

Accuracy and Same-Day Response Consistency of AI Chatbots on Complete Denture Multiple-Choice Questions

Erhan DİLBER¹ , Umut Baran SÖNMEZ¹ , Ali Vasfi AĞLARCI² , Kübra YILDIZ DOMANIÇ³ 

Abstract

Aim To compare the accuracy and same-day response consistency of four free-tier AI chatbots on single-best-answer complete denture MCQs.

Material and method Twenty-five English MCQs (A–E, single correct option) on complete denture prosthodontics were used. Four chatbots were tested via official web interfaces under default free-tier settings: ChatGPT, Claude, Gemini, and Grok. Each question was asked in a new chat to ensure zero prior context, using a standardized instruction requiring the output of only one option letter (“Answer: <A/B/C/D/E>”).

Results Fleiss’ kappa showed significant within-day agreement for all chatbots ($p < 0.001$), indicating non-random temporal consistency (ChatGPT $\kappa = 0.625$; Claude $\kappa = 0.785$; Gemini $\kappa = 0.813$; Grok $\kappa = 0.693$). Overall correct answer rates were 71% for Gemini, 67% for ChatGPT, 63% for Claude, and 53% for Grok. Correct response rates did not differ across morning/noon/evening for ChatGPT ($p = 0.607$), Claude ($p = 0.779$), or Grok ($p = 0.846$), whereas Gemini showed a significant time-of-day effect ($p = 0.039$), with higher evening accuracy (80%) than morning (68%) and noon (64%).

Conclusion All four free-tier chatbots demonstrated significant same-day response consistency, with Gemini and Claude showing the highest agreement. Accuracy was numerically highest for Gemini and lowest for Grok, although between-tool differences were not statistically significant within the same time windows. These findings suggest that both accuracy and temporal stability should be considered when using free-tier chatbots for complete denture MCQ-based learning.

Keywords Accuracy, AI, Complete denture, Consistency, MCQ

Introduction

In recent years, AI chatbots and large language models (LLMs) have gained attention in medical and dental education because they can interpret complex prompts and generate natural-language responses that may support learning and clinical reasoning (1–4). Nevertheless, concerns such as hallucinations, outdated content, and variable accuracy emphasize the need for structured, domain-specific evaluation before these systems can be relied upon in specialty education (5, 6). From a patient-information and communication perspective, chatbots may improve understanding of prosthodontic treatments and potentially support informed consent when they provide accurate and comprehensible information (7–9). However, LLM outputs can appear convincing while containing factual errors, and confidently delivered misin-

formation remains a clinically relevant limitation (10, 11). Because MCQs are widely used to assess dental knowledge, several studies have evaluated chatbot performance using exam-style, text-only multiple-choice questions and reported measurable differences across models and settings (2, 12). Most benchmarking in dentistry has historically focused on ChatGPT, whereas evidence regarding other widely accessible platforms (e.g., Gemini and Claude) particularly under free-tier conditions has been comparatively limited, supporting the rationale for multi-model comparisons (10, 13, 14). Despite its limited representation in the dental literature, Grok was included in the present study because it is an emerging and widely accessible LLM-based platform that has attracted growing public and academic attention. Its performance in dentistry-specific benchmarking, however, remains insufficiently characterized, and including it alongside more commonly studied chatbots enabled a broader comparison under free-tier conditions.

In addition, model/version selection may influence outcomes, as exam-based evaluations have shown that newer versions can outperform older ones under comparable conditions (15). Importantly, reliability is not only a matter of single-time-point accuracy; response consistency across repeated administrations is a distinct attribute, as the same question may yield different outputs even under standardized conditions (11). Accordingly, prosthodontics research has operationalized consistency by repeating identical questions within a single day (morning/afternoon/evening) and quantifying agreement using metrics such as Cronbach’s alpha (16). Longitudinal repeated-measures designs have also been used to ex-

Correspondence: Kübra YILDIZ DOMANIÇ, kubra.domanic@atlas.edu.tr

¹ Alanya Alaaddin Keykubat University, Faculty of Dentistry, Department of Prosthodontics, Antalya, Türkiye

