

Comparative evaluation of the performance of ai-powered chatbots in answering basic sciences questions of the dentistry specialization examination*

 Tolga Mavigöz,  Adem Altın,  Merve Yeniçeri Özata*

Department of Endodontics, Faculty of Dentistry, Dicle University, Diyarbakır, Türkiye

Cite this article as: Mavigöz T, Altın A, Yeniçeri Özata M. Comparative evaluation of the performance of ai-powered chatbots in answering basic sciences questions of the dentistry specialization examination. *J Dent Sci Educ.* 2026;4(1):17-22.

Received: 19.09.2025

Accepted: 15.02.2026

Published: 27.02.2026

ABSTRACT

Aims: The aim of this study is to evaluate the correctness rate/probability of answers provided by AI-powered chatbots (Gemini Advanced 2.5 Pro, ChatGPT-4 omni, ChatGPT-5 and DeepSeek v3) to single-answer, multiple-choice basic sciences questions from the Dentistry Specialization Examination (DUS) administered between 2012 and 2025.

Methods: A total of 539 multiple-choice questions from the basic sciences section of the DUS from 2012 to 2025 were used. Each question was presented directly in a new session. The rates of correct/incorrect answers were calculated based on the subject of the question, the year it was asked, and the chatbot model. The rate and probability of incorrect answers were evaluated using chi-square and binary logistic regression analyses.

Results: The rate of incorrect answers was highest in 2012, with a significant decrease observed in subsequent years ($p < 0.05$). Among the subjects, Anatomy had the highest rate of incorrect answers, while Pathology had the lowest ($p < 0.05$). When comparing chatbot models, Gemini Advanced 2.5 Pro was found to have a statistically significantly lower error rate than ChatGPT-4 omni and DeepSeek v3 ($p < 0.05$). In the regression analysis, the risk of providing an incorrect answer was statistically significantly higher for the years 2012 and 2018; for the fields of Anatomy, Physiology and Microbiology; and for the ChatGPT-4 omni and DeepSeek v3 models, compared to their respective reference groups ($p < 0.05$).

Conclusion: AI-powered chatbots provided more accurate answers to more recent questions. In the subject-based performance analysis, the rate of incorrect answers for Anatomy questions was high. In the chatbot comparison, Gemini Advanced 2.5 Pro produced more accurate answers than DeepSeek v3. While AI-powered chatbots can be a potential supplementary tool in the preparation process for DUS, their accuracy and competence across different subjects are limited.

Keywords: Basic sciences, chatbot, dentistry specialization examination

*This study was presented orally at the 'Genç Endodontistler Konuşuyor Sempozyumu'.

INTRODUCTION

Artificial intelligence (AI) is at the forefront of many aspects of our lives, transforming how we analyze information and improving decision-making processes through problem-solving, reasoning, and learning.¹ Through the creation and analysis of intelligent software and hardware, known as "intelligent agents," AI is integrating into many areas of daily life and performing various tasks.² In recent years, it has been rapidly adopted to enhance the speed and quality of the healthcare sector, becoming increasingly important.³ In dentistry, it is used in numerous areas such as diagnosing caries and predicting its progression risk,⁴ interpreting panoramic radiographs,⁵ treatment planning,⁶ detecting painful pathologies like pulpitis, trauma, cysts and tumors⁷ and determining root canal system anatomy and working length.⁸

Alongside AI, subfields like machine learning (ML) have made significant advancements in recent years, leading to groundbreaking developments.³ Machine learning is a subset of AI that improves its performance based on data provided to a general algorithm derived from experience, rather than defining rules in traditional approaches. One of the primary examples of AI systems evolved from ML is chatbots.² The term "chatbot" is a combination of the words "chat" and "robot" and it emerged initially as a text-based, AI-powered conversational system designed to simulate human language.⁹ These systems can imitate human language, provide personalized responses and engage in meaningful conversations.¹⁰ Over time, chatbots have been developed using various deep learning models and different training techniques.¹¹ They are used in many different fields. They have the potential to accelerate and enhance scientific research, thereby advancing technological progress and their problem-solving skills can be evaluated by

