Is Chatgpt a Reliable Tool For Prosthetic Dentistry?

Protetik Diş Hekimliğinde Chatgpt Güvenilir Bir Araç Mıdır?

İlknur USTA KUTLU[1], Arzu YILDIRIM[2], Işıl SARIKAYA[3]

ABSTRACT

This study aims to evaluate the information accuracy and reliability of ChatGPT's answers to frequently asked questions about prosthetic dentistry by patients and prosthodontists. A total of 40 questions (20 patient-level and 20 professional-level) were submitted to ChatGPT-based two academic prosthodontists and two clinical prosthodontists using a 5-point Likert scale (1 = very poor to 5 = very good). Inter-rater agreement was assessed using Fleiss's and Cohen's Kappa statistics. Differences between evaluator groups and question types were analyzed using Mann–Whitney U and Wilcoxon signed-rank tests. According to the mean score of all specialists, the responses to patient-level questions were significantly higher than the responses to professional-level questions ($p<0.001$).There was no significant difference $p>0.05$ between academics and clinicians in terms of patient-level questions. For professional-level questions, the mean scores evaluated by academics were significantly higher than those of clinicians. ($p<0.01$). ChatGPT indicates promising reliability in answering common patient researches and prosthodontics questions. However, for advanced clinical questions, its performance is limited and should be supplemented by expert consultation. ChatGPT can be a useful tool for patient guidance but is not yet a substitute for professional expertise in prosthodontics.

Keyword: Artificial intelligence, Chatgpt, Prosthodontics, Reliability

ÖZ

Bu çalışmanın amacı, hastalar ve protez uzmanları tarafından protez diş hekimliği hakkında sıkça sorulan sorulara ChatGPT'nin verdiği yanıtların bilgi doğruluğunu ve güvenilirliğini değerlendirmektir. ChatGPT-3.5'e toplam 40 soru (20 hasta-seviyesi ve 20 profesyonel soru) yöneltilmiştir. Yanıtlar, iki akademisyen protetik diş tedavisi uzmanı ve iki klinisyen protetik diş tedavisi uzmanı tarafından 5 puanlı Likert ölçeği (1 = çok kötü - 5 = çok iyi) kullanılarak bağımsız olarak değerlendirilmiştir. Değerlendiriciler arası uyum, Fleiss ve Cohen Kappa istatistikleri kullanılarak değerlendirilmiştir. Değerlendirici grupları ve soru tipleri arasındaki farklar, Mann-Whitney U ve Wilcoxon işaretli sıra testleri kullanılarak analiz edilmiştir. Tüm uzmanların ortalama puanlarına göre, hasta sorularına verilen yanıtlar, profesyonel sorularına verilen yanıtlardan anlamlı derecede yüksekti ($p<0.001$). Hasta odaklı sorular açısından akademisyenler ve klinisyenler arasında anlamlı bir fark yoktu ($p>0.05$). Profesyonel sorularda, akademisyenler tarafından değerlendirilen ortalama puanlar, klinisyenlerin puanlarından anlamlı derecede yüksekti ($p<0.01$). ChatGPT, yaygın hasta araştırmaları ve protetik diş hekimliği sorularını yanıtlamada umut verici bir güvenilirlik göstermektedir. Ancak, ileri düzey klinik sorular için performansı sınırlıdır ve uzman konsültasyonu ile desteklenmelidir. ChatGPT, hasta eğitimi için faydalı bir araç olabilir, ancak henüz protez diş hekimliğinde profesyonel uzmanlığın yerini tutamaz.

Anahtar Kelimeler: ChatGPT, Güvenilirlik, Protez Diş Hekimliği, Yapay zeka.

Highlights

- * ChatGPT shows higher reliability for patient-level prosthodontic queries.
- * Professional-level answers demonstrate limited accuracy and consistency.
- * Evaluations may differ between academic and clinical prosthodontists.

