential. This is especially important in fields like dental implantology, where reliable knowledge can directly impact patient outcomes. In this project, we aim to evaluate the quality and clinical relevance of AI-generated responses, exploring their potential to support learning and decision-making in dental education and practice.

Clinical decision-making integrates comprehensive patient-specific evaluations and shared decision-making models to optimize treatment outcomes [22]. Concurrently, evolving educational frameworks in dentistry increasingly utilize AI tools for enhancing knowledge dissemination and clinical training while emphasizing critical appraisal and ethical use of technology [23–25]. These

multidimensional frameworks provide the foundation for assessing the applicability and impact of zygomatic implants and AI in modern dental research and practice.

In this study, we aim to evaluate the performance and suitability of three prominent AI chatbots ChatGPT, Claude, and Gemini for providing reliable, understandable, and accurate information relevant to dentists and PhD students. By comparing these models on parameters such as content quality, readability, and trustworthiness, our research investigates their potential role as educational and clinical decision-making aids in modern dental practice.

Materials and methods

Question development and categorization

The question set used in this evaluation was collaboratively developed by four maxillofacial surgery specialists with extensive clinical expertise in zygomatic implantology. This expert-driven approach was chosen to ensure that the questions comprehensively encompass the critical domains of clinical decision-making, including indications, surgical technique, complications, prosthetic planning, and educational considerations. Drawing upon the collective experience of these specialists aligns with established practices in consensus-building methodologies, such as the Delphi technique, to formulate focused and clinically relevant evaluation items. While the number of experts was limited, their specialized knowledge aimed to maximize the content validity and relevance of the question set for assessing AI chatbot performance within this complex field.

Table1. presents the full list of queries formulated for this project. These questions were designed to comprehensively address nine critical domains of treatment outcomes and clinical decision making in zygomatic implantology, broadly aligned with the latest International Team for Implantology (ITI) Consensus framework. The ITI Consensus Workshop, held in 2023, provides evidence based clinical recommendations. By basing our question set in accordance with these consensus domains, we aimed to ensure clinical relevance and comprehensive coverage of pivotal topics in contemporary implant practice. The domains included indications, diagnostic approaches, surgical techniques, complications, prosthetic planning, success and survival rates, alternatives, educational requirements, and clinical ethics.

AI-Generated answer collection and expert evaluation

Rater Panel and Scoring Framework: Five independent raters (three board-certified oral and maxillofacial surgeons and two prosthodontists with over 10 years of clinical experience) evaluated all 38 standardized questions for each AI model. All raters scored every response independently and anonymously to ensure full data overlap and comparability across models. The evaluation employed three validated metrics: (1) DISCERN, a 15-item tool assessing the reliability and quality of health information (score range: 15–75); (2) GQS (Global Quality Scale), a 5-point Likert scale rating overall quality and educational usefulness (1 = poor, 5 = excellent); and (3) Accuracy Rubric, a 5-point expert-derived scale evaluating factual and clinical correctness (1 = incorrect or misleading, 5 = fully accurate and clinically applicable). A summary of the scoring system and interpretation for each scale has been included in Supplementary Table S1 for reference.

The AI-generated answers analyzed in this study were collected based on the standardized question set described above. The responses provided by three large language models ChatGPT-4o, Claude 3.5, and Gemini 1.5 were obtained independently through publicly available free-tier access as of June 2025. These versions were selected due to their high popularity and accessibility among dental professionals, researchers, and students, thereby providing a realistic perspective of AI performance in everyday clinical education and research contexts.

Model version and reproducibility clarification

Responses were collected in June 2025 using the publicly available free-access versions of ChatGPT-4o (OpenAI, May 2025 build), Claude 3.5 (Anthropic, June 2025), and Gemini 1.5 Pro (Google DeepMind, June 2025). Since ChatGPT-4o has now evolved into ChatGPT-5, this temporal limitation is acknowledged in the Discussion to emphasize the inherent challenge of reproducibility across evolving AI models.

Design justification

Each question was intentionally entered in a new chat session to avoid contextual contamination and to simulate independent clinical consultations. Although this design disables the models' adaptive feedback learning, it allows standardized evaluation and prevents bias from cumulative conversation memory.

The accuracy and quality of the AI-generated answers were then independently assessed by five clinical experts, including maxillofacial surgeons. These specialists evaluated the responses based on their clinical experience and current evidence-based literature, including ITI consensus treatment guidelines and best practices. The experts employed a standardized scoring methodology to objectively quantify the accuracy, comprehensiveness, and clinical applicability of each answer. This expert-driven and evidence-based evaluation ensured that the AI responses were judged not only for factual correctness

Table 1 Standardized Questions Administered to ChatGPT-4o, Claude 3.5, and Gemini 1.5

No.	Research Question
1	What are the definitive indications for zygomatic implant placement?
2	Under what circumstances should zygomatic implants be preferred over conventional implants?
3	Which classification systems are used in cases of severe maxillary atrophy?
4	Are zygomatic implants suitable for patients with a history of graft failure?
5	How do zygomatic implants provide a solution in cases of trauma, tumor resection, or congenital bone deficiency?
6	What characterizes the patient profile for whom zygomatic implant placement is contraindicated?
7	Which radiological imaging modalities should be utilized in the planning of zygomatic implants?
8	Why is the use of Cone Beam Computed Tomography (CBCT) considered essential?
9	What anatomical considerations must be taken into account when evaluating the zygomatic arch?
10	What are the current surgical techniques employed in zygomatic implant placement?
11	In which clinical scenarios should the Quad Zygoma technique be implemented?
12	Is the use of a surgical guide recommended during zygomatic implant placement?
13	What positioning strategies are employed during the placement of zygomatic implants?
14	What are the most critical anatomical risks encountered during surgery?
15	What are the most commonly observed complications associated with zygomatic implants?
16	What is the risk of developing maxillary sinusitis following zygomatic implant placement?
17	How can naso-orbital or infraorbital nerve injuries be prevented during surgery?
18	What treatment options are available in cases of zygomatic implant failure?
19	Should fixed or removable prostheses be preferred on zygomatic implants?
20	Is splinting between implants mandatory in prosthetic design?
21	Is immediate loading feasible, and under what clinical conditions is it recommended?
22	Why is the use of long angulated abutments considered important?
23	What are the long-term survival rates of zygomatic implants?
24	According to the systematic review in the ITI report, what is the five-year success rate?
25	Is there a difference in success rates among patients who undergo immediate loading?
26	Are zygomatic implants associated with higher risks compared to conventional implants?
27	How do zygomatic implants compare with bone grafting combined with conventional implants?
28	Which treatment modality is more effective in terms of patient satisfaction?
29	What are the differences in cost, treatment duration, and complication rates between approaches?
30	What level of surgical experience is required for performing zygomatic implant procedures?
31	Which disciplines should be represented in the multidisciplinary team responsible for performing zygomatic implant procedures?
32	What clinical or digital simulation tools are recommended for educational purposes in zygomatic implant training?
33	What level of evidence does the current literature provide regarding zygomatic implants?
34	Which systematic reviews form the basis of the ITI Consensus Report on zygomatic implants?
35	Which areas require further clinical research in the context of zygomatic implantology?
36	What essential elements should be included in the informed consent form for zygomatic implant procedures?
37	How should this treatment option be presented to patients to ensure informed decision-making?

but also for clinical relevance and safety considerations in real-world dental implantology practice.