*Corresponding Author: Merve Yeniçeri Özata, merveyeniceri05@hotmail.com



This work is licensed under a Creative Commons Attribution 4.0 International License.



testing them in various domains and against real professional examinations.¹²

For dentists in Türkiye to pursue specialization, they must succeed in the Dentistry Specialization Examination (DUS), organized by the Ministry of Health and the ÖSYM (Assessment, Selection, and Placement Center). The DUS, held annually since 2012, is a comprehensive exam that measures knowledge with a total of 120 multiple-choice questions, covering 8 specialty areas and basic medical sciences, with 40 questions in basic sciences and 80 in clinical sciences. The success ranking of candidates is determined by the results of this exam, which then dictates their placement into specialty programs. Furthermore, until 2024, the basic and clinical sciences sections had equal weighting coefficients, contributing equally to the total exam score.¹³ Therefore, candidates' proficiency in basic sciences is as crucial as their knowledge in clinical sciences. The basic sciences exam consists of 6 questions each from Anatomy, Physiology, Biochemistry and Microbiology, and 4 questions each from Histology-Embryology, Pathology, Pharmacology and Medical Biology-Genetics.

The Basic Sciences section of the Dental Specialty Examination (DUS) was selected as the focus of this study based on both methodological and educational considerations. This section is composed of standardized, knowledge-based questions that emphasize theoretical understanding and factual recall, rather than clinical decision-making or procedural skills. Such characteristics render the Basic Sciences section particularly appropriate for assessing the information-processing and reasoning capacities of AI-based chatbot models.

Furthermore, the scope and longitudinal integration of basic medical sciences within undergraduate dental education are generally more limited when compared with medical curricula. Consequently, dental graduates may experience difficulties in consolidating and applying foundational medical knowledge in high-level theoretical examinations. Within this framework, AI-driven chatbot systems may function as readily accessible supportive learning tools, facilitating reinforcement and clarification of fundamental concepts in basic sciences.

Additionally, questions in the Basic Sciences section are less susceptible to clinical heterogeneity or examiner-related subjectivity, enabling a more objective and reproducible evaluation of chatbot performance across multiple disciplines and examination years. Accordingly, assessing AI performance in this domain establishes a baseline for interpreting the educational utility of AI systems in dental training and exam preparation, prior to extending analyses to clinically oriented or case-based assessments.

In recent years, various companies have designed and introduced different chatbots.¹⁰ ChatGPT (OpenAI Global, San Francisco, CA, USA) is a chatbot based on the language model developed by OpenAI and has been trained for conversation using a technique called reinforced learning from human feedback.¹⁴ OpenAI launched ChatGPT-4 omni on May 13, 2024, as a multimodal chatbot. Subsequently, it introduced GPT-5 on August 7, 2025, describing it as "our best AI system ever," representing a significant leap in intelligence compared to its predecessors. The company claims this model offers state-of-the-art performance in mathematics, coding, writing, and healthcare, presenting it as versatile and reliable.¹⁵

In 2023, Gemini (Google, Mountain View, California, USA) was introduced as Google's largest and most capable AI system.¹⁶ It has three versions: Gemini Nano, Gemini Pro and Gemini Ultra, with Ultra being the most powerful in terms of problem-solving.¹⁷ DeepSeek-V1 was developed by a Chinese research group in late 2023, primarily aiming to enhance its language modeling capabilities using publicly available open-source data. On the other hand, DeepSeek-V3 was claimed to be successful enough to compete with ChatGPT-4 omni shortly after its launch in January 2025 and quickly gained worldwide popularity.¹⁸

Interest in the performance of these AI-powered chatbots in the context of medical and dental examinations is steadily increasing.¹⁹⁻²² These chatbots could facilitate the preparation process for specialization candidates when used for practice in the DUS basic sciences section. Thus, this study aims to evaluate whether AI-powered chatbots (Gemini Advanced 2.5 Pro, ChatGPT-4 omni, ChatGPT-5.0, DeepSeek v3) can correctly answer the basic sciences questions (Anatomy, Histology-Embryology, Physiology, Biochemistry, Microbiology, Pathology, Pharmacology and Medical Biology-Genetics) from the DUS administered between 2012 and 2025. By comparing the accuracy rates of the answers given by different chatbot models, the study examines whether the chatbots' knowledge level is consistent with the exam standards according to the years the exams were held and the subjects covered.