² Kastamonu University, Faculty of Medicine, Department of Biostatistics, Kastamonu, Türkiye

³ İstanbul Atlas University, Faculty of Dentistry, Department of Prosthodontics, İstanbul, Türkiye

Received: 16.03.2026 / Accepted: 20.03.2026 / Published: 30.04.2026

amine whether performance remains stable over time and across repeated sessions (17). Despite this growing literature, evidence remains limited regarding how free-tier chatbots perform specifically on complete denture domains when assessed using a standardized single-best-answer MCQ format while simultaneously capturing short-term temporal stability (5, 11, 14). Therefore, the aim of the present study is to evaluate and compare the accuracy and same-day response consistency of four widely used free-tier chatbot platforms (ChatGPT, Claude, Gemini, and Grok) when answering single-best-answer MCQs covering core complete denture prosthodontics topics.

The null hypotheses are that (i) accuracy will not differ among the evaluated chatbots and (ii) temporal consistency will not differ among the evaluated chatbots across the three-time windows.

Material and Methods

This experimental, cross-sectional study evaluated the accuracy and same-day response consistency of four AI-based chatbots when answering single-best-answer multiple-choice questions (MCQs) in complete denture prosthodontics. Ethical approval was not required because no human participants were involved and no personal data were collected. The scope and blueprint of the question set were based on core theoretical and clinical domains in complete denture therapy, guided by Çalikkocaoğlu’s textbook on conventional complete dentures (18). Within this framework, the authors drafted 25 MCQs covering key topics such as phonetics and speech, occlusal concepts, immediate dentures, the neutral zone technique, vertical dimension and wax rim principles, acrylic resin porosity, anatomical boundaries relevant to denture retention and border seal, denture hygiene and delivery/post-insertion evaluation, and impression-making concepts. Items were initially prepared and then translated into English to match the chatbot testing language. Each question consisted of one stem and five options (A–E) with only one correct answer. An answer key was generated by the authors, and the draft set was independently reviewed by two external prosthodontists (not involved in the study) for content coverage, clinical accuracy, language clarity, and appropriateness of the single-best-answer format. Ambiguous items or those potentially allowing more than one plausible correct option were revised until consensus was achieved on the final 25-question set. The final question set and answer key are provided as supplementary material.

Four free-tier chatbot platforms were evaluated via their official web interfaces under default settings: ChatGPT (<https://chatgpt.com/>), Claude (<https://claude.ai/>), Gemini (<https://gemini.google.com/>), and Grok (<https://grok.com/>). These platforms were accessed under default free-tier settings. Because explicit model/version labels were not consistently displayed in the user interfaces at the time of testing, the systems were identified and reported by their platform names rather than by underlying model versions. All platforms were accessed through a standard web browser without plug-ins, extensions, or additional custom system-level instructions. To prevent carryover effects and ensure zero prior context, a new chat session was initiated before each question. A standardized instruction prompt was appended immediately be-

fore each MCQ to constrain output format and enable objective scoring: “The following is a multiple-choice question with a single best answer among options A–E. Respond with only the single letter (A/B/C/D/E) corresponding to the option you believe is correct. Do not write the option text. Do not provide any explanation or additional text. Do not say ‘I am not sure’; select the best option. Required output (single line only): Answer: <A/B/C/D/E>.” Immediately after this instruction, the question stem and A–E options were entered. The complete 25-question set was administered three times on the same day within predefined time windows: morning (09:00–11:00), noon (13:00–15:00), and evening (17:00–19:00), yielding three responses per question per chatbot (75 responses per chatbot; 300 total responses). After each session, the selected option letter (A–E), question identifier, chatbot name, and time window were recorded in a Microsoft Excel file. For accuracy assessment, each response was compared with the answer key and coded as correct (1) or incorrect (0); accuracy was summarized as the proportion of correct answers per chatbot overall and within each time window. Differences in accuracy across time windows within each chatbot and differences across chatbots within each time window were evaluated using Cochran’s Q test ($\alpha=0.05$). Same-day response consistency was assessed using (i) full stability, defined as the percentage of questions for which the same option letter was selected across all three administrations (morning–noon–evening), and (ii) Fleiss’ kappa coefficients computed separately for each chatbot based on categorical A–E responses across the three repetitions. Kappa values were interpreted according to the Landis and Koch (19) classification as follows: $\kappa < 0.00$, poor; 0.00–0.20, slight; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, substantial; and 0.81–1.00, almost perfect agreement. All statistical analyses were performed using IBM SPSS Statistics for Windows (Version 31; IBM Corp., Armonk, NY, USA). A two-sided P value < 0.05 was considered statistically significant.

