Acıbadem Univ. Sağlık Bilim. Derg. January, February, March 2026
https://doi.org/10.31067/acusaglik.1816444

ORIGINAL ARTICLE / ARAŞTIRMA YAZISI

Clinical Sciences/ Klinik Bilimler

Performance of Large Language Models on Official Periodontology Questions: A 13-Year Analysis of the Turkish Dental Specialization Examination

Fatih Karaaslan[1], **Yaren Erişken**[1]

[1] Usak University, Faculty of Dentistry, Department of Periodontology, Usak, Türkiye

ABSTRACT

Background: This study systematically evaluated the performance of large language models (LLMs) on official periodontology questions from the Turkish Dental Specialization Examination (DUS).

Methods: A total of 180 text-based questions (159 multiple-choice (MCQs), 21 combination-type MCQs (C-MCQs)) were categorized into nine domains across 13 years (2012–2024). In April 2025, eight LLMs were tested: ChatGPT-4o, ChatGPT-4o mini (OpenAI), Gemini 1.5 Flash, Gemini 1.5 Pro, Gemini 2.0 Flash (Google DeepMind), Copilot (Microsoft), DeepSeek-V3 (DeepSeek), and Qwen 2.5-Max (Alibaba Cloud). Each question was submitted independently via official interfaces. Accuracy rates were compared across models, domains, years, and question types using Pearson's chi-square test, with Cramér's V and Phi coefficients reported for effect sizes.

Results: Accuracy differed significantly by domain ($\chi^2(8, N = 1440) = 38.20$, $p < .001$, Cramér's $V = .163$). Gemini 2.5 Pro achieved the highest performance, scoring 100% in six domains and ≥87.5% in others. ChatGPT-4o mini and Qwen 2.5-Max underperformed, particularly in Periodontium and Periodontal Treatment. Year-based analysis showed stable performance across 2012–2024 ($\chi^2(12, N = 1440) = 14.51$, $p = .269$). No difference emerged between MCQs and C-MCQs ($\chi^2(1, N = 1440) = 1.42$, $p = .233$).

Conclusion: LLM accuracy in periodontology is domain- and model-dependent. Advanced systems such as Gemini 2.5 Pro show potential as supportive tools for education and clinical decision-making, yet persistent weaknesses in reasoning- and calculation-intensive areas underscore the need for expert oversight.

Keywords: Artificial Intelligence, Dental Education, Large Language Models, Periodontics

Fatih KARAASLAN
0000-0002-9899-3316
Yaren ERİŞKEN
0009-0004-8624-2621

ÖZET

Amaç: Bu çalışma, Türk Diş Hekimliği Uzmanlık Sınavı'nın (DUS) resmi periodontoloji sorularında büyük dil modellerinin (LLM'lerin) performansını sistematik olarak değerlendirmeyi amaçlamıştır.

Yöntem: Toplam 180 metin tabanlı soru (159 çoktan seçmeli (ÇSS), 21 kombinasyon tipi ÇSS (K-ÇSS)) 13 yıllık (2012–2024) süreçte dokuz alana ayrılmıştır. Nisan 2025'te sekiz LLM test edilmiştir: ChatGPT-4o, ChatGPT-4o mini (OpenAI), Gemini 1.5 Flash, Gemini 1.5 Pro, Gemini 2.0 Flash (Google DeepMind), Copilot (Microsoft), DeepSeek-V3 (DeepSeek) ve Qwen 2.5-Max (Alibaba Cloud). Her soru resmi web arayüzleri aracılığıyla bağımsız olarak gönderilmiştir. Doğruluk oranları modeller, alanlar, yıllar ve soru türleri arasında Pearson ki-kare testiyle karşılaştırılmış, etki büyüklükleri Cramér's V ve Phi katsayılarıyla hesaplanmıştır.

Bulgular: Doğruluk oranı alanlara göre anlamlı farklılık göstermiştir ($\chi^2(8, N = 1440) = 38.20$, $p < .001$, Cramér's $V = .163$). Gemini 2.5 Pro en yüksek performansı göstermiş, altı alanda %100, diğerlerinde ≥%87.5 doğruluk elde etmiştir. ChatGPT-4o mini ve Qwen 2.5-Max, özellikle Periodonsiyum ve Periodontal Tedavi alanlarında düşük performans sergilemiştir. Yıllara göre analizde 2012–2024 arasında istikrarlı sonuçlar elde edilmiştir ($\chi^2(12, N = 1440) = 14.51$, $p = .269$). ÇSS ve K-ÇSS arasında fark bulunmamıştır ($\chi^2(1, N = 1440) = 1.42$, $p = .233$).

