Journal of Dental Sciences and Education

Original Article

DOI: 10.51271/JDSE-0082

Comparative evaluation of AI-based chatbots in answering prosthodontic questions from the Turkish Dental Specialty Exam (DUS)[a]

Bülent Kadir Tartuk

Department of Prosthodontics, Faculty of Dentistry, Bingöl University, Bingöl, Turkiye

Cite this article as: Tartuk BK. Comparative evaluation of AI-based chatbots in answering prosthodontic questions from the Turkish Dental Specialty Exam (DUS). *J Dent Sci Educ.* 2026;4(3):84-90. doi:10.51271/JDSE-0082

Received: 01.06.2026 ◆ **Accepted**: 30.07.2026 ◆ **Published**: 20.08.2026

ABSTRACT

Aims: This study compared the performance of the ChatGPT, Gemini, Copilot, and DeepSeek large language model-based chatbots in answering prosthodontic questions from the Turkish Dental Specialty Exam (DUS).

Methods: A total of 126 prosthodontic questions from DUS examinations conducted between 2012 and 2021 were included in the study after excluding visually based and canceled questions. The questions were classified as knowledge-based and case-based. Additionally, they were categorized according to topic areas as fixed prosthodontics, removable prosthodontics and implant-supported prosthodontics. The data were analyzed using the generalized estimating equations model, which accounts for repeated measurements, and differences between models were evaluated using Bonferroni-corrected post hoc tests.

Results: All chatbot models included in this study achieved an overall accuracy rate above 80% (ChatGPT-4o: 85.71%; Copilot: 83.33%; Gemini 3 Flash: 81.75%; DeepSeek-V3.2: 80.16%). No statistically significant difference was observed among the models in terms of overall performance ($p=0.619$). All models demonstrated significantly higher performance on knowledge-based questions than on case-based questions ($p<0.05$). Topic-based analysis showed numerically lower accuracy for implant-supported prosthodontic questions than for fixed and removable prosthodontic questions. However, this finding should be interpreted cautiously because the implant-supported subgroup included only four questions. No significant relationship was found between examination years and the accuracy rates of the models ($\chi^2=7.1$; $p=0.62$).

Conclusion: The examined Artificial Intelligence models demonstrated a high level of accuracy in answering prosthodontic specialty questions. These systems have the potential to serve as supportive tools in dental education and examination preparation processes. However, owing to their limitations in questions requiring clinical reasoning, particularly case-based questions, their use should be approached cautiously and critically.

Keywords: Artificial Intelligence, large language models, prosthodontics, DUS, dental education

[a]This study was presented orally at the '4th National İstanbul Aydın University Faculty of Dentistry Young Dentists and Student Congress'.

INTRODUCTION

Artificial Intelligence (AI) refers to a field of technology aimed at developing computer systems and algorithms capable of performing tasks that typically require human intelligence.[1] AI applications based on methods such as deep learning and neural networks have led to substantial transformations in numerous fields, including healthcare.[2] In recent years, these technologies have enabled major advances in healthcare services by facilitating rapid data analysis, early error detection, and support for clinical decision-making processes, thereby contributing to improved workflow efficiency, increased productivity, and reduced error rates in the healthcare sector.[3,4]

The rapid advancement of AI in the fields of health sciences and education has also significantly transformed access to information and academic learning methods.[5] One of the most notable outcomes of this transformation has been the emergence of large language models (LLMs), advanced AI systems capable of mimicking human language comprehension, interpretation, and generation.[6] These models are trained on large and heterogeneous text corpora using deep learning algorithms and are capable of producing highly human-like, contextually relevant outputs.[7,8]

LLM-based conversational systems such as ChatGPT, Gemini, Copilot, and DeepSeek, which represent some of the most widely accessible applications of this technology today, are capable of successfully performing complex tasks including text generation, summarization, translation, and question answering.[9] Nevertheless, LLMs may differ in terms of architecture, training data sources, updating mechanisms, and multimodal capabilities, which can lead to variations in accuracy and response quality across disciplines and question types.[10] Therefore, particularly in fields requiring specialized expertise such as medicine and dentistry, the accuracy and contextual reliability of the content generated by these models should be evaluated carefully and systematically.[11,12]

The integration of AI-powered chatbots into dental education and clinical practice may only be feasible to the extent