METHODS

Since the study involved no human or animal subjects, clinical interventions, or identifiable patient data, ethics committee approval was not required. This study included 539 multiple-choice questions from the Basic Sciences section of the DUS conducted between 2012 and 2025 (held once a year between 2015-2022; and as spring and autumn sessions between 2012-2014 and from 2023 onwards). The questions covered the following topics: anatomy, histology-embryology, physiology, biochemistry, microbiology, pathology, pharmacology and medical biology-genetics. The questions used were obtained from the publicly accessible website of ÖSYM and were analyzed solely for academic research purposes and were not reproduced in any way. Since ÖSYM began publishing only 10% of the DUS questions from 2022 onwards, the question counts from those years were evaluated collectively. Canceled questions (1) and questions with visual content (4) were excluded from the study.

Each question was posed to Gemini Advanced 2.5 Pro, ChatGPT-4 omni, ChatGPT-5 and DeepSeek v3 chatbots by a single operator simultaneously on September 1, 2025, with each model having only one attempt per question. Before the questions were asked, all chatbots were given the following prompt: "I will present you with a series of multiple-choice questions, each with 5 options and a single correct answer. You do not need to provide any explanation. Just tell me which option is correct." The answers were categorized and recorded as either correct or incorrect.

Statistical Analysis

The data were analyzed using IBM SPSS v23. The significance of the relationship between independent categorical variables (year, subject and chatbot model) and the accuracy of the answers (Correct/Incorrect) was examined using the Pearson



Chi-Square test. In cases where the Chi-Square test was significant, a post-hoc analysis was conducted to identify the differences between the incorrect answer rates. Column proportions were compared pairwise using the z-test and the Bonferroni correction was applied for multiple comparisons. The effect of independent variables on the dependent variable was examined using binary logistic regression analysis. Descriptive statistics for categorical variables were presented as frequency and percentage. The significance level was set at $p < 0.05$.

RESULTS

Table 1 presents the rates and analysis of correct and incorrect responses, categorized by year, subject and chatbot.

	Answer	
	Correct n (%)	Incorrect n (%)
Year		
2012	293 (91.6)	27 ^a (8.4)
2013	298 (94.3)	18 ^{ab} (5.7)
2014	309 (96.6)	11 ^{ab} (3.4)
2015	155 (96.9)	5 ^b (3.1)
2016	153 (95.6)	7 ^{ab} (4.4)
2017	159 (99.4)	1 ^b (0.6)
2018	140 (92.1)	12 ^{ab} (7.9)
2019	154 (98.7)	2 ^b (1.3)
2020	155 (96.9)	5 ^b (3.1)
2021	148 (94.9)	8 ^{ab} (5.1)
2022 and onwards	95 (99)	1 ^b (1)
Subject		
Anatomy	284 (91)	28 ^a (9)
Histology-embryology	209 (96.8)	7 ^{ab} (3.2)
Physiology	308 (95.1)	16 ^{ab} (4.9)
Biochemistry	312 (97.5)	8 ^{ab} (2.5)
Microbiology	315 (94.9)	17 ^{ab} (5.1)
Pathology	209 (98.6)	3 ^b (1.4)
Pharmacology	212 (96.4)	8 ^{ab} (3.6)
Medical biology-genetics	210 (95.5)	10 ^{ab} (4.5)
Chatbot		
Gemini advanced 2.5 pro	526 (97.6)	13 ^a (2.4)
ChatGPT-4 omni	511 (94.8)	28 ^{ab} (5.2)
ChatGPT-5	515 (95.5)	24 ^{ab} (4.5)
DeepSeek v3	507 (94.1)	32 ^b (5.9)

Pearson Chi-Square. a-b: Rows with the same letter are not significantly different from each other

Results Regarding the Distribution of Answers by Year

A statistically significant relationship was found between the year and the distribution of correct and incorrect answers ($\chi^2=31.097$, $p=0.001$). The incorrect answer rates ranged from 0.6% to 8.4% across the years. The post-hoc analysis revealed that the year 2012, with the highest incorrect answer rate (8.4%), had a statistically significantly higher error rate than the years with the lowest rates-2015, 2017, 2019, 2020 and 2022

($p < 0.05$). The error rates for the other years (2013, 2014, 2018 and 2021) did not show a statistically significant difference from these two extreme groups ($p > 0.05$).