Statistical analysis and AI response evaluation

Descriptive statistics including mean, standard deviation, median, minimum, maximum, and interquartile ranges were used to summarize the accuracy, quality, and readability metrics of AI chatbot responses across all questions, domains, and subdomains. The distribution of median accuracy scores was also calculated to provide a comprehensive view of performance variability.

Comparisons between the three AI models evaluated in this study Claude 3.5, Gemini 1.5 Pro, and ChatGPT-4o were conducted using non-parametric statistical methods due to the nature of the data distribution. The Kruskal-Wallis test was used for comparing three or more independent groups without normal distribution, followed by post hoc Bonferroni corrections to identify specific group differences [26–28]. Relationships between continuous, non-normally distributed variables were assessed via Spearman correlation coefficients. For categorical variables where expected sample.

Inter-rater reliability was calculated using the two-way random-effects intraclass correlation coefficient [ICC(2,1)] to assess agreement among raters across all scoring metrics. The resulting ICC values indicated excellent reliability for all scales (ICC = 0.86–0.91; 95% CI = 0.82–0.93; $p < 0.001$). Non-parametric analyses were maintained using Kruskal–Wallis tests followed by Bonferroni corrections, and Spearman correlation was used for ordinal data. Effect sizes (Cohen's d) were additionally reported and interpreted as small (0.2), medium (0.5), and large (0.8) to provide practical significance alongside p-values. size assumptions were unmet, Fisher's Exact test was employed.

All analyses were performed using IBM SPSS Statistics version 27 and GraphPad Prism version 10 (GraphPad Software Inc, La Jolla, CA). Statistical significance was set at $p < 0.05$.

To simulate authentic clinical use and prevent response bias, each question was entered individually into a new session or conversation window for each AI chatbot in June 2025. AI generated responses were recorded systematically for evaluation.

Three board-certified maxillofacial surgery specialists independently and anonymously scored the responses, which were presented in a randomized order to ensure evaluator blinding.

The evaluation criteria emphasized scientific adequacy, assessing the accuracy and evidence-based correctness of each response; ease of understanding, focusing on the clarity and readability for the intended dental audience; and overall reader satisfaction, reflecting the usefulness and appropriateness of the information provided. Responses were rated using a 7-point Likert scale, ranging from 1 (Strongly Disagree) to 7 (Strongly Agree).

Readability scores were computed using the Readable Pro platform (https://app.readable.com/text/), which automatically applies the Flesch Reading Ease (FRES) and Flesch–Kincaid Grade Level (FKGL) indices. Each AI-generated response was uploaded verbatim to the platform, and results were averaged across all questions. Scores above 60 for FRES and below grade 8 for FKGL were considered easily comprehensible, following standard readability benchmarks.

The FRES score rates text on a scale from 0 to 100, where higher scores indicate easier readability; values above 60 are regarded as suitable for general comprehension. The FKGL estimates the U.S. school grade level required to understand the text, with patient-directed materials ideally not exceeding an eighth-grade reading level for optimal accessibility. These indices were automatically calculated by Readable Pro based on average sentence length and syllable count per word, providing an objective measure of linguistic complexity. A summary of the distributional structure of all scoring metrics used in the study is provided in Table 2. Additional category-level evaluation parameters and scoring characteristics

Table 2 Distribution of AI Response Characteristics by Model

	Claude		Gemini		ChatGPT	
	Min.-Maks.	Ort.±S.S.	Min.-Maks.	Ort.±S.S.	Min.-Maks.	Ort.±S.S.
DISCERN	14-25	22,00±2,87	20-25	23,92±1,61	12-25	21,00±4,33
GQS	3-5	4,49±0,65	3-5	4,62±0,59	2-5	3,95±0,85
Accuracy	3-5	4,41±0,64	3-5	4,62±0,59	2-5	3,78±1,00
FRES	0-80,1	17,01±20,31	4,1-21,4	13,18±3,91	4,1-21,4	13,17±3,89
FKGL	4,1-21,4	13,19±3,91	0-80,1	17,06±20,35	0-80,1	17,09±20,38

*$p < 0,05$

that informed the analytical framework are presented in Tables 5, 6, 7, 8. These tables support the methodological description and are fully interpreted in the Result section .

Results

Descriptive statistics including count, percentage, mean, standard deviation, median, minimum, maximum, and 25th and 75th percentiles were calculated. The Kruskal-Wallis test was used to compare three or more independent groups that did not have a normal distribution. Post hoc Bonferroni corrections were applied to identify the specific group or groups causing the differences. Spearman correlation coefficients were calculated to examine relationships between continuous variables that did not follow a normal distribution. For testing relationships between categorical variables where the sample size assumption (expected count > 5) was not met, Fisher's Exact test was employed. All analyses were performed using IBM SPSS Statistics version 27. The raw data supporting the findings of this study, including detailed scoring and response characteristics for each AI model, are provided in the Supplementary Materials for transparency and further reference (Table 2).

Although Gemini consistently demonstrated higher mean scores in response quality and accuracy metrics, the absolute differences among the three AI models were modest (≤ 0.5 on 5-point scales). These variations should be interpreted as relative trends rather than definitive indicators of model superiority. Claude and ChatGPT exhibited comparable but slightly lower performance across most evaluation domains, while readability scores remained similar across all models, indicating comparable text complexity and linguistic accessibility.

The distribution of AI response characteristics across clinical and educational categories is presented in Table 3 and 4. Gemini generally achieved higher mean scores in quality, accuracy, and clarity metrics across most categories, particularly in Indications and Patient Selection, Surgical Technique and Application, and Prosthetic Planning. Claude and ChatGPT showed comparable performance in several categories, though ChatGPT often scored lower in both quality and accuracy measures. Readability scores remained consistent across all AI models and categories, indicating similar levels of text complexity and accessibility in the responses.

The distribution of response characteristics by AI model is presented in the table, and comparisons were made using Kruskal-Wallis tests. The analysis revealed statistically significant differences in DISCERN, GQS, and Accuracy scores among the AI models ($p < 0.05$). Bonferroni post hoc tests for DISCERN showed statistically significant differences between Gemini and both ChatGPT and Claude groups ($p = 0.002$ and $p = 0.011$, respectively). Gemini's scores were higher than those of ChatGPT and Claude. No statistically significant difference was found between ChatGPT and Claude ($p = 1.000$).

For GQS, Bonferroni tests revealed statistically significant differences between ChatGPT and Claude, and between ChatGPT and Gemini ($p = 0.011$ and $p < 0.001$, respectively). Both Claude and Gemini had higher scores than ChatGPT. No significant difference was found between Claude and Gemini ($p = 1.000$).

For Accuracy, Bonferroni tests found statistically significant differences between ChatGPT and Claude, and between ChatGPT and Gemini ($p = 0.020$ and $p < 0.001$, respectively). Claude and Gemini had higher scores than ChatGPT. No significant difference was observed between Claude and Gemini ($p = 0.470$).

No statistically significant differences were observed in Flesch Reading Ease Score and Flesch-Kincaid Grade Level scores among the AI models ($p > 0.05$).

The distribution of response characteristics by AI model across categories is presented in the table, and comparisons were made using Kruskal-Wallis tests. The analysis revealed statistically significant differences in DISCERN and Accuracy scores among AI models in the Indications and Patient Selection category; in DISCERN, GQS, and Accuracy scores in the Surgical Technique and Application category; in FRES and FKGL scores in the Success and Survival Rates category; and in FKGL scores in the Education and Experience Requirements category ($p < 0.05$).

