

Comparative Evaluation of Large Language Models in Prosthodontic Complication Scenarios

Protetik Komplikasyon Senaryolarında Büyük Dil Modellerinin Karşılaştırmalı Değerlendirilmesi

Rüştü Ersoy SAKARYA¹, Nuran YANIKOĞLU¹, Osayd ALAWAWDA¹

¹Atatürk Üniversitesi, Diş Hekimliği Fakültesi, Protetik Diş Tedavisi Anabilim Dalı, Erzurum, Türkiye

Abstract

Aim: This study aims to evaluate and compare the performance of 4 LLMs in responding to common prosthodontic complication scenarios.

Methods: An expert panel design was employed to assess the quality of responses generated by the four LLMs across five clinically relevant prosthodontic scenarios. Each scenario was entered 15 times into each model, yielding 300 responses in total. Seven prosthodontists independently evaluated the anonymized responses using four criteria, clinical accuracy, applicability, clarity, and comprehensiveness, on a 5-point Likert scale. Descriptive statistics, one-way ANOVA, and post-hoc pairwise comparisons were performed to detect inter-model differences. The internal consistency (reliability) of evaluators' ratings for each model was assessed using Cronbach's alpha.

Results: The models differed significantly on all evaluation criteria ($p<.05$). ChatGPT achieved the highest mean scores for clinical accuracy (3.89), clarity (3.90), and comprehensiveness (4.09), while Gemini showed the highest applicability (3.88). Cronbach's alpha coefficients indicated strong reliability across all domains (0.776–0.880), with ChatGPT showing the highest single-domain coefficient ($\alpha=0.820$ –0.880).

Conclusion: Among the evaluated models, ChatGPT and Gemini generally received higher ratings on specific criteria in pairwise comparisons; however, the four models should not be interpreted as forming a definitive overall ranking. While LLMs show potential as supportive tools in prosthodontic education and clinical decision-making, their outputs should not replace expert judgment.

Key Words: Prosthodontics, artificial intelligence, clinical decision support systems, complications, dental prostheses

Özet

Amaç: Bu çalışma, yaygın protez komplikasyon senaryolarına yanıt vermede 4 Büyük Dil Modeli'nin (BDM) performansını değerlendirmeyi ve karşılaştırmayı amaçlamaktadır.

Yöntem: Dört BDM tarafından üretilen yanıtların kalitesini beş klinik açıdan önemli protez senaryosunda değerlendirmek için bir uzman panel tasarımı kullanıldı. Her senaryo her modele 15 kez girildi ve toplamda 300 yanıt elde edildi. 7 protez uzmanı, anonimleştirilmiş yanıtları 5 puanlık Likert ölçeği kullanarak dört kritere göre bağımsız olarak değerlendirmiştir: klinik doğruluk, uygulanabilirlik, açıklık ve kapsamlılık. Modeller arası farklılıkları tespit etmek için tanımlayıcı istatistikler, tek yönlü ANOVA ve post-hoc ikili karşılaştırmalar yapılmıştır. Her bir model için değerlendiricilerin verdiği puanların iç tutarlılığı (güvenilirliği), Cronbach alfa katsayısı kullanılarak değerlendirildi.

Bulgular: Tüm değerlendirme kriterleri için modeller arasında anlamlı farklılıklar bulundu ($p<0,05$). ChatGPT, klinik doğruluk (3,89), açıklık (3,90) ve kapsamlılık (4,09) için en yüksek ortalama puanları elde ederken, Gemini en yüksek uygulanabilirliği (3,88) gösterdi. Cronbach alfa katsayıları, tüm alanlarda güçlü güvenilirliği (0,776–0,880) gösterirken, ChatGPT en yüksek iç tutarlılığı ($\alpha=0,820$ –0,880) sergiledi.