The study was presented as an oral presentation at the 48th Annual European Prosthodontic Association (EPA) Congress / 27th Scientific TPID Congress

[1] Assist. Prof. İlknur USTA KUTLU, Department of Prosthodontics, Faculty of Dentistry, Tokat Gaziosmanpaşa University, Tokat, Turkey, e-mail: dt.kutlusta_i@hotmail.com ORCID:0000-0001-7532-7954
[2] Research Assistant Arzu YILDIRIM, Department of Prosthodontics, Faculty of Dentistry, Tokat Gaziosmanpaşa University, Tokat, Turkey, e-mail: arzuyildiriim@icloud.com ORCID: 0009-0009-3363-9381
[3] Prof. Dr. Işıl SARIKAYA, Department of Prosthodontics, Faculty of Dentistry, Tokat Gaziosmanpaşa University, Tokat, Turkey, e-mail: sarikayaisil@gmail.com ORCID: 0000-0002-2172-4724

Corresponding Author: İlknur USTA KUTLU
e-mail: dt.kutlusta_i@hotmail.com

Geliş Tarihi / Received: 25.09.2025
Kabul Tarihi/Accepted: 14.12.2025

INTRODUCTION

Artificial intelligence (AI)-powered chatbots are conversational systems that interact with users through written, spoken, or visual modalities, providing human-like communication (1). In recent years, advances in natural language processing (NLP) have enabled large language models (LLMs) to be effectively utilized in healthcare domains such as information access, patient education, and clinical decision support system (2-4). In this context, ChatGPT, developed by OpenAI, has attracted attention for its ability to generate rapid, meaningful, and contextually appropriate responses to natural language queries (5-7). AI has the potential to analyze vast amounts of data, thereby offering more efficient and effective treatments, while providing a promising future in fields such as diagnosis, treatment planning, and student education (8). Moreover, it has been reported to assist clinicians by providing more predictable diagnostic and therapeutic outcomes (9). ChatGPT and similar AI tools are also gaining increasing interest in dentistry (8,10-12). They have been applied across multiple specialties, including prosthodontics (1,11), implantology (13-15), oral and maxillofacial surgery (16), periodontology (17), pediatric dentistry (18), restorative dentistry (19), endodontics (8,20), oral radiology (21,22), and orthodontics (7). These systems are reported to provide general information about treatments, answer patient queries, support clinical decision-making, and facilitate educational processes (4,6,15,23). With the increasing use of chatbots in healthcare, the accuracy and reliability of responses generated by large language models such as ChatGPT to patient-level questions have become a critical area of research (24).

Despite their widespread use in healthcare, the performance of ChatGPT in different fields of dentistry varies considerably across the literature. Several studies have suggested that the model performs better in patient-level questions but tends to lack depth or precision when addressing technical or professional queries (16,3) and reported low accuracy rates for responses regarding fixed and removable prostheses. Similarly, Sadowsky (25) highlighted ChatGPT's inconsistencies in scientific content generation, including unreliable referencing and factual inaccuracies. On the other hand, favorable outcomes have been reported for patient communication and education. Esmailpour et al. (5) emphasized that ChatGPT provided highly readable responses to frequently asked patient questions regarding dental prostheses, with good inter-rater agreement. Likewise, Praveen et al. (6) found that ChatGPT's responses to public health–related oral health questions were informative, clear, and associated with strong agreement among evaluators. These findings support the potential of ChatGPT as a patient-level educational tool.

Conversely, the accuracy and reliability of responses to technical and academic-level questions remain unclear. Previous studies have reported limited validity (15) and underperformance in prosthodontics examinations, particularly in case-based scenarios (26). Variations in the findings of existing studies in this field underscore the necessity for additional investigations.

Moreover, evaluator agreement has rarely been addressed or thoroughly reported in the literature. A considerable proportion of existing research has been conducted exclusively with examination-type questions or limited to patient perspectives. For example, Tosun and Yılmaz (26) focused solely on DUS (Dental Specialty Exam) questions to evaluate AI systems. Although Praveen et al. (6) included independent assessments by public health dentistry specialists, studies assessing multi-rater agreement at the professional level remain scarce. These limitations underscore the need for more comprehensive investigations evaluating ChatGPT's knowledge generation capacity, clarity, adequacy, and reliability in healthcare contexts.