Results Regarding the Distribution of Answers by Subject

A statistically significant relationship was detected between the subjects examined and the accuracy of the answers ($\chi^2=23.834$, $p=0.002$). The incorrect answer rates for subjects were found to range from 1.4% to 9.0%. According to the pairwise comparison results, a statistically significant difference was found between Anatomy, which had the highest incorrect answer rate (9.0%), and Pathology, which had the lowest (1.4%) ($p < 0.05$). The incorrect answer rates of the other six subjects (Histology-Embryology, Physiology, Biochemistry, Microbiology, Pharmacology and Medical Biology-Genetics) did not show a statistically significant difference from these two extremes ($p > 0.05$).

Results Regarding the Distribution of Answers by Chatbot Model

A statistically significant relationship was observed between the chatbot models used and the accuracy of the answers ($\chi^2=8.668$, $p=0.034$). The incorrect answer rates of the models varied from 2.4% to 5.9%. Post-hoc analysis showed that Gemini Advanced 2.5 Pro, with the lowest incorrect answer rate (2.4%), performed statistically significantly better than both ChatGPT-4 omni (5.2%) and DeepSeek v3 (5.9%) ($p < 0.05$). The performance of the ChatGPT-5 model (4.5%) did not show a significant difference from the other models ($p > 0.05$). No statistically significant difference was found between the performances of the ChatGPT-4 omni and DeepSeek v3 models ($p > 0.05$).

Results Regarding the Regression Model

The effect of the independent variables on the responses was examined using univariate and multiple binary logistic regression (**Table 2**).

The effect of independent variables on the answers was examined using univariate and multiple binary logistic regression (Table 2). In both the univariate and multiple models, with the years 2022 and after as the reference, the odds of chatbots giving an incorrect answer to questions from the 2012 exam were 8.754/9.946 times higher and for the 2018 exam, the odds were 8.143/9.787 times higher ($p < 0.05$). With Pathology as the reference subject, the odds of chatbots giving an incorrect answer to questions in Anatomy were 6.869/7.428 times higher ($p < 0.05$). This ratio was 3.619/3.789 times higher for Physiology and 3.76/3.968 times higher for Microbiology ($p < 0.05$). With Gemini Advanced 2.5 Pro as the reference, the odds of ChatGPT-4 omni giving an incorrect answer were 2.217/2.261 times higher and for DeepSeek v3, the odds were 2.554/2.618 times higher ($p < 0.05$).

DISCUSSION

The purpose of this study is to examine the knowledge proficiency of AI-powered chatbots in the DUS basic sciences field and their performance in assessment and evaluation processes. In this context, DUS basic sciences questions were posed to four different AI-powered chatbots; the responses received were analyzed in terms of accuracy/error rates and



Table 2. Investigation of the effect of independent variables on answers using logistic regression analysis