Sonuç: Değerlendirilen modeller arasında ChatGPT ve Gemini, ikili karşılaştırmalarda belirli kriterler bakımından genellikle daha yüksek puanlar aldı; ancak bu dört modelin kesin bir genel sıralama oluşturduğu şeklinde yorumlanmamalıdır. Büyük Dil Modelleri, protez diş hekimliği eğitiminde ve klinik karar verme süreçlerinde destekleyici araçlar olarak potansiyel göstergeler de, bu araçların sunduğu sonuçlar uzman görüşünün yerini almamalıdır.

Anahtar Kelimeler: Prostodonti, yapay zekâ, klinik karar verme destek sistemleri, komplikasyonlar, diş protezi

1. Giriş / Introduction

Large Language Models (LLMs) are a form of artificial intelligence (AI) designed to replicate human language processing capabilities. These models leverage deep learning algorithms and are trained on vast datasets to understand and generate coherent, human-like text across diverse applications such as machine translation, summarization, and interactive dialogue systems (1). As they become more sophisticated, LLMs can produce highly contextually relevant and linguistically fluent outputs, making them valuable tools in professional domains, including healthcare (1).

LLM-powered chatbots have emerged as innovative tools that deliver information and facilitate dialogue on a wide range of subjects, including medical and dental care (2). These systems can enhance patient understanding by providing information on treatment options, potential complications, prognoses, and expected outcomes (2). As AI technologies evolve, their integration into healthcare systems has accelerated, offering support in areas such as diagnostic decision-making, clinical risk assessment, and workflow efficiency (3). Notably, while many AI chat models operate using pretrained datasets, others offer real-time internet access to generate responses. However, despite their growing use, questions remain about the clinical relevance, accuracy, and readability of the medical content these systems produce, particularly when patients access it (4). Qualified professionals must verify such AI-generated health-related content, as it can influence treatment adherence, patient satisfaction, and the quality of doctor–patient communication (5).

The progression of LLM-based chatbots—marked by self-supervised learning and training on extensive textual corpora—has enabled them to closely mimic natural human interaction (6,7). Their capacity to deliver detailed and domain-specific responses renders them increasingly beneficial in clinical and educational contexts. Moreover, many of these models can be fine-tuned with human feedback to optimize performance for specialized applications, enhancing reliability and clinical utility (8).

In recent years, AI-driven chat interfaces have surged in adoption, especially among patients seeking

immediate, user-friendly access to healthcare information (9). However, the uncontrolled use of such tools without clinical validation presents risks that necessitate systematic evaluation. This is particularly relevant in fields such as prosthodontics, where accurate diagnosis and treatment planning require expert-level decision-making.

This study aimed to assess and compare the performance and reliability of 4 widely used LLMs in responding to common prosthodontic complication scenarios. The null hypothesis was that there would be no statistically significant differences in expert evaluation scores among the four LLMs on any assessed criterion.

2. Yöntem / Method

Study Design and LLM Query Protocol

This study used an expert panel design to assess the quality of clinical responses generated by 4 publicly available LLMs when presented with prosthodontic complication scenarios. To mitigate evaluator bias, the evaluation process was conducted under blinded conditions. The LLMs selected for this study were ChatGPT (GPT-4o, OpenAI), Gemini (Gemini 2.5, Google), Bing Copilot (GPT-4, Microsoft), and DeepSeek (DeepSeek-V3, DeepSeek AI).

All queries were conducted between 03 March 2026 and 10 March 2026. On each day, every scenario was queried three times (morning, afternoon, and evening), so that in five days, each scenario had been entered 15 times into each model, yielding 300 responses in total. Each query was entered as the first message of a new, separate conversation (no prior conversation history) using the standard consumer web interface for each platform (ChatGPT: chat.openai.com; Gemini: gemini.google.com; Copilot: copilot.microsoft.com; DeepSeek: chat.deepseek.com), with default settings and, where applicable, real-time internet browsing/search features disabled. The five scenarios were presented in a fixed, identical order across all repetitions and models, using the exact prompts reproduced verbatim in Table 1, with no follow-up or clarifying questions. Five realistic prosthodontic complication scenarios were developed based on current scientific literature and clinical relevance (10,11). These scenarios were designed to reflect frequently encountered challenges

in prosthodontic practice, such as denture stomatitis and implant-related complications. Three independent prosthodontic specialists (R.E.S., N.Y., and O.A.) reviewed and approved the scenarios for content validity. Table 1 presents the clinical scenarios and corresponding prompts.