Results

Temporal consistency (agreement across morning–noon–evening) was significant for all four chatbots (Fleiss’ kappa, $p<0.001$ for each), indicating non-random within-day stability of categorical responses. Google Gemini showed the highest agreement ($\kappa=0.813$), followed by Anthropic Claude ($\kappa=0.785$), Grok ($\kappa=0.693$), and ChatGPT ($\kappa=0.625$) (Table 1).

Table 1: Same-day response agreement of AI chatbots across morning, noon, and evening sessions assessed using Fleiss’ kappa

	Fleiss kappa	z	p
ChatGPT	0.625	10.5	<0.001
Anthropic Claude	0.785	12.5	<0.001
Google Gemini	0.813	13.7	<0.001
Grok	0.693	10.6	<0.001

Accuracy distributions by time window and between-tool comparisons are presented in Table 2. Within-chatbot comparisons across morning, noon, and evening did not show significant differences for Anthropic Claude ($p=0.779$), Grok ($p=0.846$), or ChatGPT ($p=0.607$). In contrast, Google Gemini exhibited a significant time-of-day effect ($p=0.039$), with higher accuracy in the

evening session (80%) compared with morning (68%) and noon (64%).

Table 2: Distribution of correct and incorrect responses across time windows and Cochran's Q test results

	Anthropic Claude		Google Gemini		Grok		Chatgpt		Cochran's Q P Value
	False	True	False	True	False	True	False	True	
Morning	9 (%36)	16 (%64)	8 (%32)	17 (%68)	12 (%48)	13 (%52)	7 (%28)	18 (%72)	0.321
Noon	9 (%36)	16 (%64)	9 (%36)	16 (%64)	11 (%44)	14 (%56)	9 (%36)	16 (%64)	0.825
Evening	10 (%40)	15 (%60)	5 (%20)	20 (%80)	12 (%48)	13 (%52)	9 (%36)	16 (%64)	0.112
Total	28 (%37)	47 (%63)	22 (%29)	53 (%71)	35 (%47)	40 (%53)	25 (%33)	50 (%67)	
Cochran's Q p value	0.779		0.039		0.846		0.607		

When chatbots were compared within the same time window, no statistically significant differences were observed in the morning (p=0.321), noon (p=0.825), or evening (p=0.112). Overall accuracy was numerically highest for Google Gemini (71%), followed by ChatGPT (67%) and Anthropic Claude (63%), whereas Grok demonstrated the lowest overall accuracy (53%) (Figure 1).

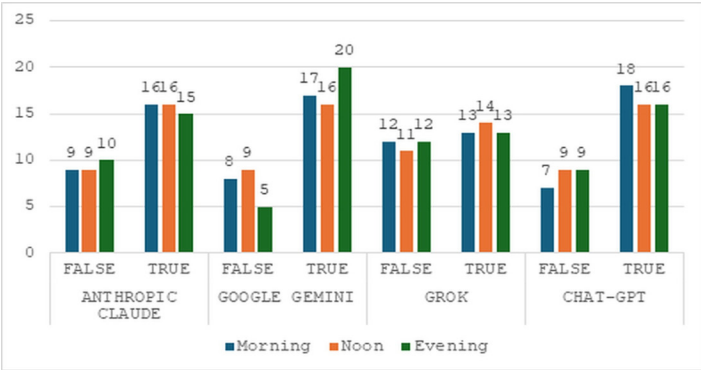


Figure 1: Distribution of correct and incorrect responses by chatbot and time window