	Answer		Univariate		Multiple	
	Correct n (%)	Incorrect n (%)	OR (95 % CI)	p	OR (95 % CI)	p
Year						
2012	293 (91.6)	27 (8.4)	8.754 (1.174-65.29)	0.034	9.946 (1.327-74.568)	0.025
2013	298 (94.3)	18 (5.7)	5.738 (0.756-43.555)	0.091	6.455 (0.846-49.249)	0.072
2014	309 (96.6)	11 (3.4)	3.382 (0.431-26.534)	0.246	3.765 (0.477-29.683)	0.208
2015	155 (96.9)	5 (3.1)	3.065 (0.353-26.631)	0.310	3.407 (0.39-29.779)	0.268
2016	153 (95.6)	7 (4.4)	4.346 (0.526-35.881)	0.172	4.857 (0.585-40.337)	0.143
2017	159 (99.4)	1 (0.6)	0.597 (0.037-9.664)	0.717	0.657 (0.04-10.682)	0.768
2018	140 (92.1)	12 (7.9)	8.143 (1.041-63.669)	0.046	9.787 (1.243-77.052)	0.030
2019	154 (98.7)	2 (1.3)	1.234 (0.11-13.792)	0.865	1.401 (0.125-15.747)	0.785
2020	155 (96.9)	5 (3.1)	3.065 (0.353-26.631)	0.310	3.407 (0.39-29.779)	0.268
2021	148 (94.9)	8 (5.1)	5.135 (0.632-41.715)	0.126	5.923 (0.725-48.412)	0.097
2022 and onwards	95 (99)	1 (1)			Reference	
Subject						
Anatomy	209 (98.6)	3 (1.4)			Reference	
Histology-embryology	284 (91)	28 (9)	6.869 (2.061-22.896)	0.002	7.428 (2.213-24.925)	0.001
Physiology	209 (96.8)	7 (3.2)	2.333 (0.595-9.146)	0.224	2.39 (0.606-9.426)	0.213
Biochemistry	308 (95.1)	16 (4.9)	3.619 (1.041-12.576)	0.043	3.789 (1.084-13.244)	0.037
Microbiology	312 (97.5)	8 (2.5)	1.786 (0.468-6.811)	0.396	1.805 (0.471-6.916)	0.389
Pathology	315 (94.9)	17 (5.1)	3.76 (1.088-12.989)	0.036	3.968 (1.142-13.788)	0.030
Pharmacology	212 (96.4)	8 (3.6)	2.629 (0.688-10.046)	0.158	2.74 (0.712-10.536)	0.142
Medical biology-genetics	210 (95.5)	10 (4.5)	3.317 (0.9-12.226)	0.072	3.473 (0.936-12.883)	0.063
Chatbot						
Gemini advanced 2.5 Pro	526 (97.6)	13 (2.4)			Reference	
ChatGPT-4 omni	511 (94.8)	28 (5.2)	2.217 (1.136-4.328)	0.020	2.261 (1.15-4.449)	0.018
ChatGPT-5	515 (95.5)	24 (4.5)	1.886 (0.95-3.744)	0.070	1.913 (0.957-3.827)	0.067
DeepSeek v3	507 (94.1)	32 (5.9)	2.554 (1.325-4.922)	0.005	2.618 (1.348-5.086)	0.004

OR (95% CI): Odds Ratio (95% Confidence Interval), CI: Confidence interval

risk, according to the years the questions were asked, the subjects and the chatbot models. The data obtained were compared to evaluate the potential of AI as a learning and assessment tool in dental education.

A study found that ChatGPT-3.5 answered pedodontics questions with 54.3% accuracy.²³ In another study, DUS prosthodontics questions were posed to ChatGPT-3.5 and Gemini, and the answers were evaluated using a Likert scale. Both models were reported to have limited capabilities, but it was also emphasized that these were the most basic versions of the models.²⁴ In our study, which used the Gemini Ultra model and GPT-5, we achieved accuracy rates above 90 %, demonstrating a remarkable increase in the performance of newer models.

In our study, the accuracy rates of AI chatbots were examined using DUS basic sciences questions from 2012 to 2025. The data revealed that chatbots generally showed a high level of accuracy, although this varied by year. However, the relatively lower success rates in some years (e.g., 2012 at 91.6% and 2018 at 92.1%) may stem from differences in the scope or difficulty level of the questions. The notably higher error rates observed in the 2012 examination can be attributed to the historical context of the DUS; 2012 marked the inaugural year of the DUS in Türkiye. Consequently, the standardization of question structure and item difficulty during this pilot period may have differed from the more refined datasets of subsequent years, posing challenges for AI pattern recognition. Similarly,

the performance drop in 2018 aligns with ÖSYM's periodic updates to the examination framework, aimed at increasing the discriminative power of the test by introducing more complex clinical scenarios or changing the distribution of topic-specific questions. These structural shifts in the exam likely presented 'out-of-distribution' data points that the tested LLMs struggled to interpret accurately. One study showed that chatbots achieved 100% accuracy for DUS questions in some years but observed a performance drop in 2018 and 2019.²⁵ Similarly, in our study, the rate of incorrect answers by chatbots increased in 2012 and 2018. This disparity might be due to the difficulty of the question content and variations in the number of questions.