Corresponding Author: Bülent Kadir Tartuk, bktartuk@bingol.edu.tr

that these systems can generate sufficiently accurate and reliable responses within relevant areas of expertise.[2] In this context, standardized examinations requiring high levels of cognitive competence provide an important benchmark for evaluating the performance of AI systems. In Turkiye, the Dental Specialty Exam (DUS) is a nationally administered, high-stakes examination used as the primary criterion for admission to postgraduate dental specialty training programs. The examination consists of multiple-choice questions designed to assess dentists' professional knowledge in both clinical dentistry and basic sciences.[2,13] Evaluating the performance of AI-based chatbots on this examination may provide valuable insight into the potential applications of these systems in dentistry.[14]

In recent years, the number of studies evaluating the performance of AI-based LLMs through the DUS and similar examinations has increased considerably. However, a substantial proportion of these studies have focused on specific specialty fields, limited question sets, or selected model comparisons. Therefore, further comparative evaluations of different models within specific dental disciplines are needed.

Prosthodontics is a complex dental specialty that encompasses biomechanical principles, materials science, occlusion, esthetic planning, and clinical decision-making processes.[15] Owing to its multidimensional nature, prosthodontics provides an appropriate testing domain for assessing the ability of LLMs to integrate theoretical knowledge with clinical reasoning. In particular, the limited number of studies investigating the comparative performance of different AI models in prosthodontics according to examination years, question types, and subtopic categories highlights the need for more comprehensive evaluations in this field.

Accordingly, this study aimed to evaluate and compare the performance of AI-powered chatbots, including ChatGPT, Gemini, Copilot, and DeepSeek, in answering prosthodontic questions from previously administered DUS examinations. The null hypotheses tested were as follows: no significant difference among the models in terms of overall accuracy rates (H_{01}), no performance difference between knowledge-based and case-based questions (H_{02}), and similar accuracy rates across removable, fixed, and implant-supported prosthodontic questions (H_{03}).

METHODS

Ethics

Ethical approval was not required for this study because all data were obtained from publicly available examination questions published on the official website of the Measurement, Selection and Placement Center (ÖSYM). Informed consent was also not required, because the study did not involve human participants, personal data, or clinical records. All procedures were carried out in accordance with the ethical rules and the principles.

Data Source and Eligibility Criteria

The DUS is a nationwide centralized examination consisting of 120 multiple-choice questions, including 40 basic sciences and 80 clinical sciences questions, 10 of which are related to prosthodontics. First administered in 2012, the DUS was conducted twice annually until 2014 and once annually between 2015 and 2021. Accordingly, a total of 13 examinations containing prosthodontic questions were included in the present study. Examinations conducted after ÖSYM discontinued the public release of DUS questions in 2021 were excluded.

The exclusion criteria comprised questions officially canceled by ÖSYM because of issues such as the presence of multiple correct answers or the absence of a valid correct answer, as well as questions containing figures, images, or graphical content. Questions with visual content were excluded because current AI models are predominantly based on text-processing architectures and still exhibit limitations in effectively interpreting complex visual information.[16]

Initially, 130 prosthodontic questions were identified during the study period. After excluding questions that met the exclusion criteria, 126 questions were included in the analysis.

Question Classification

To ensure analytical consistency, the included questions were independently evaluated by two prosthodontists and classified according to three criteria; the examination year, question type, and topic area. Based on the question type, the questions were categorized as either knowledge-based or case-based. Knowledge-based questions were defined as those requiring direct factual knowledge related to basic concepts, material properties, or theoretical information, whereas case-based questions were defined as those involving clinical scenarios requiring diagnosis, treatment planning, or clinical decision-making.

Of the 126 questions included in the overall analysis, 97 were assigned to one of the three predefined topic categories: fixed prosthodontics (n=48), removable prosthodontics (n=45), and implant-supported prosthodontics (n=4) (**Table 1**). The remaining 29 questions were excluded from the topic-based subgroup analysis because they addressed cross-cutting subjects, including impression materials, occlusion, and temporomandibular joint-related topics, that could not be assigned exclusively to a single prosthodontic treatment category. These questions were retained in the overall accuracy and question-type analyses.