Bonferroni post hoc tests for the Indications and Patient Selection category showed statistically significant differences in DISCERN scores between Gemini and both Claude and ChatGPT groups ($p = 0.049$ and $p = 0.018$, respectively). Gemini's scores were higher than those of ChatGPT and Claude. For Accuracy in this category, a significant difference was found between ChatGPT and Gemini ($p = 0.023$), with Gemini scoring higher.

In the Surgical Technique and Application category, Bonferroni tests revealed significant differences in DISCERN scores between ChatGPT and Gemini ($p = 0.049$), with Gemini scoring higher. Significant differences were also observed in GQS ($p = 0.006$) and Accuracy ($p = 0.005$) scores between ChatGPT and Gemini, again with Gemini outperforming ChatGPT.

For the Success and Survival Rates category, Bonferroni tests showed statistically significant differences in FRES scores between Claude and both Gemini and ChatGPT ($p = 0.018$ for both), with Claude scoring higher. Similarly, FKGL scores in this category were significantly different between Claude and the other two groups ($p = 0.018$ for both), with Claude having higher scores.

In the Education and Experience Requirements category, Bonferroni tests indicated significant differences

Table 3 Distribution of AI Response Metrics across Clinical and Educational Domains (DISCERN, GQS, Accuracy, FRES, FKGL)

Category	Model	DISCERN Min–Max	DISCERN Mean±SD	GQS Min–Max	GQS Mean±SD	Accuracy Min–Max	Accuracy Mean±SD	FRES Min–Max	FRES Mean±SD	FKGL Min–Max	FKGL Mean±SD
Indications & Patient Selection	Claude	19–25	21.5±2.35	4–5	4.5±0.55	4–5	4.33±0.52	0–80.1	33±28.54	4.1–14.4	10.17±3.69
Indications & Patient Selection	Gemini	23–25	24.33±1.03	4–5	4.67±0.52	4–5	4.67±0.52	4.1–14.4	10.17±3.69	0–80.1	33±28.54
Indications & Patient Selection	ChatGPT	18–25	20.83±2.79	3–5	3.83±0.75	3–4	3.67±0.52	4.1–14.4	10.17±3.69	0–80.1	33±28.54
Diagnostic & Imaging Methods	Claude	25–25	25±0	5–5	5±0	5–5	5±0	0–24.6	13.67±12.53	10.9–14.3	12.4±1.73
Diagnostic & Imaging Methods	Gemini	22–25	24±1.73	4–5	4.67±0.58	4–5	4.67±0.58	10.9–14.3	12.4±1.73	0–24.6	13.67±12.53
Diagnostic & Imaging Methods	ChatGPT	21–25	22.67±2.08	4–5	4.33±0.58	4–5	4.33±0.58	10.9–14.3	12.4±1.73	0–24.6	13.67±12.53
Surgical Technique & Application	Claude	19–24	21.6±1.95	4–5	4.6±0.55	4–5	4.4±0.55	0–41.5	13.38±18.99	8.6–15.5	12.8±2.96
Surgical Technique & Application	Gemini	23–25	24.6±0.89	5–5	5±0	5–5	5±0	8.6–15.5	12.8±2.96	0–41.5	13.38±18.99
Surgical Technique & Application	ChatGPT	16–24	21±3.16	3–4	3.8±0.45	3–4	3.6±0.55	8.6–15.5	12.8±2.96	0–42.5	13.58±19.36
Complications & Risks	Claude	20–23	21.25±1.26	4–5	4.5±0.58	4–5	4.25±0.5	0–16.4	4.1±8.2	12–18.8	15.38±2.78
Complications & Risks	Gemini	20–24	22.25±2.06	4–5	4.5±0.58	4–5	4.5±0.58	12–18.8	15.38±2.78	0–16.4	4.1±8.2
Complications & Risks	ChatGPT	14–25	20±5.83	3–5	4±1.15	2–5	3.5±1.73	12–18.8	15.38±2.78	0–16.4	4.1±8.2
Prosthetic Planning	Claude	18–25	21±3.16	3–5	4±0.82	3–5	4±0.82	0–32	8±16	10.2–19.1	16.23±4.18
Prosthetic Planning	Gemini	20–25	23.5±2.38	4–5	4.5±0.58	4–5	4.5±0.58	10.2–19.1	16.23±4.18	0–32	8±16
Prosthetic Planning	ChatGPT	14–24	20.75±4.72	3–4	3.75±0.5	2–4	3.5±1	10.2–19.1	16.1±4.17	0–32	8±16
Success & Survival Rates	Claude	20–25	23±2.45	4–5	4.5±0.58	4–5	4.5±0.58	21–66.7	34.88±21.51	5.1–12.3	10.08±3.35
Success & Survival Rates	Gemini	21–25	24±2	3–5	4.5±1	3–5	4.5±1	5.1–12.3	10.08±3.35	21–66.7	34.88±21.51
Success & Survival Rates	ChatGPT	14–25	21±4.97	3–4	3.75±0.5	2–4	3.5±1	5.1–12.3	10.08±3.35	21–66.7	34.88±21.51
Alternative Treatments & Comparison	Claude	15–23	20±4.36	3–5	4.33±1.15	3–5	4.33±1.15	13.7–50.1	31.33±18.23	7.4–13.1	10.2±2.85
Alternative Treatments & Comparison	Gemini	22–25	24±1.73	4–5	4.67±0.58	4–5	4.67±0.58	7.4–13.1	10.2±2.85	13.7–50.1	32±18.2
Alternative Treatments & Comparison	ChatGPT	13–25	19.67±6.11	3–5	4±1	3–5	4±1	7.4–13.1	10.2±2.85	13.7–50.1	32±18.2
Education & Experience Requirements	Claude	14–25	21.33±6.35	3–5	4.33±1.15	3–5	4.33±1.15	0–0	0±0	14.9–21.4	17.97±3.27
Education & Experience Requirements	Gemini	23–25	24±1	4–5	4.33±0.58	4–5	4.33±0.58	14.9–21.4	17.97±3.27	0–0	0±0
Education & Experience Requirements	ChatGPT	24–25	24.67±0.58	4–5	4.67±0.58	4–5	4.67±0.58	14.9–21.4	17.97±3.27	0–0	0±0
Clinical Decision Ethics & Evidence Level	Claude	21–25	23.4±2.19	4–5	4.6±0.55	4–5	4.6±0.55	0–15	8.28±7.59	12.2–19.2	14.82±2.87
Clinical Decision Ethics & Evidence Level	Gemini	21–25	24.2±1.79	3–5	4.6±0.89	3–5	4.6±0.89	12.2–19.2	14.82±2.87	0–15	8.3±7.61
Clinical Decision Ethics & Evidence Level	ChatGPT	12–25	19.8±7.12	2–5	3.8±1.64	2–5	3.8±1.64	12.6–19.2	14.9±2.78	0–15	8.3±7.61

Values are presented as minimum–maximum and mean ± standard deviation

Table 4 Comparative Distribution of AI Response Characteristics by Model

	Claude	Gemini	ChatGPT		
	M.(%25-%75Ç.)	M.(%25-%75Ç.)	M.(%25-%75Ç.)	Test İstatistiği	p
DISCERN	22(21-25)	25(23-25)	22(18-25)	13,256	0,001*
GQS	5(4-5)	5(4-5)	4(3-5)	15,583	<0,001*
Accuracy	4(4-5)	5(4-5)	4(3-4)	17,602	<0,001*
FRES	13,3(0-29,2)	13,2(10,9-15,5)	13,2(10,9-15,5)	0,073	0,964
FKGL	13,2(10,9-15,5)	13,4(0-29,2)	13,4(0-29,2)	0,077	0,962

in FKGL scores between Claude and both Gemini and ChatGPT groups ($p = 0.017$ for both), with Claude scoring higher.