Sonuç: Periodontolojide LLM doğruluğu modele ve alana bağlıdır. Gemini 2.5 Pro gibi ileri sistemler eğitim ve klinik karar desteğinde umut vaat etmektedir; ancak akıl yürütme ve hesaplama gerektiren alanlardaki zayıflıklar uzman gözetiminin gerekliliğini göstermektedir.

Anahtar Kelimeler: Büyük Dil Modelleri, Diş Hekimliği Eğitimi, Periodontoloji, Yapay Zekâ

Correspondence: Yaren Erişken
Usak University, Faculty of Dentistry, Department of Periodontology, Usak, Türkiye
Phone: +90 531 429 60 48
E-mail: yaren.erisken@usak.edu.tr

Received: 03.11.2025
Accepted: 11.12.2025

Copyright © 2025 the Author(s). Published by Acibadem University. This is an open access article licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives (CC BY-NC-ND 4.0) International License, which is downloadable, re-usable and distributable in any medium or format in unadapted form and for noncommercial purposes only where credit is given to the creator and publishing journal is cited properly. The work cannot be used commercially without permission from the journal.

Large Language Models (LLMs) are advanced artificial intelligence (AI) systems that simulate human language processing through deep learning and neural networks. Trained on vast corpora derived from books, scientific literature, and online sources, these models are increasingly integrated into medical and dental fields for purposes such as clinical decision support, administrative efficiency, patient communication, scientific writing, and multilingual interaction (1,2). Their growing adoption highlights both opportunities and challenges-particularly regarding their role as complementary, rather than substitutive, tools for expert judgment (3).

Periodontology, a core discipline of dentistry, focuses on the prevention, diagnosis, and treatment of diseases affecting the tooth-supporting structures. It encompasses non-surgical and surgical therapies, regenerative procedures, implantology, and systemic interactions. Accurate diagnosis and treatment planning in this field require not only factual recall but also higher-order clinical reasoning (4). In recent years, AI has shown promising applications in periodontology, including radiographic assessment of alveolar bone loss, diagnosis of periodontal and peri-implant diseases, detection of halitosis, and prediction of tooth prognosis (5,6). These developments suggest that LLMs—by synthesizing evidence-based knowledge and clinical guidelines—could contribute to periodontal education, research, and clinical decision-making.

In Turkey, postgraduate specialization in dentistry requires success in the Dentistry Specialty Examination (DUS), administered by the Student Selection and Placement Center (OSYM) since 2012. The exam consists of 120 items, of which 80 address clinical sciences across eight specialties, including periodontology. Because the DUS draws its content from internationally recognized textbooks designated by OSYM, it reflects globally accepted standards of dental knowledge and reasoning (7,8).

Although AI applications have achieved notable progress in diagnostic imaging and clinical decision support (9,10), their effectiveness in discipline-specific reasoning tasks remains insufficiently explored. Prior studies have demonstrated that LLM accuracy varies with model architecture, training data, question complexity, and disciplinary context (1,11,12). Moreover, little is known about their performance across distinct exam formats. While standard multiple-choice questions (MCQs) primarily test factual knowledge, combination multiple-choice questions (C-MCQs) require the simultaneous evaluation of multiple statements and therefore better reflect the integrative reasoning necessary in real-world periodontal decision-making. Assessing LLM performance in C-MCQs thus provides a rigorous test of their reasoning capabilities (15,13).

Although recent investigations have assessed AI models in endodontics, radiology, orthodontics, and oral surgery (11,12,14), no study to date has systematically evaluated LLM performance using official periodontology questions from the DUS. Furthermore, prior research has not clarified whether model performance varies depending on exam year or topic distribution, nor whether question-selection approaches can influence benchmarking outcomes—issues particularly relevant when analyzing longitudinal standardized examinations (1, 9-12). This study aims to address this gap by analyzing 180 periodontology questions (2012–2024) across multiple models, years, subject domains, and question types. Specifically, it investigates whether (1) performance differs between MCQs and C-MCQs, (2) accuracy varies across topics and examination years, and (3) certain models demonstrate consistent superiority. Ultimately, the study evaluates whether LLMs can be regarded as reliable knowledge sources in periodontal education and clinical decision-making, or whether their role should remain strictly supportive under expert oversight.