Discussion

This experimental, cross-sectional benchmark study compared the accuracy and same-day short-term response consistency of four widely used free-tier AI chatbots (ChatGPT, Claude, Gemini, and Grok) on 25 single-best-answer multiple-choice questions (MCQs) covering core theoretical and clinical concepts in complete denture prosthodontics. Across repeated administrations within the same day (morning–noon–evening), all chatbots demonstrated statistically significant agreement in their categorical responses (Fleiss' κ : ChatGPT 0.625; Claude 0.785; Gemini 0.813; Grok 0.693; all $p < 0.001$), indicating that outputs were not randomly distributed across time. In contrast, when accuracy was compared between chatbots within the same time window, no statistically significant differences were observed for morning ($p = 0.321$), noon ($p = 0.825$), or evening ($p = 0.112$). Collectively, these findings support the null hypothesis that accuracy does not differ between chatbots within the same time window, while consistency

should be interpreted as generally high yet not identical across models, given the observed variability in κ magnitudes.

An important practical observation is that numerical differences in accuracy (Gemini highest; Grok lowest) did not translate into statistically significant between-chatbot differences within time windows. This pattern is common in MCQ-based benchmarking and may relate to the limited item count, heterogeneous item difficulty, and the nature of the single-best-answer format, which can make small performance differences difficult to detect with conventional hypothesis testing. Therefore, while descriptive accuracy differences may still be educationally meaningful, the present results suggest that larger item pools or higher-powered designs may be required to robustly discriminate performance among free tier chatbots.

Our findings are consistent with the MCQ-based intra-oral maxillofacial prosthodontics study by Ayata and Albayrak (20), which used a comparable repeated-measures protocol and likewise reported no significant differences in accuracy among the evaluated chatbots; that work also highlighted the importance of question identity (item difficulty) as a contributor to performance variance. In our study, Copilot was replaced by Grok, thereby expanding the comparative scope under free-tier conditions. However, the broader literature also demonstrates that model/version selection can substantially influence measured performance. For example, Salem et al. (15) reported significantly higher performance of ChatGPT-4.0 compared with ChatGPT-3.5 on an implant certification examination–derived question set, with ChatGPT-4.0 approaching dentist-level performance in that context. Likewise, in an implant-supported prostheses benchmark, Yilmaz et al. (21) found that ChatGPT-4.0 achieved higher accuracy than Gemini and that pre-prompting enhanced performance, with the most pronounced improvement observed for Gemini. Such discrepancies across studies likely reflect differences in access tier (free vs paid), prompting strategy (pre-prompt vs none), test language, and question format, each of which may meaningfully modulate observed outcomes.

Performance heterogeneity across prosthodontic subdomains has also been repeatedly emphasized. Freire et al. (5) reported that ChatGPT-4 could show relatively limited accuracy and weaker consistency for removable and tooth-supported fixed prosthodontic questions, suggesting that difficulty may shift depending on domain and cognitive demands. Likewise, Eraslan et al. (4) highlighted variability in chatbot performance across prosthodontic subtopics, underscoring that areas integrating biological and mechanical determinants may be more challenging for LLM-based systems. In line with this, Shirani (17) observed that although repeatability can be acceptable in repeated testing, overall accuracy may remain suboptimal and performance profiles may differ by chatbot. Together, these studies support a key interpretation of our results: high same-day stability can coexist with moderate accuracy, and performance is sensitive to both the prosthodontic subdomain and the assessment design.

Patient-question-oriented and rubric-based evaluations further contextualize the present findings. In implant dentistry, Taymour et al. (22) reported higher “reliability/usefulness” ratings for Gemini compared with ChatGPT versions, highlighting that some comparative conclusions depend on the endpoint used.