Therefore, the present study aimed to comprehensively assess the information quality of responses generated by ChatGPT-

3.5 in the field of prosthodontics for both patient- and professional-level queries. In addition, the study sought to compare the ratings provided by academicians and clinicians, while also analyzing inter-rater agreement to gain further insights into the consistency and reliability of the model's outputs.

The null hypothesis of the study was that ChatGPT-3.5 responses would be generally informative and acceptable for use, with patient-level queries expected to receive higher scores than professional-level ones.

Furthermore, it was anticipated that clinicians would rate the model's responses more favorably compared to academicians, reflecting potential differences in evaluation perspectives.

MATERIALS AND METHODS

Study Design and Ethical Considerations

This research was designed as a descriptive, cross-sectional study. Since only publicly available digital data were used and no biological material or personal data were collected from individuals, approval from an institutional ethics committee was not required. The study was conducted in accordance with the ethical principles of the Declaration of Helsinki.

Question Development and Data Collection

The reliability of information generated by the AI-based language model ChatGPT-3.5 (OpenAI, 2022) was evaluated specifically within the scope of prosthodontics, a specialty where accurate and evidence-based knowledge is critical for both patient care and clinical decision-making. For this purpose, a total of 40 questions were prepared, consisting of 20 patient-level questions that reflected common concerns of individuals seeking dental treatment and 20 professional-level questions designed to address technical and clinical issues relevant to prosthodontists.

Patient-level questions were compiled from a variety of sources, including online patient information platforms, health-related websites, and commonly observed user search trends, in order to reflect real concerns frequently encountered in everyday practice (27,28,29,30). Professional-level questions, on the other hand, were formulated based on clinical practice experience and domain expertise, ensuring that they addressed issues relevant to decision-making and technical knowledge. None of the questions were categorized under specific subtopics, as the intention was to design them in an open-ended and content-focused manner, allowing for broader evaluation of the responses.

Data were collected on April 29, 2025, using a newly created ChatGPT account accessed via https://chat.openai.com. Prior to data collection, cookies and browsing history were cleared, and a new chat window was opened for each question. Only the first response generated by ChatGPT was included in the analysis; no additional responses were requested.

Evaluation Process

ChatGPT responses were independently evaluated by four experienced prosthodontists (two academicians and two clinicians), who were blinded to each other's scores to prevent assessment bias. Each response was rated using a 5-point Likert scale, based on accuracy, adequacy, and clinical relevance:

1- Very poor: The response is scientifically incorrect or misleading.
2- Poor: The response includes relevant content but contains major inaccuracies or lacks essential detail.
3- Acceptable: The response is partially correct but incomplete or lacks sufficient explanation.
4- Good: The response is accurate and mostly complete, with minor deficiencies.
5- Very good: The response is comprehensive, scientifically accurate, and clinically relevant.

Statistical Analysis

Data analysis was performed using IBM SPSS Statistics 22 (IBM Corporation, USA), which allowed for both descriptive and inferential statistical evaluations. Descriptive statistics included the calculation of mean, standard deviation, median, minimum, and maximum values to summarize the distribution of the data. Inter-rater agreement across all evaluators was analyzed using Fleiss's Kappa, providing an overall measure of consistency, whereas pairwise agreement between academicians and clinicians was specifically assessed using Cohen's Kappa. For comparisons between groups, the non-parametric Mann–Whitney U test was employed, as the data did not meet the assumptions of normal distribution. A significance level of p<0.05 was considered statistically meaningful.