Table 1. Classification of DUS prosthodontic questions according to topic area

Topic	Definition/scope
Fixed prosthesis	Questions related to diagnosis, treatment planning, tooth preparation principles, material selection, and occlusal considerations for fixed prosthetic restorations
Removable prosthesis	Questions concerning the planning, design, biomechanics, and clinical application of removable prosthetic treatments
Implant-supported prosthesis	Questions related to prosthetic rehabilitation using dental implants, including treatment planning, prosthetic design, and implant-supported restorations

DUS: Turkish Dental Specialty Exam

Query Protocol and Response Evaluation

All chatbot interactions were conducted on January 13, 2026, in a desktop environment using the Google Chrome browser (version 124.0; Google LLC, Mountain View, CA, USA) on the Windows 10 operating system (Microsoft Corp., Redmond, WA, USA). The final question set was presented in Turkish under standardized conditions using the publicly

accessible versions of ChatGPT-4o (OpenAI, San Francisco, CA, USA; https://chat.openai.com/), Gemini 3 Flash (Google LLC, Mountain View, CA, USA; https://gemini.google.com/), Copilot (Microsoft Corp., Redmond, WA, USA; https://copilot.microsoft.com/), and DeepSeek-V3.2 (DeepSeek AI, Hangzhou, China; https://chat.deepseek.com/) available on that date. All systems were accessed through their official browser-based web platforms rather than application programming interfaces (APIs).

New accounts were created for each of the chatbots. In addition, a separate new session was initiated for each question to prevent the previous conversational context from influencing the AI system's responses. Before each question, the following standardized Turkish prompt was provided to all models: "assume that you are a candidate taking the DUS examination. Select the most appropriate answer from the options provided in the following multiple-choice questions. Each question contains five options, and only one option is correct."

Only the first response generated by each chatbot was included in the evaluation, and no regeneration of responses was performed. Responses were considered "correct" if they matched the official answer key published by the ÖSYM. The accuracy rate for each chatbot was calculated as the percentage obtained by dividing the number of correctly answered questions by the total number of questions answered.

Inter-rater reliability for the classification of question type, topic area, and response accuracy was assessed using Cohen's kappa (κ) coefficient. The Cohen's kappa value was 0.91, indicating a high level of agreement between evaluators. In cases of disagreement, a consensus was reached between the evaluators.

Statistical Analysis

All data analyses were performed using SPSS statistical software version 27.0 (IBM Corp., Armonk, NY, USA). Descriptive statistics were used to summarize the accuracy rates of the AI models. The dependent variable was defined as response accuracy (correct/incorrect). Because the responses were binary and each question was answered by all models, the generalized estimating equations (GEE) approach was employed to account for the repeated-measures structure of the data. A binomial distribution with a logit link function was used in the analyses, and the correlation structure was specified as exchangeable. The effects of model type, examination year, question type, and topic area on the probability of a correct response were evaluated. Bonferroni-corrected pairwise comparisons were performed for statistically significant findings. A p-value of <0.05 was considered statistically significant.

RESULTS

The distributions of correct and incorrect responses of the chatbot models according to question type are presented in **Table 2**. Across all models, the accuracy rates for knowledge-based questions were higher than those for case-based questions. Among the knowledge-based questions, ChatGPT demonstrated the highest accuracy rate (90.48%), followed by Copilot (87.62%), Gemini (86.67%), and DeepSeek (85.71%). For case-based questions, ChatGPT and Copilot achieved identical accuracy rates (61.90%). In contrast, Gemini (57.14%) and DeepSeek (52.38%) demonstrated lower accuracy rates for case-based questions.

Table 2. Accuracy rates of chatbot models according to question type

Model	Knowledge (n=105)		Case-based (n=21)		
	True	False	True	False	p-value
ChatGPT	90.48% (95)	9.52% (10)	61.90% (13)	38.10% (8)	0.031*
Gemini	86.67% (91)	13.33% (14)	57.14% (12)	42.86% (9)	0.039*
Copilot	87.62% (92)	12.38% (13)	61.90% (13)	38.10% (8)	0.025*
DeepSeek	85.71% (90)	14.29% (15)	52.38% (11)	47.62% (10)	0.028*

*Statistically significant

When all questions were evaluated together, ChatGPT achieved the highest overall accuracy rate (85.71%), followed by Copilot (83.33%), Gemini (81.75%) and DeepSeek (80.16%). However, no statistically significant differences were observed among the models in terms of knowledge-based questions, case-based questions, or overall accuracy rates (p=0.684, p=0.892, and p=0.619, respectively) (**Table 3**).