Methods

As no human participants or animals were involved and only publicly available examination data were analyzed, ethical approval was not required. This approach aligns with prior studies that evaluated LLMs on standardized test items, where institutional review was deemed unnecessary (1,9). Nevertheless, the study was conducted in accordance with the principles of the Declaration of Helsinki.

Dataset

A total of 180 official periodontology questions from the DUS, administered between 2012 and 2024, were included. Because each DUS examination contains exactly 10 periodontology questions, all available questions from each year were included without the need for randomization. Thus, the full set of periodontology items over the 13-year period was analyzed, ensuring complete

and unbiased longitudinal coverage. All questions were text-based and presented in Turkish, the original examination language. Among these, 159 were MCQs and 21 were C-MCQs.

Classification Procedure

To enable domain-specific performance analysis, all 180 questions underwent a structured classification process. Two periodontology specialists independently assigned each question to one of nine predefined categories based on its primary content focus. Discrepancies were resolved through consensus discussion. The final categories were:

(a) Periodontium (n = 29),
(b) Periodontal diseases and classification (n = 23),
(c) Periodontal pathology (n = 32),
(d) Etiology (n = 12),
(e) Systemic diseases (n = 8),
(f) Diagnosis and treatment planning (n = 13),
(g) Periodontal treatment (n = 28),
(h) Implantology (n = 12),
(i) Devices, materials, and drugs (n = 23).

Models and Evaluation Protocol

In April 2025, all questions were presented to eight state-of-the-art LLMs: ChatGPT-4o and ChatGPT-4o mini (OpenAI, San Francisco, CA, USA); Gemini 1.5 Flash, Gemini 1.5 Pro, and Gemini 2.0 Flash (Google DeepMind, London, UK); Copilot (Microsoft, Redmond, WA, USA); DeepSeek-V3 (DeepSeek, Beijing, China); and Qwen 2.5-Max (Alibaba Cloud, Hangzhou, China). Interactions were conducted exclusively through the official web-based graphical user interfaces of each platform, accessed via Google Chrome (Google LLC, Mountain View, CA, USA).

Prompt Standardization

To ensure uniformity across models, a strict prompt standardization protocol was applied. Each question was presented using an identical instruction template: "Please answer the following DUS periodontology question by indicating only the correct option. Do not provide any explanation." This standardized prompt was used for every model and every question without modification. No examples, clarifications, or chain-of-thought prompts were provided.

Response-Format Protocol

All models were required to produce a single-letter answer (A–E for MCQs; combinations such as AB, BCE for C-MCQs). Responses containing explanations, additional text, or ambiguous formatting were trimmed to extract only the final selected option(s). If a model produced multiple answers, only the first explicit choice was recorded.

Context Isolation

To prevent contextual carryover, each question was submitted in a new chat session, and the conversation history was manually cleared before entering the next question This ensured isolated evaluation of each item. This protocol mirrors established benchmarking approaches emphasizing independence of inputs (11,12).

Statistical Analysis

All analyses were performed using IBM SPSS Statistics, version 23.0 (IBM Corp., Armonk, NY, USA). Descriptive statistics were expressed as frequencies (n) and percentages (%). Accuracy rates were compared across models, question types, years, and subject domains using Pearson's chi-square (χ^2) test. For 2×2 contingency tables, Yates' continuity correction was applied. Post hoc multiple comparisons were performed using Bonferroni-adjusted z-tests to minimize Type I error. Effect sizes were reported using the Phi (Φ) coefficient for dichotomous variables and Cramér's V for multi-category variables, interpreted according to Cohen's thresholds ($|\Phi| \approx 0.10$ = small, $|\Phi| \approx 0.30$ = moderate, $|\Phi| \geq 0.50$ = large). Statistical significance was set at $p < 0.05$.