The strength of inter-rater agreement was interpreted according to the commonly accepted classification of Cohen's Kappa values: (31)

< 0.20 = very poor/minimal agreement,

0.21–0.40 = weak agreement,

0.41–0.60 = moderate agreement,

0.61–0.80 = good agreement, and

0.81–1.00 = very good/excellent agreement

RESULTS

A total of 40 ChatGPT-generated responses, comprising 20 patient-level and 20 professional-level queries, were subjected to independent evaluation by four prosthodontists, including two academicians and two practicing clinicians. Each response was rated using a standardized 5-point Likert scale, with the evaluation criteria focusing primarily on the accuracy of the information provided and the adequacy of the content in addressing the query.

When all responses were analyzed regardless of question type or evaluator, the overall mean score was 3.92, indicating that the model generally delivered good-level information. Considering question types separately, the mean score assigned to patient-level questions (4.34 ± 0.71) was significantly higher than that of professional-level questions (3.50 ± 1.15) (p<0.001) (Table 1). ChatGPT performs better in patient information delivery.

Table 1. Comparison of Mean Scores Given by All Raters for Patient level and Professional level Queries

	Patient level queries					Professional level queries				
	Mean	Standard Deviation	Median	Minimum	Maximum	Mean	Standard Deviation	Median	Minimum	Maximum
Score	4.34	.71	4.00	3.00	5.00	3.50	1.15	4.00	1.00	5.00
Sig. *P*										**<0.001***

(Mann–Whitney U test results; n=20 N=160)

When evaluator groups (academicians vs. clinicians) were compared independent of question type, academicians had a mean score of 4.07 ± 0.96, whereas clinicians had 3.76 ± 1.09 (p>0.05) (Table 2). When the question types are taken into consideration, academicians scored patient questions with a mean of 4.33 ± 0.66 and professional questions with 3.83 ± 1.15. Clinicians scored patient questions with a mean of 4.35 ± 0.77 but professional questions significantly lower at 3.18 ± 1.06 (Table 2). Thus, while no difference was observed between groups in patient-level questions (p>0.05) academicians rated professional-level responses significantly higher than clinicians did (p<0.05). Clinicians evaluated ChatGPT more critically in professional contexts.

Table 2. Comparison of Mean Scores Given by Academicians and Clinicians for Patient level and Professional level Queries

Academician						Clinician					
	Mean	Standard Deviation	Median	Minimum	Maximum	Mean	Standard Deviation	Median	Minimum	Maximum	*p*
Patient Level queries	4.33	.66	4.00	3.00	5.00	4.35	.77	5.00	3.00	5.00	0.876
Professional level queries	3.83	1.15	4.00	1.00	5.00	3.18	1.06	3.00	1.00	5.00	**0.010***
Overall queries	4.07	.96	4.00	1.00	5.00	3.76	1.09	4.00	1.00	5.00	0.057

(Mann–Whitney U test results; n=20 N=160)

Kappa statistics revealed that inter-rater agreement for patient-level questions was minimal among academicians (Cohen's Kappa = 0.209; Table 3), weak among clinicians (Cohen's Kappa = 0.365; Table 3), and weak overall among all evaluators (Fleiss's Kappa = 0.229). For professional-level questions, there was no agreement between academicians (Cohen's Kappa = 0.064; Table 4), minimal agreement among clinicians (Cohen's Kappa = 0.192; Table 4), and weak agreement across all evaluators (Fleiss's Kappa = 0.171; Table 5).

Table 3. Intra-rater Kappa Agreement: Academicians versus Clinicians (Patient level Queries)

Academician					Clinician			
	Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance	Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement Kappa	**.209**	.160	1.420	.156	**.365**	.150	2.456	.014
N of Valid Cases	20				20			

(Kappa agreement interpretation: <0.20 = very poor/minimal agreement; 0.21–0.40 = weak agreement; 0.41–0.60 = moderate agreement; 0.61–0.80 = good agreement; 0.81–1.00 = very good/excellent agreement.)