The distribution of GQS and Accuracy scores by AI model is presented in the table, and their associations were analyzed using Fisher's Exact tests. The analysis revealed statistically significant relationships between AI models and both GQS and Accuracy scores ($p < 0.05$). The proportion of responses with a GQS score of 5 was higher in Claude (56.8%) and Gemini (67.6%) models, while this rate was lower in ChatGPT (27%). Responses with an Accuracy score of 2 were observed only in ChatGPT (16.2%) and not in the other models. The proportion of responses with an Accuracy score of 5 was higher in Gemini (67.6%) and Claude (48.6%) models, while it was lower in ChatGPT (24.3%).

The distribution of GQS and Accuracy scores by AI model across categories is presented in the table, and their associations were analyzed using Fisher's Exact tests. The analysis revealed statistically significant relationships between AI models and both GQS and Accuracy scores in the Surgical Technique and Application category ($p < 0.05$). The proportion of responses with a GQS score of 5 was 100% in Gemini, 60% in Claude, and 0% in ChatGPT. The proportion of responses with an Accuracy score of 5 was 100% in Gemini, 40% in Claude, and 0% in ChatGPT.

The relationships among quality-related scores for AI types were examined using Spearman Correlation tests.

For Claude, analyses revealed positive and strong correlations between DISCERN and both GQS and Accuracy ($r = 0.782$ and $r = 0.782$; $p < 0.05$). A positive and strong correlation was also found between GQS and Accuracy ($r = 0.881$; $p < 0.05$). A very high negative correlation was calculated between FRES and FKGL ($r = -0.964$; $p < 0.05$).

For Gemini, analyses showed statistically significant, positive, and strong correlations between DISCERN and both GQS and Accuracy ($r = 0.883$ and $r = 0.883$; $p < 0.05$). A statistically significant, positive, and perfect correlation was found between GQS and Accuracy ($r = 1.000$, $p < 0.05$). A statistically significant, negative, and strong correlation was observed between Flesch Reading Ease Score (FRES) and Flesch-Kincaid Grade Level (FKGL) ($r = -0.964$, $p < 0.05$) (Figs. 1, 2, 3, 4).

For ChatGPT, analyses showed statistically significant, positive correlations at high/moderate levels between DISCERN and GQS, Accuracy, and FRES ($r = 0.933$; $r = 0.882$; $r = 0.417$; $p < 0.05$). A statistically significant, negative, and moderate correlation was found between DISCERN and FKGL ($r = -0.437$; $p < 0.05$). Positive and strong correlations were discovered between GQS and both Accuracy and FRES; while FKGL had a negative and moderate correlation ($r = 0.954$; $r = 0.391$; $r = -0.420$; $p < 0.05$).

Statistically significant, positive, and moderate correlations were found between Accuracy and FRES; and negative, moderate correlations with FKGL ($r = 0.351$; r

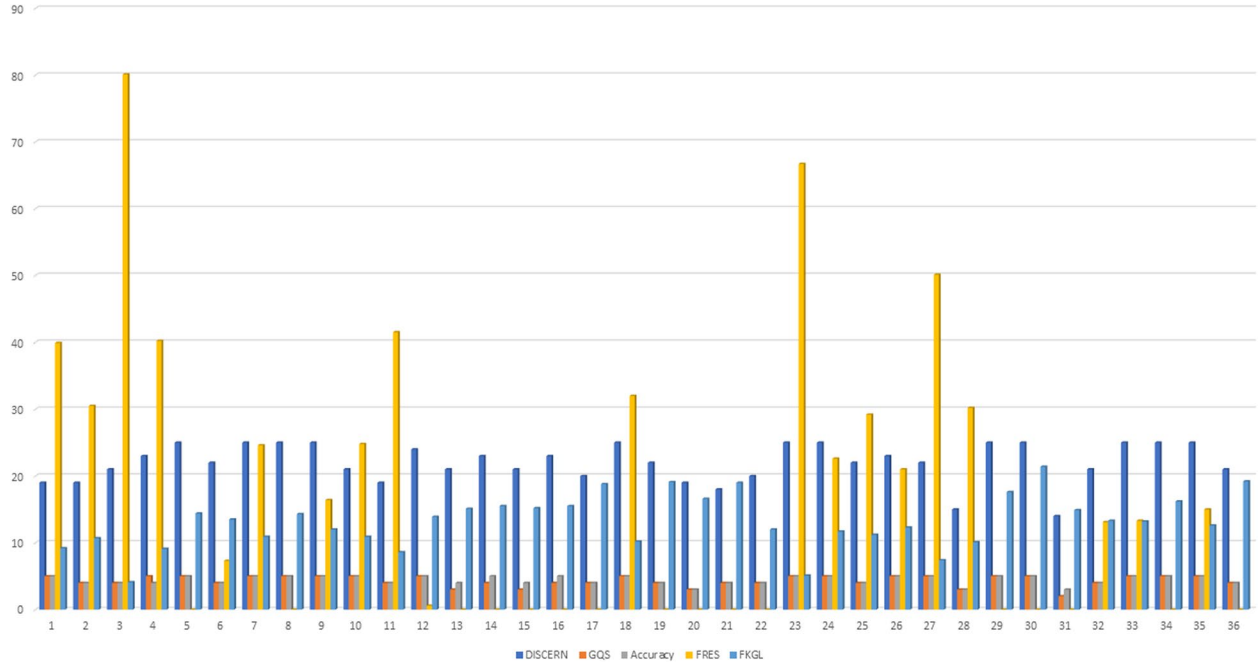

Fig. 1 Overall performance scores of Claude in dental implantology

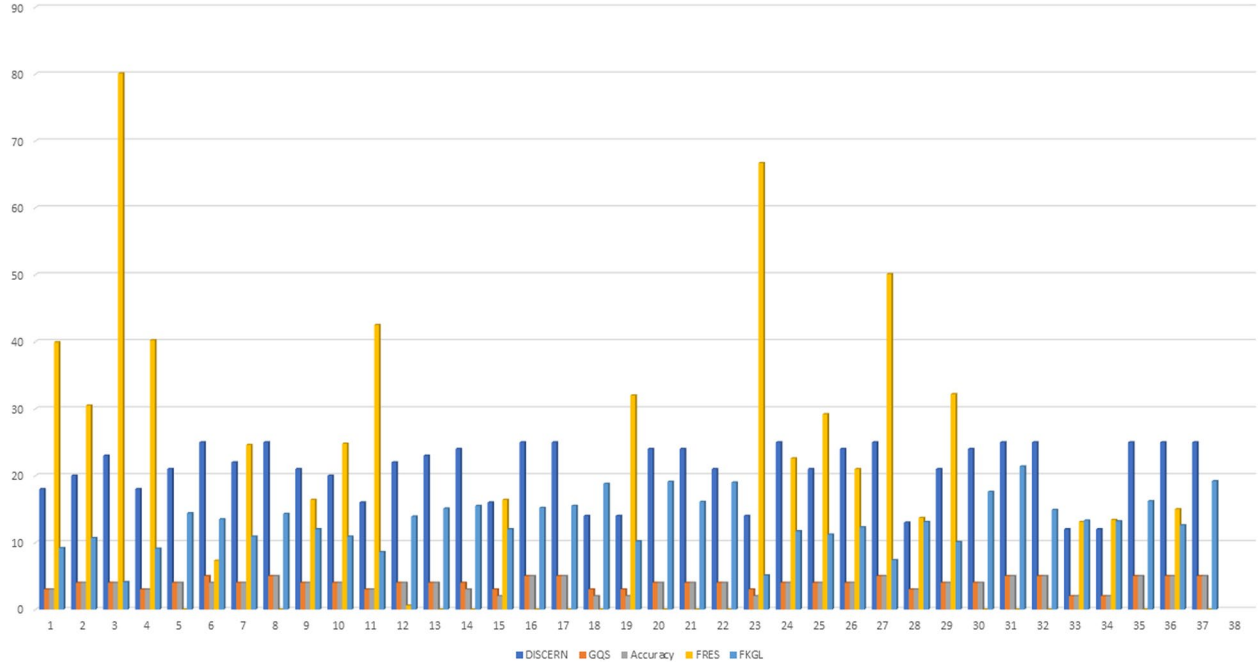

Fig. 2 Overall performance scores of ChatGPT in dental implantology

= −0.376; $p < 0.05$). A statistically significant, high-level, negative correlation was calculated between FRES and FKGL ($r = -0.965$; $p < 0.001$).