When the performance of chatbots in the basic sciences of dentistry was evaluated by subject, accuracy rates were seen to range from 91% to 98.6%. The highest accuracy was in Pathology (98.6%), and the lowest was in Anatomy (91%). This result suggests that chatbots are more successful in areas based on conceptual knowledge and recognition (e.g., Pathology, Biochemistry and Histology) but are relatively limited in fields requiring visual memory and three-dimensional relational skills (e.g., Anatomy). The observed variation in error rates across basic science disciplines may be attributed to inherent differences in content structure and cognitive demands. Anatomy questions typically require strong spatial reasoning, three-dimensional visualization, and precise interpretation of anatomical relationships,



which may pose challenges for AI-based chatbot models that primarily rely on text-based knowledge representations. In contrast, Pathology questions are more frequently concept-driven and terminology-oriented, often emphasizing pattern recognition, definitions, and disease mechanisms, which may be more compatible with large language model architectures trained on extensive biomedical text corpora. High success in pathology and biochemistry can be explained by the models' proficiency with textual and theoretical information, whereas the lower accuracy in anatomy questions may be due to the models' limitations in processing visuospatial data and the fact that anatomical knowledge is often learned through images rather than text.

The performance comparison of chatbots showed that all four models achieved high accuracy rates on DUS basic sciences questions. The highest accuracy was recorded for Gemini Advanced 2.5 Pro at 97.6%, followed by ChatGPT-5 (95.5%), ChatGPT-4 omni (94.8%), and DeepSeek v3 (94.1%). Furthermore, no significant difference was found between Gemini Advanced 2.5 Pro, ChatGPT-5 and ChatGPT-4 omni. This indicates that the three models exhibit a similarly reliable level of performance on basic sciences questions. In a study similar to ours,²⁵ ChatGPT-5 achieved 93 % accuracy and Gemini 2.5 Pro achieved 96.9 % on DUS oral and maxillofacial surgery questions. Another study reported that Gemini Advanced showed the best performance among seven different chatbots in answering endodontics questions.²⁶ However, a 2025 study that had ChatGPT-4 omni and Gemini Advanced solve DUS questions from 2020 and 2021 reported that ChatGPT-4 omni performed better than Gemini Advanced.²⁰ This variance could be due to the fact that only two years of DUS questions were used or due to a different version of Gemini.

A study evaluating the accuracy of answers from ChatGPT-4 omni, ChatGPT-4, Gemini 1.5 Flash, Gemini 1.5 Pro, Gemini 2.0 Flash, Copilot, Deepseek-V3 and Qwen2.5-Max on DUS endodontics questions showed that DeepSeek v3 had a lower accuracy performance than ChatGPT-4 omni.²¹ In a study where UK medical board exam questions were posed to seven chatbots (ChatGPT-3.5, ChatGPT 4, Bard, Perplexity, Claude, Bing and Claude Instant), the accuracy varied significantly, with ChatGPT-4 generally performing the best. The authors suggested ChatGPT-4 as a secondary learning resource for medical students.¹⁹ However, another recent study indicated that DeepSeek-R1 could be a valuable aid in diagnosing oral diseases, outperforming ChatGPT-4o in this regard.²⁷ This shows that different versions of DeepSeek may have varying performance in different fields.

In a study evaluating the responses of chatbots [ChatGPT-3.5, ChatGPT-4 omni, Google Bard (now Gemini), Microsoft Copilot] to oral radiology questions from the DUS between 2012-2021, ChatGPT-4 omni (86.1%) was reported to have a higher accuracy rate than Gemini (61.8%).²² In our study, across all basic sciences, ChatGPT-4 omni achieved 94.8% and Gemini Advanced 2.5 Pro achieved 97.6% accuracy. The accuracy rate for both chatbots was significantly higher. This might be because oral radiology is more specific to dentistry, whereas basic medical sciences receive intensive attention from all medical-related fields, and chatbots may be more extensively trained on this broader subject matter.