DISCUSSION

The rapid advancement of artificial intelligence (AI) technologies, particularly those based on large language models, has opened new opportunities in healthcare for information access, patient communication, and clinical support applications (21,32). Tools such as ChatGPT are increasingly being recognized as practical, fast, and accessible sources of information for both healthcare professionals and patients (12,21,33). In this context, the present study evaluated the information quality of ChatGPT-3.5's responses to both patient-level and professional-level questions in prosthodontics. The results revealed that patient-level responses were rated significantly higher than professional-level responses. Although no difference was observed between evaluator groups when question type was disregarded, academicians tended to assign higher scores to professional-level responses compared to clinicians. Therefore, the hypothesis of the study was partially rejected.

There are notable differences in how ChatGPT has been evaluated in the literature. Some studies have focused solely on accuracy rates (2) whereas others have assessed responses exclusively from the patient's perspective (5). By contrast, the present study encompassed both patient and professional-level questions, included evaluators with different levels of expertise, and provided a bidirectional analysis.

Our findings are consistent with those reported in previous research. Balel (16) demonstrated that ChatGPT's responses to patient-level questions in oral and maxillofacial surgery were significantly higher rated than technical ones. Similarly, Bayraktar et al. (18) reported that responses directed to children and their parents received higher scores compared to academic questions. Likewise, in another study evaluating chatbot responses regarding the All-on-Four implant concept, patient-level responses were rated significantly higher than dentist-centered responses (13). The higher ratings and stronger inter-rater agreement observed for patient-level questions suggest that ChatGPT performs better in the context of patient communication. This observation aligns with Esmailpour et al. (5) , who found that ChatGPT's answers to frequently asked prosthodontics-related patient questions were largely accurate and informative. Additionally, Praveen et al. (6) reported that ChatGPT's responses to public health-related oral health questions were clear, understandable, and satisfactory in terms of content, with strong inter-rater agreement. Similarly, in the present study, patient-centered responses achieved higher scores and demonstrated stronger evaluator consistency.

While this pattern aligns with existing literature, the lower scores for professional-level responses require interpretation. Expert prosthodontic questions demand precise terminology, evidence-based justification, and reference to clinical protocols. However, ChatGPT-3.5 often provides generalized statements without citing sources or detailing procedural nuances. This simplification of complex concepts likely reduced perceived reliability among evaluators, particularly academicians. In contrast, patient-level queries primarily require clear and accessible communication.

In line with previous literature, these findings reinforce the idea that AI-powered chatbots should be regarded primarily as supportive tools in healthcare, requiring human oversight in domains that demand specialized expertise (34,35). The development of "ChatGPT-Academic" or similar systems, integrated with scientific databases and capable of source verification, is therefore recommended.

Interestingly, although no overall difference was observed between evaluator groups, academicians assigned higher scores to professional-level responses compared to clinicians. The inter-rater agreement data support this finding. According to Kappa values, clinicians demonstrated greater consistency than academicians in evaluating both patient and professional-level responses. The unexpectedly higher ratings given by academicians for professional-level responses may be explained by the lower consistency observed among academicians themselves compared to clinicians (Cohen's Kappa 0.064–0.192; Table 4).

Table 4. Intra-rater Kappa Agreement: Academicians versus Clinicians (Professional level Questions)

Academician					Clinician			
	Value	Asymptotic Standard Error^a	Approximate T^b	Approximate Significance	Value	Asymptotic Standard Error^a	Approximate T^b	Approximate Significance
Measure of Agreement Kappa	**.064**	.133	.548	.583	**.192**	.127	1.642	.101
N of Valid Cases	20				20			

(Kappa agreement interpretation: <0.20 = very poor/minimal agreement; 0.21–0.40 = weak agreement; 0.41–0.60 = moderate agreement; 0.61–0.80 = good agreement; 0.81–1.00 = very good/excellent agreement.)