Table 1. Clinical prosthodontic scenarios and corresponding questions used in the evaluation

Scenario No.	Clinical Scenario Description	Question Asked to the LLMs
1	A 65-year-old patient wearing a complete maxillary denture complains of redness and a burning sensation in the mucosa beneath the denture. The patient wears it during sleep.	What is the most likely diagnosis in this case, and what are the appropriate treatment options?
2	A newly fabricated complete maxillary denture was reported to fall during speech and lacks adequate retention.	What clinical and laboratory steps can be taken to improve retention in this case?
3	A patient using a clasp-retained removable partial denture (RPD) has developed mobility and tenderness in the abutment tooth.	How should this RPD-related complication be evaluated and managed?
4	A patient with an implant-supported fixed prosthesis reports a fracture in the prosthesis and pain around the implant site.	What complications should be considered in this situation, and how should the treatment be planned?
5	A mandibular complete denture frequently displaces and causes pain during mastication.	What techniques or modifications are recommended to enhance the stability of mandibular complete dentures?

Each of the five scenarios was entered into the four LLMs, generating a total of 300 responses (5 scenarios × 4 models × 15 repetitions). All generated responses were anonymized and randomly coded as Model A, B, C, or D to maintain evaluator blinding. Seven experienced prosthodontists served as independent evaluators.

Each evaluator assessed the AI-generated responses according to four predefined criteria: clinical accuracy, applicability in clinical practice, clarity and comprehensibility, and comprehensiveness. Each criterion was scored using a 5-point Likert scale, adapted from the Global Quality Score (GQS) (12,13). A score of 5 indicated a fully correct and comprehensive answer; 4 represented a correct response with minor inaccuracies or missing details; 3 denoted a partially correct answer with substantial missing or irrelevant information; 2 indicated a mostly incorrect response containing limited correct elements; and 1 corresponded to a completely incorrect or irrelevant answer. The four domains were operationally defined as follows: clinical accuracy referred to the factual and diagnostic correctness of the response relative to current prosthodontic evidence and guidelines; applicability referred to the extent to which the recommended management could

be feasibly implemented in routine clinical practice; clarity and comprehensibility referred to how clearly and unambiguously the response was expressed for the intended reader; and comprehensiveness referred to the degree to which the response addressed all clinically relevant aspects of the scenario (diagnosis, etiology, and management) rather than a partial subset.

Descriptive statistics, including mean and standard deviation, were calculated for each evaluation criterion across all responses. One-way analysis of variance (ANOVA) was used to compare scores among the models. For each of the four LLMs, Cronbach’s alpha was calculated across the four evaluation domains (clinical accuracy, applicability, clarity, and comprehensiveness) to assess the internal consistency of evaluators’ ratings — that is, the extent to which the seven evaluators rated a given model’s responses in a consistent manner across domains. The resulting coefficients were interpreted according to Gwet’s benchmark classification. In this classification, alpha values below 0.0 indicated poor reliability; values between 0.0 and 0.2 indicated slight reliability; values between 0.2 and 0.4 indicated fair reliability; values between 0.4 and 0.6 indicated moderate reliability; values between 0.6 and 0.8 were considered substantial; and values between 0.8 and

1.0 reflected almost perfect reliability. All statistical analyses were performed using IBM SPSS Statistics (version 27; IBM Corp., Armonk, NY, USA).

Ethical approval of the study was received from the Committee for Ethics Research at the University of Atatürk (Date: 25/02/2026, Number: 15).

3. Bulgular/Results

Table 2 presents descriptive statistics for evaluator ratings across the four LLMs. Overall, mean scores

for all evaluation criteria ranged from 3.70 to 4.09, indicating moderate-to-high performance across models. Among the models, ChatGPT (GPT-4o) generally achieved the highest mean scores for clinical accuracy (3.89), clarity (3.90), and comprehensiveness (4.09), whereas Gemini obtained the highest mean score for applicability (3.88). Across all criteria, Copilot and DeepSeek performed similarly, with mean values slightly below ChatGPT and Gemini; the pairwise comparisons below explore these differences further.