Results

Model performance differed significantly across periodontal domains ($\chi^2(8, N = 1440) = 38.20$, $p < .001$, Cramér's $V = .163$), indicating a small-to-moderate association between subject area and accuracy (Table 1). Gemini 2.5 Pro achieved the highest overall performance, with 100% accuracy in six of nine categories and strong performance (87.5–89.3%) in the remaining domains. In contrast, ChatGPT-4o mini and Qwen 2.5-Max consistently showed lower accuracies, particularly in Periodontium and Periodontal Treatment, underscoring the domain-specific variability of LLM capabilities.

Table 1. Accuracy rates of LLM responses across different topics

Topic	Model	True	False	Test Statistics	p
Periodontium	ChatGPT-4o mini	18 (62.1)	11 (37.9)	18.694	0.009
	Copilot	24 (82.8)	5 (17.2)		
	DeepSeek-V3	23 (79.3)	6 (20.7)		
	GPT-4o	26 (89.7)	3 (10.3)		
	Gemini 2.0 Flash	26 (89.7)	3 (10.3)		
	Gemini 2.5 Flash	24 (82.8)	5 (17.2)		
	Gemini 2.5 Pro	29 (100.0)	NA		
	Qwen 2.5-Max	21 (72.4)	8 (27.6)		
Periodontal diseases and classification	ChatGPT-4o mini	16 (69.6)	7 (30.4)	17.054	0.017
	Copilot	22 (95.7)	1 (4.3)		
	DeepSeek-V3	19 (82.6)	4 (17.4)		
	GPT-4o	21 (91.3)	2 (8.7)		
	Gemini 2.0 Flash	20 (87.0)	3 (13.0)		
	Gemini 2.5 Flash	23 (100.0)	NA		
	Gemini 2.5 Pro	23 (100.0)	NA		
	Qwen 2.5-Max	20 (87.0)	3 (13.0)		
Periodontal pathology	ChatGPT-4o mini	30 (93.8)	2 (6.2)	21.612	0.003
	Copilot	31 (96.9)	1 (3.1)		
	DeepSeek-V3	29 (90.6)	3 (9.4)		
	GPT-4o	32 (100.0)	NA		
	Gemini 2.0 Flash	28 (87.5)	4 (12.5)		
	Gemini 2.5 Flash	32 (100.0)	NA		
	Gemini 2.5 Pro	32 (100.0)	NA		
	Qwen 2.5-Max	25 (78.1)	7 (21.9)		
Etiology	ChatGPT-4o mini	9 (75.0)	3 (25.0)	4.826	0.681
	Copilot	11 (91.7)	1 (8.3)		
	DeepSeek-V3	11 (91.7)	1 (8.3)		
	GPT-4o	11 (91.7)	1 (8.3)		
	Gemini 2.0 Flash	10 (83.3)	2 (16.7)		
	Gemini 2.5 Flash	11 (91.7)	1 (8.3)		
	Gemini 2.5 Pro	12 (100.0)	NA		
	Qwen 2.5-Max	10 (83.3)	2 (16.7)		
Systemic diseases	ChatGPT-4o mini	8 (100.0)	NA	8.533	0.288
	Copilot	8 (100.0)	NA		
	DeepSeek-V3	7 (87.5)	1 (12.5)		
	GPT-4o	8 (100.0)	NA		
	Gemini 2.0 Flash	8 (100.0)	NA		
	Gemini 2.5 Flash	8 (100.0)	NA		
	Gemini 2.5 Pro	7 (87.5)	1 (12.5)		
	Qwen 2.5-Max	6 (75.0)	2 (25.0)		
Diagnosis and treatment plan	ChatGPT-4o mini	8 (61.5)	5 (38.5)	4.381	0.735
	Copilot	10 (76.9)	3 (23.1)		
	DeepSeek-V3	11 (84.6)	2 (15.4)		
	GPT-4o	11 (84.6)	2 (15.4)		
	Gemini 2.0 Flash	11 (84.6)	2 (15.4)		
	Gemini 2.5 Flash	11 (84.6)	2 (15.4)		
	Gemini 2.5 Pro	11 (84.6)	2 (15.4)		
	Qwen 2.5-Max	9 (69.2)	4 (30.8)		