Overall, the fact that all models demonstrated over 90% accuracy indicates that AI-powered chatbots have a very high potential for understanding the level of knowledge in dentistry. The findings suggest that AI-powered models can be a potential supplementary tool for DUS preparation. In particular, Gemini Advanced 2.5 Pro, ChatGPT-5 and ChatGPT-4 omni, with their high accuracy rates, can support students in topic review, concept clarification and question-solving practice. However, it should be noted that these systems cannot always correctly interpret the context of exam questions and can make conceptual errors in some subjects. Therefore, information obtained from these models should not be used alone without human supervision and caution should be exercised, especially in areas with high terminological complexity.

Limitations

One of the strengths of this study is its statistical analysis of chatbot performance on a yearly and subject-specific basis. This approach allowed for the measurement of the AI systems' knowledge level not just by overall accuracy but also according to different variables. Nevertheless, certain limitations of our study should be considered. Only text-based questions were evaluated, excluding those with images, and the models were not asked to provide justifications for their answers. This limited a detailed analysis of the models' reasoning processes and the causes of incorrect inferences. Additionally, the non-standardized number of questions across years and subjects directly impacted the analysis results. In this study, only 10% of the exam questions from 2022 onwards were published, which led to the evaluation of these exams together and prevented a separate analysis of them. Furthermore, asking the questions in Turkish instead of English may have contributed to the models providing incorrect answers.

CONCLUSION

In summary, AI-powered chatbots, particularly Gemini Advanced, ChatGPT-5 and ChatGPT-4 omni, hold the potential to be useful tools in basic sciences education for dentistry, especially in preparing for the DUS. However, to enhance the effectiveness of these systems, it is necessary to improve both the currency of their training data and their capabilities to include clinical and visual components. With such advancements, it is believed that AI models have the potential to evolve from merely assessing exam performance to becoming reliable educational assistants that support the learning process. Furthermore, asking correct and well-structured questions is vital for users to obtain reliable and accurate answers from these chatbots.

ETHICAL DECLARATIONS

Ethics Committee Approval

Since the study involved no human or animal subjects, clinical interventions, or identifiable patient data, ethics committee approval was not required.

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 authors declare no conflicts of interest related to this study.

Financial Disclosure

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

Author Contributions

Author Contributions Concept: M.Y.Ö.; Design: T.M., A.A., M.Y.Ö.; Control: M.Y.Ö.; Data Collection and/or Processing: T.M., A.A.; Analysis and/or Interpretation: M.Y.Ö.; Literature Review: T.M.; Article Writing: T.M.; Critical Review: M.Y.Ö.

Acknowledgments

This study was presented orally at the 'Genç Endodontistler Konuşuyor Sempozyumu'.