Table 3. Comparison of chatbot models according to question type and overall accuracy

		ChatGPT		Gemini		Copilot		DeepSeek		
Question type	n	True	False	True	False	True	False	True	False	p-value
Knowledge	105	90.48% (95)	9.52% (10)	86.67% (91)	13.33% (14)	87.62% (92)	12.38% (13)	85.71% (90)	14.29% (15)	0.684
Case-based	21	61.90% (13)	38.10% (8)	57.14% (12)	42.86% (9)	61.90% (13)	38.10% (8)	52.38% (11)	47.62% (10)	0.892
Total	126	85.71% (108)	14.29% (18)	81.75% (103)	18.25% (23)	83.33% (105)	16.67% (21)	80.16% (101)	19.84% (25)	0.619

In the Bonferroni-corrected pairwise comparison analyses, no statistically significant differences were detected among the chatbot models for either knowledge-based or case-based questions (p>0.05). This finding indicates that the models demonstrated comparable performance levels regardless of the question type (**Table 4**).

Table 4. Pairwise comparisons of chatbot performance according to question type

Chatbots	Knowledge (p)	Case-based (p)
ChatGPT-Gemini	0.342	0.743
ChatGPT-Copilot	0.514	1.000
ChatGPT-DeepSeek	0.287	0.612
Gemini-Copilot	0.621	0.721
Gemini-DeepSeek	0.418	0.534
Copilot-DeepSeek	0.552	0.674

In the topic-based evaluation, all the chatbot models showed similar accuracy rates for fixed and removable prosthodontic questions, whereas lower accuracy rates were observed for implant-supported prosthodontic questions. Within-model

comparisons indicated a statistically significant effect of topic area on accuracy rates ($p<0.05$). However, the implant-supported prosthodontics subgroup comprised only four questions; therefore, the corresponding percentage estimates and subgroup comparisons should be regarded as unstable and interpreted cautiously (**Table 5**).

Table 5. Accuracy rates of chatbot models according to prosthodontic topic categories

	Fixed prosthesis (n=48)		Removable prosthesis (n=45)		Implant-supported prosthesis (n=4)		
Model	**True**	**False**	**True**	**False**	**True**	**False**	**p-value**
ChatGPT	87.50% (42)	12.50% (6)	86.67% (39)	13.33% (6)	50.00% (2)	50.00% (2)	0.021*
Gemini	85.42% (41)	14.58% (7)	84.44% (38)	15.56% (7)	50.00% (2)	50.00% (2)	0.027*
Copilot	83.33% (40)	16.67% (8)	82.22% (37)	17.78% (8)	25.00% (1)	75.00% (3)	0.018*
DeepSeek	81.25% (39)	18.75% (9)	80.00% (36)	20.00% (9)	25.00% (1)	75.00% (3)	0.015*

*Statistically significant

However, Bonferroni-corrected pairwise comparison analyses revealed no statistically significant differences among the chatbot models in any of the fixed, removable, or implant-supported prosthodontic categories ($p>0.05$) (**Table 6**).

Table 6. Pairwise comparisons of chatbot performance according to prosthodontic topic categories

Chatbots	Fixed prosthesis (p)	Removable prosthesis (p)	Implant-supported prosthesis (p)
ChatGPT-Gemini	0.79	0.82	1.00
ChatGPT-Copilot	0.56	0.59	0.50
ChatGPT-DeepSeek	0.34	0.39	0.50
Gemini-Copilot	0.71	0.74	0.50
Gemini-DeepSeek	0.49	0.54	0.50
Copilot-DeepSeek	0.76	0.79	1.00

The accuracy rates of the chatbot models according to the examination years are presented in **Figure**. Although some fluctuations in accuracy rates were observed across the years, statistical analysis demonstrated that the accuracy rates did not differ significantly according to the examination year ($\chi^2=7.1$; $p=0.62$). This finding suggests that chatbot performance remained consistent throughout the study period.