Table 2. Descriptive statistics for evaluator scores across the four LLMs

Evaluation Criterion	Model	Mean	SD
Clinical Accuracy	DeepSeek	3.69	0.99
	Gemini	3.77	0.82
	ChatGPT	3.89	0.89
	Copilot	3.81	0.97
	Total	3.79	0.92
Applicability	DeepSeek	3.70	1.04
	Gemini	3.88	0.88
	ChatGPT	3.70	0.82
	Copilot	3.81	0.92
	Total	3.77	0.92
Clarity	DeepSeek	3.82	1.17
	Gemini	3.56	1.09
	ChatGPT	3.90	1.11
	Copilot	3.63	1.05
	Total	3.73	1.11
Comprehensiveness	DeepSeek	3.89	1.05
	Gemini	3.95	0.86
	ChatGPT	4.09	0.93
	Copilot	3.87	0.87
	Total	3.95	0.93

SD: Standard Deviation

One-way analysis of variance (ANOVA) revealed significant differences among the four LLMs for all evaluation criteria ($p < .05$). Post-hoc pairwise comparisons (Table 3) identified specific differences between models. For clinical accuracy, the comparison indicated that ChatGPT produced significantly more accurate responses than DeepSeek ($p = .002$). No other pairwise differences were statistically significant. For applicability, Gemini achieved significantly higher scores than both DeepSeek ($p = .013$) and ChatGPT

($p = .008$). Regarding clarity, multiple significant differences were observed. DeepSeek produced clearer responses than Gemini ($p = .001$) and Copilot ($p = .027$), whereas ChatGPT outperformed both Gemini ($p < .001$) and Copilot ($p < .001$). In terms of comprehensiveness, ChatGPT provided significantly more detailed and complete answers than DeepSeek ($p = .003$) and Copilot ($p = .001$).

**Table 3. Post-hoc pairwise comparisons (mean differences) among models for each evaluation criterion**

Dependent Variable	Model (I)	Model (J)	Mean Difference (I-J)	Std. Err.	Sign.
Clinical Accuracy	DeepSeek	Gemini	-0.086	0.057	0.431
		ChatGPT	-0.200*	0.057	0.002
		Copilot	-0.124	0.057	0.128
	Gemini	ChatGPT	-0.114	0.057	0.183
		Copilot	-0.038	0.057	0.908
		ChatGPT	0.076	0.057	0.535
Applicability	DeepSeek	Gemini	-0.171*	0.057	0.013
		ChatGPT	0.010	0.057	0.998
		Copilot	-0.105	0.057	0.251
	Gemini	ChatGPT	0.181*	0.057	0.008
		Copilot	0.067	0.057	0.642
		ChatGPT	-0.114	0.057	0.182
Clarity	DeepSeek	Gemini	0.257*	0.068	0.001
		ChatGPT	-0.086	0.068	0.592
		Copilot	0.190*	0.068	0.027
	Gemini	ChatGPT	-0.343*	0.068	<0.001
		Copilot	-0.067	0.068	0.763
		ChatGPT	0.276*	0.068	<0.001
Comprehensiveness	DeepSeek	Gemini	-0.067	0.057	0.649
		ChatGPT	-0.200*	0.057	0.003
		Copilot	0.020	0.057	0.985
	Gemini	ChatGPT	-0.133	0.057	0.092
		Copilot	0.087	0.057	0.428
		ChatGPT	0.220*	0.057	0.001

* Indicates significant difference between the groups ($p < .05$). Std. Err= Standard Error, Sign= Significance

The internal consistency of the evaluation domains was assessed using Cronbach's alpha coefficients for each artificial intelligence model (Table 4). All models showed acceptable to excellent reliability, with α values ranging between 0.776 and 0.880 across the four domains (Table 4). ChatGPT's coefficients ranged from 0.820 to 0.880, Gemini's from 0.795 to

0.860, Copilot's from 0.800 to 0.864, and DeepSeek's from 0.776 to 0.842. Because these ranges overlap considerably, the models should not be interpreted as forming a strict overall ranking of reliability; ChatGPT recorded the highest domain-level coefficient (Applicability, $\alpha=0.880$), while DeepSeek recorded the lowest (Clinical Accuracy, $\alpha=0.776$).