Moreover, for professional questions, Fleiss's Kappa across all evaluators was only 0.171, indicating very low agreement (Table 5). These results highlight that subjective judgments play a significant role in evaluating ChatGPT's responses, leading to considerable variability among evaluators. This underscores the need for future studies to include larger numbers of evaluators in order to achieve more reliable assessments

Table 5. Fleiss' Kappa Agreement of All Raters for Patient level Questions and Professional level Questions

Patient level queries							Professional level queries					
	Asymptotic				Asymptotic 95% Confidence Interval		Asymptotic				Asymptotic 95% Confidence Interval	
	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
Overall Agreement	**.229**	.070	3.279	.001	.092	.366	**.171**	.050	3.403	.001	.073	.270

(Kappa agreement interpretation: <0.20 = very poor/minimal agreement; 0.21–0.40 = weak agreement; 0.41–0.60 = moderate agreement; 0.61–0.80 = good agreement; 0.81–1.00 = very good/excellent agreement.)

Freire et al. (3) reported that, despite being a newer version, ChatGPT-4 achieved only a 25.6% accuracy rate in prosthodontics-related questions. Similarly, Dashti et al. (2) found that GPT-3.5 achieved 57.9% accuracy in U.S. prosthodontics board examinations, while GPT-4 reached 73.6%; however, even GPT-4 achieved only a modest inter-rater Kappa value of 0.39. In academic contexts, not only accuracy but also reliability and reproducibility are of great importance. Sadowsky (25) raised concerns about ChatGPT's reliability in academic content generation, citing issues such as reference fabrication, lack of source attribution, and content inconsistency. Collectively, these findings indicate that even advanced versions of ChatGPT face notable limitations in terms of information integrity and reliability. Despite these shortcomings, the present study employed ChatGPT-3.5, as it remains a more accessible and freely available option for users.

Several limitations of the present study should be acknowledged. ChatGPT-3.5 does not provide references for its responses, and the accuracy of its content was not independently verified. Furthermore, evaluations were based solely on written text and did not assess the applicability of responses in clinical practice. The focus on a single AI model without comparison to other systems, the subjective nature of evaluator scoring, and the potential lag behind rapidly evolving AI technology all represent additional limitations.

Future research should evaluate AI systems' applicability in medical and dental contexts using larger and more diverse datasets, advanced model versions, and standardized multicenter question sets. Verification of sources for content validity, assessment of impacts across different patient populations, and domain-specific evaluations will also be essential to ensure more comprehensive and robust conclusions (17,36).

CONCLUSION AND RECOMMENDATIONS

Based on the results of the present study, in which patient-level questions were rated higher than professional-level questions, it can be concluded that ChatGPT-3.5 performs better in patient education and communication, while remaining limited in technical and Professional domains. Although inter-rater agreement among evaluators influenced the data, the evaluator's background—whether academician or clinician prosthodontist—did not have a statistically significant effect on the evaluation scores.

REFERENCES

1. Gheisarifar M, Shembesh M, Koseoglu M, Fang Q, Afshari FS, Yuan JC-C, et al. Evaluating the validity and consistency of artificial intelligence chatbots in responding to patients' frequently asked questions in prosthodontics. J Prosthet Dent. 2025; 134(1):199-206. https://doi: 10.1016/j.prosdent.2025.03.009.

2. Dashti M, Khosraviani F, Azimi T, Hefzi D, Ghasemi S, Fahimipour A, et al. Assessing ChatGPT-4's performance on the

US prosthodontic exam: impact of fine-tuning and contextual prompting vs. base knowledge, a cross-sectional study. BMC Med Educ. 2025; 25(1):761. https://doi: 10.1186/s12909-025-07371-9.

3. Freire Mancebo Y, Santamaría Laorden A, Orejas Pérez J, Gómez Sánchez M, Díaz-Flores García V, Suárez García A. ChatGPT performance in prosthodontics: Assessment of accuracy and repeatability in answer generation. 2024; 131(4):659.e1-659.e6. https://doi: 10.1016/j.prosdent.2024.01.018.