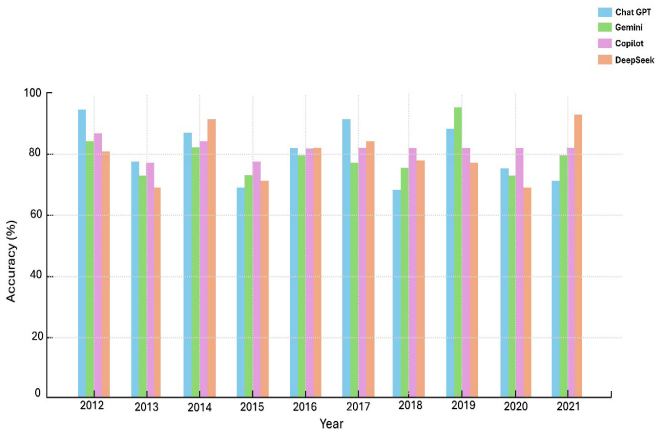

Figure. Accuracy rates of chatbot models across DUS examination years (2012-2021)

DUS: Turkish Dental Specialty Exam

DISCUSSION

In the present study, the accuracy of the responses generated by different AI-based LLMs to prosthodontic questions from the DUS was evaluated and compared. The findings demonstrated no statistically significant differences among the models in terms of overall accuracy. In addition, higher accuracy rates were observed for knowledge-based questions than for case-based questions. Topic-based evaluation revealed similar performance in fixed and removable prosthodontics, whereas numerically lower accuracy was observed in the implant-supported subgroup. Accordingly, the null hypothesis regarding overall performance differences among the models (H_{01}) was not rejected, whereas the null hypotheses concerning differences between knowledge-based and case-based questions (H_{02}) and differences according to topic area (H_{03}) were rejected.

The impact of AI has become increasingly prominent in healthcare fields such as medicine and dentistry, with applications in numerous processes including diagnosis, treatment planning, and patient follow-up.[17,18] Parallel to the growing adoption of AI-based applications in healthcare services, new opportunities have emerged in education to support teaching and learning processes for both students and educators.[19] In a previous study evaluating the performance of AI-based LLMs in medical education, the system was reported to achieve near-passing performance on the United States Medical Licensing Examination (USMLE) without specialized medical training or additional reinforcement.[20]

All models achieved overall accuracy rates above 80%, indicating potential value for examination preparation and information access. Nevertheless, these systems were not specifically developed for dental education or clinical decision-making, and high-test accuracy should not be interpreted as equivalent to domain-specific expertise.[2]

Consistent with our findings, previous studies have shown that LLMs such as ChatGPT can achieve passing-level or high accuracy rates on medical licensing examinations.[21] Recent investigations evaluating examination performance in various dental disciplines have further demonstrated that AI models can, in many cases, produce results comparable to human performance and occasionally even outperform human participants. These findings indicate that AI models may be effectively utilized to support educational materials and examination preparation processes.[2,22]

However, all chatbot models demonstrated higher accuracy for knowledge-based questions than for case-based questions, indicating that the models performed better on theoretical content than on clinically oriented reasoning. This finding is consistent with previous studies reporting that, although LLMs can achieve high levels of factual accuracy, they may have limitations in case interpretation and practice-based decision-making processes.[22] No statistically significant differences were detected across examination years, suggesting generally stable performance across the evaluated periods.

Ekici et al.[23] reported in their evaluation of DUS questions that AI-based LLMs performed better on knowledge-based questions, such as those in basic sciences, whereas their performance was lower on questions requiring not only factual knowledge but also clinical reasoning, discussion, and interpretation, such as those in the clinical sciences.

Similarly, Haberal et al.[24] evaluated the performance of different AI chatbots using restorative dentistry questions from the DUS and reported that all models achieved high accuracy levels. However, no statistically significant differences were found among the models in terms of overall accuracy rates. The study also noted a decrease in accuracy, particularly for complex questions that require clinical interpretation. This finding supports the view that LLMs may have limitations in situations requiring clinical decision-making and advanced reasoning.

The lower accuracy observed in case-based questions suggests that LLMs may have limitations in multistep decision-making processes that require clinical reasoning skills. Case-based questions require not only the recall of factual knowledge but also the integrated interpretation of clinical findings, formulation of differential diagnoses, and development of context-appropriate treatment plans. Previous studies have reported that LLMs often generate responses through probabilistic pattern matching based on training data rather than genuine clinical reasoning, and may demonstrate limited flexible reasoning, particularly in complex clinical scenarios. Moreover, these models have been reported to produce overconfident yet incorrect responses and to exhibit a tendency toward hallucination during clinical problem-solving. This may explain the reduced performance observed in the case-based questions.[25]

Numerically lower accuracy was observed for implant-supported prosthodontic questions. Implant dentistry involves multidisciplinary considerations, including prosthetic design, biomechanical assessment, anatomical risk analysis, and long-term complication management.[26-28] However, this finding should be interpreted cautiously because only four implant-supported prosthodontic questions were included in the present study. With such a small denominator, each response changed the subgroup accuracy estimate by 25 percentage points. Therefore, the observed difference may reflect sampling variability rather than a robust topic-specific performance deficit. Larger and more balanced question sets are required before firm conclusions can be drawn regarding the performance of LLMs in implant-supported prosthodontics.