Table 4. Cronbach's Alpha coefficients for each AI model across the evaluation domains

Model	Clinical Accuracy	Applicability	Clarity	Comprehensiveness
DeepSeek	0.776	0.842	0.793	0.825
Gemini	0.825	0.860	0.834	0.795
ChatGPT	0.820	0.880	0.875	0.873
Copilot	0.800	0.864	0.820	0.808

4. Tartışma/Discussion

The present study aimed to evaluate and compare the performance of 4 LLMs in responding to common prosthodontic complication scenarios. The analysis focused on 4 key domains: clinical accuracy, practical applicability, clarity, and comprehensiveness. Based on the statistical outcomes, significant differences were found among the models for all evaluation criteria ($p < .05$); therefore, the null hypothesis was rejected. ChatGPT achieved significantly higher scores than DeepSeek in clinical accuracy and comprehensiveness, and significantly higher scores than both Gemini and Copilot in clarity, while Gemini achieved significantly higher applicability scores than both DeepSeek and ChatGPT. Because several pairwise comparisons were not statistically significant (Table 3), these findings do not support a definitive overall ranking of the four models.

Previous studies have highlighted the inconsistent accuracy of chatbot responses to frequently asked questions in prosthodontics, emphasizing the importance of cautious clinical application (12). Some comparative studies have shown that Gemini and Copilot are more readable than ChatGPT 3.5 and ChatGPT 4, particularly in contexts such as refractive surgery. Among these, Gemini provided the most relevant and readable content (14). Similarly, despite its widespread use, ChatGPT 3.5 occasionally produced misleading or inaccurate answers in cases involving traumatic dental injuries, whereas ChatGPT 4.0 and Google Gemini delivered more accurate and comprehensive responses (15).

Beyond dentistry-specific findings, other studies evaluating chatbot responses across various medical domains have consistently reported challenges in readability, with many responses requiring comprehension levels equivalent to a high school or college education (16–18). Furthermore, current evidence indicates that ChatGPT still has limited capability in addressing issues related to removable partial dentures and tooth-supported fixed dental prostheses, reaffirming that such tools cannot substitute for professional dental expertise. Accordingly, dental professionals utilizing these tools must remain fully cognizant of their limitations (19). Given the evolving nature of AI technologies,

continued research and iterative development are essential to unlock their full potential within dental healthcare. While LLMs offer considerable advantages, it is equally important to address the associated risks. Implementing effective quality control frameworks is crucial for monitoring and validating the accuracy of AI-generated medical content. Furthermore, strict ethical and regulatory standards must govern the deployment of LLMs in dental settings to ensure patient safety and uphold professional integrity (19). To ensure reliable LLM outputs, training datasets must be current, comprehensive, and scientifically vetted. Only access to validated and up-to-date data enables LLMs to provide balanced, evidence-based responses while minimizing the risk of misinformation (20,21). Additionally, clinical validation and continuous model refinement are critical prerequisites for safely and effectively integrating LLMs into dental practice. Dental professionals should remain aware of these tools' capabilities and limitations when incorporating them into clinical or educational workflows (22).

In general, LLMs continue to have several inherent limitations, including issues related to accuracy, information recency, coherence, transparency and interpretability, and ethical considerations (23). This study also presents certain limitations that should be acknowledged. First, it was restricted to five predefined prosthodontic scenarios, which may not fully capture the breadth and complexity of real-world clinical situations. Second, the evaluation focused exclusively on textual responses and did not account for interactive or iterative dialogue, which more closely reflects typical user engagement with chatbot systems. Third, only English-language outputs were analyzed, potentially limiting the generalizability of the findings to non-English-speaking populations. Finally, the models were evaluated using standard consumer web interfaces with real-time search disabled; these interfaces do not permit direct control of generation parameters such as temperature or top-p. Accordingly, although repeated sampling across 15 iterations was used to mitigate and characterize output variability, direct parameter-level control of model randomness was not possible.