Similarly, in patient-question settings, Gemini has shown strong performance in some scenarios; however, these studies typically assess multidimensional outcomes (e.g., response quality, readability, and perceived usefulness) rather than MCQ accuracy alone (11, 23, 24). In contrast, broader prosthodontic commentaries emphasize that LLM outputs should not be considered clinically authoritative when used in isolation and that performance is sensitive to question type, context, and the chosen evaluation approach (25, 26). Our results align with these cautions: fluent and stable outputs do not guarantee correctness, and independent validation remains necessary, particularly for clinically nuanced topics.

A further contribution of this study is its inclusion of Claude and Grok alongside ChatGPT and Gemini. The current dental literature has largely concentrated on ChatGPT and Gemini, with fewer studies incorporating Claude, and even fewer directly evaluating Grok in dentistry-specific benchmarking. A multi-model educational comparison incorporated Claude (Pro); however, it was not conducted within a complete-denture-focused MCQ framework (4). Pediatric dentistry studies have included Claude variants (e.g., Claude 2/Claude-instant), but these are outside the complete denture domain (16). Regarding Grok, it has been mentioned among newer/emerging models in the literature; however, direct prosthodontic patient-question comparisons have generally been limited to ChatGPT and Gemini (24). Therefore, by reporting both accuracy and same-day consistency for Grok and Claude in a complete denture MCQ benchmark, our work extends multi-platform comparison under free-tier conditions and provides domain-specific baseline data for future studies.

Item-level analysis suggested that errors were not randomly distributed; instead, incorrect responses tended to cluster in specific topic areas. All chatbots failed the phonation/phonetics-related item (Question 6), which may reflect the need for fine-grained clinical distinctions in speech assessment and the high “distractor proximity” typical of single-best-answer formats. Increased error frequency was also observed in topics requiring multi-factor process knowledge (e.g., acrylic resin porosity; impression materials such as zinc oxide–eugenol; and impression techniques) and in questions dependent on mandibular anatomy, including an item with visual content. Such items require integrated anatomical–mechanical reasoning and, in the visual case, may additionally depend on multimodal capabilities that can be limited or inconsistently available in free-tier interfaces. Similarly, questions on factors improving impression tray retention involve practical, detail-rich knowledge where multiple options may appear plausible, increasing the likelihood of incorrect selection in single-best-answer MCQs. Importantly, the combination of high consistency and recurrent errors implies that stable outputs may also represent stable reproduction of incorrect choices—reinforcing the need to avoid equating repeatability with correctness in educational use.

The same-day, three-time-window design is a methodological strength because it captures short-term stability under standardized conditions; nevertheless, the observation of time-dependent variability—most notably the improvement in Gemini’s evening accuracy—suggests that performance may remain dynamic even within a single day. Potential contributors include service load, background model updates, and inherent probabilistic output generation. These considerations support repeated-measures ap-

proaches as more realistic than single-time-point testing for evaluating educational reliability in real-world usage settings.

This study has limitations. First, the evaluation was restricted to a 25-item single-best-answer MCQ set; thus, generalizability is constrained by item coverage and difficulty distribution. Second, testing was limited to free-tier web interfaces, and model labels, server-side updates, or access conditions may change over time, potentially affecting reproducibility across study periods. Third, constraining responses to a single-letter output improves scoring objectivity but prevents assessment of explanatory quality, reasoning depth, safety warnings, and citation behavior. Fourth, temporal consistency was assessed only within a single day; longer-term stability (days/weeks) and the effects of iterative model updates were not examined. Fifth, the limited number of visual items restricts conclusions regarding multimodal performance. Furthermore, the question set was developed by a single research team and was not externally validated using independent question banks from other institutions; therefore, its representativeness beyond the present academic context should be interpreted with caution. Future studies should (i) compare free-tier versus paid tiers (e.g., GPT-4-level access) and dentistry-adapted models; (ii) evaluate the effects of structured pre-prompts and prompt-engineering strategies; (iii) adopt longer longitudinal designs to capture stability over weeks/months; and (iv) directly compare free-tier and paid-tier chatbot versions using case-based, visual, and multimodal question sets to better reflect real-world educational and clinical scenarios, particularly for anatomically and visually demanding complete denture tasks. In addition to MCQ accuracy, systematic rubrics assessing response quality and safety—such as transparency, uncertainty communication, contraindication emphasis, and evidence referencing—may provide a more comprehensive framework for educational and clinical decision-support contexts.