Turan Gökduman et al.[29] evaluated DUS endodontic questions and reported accuracy rates of 82.8%, 83.6%, and 77.9% for ChatGPT, Gemini, and Copilot, respectively. No statistically significant overall difference was found among the models. In contrast to the present study, Copilot demonstrated a significantly higher accuracy rate during 2012–2015 than during the 2016–2021 period. These findings suggest that the performance of AI models across different dental disciplines within the DUS may be broadly similar in terms of overall accuracy, although performance fluctuations may occur in specific models depending on the examination year and question structure.

A study evaluating oral radiology questions from the DUS in Turkiye compared the performance of ChatGPT and Copilot and reported that ChatGPT achieved a markedly higher accuracy than Copilot. In that study, ChatGPT reached an accuracy rate of 86.1%, whereas Copilot achieved 41.5%. This finding suggests that ChatGPT may outperform Copilot in answering multiple-choice examination questions in dentistry.[30]

Dündar Sarı[16] found no significant difference between ChatGPT and Gemini in terms of overall accuracy on DUS questions; however, ChatGPT demonstrated significantly superior performance in clinical dental sciences. Notably, its higher accuracy in prosthodontics than that of Gemini indicates that model performance may vary according to the dental discipline evaluated.

When these studies are considered collectively, the discrepancies among their findings may be attributable to several factors. First, differences in question-set scope, question type, and sample size used may directly influence the performance of AI models. In addition, variations in the dental disciplines examined, differences in model versions and update timing may also contribute to differences in accuracy rates. Collectively, these discrepancies may reflect differences in dental discipline, question-set composition, sample size, model version, training data, and evaluation timing.[31]

These findings suggest that LLMs may serve as supportive tools in dental education, particularly for examination preparation, rapid access to information, and self-directed learning processes. The high accuracy rates observed especially in theoretically oriented questions indicate that LLMs may contribute to students' knowledge reinforcement and revision processes. However, several important limitations related to reliability and patient safety remain in the clinical and educational use of LLMs. Previous studies have reported that LLMs may present non-existent or unverifiable information in a fluent and convincing manner, which may lead to erroneous diagnostic or therapeutic guidance, particularly in clinical settings.[32] Evidence from dental education further indicates that LLMs may generate misinformation, provide inaccurate or fabricated references, show a limited ability to interpret visual content, and contribute to excessive dependence among students.[33] Therefore, LLM-based systems may be more appropriately used as complementary support tools supervised by educators or clinicians rather than as independent decision-making systems in dental education and clinical practice.

Although LLM-based AI models can generate correct answers, they often exhibit limitations in terms of explanatory depth, reliability of reference support, and clinical reasoning.[8] Similar concerns have been raised in studies focusing on dentistry, reporting that although LLMs may provide some correct answers to dental questions, their overall performance may remain limited, their responses may be debatable, and they may sometimes present incorrect answers with seemingly convincing reasoning.[34] Therefore, evaluating AI performance solely based on accuracy rates may not fully reflect the educational and clinical reliability of these systems.

Only the first response generated by each model was analyzed, although LLM outputs may vary across repeated queries. All questions were presented in Turkish, which may limit cross-language generalizability.[35]

Limitations

Another important limitation is the small number of questions available for each examination year. The inclusion of only 9-10 questions in each yearly subgroup may have limited the statistical power of the year-based analyses. Consequently, small changes in the number of correct responses may produce large proportional differences and distort apparent between-model differences. The implant-supported prosthodontics subgroup included only four questions, which limited the precision and generalizability of the topic-based analysis. Accordingly, these findings should be regarded as exploratory. Although generalized linear models were used in the present study, the underlying reasoning, reference support, and clinical explanations provided by the models were not analyzed in detail. Therefore, future studies should comprehensively evaluate not only accuracy rates but also the reasoning processes, source use, and clinical judgment demonstrated by AI systems. Finally, because LLMs are dynamic systems that are continuously updated, the performance results obtained in this study may be specific to the time at which the evaluation was conducted and may change with future updates.