Future research should aim to broaden the scope of evaluated clinical cases to encompass a wider variety of prosthodontic conditions. Moreover, assessing LLM



performance in multilingual and interactive contexts, and within authentic clinical or educational settings, could yield more comprehensive insights into their reliability, adaptability, and pedagogical value. Studies conducted through API-based environments could additionally incorporate controlled manipulation of generation parameters, such as temperature and top-p, to systematically investigate their effects on response variability, consistency, and clinical performance.

5. Sonuç/Conclusion

Based on the findings of this study, the following conclusions were drawn:

1. This study demonstrated statistically significant differences among LLMs in specific, pairwise-compared aspects of their responses to prosthodontic complication scenarios.
2. In the pairwise comparisons that reached statistical significance, ChatGPT outperformed one or more of the other models in accuracy, clarity, and comprehensiveness, while Gemini outperformed DeepSeek and ChatGPT in applicability; these results should be interpreted as domain-specific pairwise findings rather than a definitive overall ranking of the four models.
3. Despite these encouraging results, none of the models achieved perfect consistency or accuracy, underscoring the continuing limitations of current AI systems in specialized dental applications.

Kaynaklar / References

1. Deiana G, Dettori M, Arghittu A, Azara A, Gabutti G, Castiglia P. Artificial Intelligence and Public Health: Evaluating ChatGPT Responses to Vaccination Myths and Misconceptions. *Vaccines*. 2023 Jul 7;11(7):1217. doi: 10.3390/vaccines11071217.
2. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med*. 2023 Jun 1;183(6):589–596. doi: 10.1001/jamainternmed.2023.1838.
3. Muthukrishnan N, Maleki F, Ovens K, Reinhold C, Forghani B, Forghani R. Brief History of Artificial Intelligence. *Neuroimaging Clin N Am*. 2020 Nov;30(4):393–399. doi: 10.1016/j.nic.2020.07.004.
4. Behers BJ, Vargas IA, Behers BM, Rosario MA, Wojtas CN, Deevers AC, et al. Assessing the Readability of Patient Education Materials on Cardiac Catheterization From Artificial Intelligence Chatbots: An Observational Cross-Sectional Study. *Cureus*. 2024 Jul 4;16(7):e63865. doi:10.7759/cureus.63865.
5. ACAR AH. Can natural language processing serve as a consultant in oral surgery? *J Stomatol Oral Maxillofac Surg*. 2024 Jun;125(3):101724. doi: 10.1016/j.jormas.2023.101724.
6. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PL, et al. Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank. *Neurosurgery*. 2023 Nov 1;93(5):1090–1098. doi: 10.1227/neu.0000000000002551.
7. Bernstein IA, Zhang Y, Govil D, Majid I, Chang RT, Sun Y, et al. Comparison of Ophthalmologist and Large Language Model Chatbot Responses to Online Patient Eye Care Questions. *JAMA Netw Open*. 2023 Aug 1;6(8):e2330320. doi: 10.1001/jamanetworkopen.2023.30320.
8. Liu J, Wang C, Liu S. Utility of ChatGPT in Clinical Practice. *J Med Internet Res*. 2023 Jun 28;25:e48568. doi: 10.2196/48568.
9. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of Artificial Intelligence Chatbot Responses to Top Searched Queries about Cancer. *JAMA Oncol*. 2023 Oct 1;9(10):1437–1440. doi: 10.1001/jamaoncol.2023.2947.
10. Felton D, Cooper L, Duqum I, Minsley G, Guckes A, Haug S, et al. Evidence-Based Guidelines for the Care and Maintenance of Complete Dentures: A Publication of the American College of Prosthodontists. *J Prosthodont*. 2011 Feb;20 Suppl 1:S1-S12. doi: 10.1111/j.1532-849X.2010.00683.x.
11. Tallarico M, Lee SY, Cho YJ, Noh KT, Chikahiro O, Aguirre F, et al. Prosthetic Guidelines to Prevent Implant Fracture and Peri-Implantitis: A Consensus Statement from the Osstem Implant Community. *Prosthesis*. 2025 Jun 16;7(3):65. https://doi.org/10.3390/prosthesis7030065
12. Gheisarifar M, Shembesh M, Koseoglu M, Fang Q, Afshari FS, Yuan JCC, et al. Evaluating the validity and consistency of artificial intelligence chatbots in responding to patients' frequently asked questions in prosthodontics. *J Prosthet Dent*. 2025 Jul;134(1):199–206. doi: 10.1016/j.prosdent.2025.03.009.
13. Mohammad-Rahimi H, Ourang SAH, Pourhoseingholi MA, Dianat O, Dummer PMH, Nosrat A. Validity and reliability of artificial intelligence chatbots as public sources of information on endodontics. *Int Endod J*. 2024 Mar;57(3):305–314. doi: 10.1111/iej.14014.
14. Aydın FO, Aksoy BK, Ceylan A, Akbaş YB, Ermiş S, Kepez Yıldız B, et al. Readability and Appropriateness of Responses Generated by ChatGPT 3.5, ChatGPT 4.0, Gemini, and Microsoft Copilot for FAQs in Refractive Surgery. *Türk J Ophthalmol*. 2024 Dec 31;54(6):313–317. doi: 10.4274/tjo.galenos.2024.28234.