4. Jin HK, Lee HE, Kim E. Performance of ChatGPT-3.5 and GPT-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: a systematic review and meta-analysis. BMC Med Educ. 2024; 24(1):1013. https://doi: 10.1186/s12909-024-05944-8.

5. Esmailpour H, Rasaie V, Babaee Hemmati Y, Falahchai M. Performance of artificial intelligence chatbots in responding to the frequently asked questions of patients regarding dental prostheses. BMC Oral Health. 2025; 25(1):574. https://doi: 10.1186/s12903-025-05965-9.

6. Praveen G, Poornima U, Akkaloori A, Bharathi V: ChatGPT as a Tool for Oral Health Education. A Systematic Evaluation of ChatGPT Responses to Patients' Oral Health-related Queries. J Nat Sci Med. 2024; 7(3):154-157. https://doi: 10.4103/jnsm.jnsm_208_23.

7. 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):263-272. https://doi: 10.2319/071123-484.1.

8. Arılı Öztürk E, Turan Gökduman C, Çanakçi BC. Evaluation of the Performance of ChatGPT-4 and ChatGPT-4o as a Learning Tool in Endodontics. Int Endod J. 2025;00:1-13 https.//doi: 10.1111/iej.14217.

9. Ji K, Wu Z, Han J, Zhai G, Liu J. Evaluating ChatGPT-4's performance on oral and maxillofacial queries: Chain of Thought and standard method. Front Oral Health. 2025; 6:1541976. https://doi: 10.3389/froh.2025.1541976. eCollection 2025.

10. Giannakopoulos K, Kavadella A, Aaqel Salim A, Stamatopoulos V, Kaklamanos EG. Evaluation of the performance of generative AI large language models ChatGPT, Google Bard, and Microsoft Bing Chat in supporting evidence-based dentistry: comparative mixed methods study. J Med Internet Res. 2023; 25:e51580. https://doi: 10.2196/51580.

11. Schwendicke F, Rahimi HM, Tichy A. Artificial intelligence in prosthodontics. Dental Clinics 2025; 69(2):315-326. https://doi: 10.1016/j.cden.2024.11.009.

12. Buldur M, Sezer B. Evaluating the accuracy of Chat Generative Pre-trained Transformer version 4 (ChatGPT-4) responses to United States Food and Drug Administration (FDA) frequently asked questions about dental amalgam. BMC Oral Health. 2024; 24(1):605. https.//doi: 10.1186/s12903-024-04358-8.

13. Akpınar H. Comparison of responses from different artificial intelligence-powered chatbots regarding the All-on-four dental implant concept. BMC Oral Health. 2025; 25(1):922. https://doi: 10.1186/s12903-025-06294-7.

14. Aseri AA. Exploring the Role of Artificial Intelligence in Dental Implantology: A Scholarly Review. J Pharm Bioallied Sci. 2025; 17(Suppl 1):S102-S104. https://doi: 10.4103/jpbs.jpbs_442_25.

15. Binaljadm TM, Alqutaibi AY, Halboub E, Zafar MS, Saker S. Artificial Intelligence Chatbots as Sources of Implant Dentistry Information for the Public: Validity and Reliability Assessment. Eur J Dent. 2025; https://doi: 10.1055/s-0045-1809155.

16. Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? J Stomatol Oral Maxillofac Surg. 2023; 124(5):101471. https:// doi: 10.1016/j.jormas.2023.101471.

17. Chatzopoulos GS, Koidou VP, Tsalikis L, Kaklamanos EG. Evaluation of Large Language Model Performance in Answering Clinical Questions on Periodontal Furcation Defect Management. Dentistry Journal. 2025; 13(6):271. https://doi: 10.3390/dj13060271.

18. Bayraktar Nahir C: Can ChatGPT be guide in pediatric dentistry? BMC Oral Health. 2025; 25(1):9. https://doi: 10.1186/s12903-024-05393-1.

19. Ozdemir ZM, Yapici E. Evaluating the Accuracy, Reliability, Consistency, and Readability of Different Large Language Models in Restorative Dentistry. J J Esthet Restor Dent. 2025; 37(7):1740-1752. https://doi: 10.1111/jerd.13447.