Conclusion

Within the limits of this study, four free-tier AI chatbots demonstrated similar accuracy within the same time windows and generally high same-day response consistency on complete denture MCQs. However, the clustering of incorrect responses in clinically nuanced topics and difficulties in visually/anatomically demanding items suggest that these tools should not be used as stand-alone authorities in complete denture education. Instead, they are best positioned as adjunctive resources used alongside expert supervision and up-to-date evidence-based references.

Declarations

Ethics Committee Approval: Ethical approval was not required because this study did not involve human participants, biological materials, or identifiable personal data.

Informed Consent: There are no participants in the study

Peer Review: Externally peer-reviewed.

Author Contributions: Conception/Design of Study- E.D.; Data Acquisition- U.B.S.; Data Analysis/Interpretation- A.V.A.; Drafting

Manuscript- E.D., U.B.S., K.Y.D.; Critical Revision of Manuscript- E.D.; Final Approval and Accountability- E.D.; Material and Technical Support- A.V.A.; Supervision- E.D.

Conflict of Interest: Authors declared no conflict of interest.

Financial Disclosure: Authors declared no financial support.

REFERENCES

- Hanci, V., Ergun, B., Gul, S., Uzun, O., Erdemir, I., & Hanci, F. B. (2024). Assessment of readability, reliability, and quality of ChatGPT, Bard, Gemini, Copilot, and Perplexity responses on palliative care. *Medicine (Baltimore)*, 103(33), e39305.
- Revilla-Leon, M., Barmak, B. A., Sailer, I., Kois, J. C., & Att, W. (2024). Performance of an artificial intelligence-based chatbot (ChatGPT) answering the European certification in implant dentistry exam. *The International Journal of Prosthodontics*, 37(2), 221–224.
- Schwendicke, F., Samek, W., & Krois, J. (2020). Artificial intelligence in dentistry: Chances and challenges. *Journal of Dental Research*, 99(7), 769–774.
- Eraslan, R., Ayata, M., Yagci, F., & Albayrak, H. (2025). Exploring the potential of artificial intelligence chatbots in prosthodontics education. *BMC Medical Education*, 25(1), 321.
- Freire, Y., Santamaria Laorden, A., Orejas Perez, J., Gomez Sanchez, M., Diaz-Flores Garcia, V., & Suarez, A. (2024). ChatGPT performance in prosthodontics: Assessment of accuracy and repeatability in answer generation. *The Journal of Prosthetic Dentistry*, 131(4), 659.e1–659.e6.
- Suarez, A., Jimenez, J., Llorente de Pedro, M., Andreu-Vazquez, C., Diaz-Flores Garcia, V., Gomez Sanchez, M., et al. (2024). Beyond the scalpel: Assessing ChatGPT's potential as an auxiliary intelligent virtual assistant in oral surgery. *Computational and Structural Biotechnology Journal*, 24, 46–52.
- Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., et al. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*, 183(6), 589–596.
- Safi, Z., Abd-Alrazaq, A., Khalifa, M., & Househ, M. (2020). Technical aspects of developing chatbots for medical applications: Scoping review. *Journal of Medical Internet Research*, 22(12), e19127.
- Suarez, A., Diaz-Flores Garcia, V., Algar, J., Gomez Sanchez, M., Llorente de Pedro, M., & Freire, Y. (2024). Unveiling the ChatGPT phenomenon: Evaluating the consistency and accuracy of endodontic question answers. *International Endodontic Journal*, 57(1), 108–113.
- Danesh, A., Pazouki, H., Danesh, K., Danesh, F., & Danesh, A. (2023). The performance of artificial intelligence language models in board-style dental knowledge assessment: A preliminary study on ChatGPT. *Journal of the American Dental Association*, 154(11), 970–974.