15. Guven Y, Ozdemir OT, Kavan MY. Performance of Artificial Intelligence Chatbots in Responding to Patient Queries Related to Traumatic Dental Injuries: A Comparative Study. *Dent Traumatol*. 2025 Jun;41(3):338-347. doi: 10.1111/edt.13020.
16. Shen SA, Perez-Heydrich CA, Xie DX, Nellis JC. ChatGPT vs. web search for patient questions: what does ChatGPT do better? *Eur Arch Otorhinolaryngol*. 2024 Jun;281(6):3219-3225. doi: 10.1007/s00405-024-08524-0.
17. Ghanem YK, Rouhi AD, Al-Houssan A, Saleh Z, Moccia MC, Joshi H, et al. Dr. Google to Dr. ChatGPT: assessing the content and quality of artificial intelligence-generated medical information on appendicitis. *Surg Endosc*. 2024 May;38(5):2887-2893. doi: 10.1007/s00464-024-10739-5.
18. Onder CE, Koc G, Gokbulut P, Taskaldiran I, Kuskonmaz SM. Evaluation of the reliability and readability of ChatGPT-4 responses regarding hypothyroidism during pregnancy. *Sci Rep*. 2024 Jan 2;14(1):243. doi: 10.1038/s41598-023-50884-w.
19. Freire Y, Santamaría Laorden A, Orejas Pérez J, Gómez Sánchez M, Díaz-Flores García V, Suárez A. ChatGPT performance in prosthodontics: Assessment of accuracy and repeatability in answer generation. *J Prosthet Dent*. 2024 Apr;131(4):659.e1-659.e6. doi: 10.1016/j.prosdent.2024.01.018.
20. Patcas R, Bornstein MM, Schätzle MA, Timofte R. Artificial intelligence in medico-dental diagnostics of the face: a narrative review of opportunities and challenges. *Clin Oral Investig*. 2022 Dec;26(12):6871-6879. doi: 10.1007/s00784-022-04724-2.
21. Zhu JJ, Jiang J, Yang M, Ren ZJ. ChatGPT and Environmental Research. *Environ Sci Technol*. 2023 Nov 21;57(46):17667-17670. doi: 10.1021/acs.est.3c01818.
22. Giannakopoulos K, Kavadella A, Salim AA, 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 Dec 28;25:e51580. doi: 10.2196/51580.
23. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023 Aug;29(8):1930-1940. doi: 10.1038/s41591-023-02448-8.