Discussion

Artificial intelligence-driven chatbots have established a presence across diverse sectors including finance, education, and customer support, and their application in healthcare is now gaining significant research attention [25].

People are increasingly turning to the Internet and social media platforms to seek medical information due to ease of accessibility, lower costs, and a desire to become more informed [23]. Patients often consult online resources and social media before seeking professional healthcare advice, which raises important

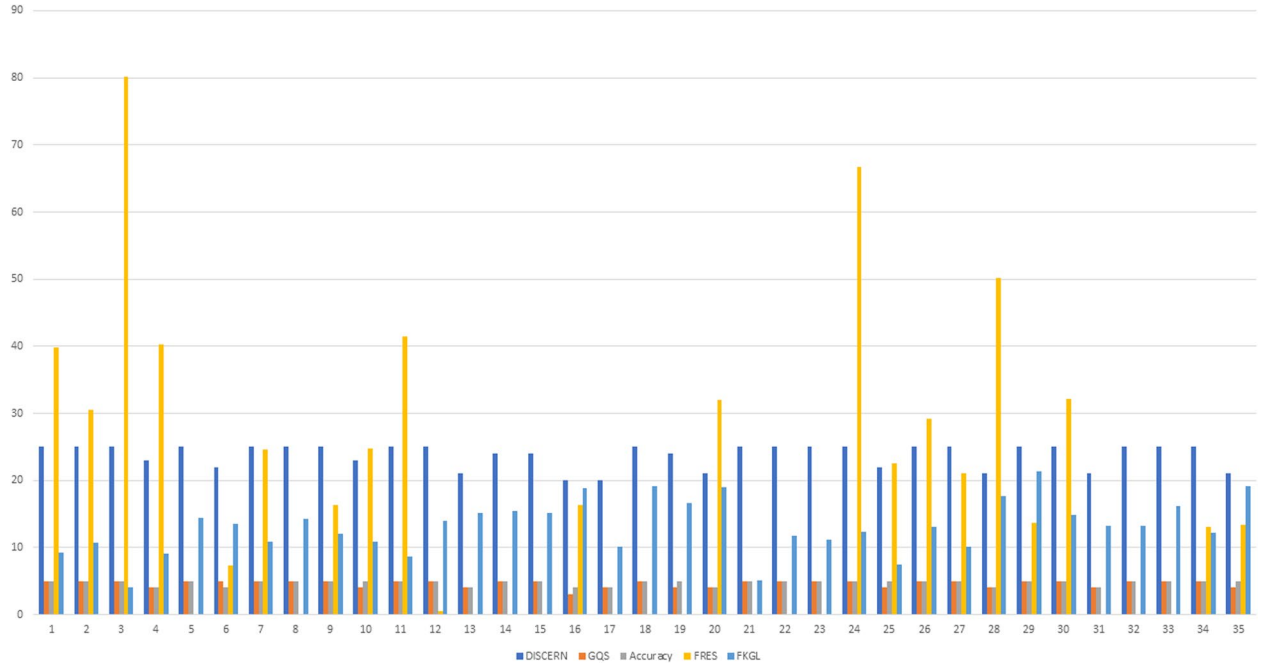

Fig. 3 Overall performance scores of Gemini in dental implantology

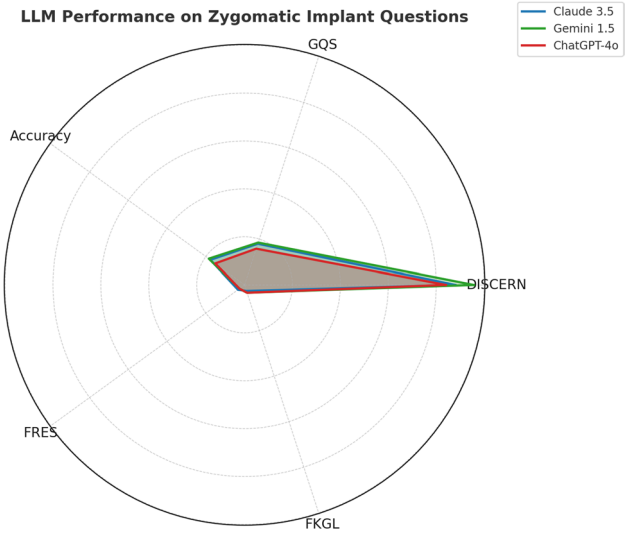

Fig. 4 Radar Plot Comparing Claude 3.5, Gemini 1.5, and ChatGPT-4o Across DISCERN, GQS, Accuracy, FRES, and FKGL Metrics for Zygomatic Implant Queries

questions about the reliability and accuracy of the information they receive. In this context, AI powered chatbots like ChatGPT have gained attention as potential tools for patient education and clinical decision support. This study highlights that while AI models such as ChatGPT-4 demonstrate improved performance over earlier versions like ChatGPT-3, there are notable limitations particularly regarding accuracy, detail orientation, and the ability to interpret or analyze images. Previous literature has noted that the older ChatGPT-3 model [25], trained on data up to 2021, lacked the capability to process visual inputs and often provided generalized or incomplete information, which could contribute to misinformation.

Although newer AI models such as ChatGPT-4 have introduced advanced features including potential image analysis capabilities, their current application remains largely focused on generating descriptive natural language responses. In highly specialized fields like oral and maxillofacial surgery, these models still face challenges regarding consistency, depth of understanding, and clinical nuance. Our findings align with broader reports highlighting variability in AI-generated content quality. Given these considerations and inherent limitations, it is premature to assert that AI currently offers transformative advantages in education or clinical decision-making. Careful evaluation and cautious implementation remain essential to mitigate risks of misinformation and medico-legal concerns.

The evaluation of AI generated responses on healthcare questions provides valuable insights into the capabilities and limitations of these models as potential clinical decision aids and educational tools. In this study, over 90% of responses from ChatGPT yielded ratings above the "poor" level, with many classified as "fair" or "good," suggesting a moderate level of accuracy and clinical relevance. These findings align with prior research assessing ChatGPT's utility in contexts such as otosclerosis surgery, where the majority of AI responses were deemed satisfactory by expert reviewers. For example, Sahin et al. observed that ChatGPT generally produces

effective advice, though some responses lack detailed contextualization.

Similarly, our results indicate that while AI can offer generally reliable information, certain answers particularly those related to postoperative care may lack specificity and depth, highlighting important areas for ongoing improvement and refinement. As such, these findings should be interpreted cautiously, recognizing both the potential and current limitations of AI systems in complex clinical settings.