REFERENCES

- Xu L, Sanders L, Li K, et al. Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review. *JMIR Cancer*. 2021;7(4):e27850. doi:10.2196/27850
- Adamopoulou E, Moussiades L. An overview of chatbot technology. IFIP international conference on artificial intelligence applications and innovations. *Cham: Springer International Publishing*. 2020;373-383. doi:10.1007/978-3-030-49186-4_31
- Alhejaily A-MG. Artificial intelligence in healthcare. *Biomed Rep*. 2025; 22(1):1-8. doi:10.3892/br.2024.1889
- Reyes LT, Knorst JK, Ortiz FR, et al. Machine learning in the diagnosis and prognostic prediction of dental caries: a systematic review. *Caries Res*. 2022;56(3):161-70. doi:10.1159/000524167
- Wang YCC, Chen TL, Vinayahalingam S, et al. Artificial intelligence to assess dental findings from panoramic radiographs-a multinational study. *arXiv preprint arXiv*. 2025;250210277. doi:10.48550/arXiv.2502.10277
- Tyagi M, Jain S, Ranjan M, et al. Artificial intelligence tools in dentistry: a systematic review on their application and outcomes. *Cureus*. 2025; 17(5):e85062. doi:10.7759/cureus.85062
- Farook TH, Jamayet NB, Abdullah JY, et al. Machine learning and intelligent diagnostics in dental and orofacial pain management: a systematic review. *Pain Res Manag*. 2021;2021(1):6659133. doi:10.1155/2021/6659133
- Karobari MI, Adil AH, Basheer SN, et al. Evaluation of the diagnostic and prognostic accuracy of artificial intelligence in endodontic dentistry: a comprehensive review of literature. *Comput Math Methods Med*. 2023;2023(1):7049360. doi:10.1155/2023/7049360
- Al-Amin M, Ali MS, Salam A, et al. History of generative artificial intelligence (AI) chatbots: past, present, and future development. *arXiv preprint arXiv*. 2024;2402.05122. doi:10.48550/arXiv.2402.05122
- Esmailpour H, Rasaie V, Babaee Hemmati Y, et al. Performance of artificial intelligence chatbots in responding to the frequently asked questions of patients regarding dental prostheses. *BMC Oral Health*. 2025;25(1):574. doi:10.1186/s12903-025-05965-9
- Lin C-C, Huang AYQ, Yang SJH. A review of AI-driven conversational chatbots implementation methodologies and challenges (1999-2022). *Sustainability*. 2023;15(5):4012. doi:10.3390/su15054012
- Plevris V, Papazafeiropoulos G, Jiménez Rios A. Chatbots put to the test in math and logic problems: a comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard. *AI*. 2023;4(4):949-69.
- ÖSYM. Dentistry Specialization Education Entrance Exam. © Presidency of the Republic of Türkiye measuring, selection and placement center. 2025.
- Zhai X. ChatGPT user experience: Implications for education. Available at SSRN 2022;4312418. doi:10.2139/ssrn.4312418
- OpenAI. Introducing GPT-5. 2025.
- Mihalache A, Grad J, Patil NS, et al. Google Gemini and Bard artificial intelligence chatbot performance in ophthalmology knowledge assessment. *Eye (Lond)*. 2024;38(13):2530- 2535. doi:10.1038/s41433-024-03067-4
- Team G, Anil R, Borgeaud S, Alayrac J-B, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv*. 2023;231211805. doi:10.48550/arXiv.2312.11805
- Conroy G, Mallapaty S. How China created AI model DeepSeek and shocked the world. *Nature*. 2025;638(8050):300-301. doi:10.1038/d41586-025-00259-0
- Sadeq MA, Ghorab RMF, Ashry MH, et al. AI chatbots show promise but limitations on UK medical exam questions: a comparative performance study. *Sci Rep*. 2024;14(1):18859. doi:10.1038/s41598-024-68996-2
- Şismanoğlu S, Çapan 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
- Çekiç EC, Tavşan O. Evaluating large language models using national endodontic specialty examination questions: are they ready for real-world dentistry? *BMC Med Educ*. 2025;25(1):1308. doi:10.1186/s12909-025-07896-z
- Taşsöker 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
- Aşık A, Kuru E. Analysis of ChatGPT's answers to pedodontics questions asked in the dentistry specialization training entrance exam: cross-sectional study. *Türkiye Klinikleri J Dent Sci*. 2025;31(3):401-406. doi:10.5336/dentalsci.2024-107488
- Bilgin AD, Ertan A. A comparative study of ChatGPT-3.5 and Gemini's performance of answering the prosthetic dentistry questions in dentistry specialty exam: cross-sectional study. *Türkiye Klinikleri J Dent Sci*. 2024;30(4):668-673. doi:10.5336/dentalsci.2024-104610
- Çetiner EY. Comparative evaluation of ChatGPT-5 and Gemini 2.5 Pro in answering oral and maxillofacial surgery questions from dentistry specialization exams: a cross-sectional study. *EurAsian J Oral Maxillofac Surg*. 4(3):59-65.
- Jalali P, Mohammad-Rahimi H, Wang F-M, et al. Performance of seven artificial intelligence chatbots on board-style endodontic questions. *J Endod*. 2025;51(10):1413-1419. doi:10.1016/j.joen.2025.06.014
- Diniz-Freitas M, Diz-Dios P. DeepSeek: another step forward in the diagnosis of oral lesions. *J Dent Sci*. 2025;20(3):1904-1907. doi:10.1016/j.jds.2025.02.0