CONCLUSION

Within the limitations of this study, the evaluated LLMs demonstrated high accuracy in answering multiple-choice prosthodontic examination questions and may have the potential to serve as complementary tools in dental education, particularly in examination preparation and information access. However, the limitations observed in scenarios requiring clinical reasoning and complex decision-making indicate that current LLMs should be regarded as supervised support systems in educational and clinical settings rather than replacements for human expertise.

ETHICAL DECLARATIONS

Ethics Committee Approval

This article does not require ethics committee approval, as it does not involve human or animal studies.

Informed Consent

Since the study was conducted without the participation of any living being, no written consent form was obtained.

Peer Review Process

This manuscript was subject to external peer review.

Conflict of Interest

The author declare no conflicts of interest related to this study.

Financial Disclosure

The author received no financial support for the conduct or publication of this research.

Author Contributions

The author is solely responsible for the entirety of conception, execution, analysis, and writing of the manuscript.

REFERENCES

1. Hoch CC, Wollenberg B, Lüers JC, et al. ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. *Eur Arch Otorhinolaryngol*. 2023;280(9):4271-4278. doi:10.1007/s00405-023-08051-4
2. 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. doi:10.1186/s12909-024-06389-9
3. Fogel AL, Kvedar JC. Artificial Intelligence powers digital medicine. *NPJ Digit Med*. 2018;1:5. doi:10.1038/s41746-017-0012-2
4. Cicciù M, Fiorillo L, D'Amico C, et al. 3D digital impression systems compared with traditional techniques in dentistry: a recent data systematic review. *Materials*. 2020;13(8):1982. doi:10.3390/ma13081982
5. Sezer B, Aydoğdu T. Performance of advanced Artificial Intelligence models in pulp therapy for immature permanent teeth: a comparison of ChatGPT-4 Omni, DeepSeek, and Gemini Advanced in accuracy, completeness, response time, and readability. *J Endod*. 2025;51(11):1675-1684. doi:10.1016/j.joen.2025.08.011
6. Iannantuono GM, Bracken-Clarke D, Floudas CS, Roselli M, Gulley JL, Karzai F. Applications of large language models in cancer care: current evidence and future perspectives. *Front Oncol*. 2023;13:1268915. doi:10.3389/fonc.2023.1268915
7. Zhang P, Kamel Boulos MN. Generative AI in medicine and healthcare: promises, opportunities and challenges. *Future Internet*. 2023;15(9):286. doi:10.3390/fi15090286
8. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare*. 2023;11(22):2955. doi:10.3390/healthcare11222955
9. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020;33:1877-1901. doi:10.48550/arXiv.2005.14165
10. 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. doi:10.1186/s12909-025-06849-w
11. Dave M, Patel N. Artificial Intelligence in healthcare and education. *Br Dent J*. 2023;234(10):761-764. doi:10.1038/s41415-023-5845-2
12. Tartuk BK, Altıntaş E. Evaluation of accuracy, clinical reliability and readability of LLM-based chatbot responses in prosthetic dentistry FAQs. *Anatolian Curr Med J*. 2026;8(3):513-523. doi:10.38053/acmj.1905988
13. Çulhaoğlu AK, Kılıçarslan MA, Denız KZ. Diş Hekimliğinde Uzmanlik Sınavının farklı eğitim seviyelerdeki algı ve tercih durumlarının değerlendirilmesi. *Ata Diş Hek Fak Derg*. 2021;31(3):420-426. doi:10.17567/ataunidfd.911839
14. 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. doi:10.1016/j.jds.2025.01.025
15. Venkatesan S, Krishnamoorthi D, Raju R, Mohan J, Thomas PA, Rubasree B. Evidence-based prosthodontics. *J Pharm Bioallied Sci*. 2022;14(Suppl 1):S50-S59. doi:10.4103/jpbs.jpbs_149_22
16. Dundar Sari MB, Sezer B. Comparative performance evaluation of ChatGPT-4 Omni and Gemini Advanced in the Turkish Dentistry Specialization Exam. *BMC Med Educ*. 2026;26(1):251. doi:10.1186/s12909-026-08621-0
17. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of Artificial Intelligence in clinical practice. *BMC Med Educ*. 2023;23(1):689. doi:10.1186/s12909-023-04698-z
18. Erdal SG, Güdül A, Köroğlu A. The effectiveness of large language models in dental specialty questions: a comparative study in the field of prosthodontics. *BMC Med Educ*. 2026;26(1):455. doi:10.1186/s12909-026-08808-5
19. Yilmaz BE, Gokkurt Yilmaz BN, Ozbey F. Artificial Intelligence performance in answering multiple-choice oral pathology questions: a comparative analysis. *BMC Oral Health*. 2025;25(1):573. doi:10.1186/s12903-025-05926-2
20. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198
21. Kasagga A, Sapkota A, Changaramkumarath G, et al. Performance of ChatGPT and large language models on medical licensing exams worldwide: a systematic review and network meta-analysis with meta-regression. *Cureus*. 2025;17(10):e94300. doi:10.7759/cureus.94300
22. Temiz M, Güzel C. Assessing the performance of ChatGPT on dentistry specialization exam questions: a comparative study with DUS examinees. *Medical Records*. 2025;7(1):162-166. doi:10.37990/medr.1567242
23. Ekici Ö. Comparative evaluation of four large language models in Turkish Dentistry Specialization Exam. *Selcuk Dent J*. 2025;12(4):6-10. doi:10.15311/selcukdentj.1674113