The readability analysis revealed that despite reasonable accuracy, the complexity of the generated texts may present comprehension challenges for certain patient groups or individuals with limited health literacy. This aligns with prior research emphasizing the need to simplify AI-generated outputs to improve accessibility without compromising informational integrity. Our findings, consistent with previous studies in dental and maxillofacial contexts, highlight variability in performance both across AI models and within different clinical domains such as indications, surgical techniques, and patient education. While models like Gemini tend to demonstrate somewhat better accuracy and reliability in our evaluation, these differences are often modest and underscore the ongoing need for refinement to enhance AI's practical utility in diverse educational settings.

In this study, we aimed to evaluate the performance of AI chatbots in providing reliable and clinically relevant information related to dental implantology. While tools like Copilot are notable in the AI landscape, especially for coding and technical tasks, their applicability in medical education and clinical decision making remains limited. Therefore, we did not include Copilot in our evaluation. Instead, we selected models specifically designed or fine tuned for medical and educational purposes namely, ChatGPT, Gemini, and Claude due to their widespread popularity, accessibility, and documented capabilities in healthcare related content [17].

ChatGPT is extensively used and well known for its usability and vast data training, which makes it a common reference point (as seen, for example, in studies like) [10, 12, 19]. However, recent advancements and reports highlight the superior performance of newer models such as Gemini and Claude, particularly in medical and educational domains. These models have demonstrated significant strengths in understanding complex clinical contexts, providing more accurate, comprehensive, and nuanced responses advantages that are critically important in dental and medical education.

Consistent with previous literature, our results indicate that the Gemini and Claude models generally provided responses rated higher than those of ChatGPT across multiple metrics, including accuracy, comprehensiveness, and clinical relevance. As presented in Tables 4 and 5, Gemini and Claude achieved statistically significantly higher median scores in DISCERN, GQS, and accuracy evaluations compared to ChatGPT. For instance, within the domain of indications and patient selection, Gemini obtained a median DISCERN score of 25, which was significantly greater than ChatGPT's 20.5 ($p < 0.05$). Similarly, in categories such as surgical techniques and educational content, Gemini and Claude consistently outperformed ChatGPT with statistically significant differences ($p < 0.05$) (Tables 6, 7, 8).

However, it is important to interpret these differences with caution. The absolute differences in mean or median Likert scale scores between models were often modest, frequently less than 0.5 on a 5-point scale. While statistically significant, the clinical relevance of these small effect sizes may be limited. Incorporating effect size measures such as Cohen's d and confidence intervals into future analyses would enhance the meaningfulness of these results for clinical audiences by quantifying the magnitude and precision of the observed differences.

Likert-type scales, commonly used in medical and dental research for subjective evaluations, produce ordinal data where the numeric differences do not always correspond to equal intervals of meaningful change. Therefore, modest numeric differences, even if statistically significant, may not translate directly into a clinically important improvement or meaningful difference in AI chatbot performance.

In this context, incorporation of effect size measures (e.g., Cohen's d) and confidence intervals alongside p-values would provide a more comprehensive understanding of the magnitude and precision of the differences. This additional statistical reporting would help clinicians and educators better interpret whether improvements reflect substantive advances in AI quality or merely subtle variations detectable in a controlled study population.

Despite demonstrating a trend for higher performance by the Gemini model, the fluctuating accuracy and variation within domains suggest that no AI model currently offers a definitive advantage across all clinical decision making aspects in zygomatic implantology. This variability highlights ongoing challenges in applying AI as reliable, standalone educational or clinical decision support tools. Consequently, the role of AI in dental education and practice should be considered supplementary and supportive, contingent on rigorous validation, continual assessment, and clear guideline development.

Despite these encouraging results, significant gaps remain in the current capabilities of AI models. Limitations such as restricted question scope, lack of multimodal input processing (e.g., imaging), and the absence of longitudinal clinical validation reduce their immediate applicability for autonomous clinical decision-making. The complexity and variability of dental implant scenarios

Table 5 Comparative Analysis of AI Response Metrics across Clinical and Educational Categories (Only representative results. Full statistical analysis in Supplementary Table S5.)

Domain	Metric	Claude (M, IQR)	Gemini (M, IQR)	ChatGPT (M, IQR)	p-value
Indications & Patient Selection	DISCERN	21.5 (19–23)	25 (23–25)	20.5 (18–23)	0.041*
	Accuracy	4 (4–5)	5 (4–5)	4 (3–4)	0.026*
Surgical Technique & Application	GQS	5 (4–5)	5 (5–5)	4 (4–4)	0.008*
	Accuracy	4 (4–5)	5 (5–5)	4 (3–4)	0.007*
Success & Survival Rates	FRES	25.9 (21.8–48.0)	11.5 (8.2–12.0)	11.5 (8.2–12.0)	0.024*
Education & Experience	FKGL	17.6 (14.9–21.4)	0 (0–0)	0 (0–0)	0.022*

*Values are presented as median (25th–75th percentile). Significant results are marked with $p < 0.05$. Full detailed results across all categories are provided in Supplementary Table S5

Table 6 Distribution and Association of GQS and Accuracy Scores by AI Model

		Claude			Gemini			ChatGPT				
		n	%	%Ai.	n	%	%Ai.	n	%	%Ai.	Test Statistic	p
GQS	2	0a	0,0	0,0	0a	0,0	0,0	2a	100,0	5,4	15,443	0,007*
	3	3a	23,1	8,1	2a	15,4	5,4	8a	61,5	21,6		
	4	13a	32,5	35,1	10a	25,0	27,0	17a	42,5	45,9		
	5	21a	37,5	56,8	25a	44,6	67,6	10b	17,9	27,0		
Accuracy	2	0a	0,0	0,0	0a	0,0	0,0	6b	100,0	16,2	20,259	<0,001*
	3	3a	30,0	8,1	2a	20,0	5,4	5a	50,0	13,5		
	4	16a	37,2	43,2	10a	23,3	27,0	17a	39,5	45,9		
	5	18a	34,6	48,6	25b	48,1	67,6	9a	17,3	24,3		

*p < 0.05; %: Row percentage and %AI: Column percentage for AI types; different letters indicate statistically significant differences between column percentages within the same row

Table 7 Category-Wise Distribution and Correlations of GQS and Accuracy Scores by AI Model(Only representative results. Full statistical analysis in Supplementary Table S7.)

Domain	Metric	Claude	Gemini	ChatGPT	Test Statistic	p-value
Surgical Technique and Application	GQS	Scores mainly 4–5; balanced	Predominantly 5	Lower (3–4)	10.521	**0.010***
	Accuracy	4–5, balanced	Predominantly 5	Lower, some 3s	10.788	**0.009***
Success and Survival Rates	Accuracy	High (4–5)	Highest (5)	Lower (3–4)	9.088	0.085 (trend)
Education & Experience Requirements	FKGL (readability)	17.6 (14.9–21.4)	Lower (0)	Lower (0)	7.624	**0.022***
Clinical Decision Ethics	Accuracy	High (mostly 5)	Highest (5)	Wider variability (2–5)	7.287	0.121 (trend)

*p < 0.05; %: Row percentage and %AI: Column percentage for AI types; different letters indicate statistically significant differences between column percentages within the same row

Table 8 Pearson Correlation Coefficients (r) and Significance (p) Between DISCERN, GQS, Accuracy, FRES, and FKGL Scores Across AI Models