- Gheisarifar, M., Shembesh, M., Koseoglu, M., Fang, Q., Afshari, F. S., Yuan, J. C., et al. (2025). Evaluating the validity and consistency of artificial intelligence chatbots in responding to patients' frequently asked questions in prosthodontics. *The Journal of Prosthetic Dentistry*, 134(1), 199–206.
- Warwas, F. B., & Heim, N. (2025). Performance of GPT-4 in oral and maxillofacial surgery board exams: Challenges in specialized questions. *Oral and Maxillofacial Surgery*, 29(1), 113.
- Ittarat, M., Cheungpasitporn, W., & Chansangpetch, S. (2023). Personalized care in eye health: Exploring opportunities, challenges, and the road ahead for chatbots. *Journal of Personalized Medicine*, 13(12).
- Tosun, B., & Yilmaz, Z. S. (2025). Comparison of artificial intelligence systems in answering prosthodontics questions from the dental specialty exam in Turkey. *Journal of Dental Sciences*, 20(3), 1454–1459.
- Salem, M., Karasan, D., Revilla-Leon, M., Barmak, A. B., & Sailer, I. (2025). Performance of artificial intelligence-based chatbots (ChatGPT-3.5 and ChatGPT-4.0) answering the International Team of Implantology exam questions. *Journal of Esthetic and Restorative Dentistry*, 37(11), 2412–2416.
- Rokhshad, R., Zhang, P., Mohammad-Rahimi, H., Pitchika, V., Entezari, N., & Schwendicke, F. (2024). Accuracy and consistency of chatbots versus clinicians for answering pediatric dentistry questions: A pilot study. *Journal of Dentistry*, 144, 104938.
- Shirani, M. (2025). Comparing the performance of ChatGPT-4o, DeepSeek R1, and Gemini 2 Pro in answering fixed prosthodontics questions over time. *The Journal of Prosthetic Dentistry*. Advance online publication.
- Çalikkocaoğlu, S. (2013). *Prosthetic treatment of edentulous patients: Conventional complete dentures (6th ed.)*. İstanbul, Türkiye: Quintessence Yayıncılık.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.
- Ayata, M., & Albayrak, H. (2025). İntraoral maksillofasiyal protez sorularında yapay zeka tabanlı sohbet robotlarının doğruluk ve tutarlılığının değerlendirilmesi. *Akdeniz Diş Hekimliği Dergisi*, 4(3), 204–211.
- Yılmaz, D., & Çolpak, E. D. (2025). ChatGPT vs. Google Gemini: Assessment of performance regarding the accuracy and repeatability of responses to questions in implant-supported prostheses. *European Annals of Dental Sciences*, 52(2), 71–78.
- Taymour, N., Fouda, S. M., Abdelrahman, H. H., & Hassan, M. G. (2025). Performance of ChatGPT-3.5, ChatGPT-4, and Google Gemini large language models in responding to dental implantology inquiries. *The Journal of Prosthetic Dentistry*, 134(6), 2427–2434.
- Esmailpour, H., Rasaie, V., Babaee Hemmati, Y., & Falahchai, M. (2025). Performance of artificial intelligence chatbots in responding to frequently asked questions of patients regarding dental prostheses. *BMC Oral Health*, 25(1), 574.
- Francis, L., S. D. V., Viswambharan, P., & Nair, V. V. (2025). Comparative evaluation of accuracy, completeness, and readability of common patient queries related to prosthodontic treatment by two artificial intelligence models (ChatGPT-4o and Gemini). *Cureus*, 17(12), e98458.
- Revilla-Leon, M., Gomez-Polo, M., Vyas, S., Barmak, A. B.,

-
- Gallucci, G. O., Att, W., et al. (2023). Artificial intelligence models for tooth-supported fixed and removable prosthodontics: A systematic review. *The Journal of Prosthetic Dentistry*, 129(2), 276–292.
26. Sadowsky, S. J. (2025). Can ChatGPT be trusted as a resource for a scholarly article on treatment planning implant-supported prostheses? *The Journal of Prosthetic Dentistry*, 134(2), 438–443.