24. Haberal M, Hançerlioğulları D. Can Artificial Intelligence chatbots think like dentists? A comparative analysis based on dental specialty examination questions in restorative dentistry. *BMC Oral Health*. 2026; 26(1):231. doi:10.1186/s12903-025-07612-9
25. Kim J, Podlasek A, Shidara K, Liu F, Alaa A, Bernardo D. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. *Sci Rep*. 2025;15(1):39426. doi:10.1038/s41598-025-22940-0
26. Karvekar S, Anand V, Singh D, Chaudhary VK, Fatima S, Mathar MI. Digital dentistry and Artificial Intelligence: a systematic review on innovations in diagnosis, treatment planning, and prosthodontics. *Cureus*. 2026;18(2):e104422. doi:10.7759/cureus.104422
27. Alfaraj A, Limones Á, Ahmad S, et al. Harnessing AI in prosthodontics and implant dentistry: an umbrella review of systematic evidence. *J Prosthodont*. 2026;35(2):127-142. doi:10.1111/jopr.70091
28. Revilla-León M, Gómez-Polo M, Vyas S, et al. Artificial Intelligence applications in implant dentistry: a systematic review. *J Prosthet Dent*. 2023;129(2):293-300. doi:10.1016/j.prosdent.2021.05.008
29. Turan Gökduman C, Arılı Öztürk E, Aktaş Ş, Çanakçi BC. Comparison of chatbots' accuracy in endodontics questions in dentistry specialization exam in Turkiye: ChatGPT-4o, Gemini Advanced, Copilot, and Claude. *BMC Oral Health*. 2025;26(1):28. doi:10.1186/s12903-025-07346-8
30. Tassoker M. ChatGPT-4 Omni's superiority in answering multiple-choice oral radiology questions. *BMC Oral Health*. 2025;25(1):173. doi:10.1186/s12903-025-05554-w
31. Çetin SG, Karadağ GG. Comparative evaluation of Artificial Intelligence models in answering pediatric dentistry questions of the Turkish Dental Specialization Exam (DUS). *Dicle Dent J*. 2025;26(2):31-37.
32. Kong M, Fok EHW, Yiu CKY. A Scoping review of large language models in dental education: applications, challenges, and prospects. *Int Dent J*. 2025;75(6):103854. doi:10.1016/j.identj.2025.103854
33. Roustan D, Bastardot F. The clinicians' guide to large language models: a general perspective with a focus on hallucinations. *Interact J Med Res*. 2025;14:e59823. doi:10.2196/59823
34. Aşık A, Kuru E. Analysis of ChatGPT's answers to pedodontics questions asked in the dentistry specialization training entrance exam: cross-sectional study. *Turkiye Klinikleri J Dental Sci*. 2025;31(3):401-406. doi:10.5336/dentalsci.2024-107488
35. Mavrych V, Ganguly P, Bolgova O. Using large language models (ChatGPT, Copilot, PaLM, Bard, and Gemini) in gross anatomy course: comparative analysis. *Clin Anat*. 2025;38(2):200-210. doi:10.1002/ca.24244