			GQS	Accuracy	FRES	FKGL
Claude	DISCERN	r	0,782	0,782	-0,098	0,085
		p	<0,001*	<0,001*	0,564	0,618
	GQS	r		0,881	0,098	-0,118
		p		<0,001*	0,564	0,485
	Accuracy	r			0,071	-0,078
		p			0,677	0,646
	FRES	r				-0,964
		p				<0,001*
Gemini	DISCERN	r	0,883	0,883	0,090	-0,054
		p	<0,001*	<0,001*	0,595	0,750
	GQS	r		1,000	0,050	-0,032
		p		<0,001*	0,769	0,853
	Accuracy	r			0,050	-0,032
		p			0,769	0,853
	FRES	r				-0,964
		p				<0,001*
ChatGPT	DISCERN	r	0,933	0,882	0,417	-0,437
		p	<0,001*	<0,001*	0,010*	0,007*
	GQS	r		0,954	0,391	-0,420
		p		<0,001*	0,017*	0,010*
	Accuracy	r			0,351	-0,376
		p			0,033*	0,022*
	FRES	r				-0,965
		p				<0,001*

*p < 0.05, r: Correlation coefficient, ** Pearson Correlation

require further refinement of AI educational models, as well as integration with expert oversight to ensure safe, reliable, and contextually appropriate guidance.

The representativeness of the question set relative to the full spectrum of clinical decision making remains uncertain due to the absence of a validated, standardized question bank specific to zygomatic implantology. The complexity of these cases and the evolving nature of treatment protocols contribute to this challenge. Additionally, current frameworks such as the recent International Team for Implantology (ITI) consensus provide structured guidance but do not offer comprehensive question repositories for systematic assessment.

Therefore, the external validity and generalizability of the findings are constrained to the scope of the selected queries. Future research would benefit from the development and utilization of larger, validated question repositories that can better capture the breadth of clinical scenarios and decision points. Such repositories would enhance the evaluation of AI tools across diverse clinical cases and improve the robustness of performance assessments.

Future research should focus on expanding AI evaluation frameworks, enhancing model adaptability to diverse clinical cases, and developing standardized assessment protocols that include multimodal data. With continued improvement and rigorous validation, AI models have the potential to become valuable, trusted adjuncts within dental education and clinical workflows, augmenting practitioner expertise and supporting optimal patient outcomes. To contextualize these findings, recent literature has addressed the growing role of AI-driven tools in implant planning and execution. For example, Marini et al. (2024) provided a comprehensive systematic review on the use of artificial intelligence in computerized implant placement, highlighting how AI-enhanced CBCT analysis can improve surgical outcome prediction and optimize prosthetic load distribution (Marini et al., 2024). Including this evidence aligns the present study with current advancements in predictive and computer-guided implantology, reinforcing the potential for AI-based systems to inform both preoperative assessment and intraoperative decision-making [20].

While this study situates the evaluation of AI chatbots within recognized frameworks such as the ITI Consensus on zygomatic implants, there is a limitation in fully integrating AI response assessments with the broader, nuanced body of implantology literature and clinical practice standards. Specifically, our analysis primarily focuses on comparing AI-generated answers against selected expert questions and consensus guidelines without deeply exploring how chatbot responses align or diverge from the established surgical protocols, complication management strategies, and real-world clinical decision-making intricacies in zygomatic implant therapy.

Recent advances in artificial-intelligence-based surgical planning, particularly in CBCT-guided and 3D model-driven zygomatic implant placement, demonstrate the growing relevance of AI in clinical decision support. AI-assisted systems now incorporate predictive algorithms for surgical trajectory optimization, bone-density mapping, and risk assessment, contributing to more accurate implant positioning and complication prevention. Integration of these algorithms into digital CAD/CAM workflows can facilitate pre-operative visualization, enhance prosthetic accuracy, and support individualized treatment planning. Such developments emphasize the translational bridge between educational language-model evaluations and real-world clinical applications of AI in maxillofacial surgery.

To illustrate model behavior in a complex scenario, the authors simulated an edentulous maxilla with previous sinus graft failure and pneumatized sinuses. The prompt "Outline a treatment plan using zygomatic implants" was submitted to all three AI models. Gemini proposed both intra- and extra-sinus trajectories with appropriate caution on posterior cantilever length; Claude offered partial guidance on sinus access and bone anchorage; ChatGPT produced a generalized response lacking risk stratification. This example highlights the variable depth and clinical precision of the models, reinforcing the need for expert oversight in interpretation.

To further assess educational value, responses were evaluated on three structural indicators: (1) presence of a pre-operative checklist (CBCT landmarks, sinus status, and zygomatic window planning), (2) intra-operative sequence with contingency decisions (e.g., handling soft bone at the malar buttress), and (3) post-operative surveillance guidance. Mean educational rubric scores were 2.4 ± 0.5 for Gemini, 2.1 ± 0.6 for Claude, and 1.7 ± 0.4 for ChatGPT. These findings suggest that while readability is comparable, the structural clarity and didactic organization of responses differ across models.

This gap limits the ability to fully appraise the clinical relevance and potential practical impact of AI assistance. Future studies should aim for a more comprehensive synthesis that contextualizes AI output within the rich, evolving implantology literature, including variable clinical scenarios, patient specific factors, and multidisciplinary treatment considerations. Such integration would enhance the translational value of AI assessments from theoretical performance metrics toward validated clinical utility.

This study's statistical approach aligns with best practices in medical education research, but further refinement in reporting and interpretation of clinical relevance is warranted for broader applicability of findings. Future

work should also consider user centered outcome measures and real-world impact assessments to better contextualize the importance of observed score differences.

AI-generated responses must not be used for autonomous surgical planning, unsupervised patient-specific postoperative advice, or imaging interpretation. These tools are intended solely as educational and supportive aids under qualified clinician supervision. Departments may adapt this "safe-use" framework as a policy statement to guide responsible application of AI chatbots in dental and surgical training.

Conclusions

While numerical performance differences between models were small, their qualitative impact becomes clearer when analyzed through real surgical examples. Translating statistical results into clinical language such as omission of sinus-care steps or incomplete prosthetic guidance makes the findings directly relevant to educators and implant surgeons.

In summary, this study highlights differences in diagnostic and informational performance among the AI models Gemini, Claude, and ChatGPT in the context of dental implantology, with a specific focus on complex cases involving zygomatic implants While the Gemini model generally demonstrates a trend toward higher scores in accuracy and overall response quality compared to Claude and ChatGPT, many of the observed differences on 5-point Likert scales are relatively small and fall within close ranges across all evaluated AI models. Therefore, it is more appropriate to state that Gemini *appears* to exhibit somewhat higher performance rather than asserting definitive superiority. This nuanced interpretation acknowledges the overlapping confidence in performance metrics and encourages cautious consideration of these findings, especially given the inherent variability in subjective scoring and the complex nature of clinical information evaluation.

The current analysis evaluated the free, publicly accessible versions of ChatGPT-4o, Claude 3.5, and Gemini 1.5 at a single time point. While Gemini demonstrated superior performance in accuracy, quality, and clinical relevance across multiple assessment domains, it is important to acknowledge the dynamic and rapidly evolving nature of AI language models. Frequent updates, changes in access tiers, and ongoing model retraining can substantially affect AI outputs over time. Consequently, the reproducibility and generalizability of these findings are inherently limited.

This study highlights the potential for AI chatbots to support dental education and clinical decision-making in the future. However, thoughtful integration into curricula and practice is essential, relying on robust validation, continuous evaluation, and clear guidelines. Considerable variability between models and clinical topics indicates the need for further refinement, research, and development before AI tools can be reliably adopted in real-world settings.