20. Durmazpinar PM, Ekmekci E. Comparing diagnostic skills in endodontic cases: dental students versus ChatGPT-4o. BMC Oral Health. 2025; 25(1):1-8. https://doi: 10.1186/s12903-025-05857-y.

21. Mago J, Sharma M. The potential usefulness of ChatGPT in oral and maxillofacial radiology. Cureus. 2023; 15(7) :e42133. https://doi: 10.7759/cureus.42133.

22. Tassoker M. ChatGPT-4 Omni's superiority in answering multiple-choice oral radiology questions. BMC Oral Health. 2025; 25(1):173. https://doi: 10.1186/s12903-025-05554-w.

23. Ali K, Barhom N, Tamimi F, Duggal M. ChatGPT—A double-edged sword for healthcare education? Implications for assessments of dental students. Eur J Dent Educ. 2024; 28(1):206-211. https://doi: 10.1111/eje.12937.

24. Çitir M. ChatGPT and oral cancer: a study on informational reliability. BMC Oral Health. 2025; 25(1):86. https://doi: 10.1186/s12903-025-05479-4.

25. Sadowsky SJ. Can ChatGPT be trusted as a resource for a scholarly article on treatment planning implant-supported prostheses? J Prosthet Dent. 2025; 134(2):438-443. https://doi: 10.1016/j.prosdent.2025.03.025.

26. Tosun B, Yilmaz ZS. Comparison of artificial intelligence systems in answering prosthodontics questions from the dental specialty exam in Turkey. J Dent Sci. 2025; 20(3):1454-1459. https://doi: 10.1016/j.jds.2025.01.025.

27. Memorial Health Group. Dental prostheses treatment methods. Internet. 2025. https://www.memorial.com.tr/tedavi-yontemleri/dis-protezleri

28. Yeditepe University Dental Hospital. Frequently asked questions about dental prostheses. Internet. 2025. https://www.yeditepedishastanesi.com/sss?c=23

29. PlusDent Dental Clinic. Frequently asked questions about dental prostheses. Internet. 2025. https://www.plusdent.com.tr/protez-sik-sorulan-sorular/

30. Metco Dental. Frequently asked questions about dental prostheses. Internet. 2025. https://metcodental.com/blog/agiz-ve-dis-sagligi/protez-dis-hakkinda-sik-sorulan-sorular/

31. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–174. https://doi.org/10.2307/2529310

32. Alhaidry HM, Fatani B, Alrayes JO, Almana AM, Alfhaed NK, Alhaidry H, et al. ChatGPT in dentistry: a comprehensive review. Cureus. 2023; 15(4). https://doi: 10.7759/cureus.38317.

33. Topdağı B, Kavaz T. Assessment of information quality in contemporary artificial intelligence systems for digital smile design: A comparative analysis. J Prosthet Dent. 2025; 6:S0022-3913(25)00556-6. https://doi: 10.1016/j.prosdent.2025.06.030.

34. Alsayed AA, Aldajani MB, Aljohani MH, Alamri H, Alwadi MA, Alshammari BZ, et al. Assessing the quality of AI information from ChatGPT regarding oral surgery, preventive dentistry, and oral cancer: An exploration study. Saudi Dent J. 2024; 36(11):1483-1489. https://doi: 10.1016/j.sdentj.2024.09.009.

35. Sismanoglu S, Capan BS. Performance of artificial intelligence on Turkish dental specialization exam: can ChatGPT-4.0 and

gemini advanced achieve comparable results to humans? BMC Med Educ. 2025; 25(1):214. https://doi: 10.1186/s12909-024-06389-9.

36. Eraslan R, Ayata M, Yagci F, Albayrak H. Exploring the potential of artificial intelligence chatbots in prosthodontics education. BMC Med Educ. 2025; 25(1):321. https://doi: 10.1186/s12909-025-06849-w.