Future research should track AI model performance over time across versions and access tiers, with transparent version control and reporting to ensure reproducibility. Until such comprehensive validation exists, interpretations based on single-time-point assessments should be made cautiously. Continuous monitoring and rigorous validation are critical to safely incorporating AI into educational and clinical environments.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12903-025-07426-9.

Supplementary Material 1.

Acknowledgements

.

Authors' contributions

S.M.U.: Conceptualization, data collection, data analysis, manuscript drafting, corresponding author.S.G.: Study design, critical revision of the manuscript, approval of final version, data analysis, second corresponding authot.R.S.: Assistance in manuscript preparation, proofreading, data analysis, approval of final version.

Funding

The authors received no specific funding for this work.

Data availability

The full dataset is provided in the Supplementary Material.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent to publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 13 September 2025 / Accepted: 24 November 2025
Published online: 17 December 2025

References

1. Muñoz DG, Aldover CO, Zubizarreta-Macho Á, et al. Survival rate and prosthetic and sinus complications of zygomatic dental implants for the rehabilitation of the atrophic edentulous maxilla: a systematic review and meta-analysis. Biology. 2021. https://doi.org/10.3390/biology10070601.
2. Brennand Roper M, Vissink A, Dudding T, et al. Long-term treatment outcomes with zygomatic implants: a systematic review and meta-analysis. Int J Implant Dent. 2023;9(1):21. https://doi.org/10.1186/s40729-023-00479-x.
3. Kämmerer PW, Fan S, Aparicio C, et al. Evaluation of surgical techniques in survival rate and complications of zygomatic implants for the rehabilitation of the atrophic edentulous maxilla: a systematic review. Int J Implant Dent. 2023;9(1):11. https://doi.org/10.1186/s40729-023-00478-y.
4. Aalam AA, Krivitsky-Aalam A, Kurtzman GM, Mahesh L. The severely atrophic maxilla: decision making with zygomatic and pterygoid dental implants. J

Oral Biol Craniofac Res. 2023;13(2):202–6. https://doi.org/10.1016/j.jobcr.2023.01.008.
5. Al-Nawas B, Aghaloo T, Aparicio C, et al. ITI consensus report on zygomatic implants: indications, evaluation of surgical techniques and long-term treatment outcomes. Int J Implant Dent. 2023. https://doi.org/10.1186/s40729-023-00489-9.
6. Sailer I, Mühlemann S, Zwahlen M, Hämmerle CHF, Schneider D. Cemented and screw-retained implant reconstructions: a systematic review of the survival and complication rates. Clin Oral Implants Res. 2012;23(s6):163–201. https://doi.org/10.1111/j.1600-0501.2012.02538.x.
7. Chrcanovic BR, Albrektsson T, Wennerberg A. Survival and complications of zygomatic implants: an updated systematic review. J Oral Maxillofac Surg. 2016;74(10):1949–64. https://doi.org/10.1016/j.joms.2016.06.166.
8. Kämmerer PW, Al-Nawas B. Bone reconstruction of extensive maxillomandibular defects in adults. Periodontol 2000. 2023;93(1):340–57. https://doi.org/10.1111/prd.12499.
9. Sahin S, Erkmen B, Duymaz YK, Bayram F, Tekin AM, Topsakal V. Evaluating ChatGPT-4's performance as a digital health advisor for otosclerosis surgery. Front Surg. 2024. https://doi.org/10.3389/fsurg.2024.1373843.
10. 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. https://doi.org/10.1016/j.oooo.2025.01.001.
11. Araújo ALD, da Silva VM, Kudo MS, et al. Machine learning concepts applied to oral pathology and oral medicine: a convolutional neural networks' approach. J Oral Pathol Med. 2023;52(2):109–18. https://doi.org/10.1111/jop.13397.
12. Dave M, Patel N. Artificial intelligence in healthcare and education. Br Dent J. 2023;234(10):761–4. https://doi.org/10.1038/s41415-023-5845-2.
13. Tokgöz Kaplan T, Cankar M. Evidence-Based potential of generative artificial intelligence large Language models on dental avulsion: ChatGPT versus gemini. Dent Traumatol. 2025;41(2):178–86. https://doi.org/10.1111/edt.12999.
14. Zhang Q, Wu Z, Song J, Luo S, Chai Z. Comprehensiveness of large Language models in patient queries on gingival and endodontic health. Int Dent J. 2025;75(1):151–7. https://doi.org/10.1016/j.identj.2024.06.022.
15. Dursun D, Bilici Geçer R. Can artificial intelligence models serve as patient information consultants in orthodontics? BMC Med Inform Decis Mak. 2024. https://doi.org/10.1186/s12911-024-02619-8.
16. Uribe SE, Maldupa I, Kavadella A, et al. Artificial intelligence chatbots and large language models in dental education: worldwide survey of educators. Eur J Dent Educ. 2024. https://doi.org/10.1111/eje.13009.
17. Yilmaz BE, Gokkurt Yilmaz BN, Ozbey F. Artificial intelligence performance in answering multiple-choice oral pathology questions: a comparative analysis. BMC Oral Health. 2025. https://doi.org/10.1186/s12903-025-05926-2.
18. Arqub SA, Al-Moghrabi D, Allareddy V, Upadhyay M, Vaid N, Yadav S. Content analysis of AI-generated (ChatGPT) responses concerning orthodontic clear aligners. Angle Orthod. 2024;94(3):263–72. https://doi.org/10.2319/071123-484.1.
19. Uranbey Ö, Özbey F, Kaygısız Ö, Ayrancı F. Assessing chatgpt's diagnostic accuracy and therapeutic strategies in oral pathologies: A Cross-Sectional study. Cureus Published Online April. 2024;19. https://doi.org/10.7759/cureus.58607.
20. Macrì M, D'Albis V, D'Albis G, et al. The role and applications of artificial intelligence in dental implant planning: a systematic review. Bioengineering. 2024;11(8):778. https://doi.org/10.3390/bioengineering11080778.
21. Tam W, Huynh T, Tang A, Luong S, Khatri Y, Zhou W. Nursing education in the age of artificial intelligence powered Chatbots (AI-Chatbots): are we ready yet? Nurse Educ Today. 2023;129:105917. https://doi.org/10.1016/J.NEDT.2023.105917.
22. Lingen MW, Abt E, Agrawal N, et al. Evidence-based clinical practice guideline for the evaluation of potentially malignant disorders in the oral cavity: a report of the American Dental Association. J Am Dent Assoc. 2017;148(10):712-727.e10. https://doi.org/10.1016/j.adaj.2017.07.032.
23. 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. https://doi.org/10.1016/j.oooo.2024.04.010.
24. Kılınç DD, Mansız D. Examination of the reliability and readability of chatbot generative pretrained transformer's (ChatGPT) responses to questions about orthodontics and the evolution of these responses in an updated version. Am J Orthod Dentofacial Orthop. 2024;165(5):546–55. https://doi.org/10.1016/j.ajodo.2023.11.012.
25. Mago J, Sharma M. The potential usefulness of ChatGPT in oral and maxillofacial radiology. Cureus Published Online July. 2023;19. https://doi.org/10.7759/cureus.42133.
26. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the united States medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023;9:e45312. https://doi.org/10.2196/45312.
27. Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? J Stomatol Oral Maxillofac Surg. 2023. https://doi.org/10.1016/j.jormas.2023.101471.
28. Kim SG. Using ChatGPT for language editing in scientific articles. Maxillofac Plast Reconstr Surg. 2023;45(1):13. https://doi.org/10.1186/s40902-023-00381-x.

Publisher's Note