Periodontal treatment	ChatGPT-4o mini	16 (57.1)	12 (42.9)	16.88	0.018
	Copilot	21 (75.0)	7 (25.0)		
	DeepSeek-V3	24 (85.7)	4 (14.3)		
	GPT-4o	26 (92.9)	2 (7.1)		
	Gemini 2.0 Flash	21 (75.0)	7 (25.0)		
	Gemini 2.5 Flash	25 (89.3)	3 (10.7)		
	Gemini 2.5 Pro	25 (89.3)	3 (10.7)		
	Qwen 2.5-Max	21 (75.0)	7 (25.0)		
Implantology	ChatGPT-4o mini	8 (66.7)	4 (33.3)	13.333	0.064
	Copilot	8 (66.7)	4 (33.3)		
	DeepSeek-V3	8 (66.7)	4 (33.3)		
	GPT-4o	10 (83.3)	2 (16.7)		
	Gemini 2.0 Flash	7 (58.3)	5 (41.7)		
	Gemini 2.5 Flash	12 (100.0)	NA		
	Gemini 2.5 Pro	12 (100.0)	NA		
	Qwen 2.5-Max	7 (58.3)	5 (41.7)		
Devices / materials / drugs	ChatGPT-4o mini	17 (73.9)	6 (26.1)	13.713	0.057
	Copilot	21 (91.3)	2 (8.7)		
	DeepSeek-V3	17 (73.9)	6 (26.1)		
	GPT-4o	21 (91.3)	2 (8.7)		
	Gemini 2.0 Flash	19 (82.6)	4 (17.4)		
	Gemini 2.5 Flash	22 (95.7)	1 (4.3)		
	Gemini 2.5 Pro	22 (95.7)	1 (4.3)		
	Qwen 2.5-Max	16 (69.6)	7 (30.4)		

Table 2. Accuracy rates of LLM responses cross Years (2012-2024)

Year	GPT-4o	DeepSeek-V3	Copilot	Gemini 2.0 Flash	Qwen 2.5-Max	ChatGPT-4o mini	Gemini 2.5 Flash	Gemini 2.5 Pro	Total Questions
2012	90.0	85.0	95.0	85.0	80.0	90.0	95.0	100.0	20
2013	88.9	83.3	88.9	83.3	77.8	88.9	94.4	94.4	36
2014	87.5	81.3	87.5	81.3	75.0	87.5	93.8	93.8	32
2015	90.0	90.0	90.0	90.0	90.0	90.0	90.0	90.0	20
2016	88.9	88.9	77.8	77.8	77.8	88.9	88.9	88.9	18
2017	90.0	90.0	90.0	80.0	80.0	90.0	90.0	90.0	20
2018	87.5	87.5	87.5	87.5	87.5	87.5	87.5	87.5	16
2019	90.0	85.0	85.0	80.0	80.0	85.0	90.0	95.0	20
2020	90.0	80.0	80.0	80.0	80.0	80.0	90.0	90.0	20
2021	88.9	88.9	77.8	77.8	77.8	88.9	88.9	88.9	18
2022	87.5	75.0	87.5	75.0	75.0	87.5	87.5	87.5	16
2023	85.7	78.6	78.6	78.6	78.6	85.7	92.9	92.9	28
2024	88.2	82.4	82.4	76.5	76.5	88.2	94.1	94.1	34

Note. All values represent percentage accuracy rates

Across the 2012–2024 examination period, accuracy rates ranged from 75% to 100% depending on the model (Table 2). Gemini 2.5 Pro consistently outperformed all other systems, maintaining accuracy above 90% annually and achieving perfect scores in selected years. Minor fluctuations were observed in Copilot, Qwen 2.5-Max, and ChatGPT-4o mini; however, these differences were not statistically significant ($\chi^2(12, N = 1440) = 14.51$, $p = .269$, Cramér's $V = .28$). Overall, the longitudinal analysis suggests that LLM performance remained stable over time, with year-to-year variability exerting minimal influence compared to intrinsic model architecture and training.

Table 3. The accuracy rates of LLM responses based on different question types

Question Style	Model	True (n, %)	False (n, %)	Test Statistics	p
MCQs	ChatGPT-4o mini	15 (71.4)	6 (28.6)	7.629	0.366
	Copilot	17 (81.0)	4 (19.0)		
	DeepSeek-V3	17 (81.0)	4 (19.0)		
	GPT-4o	20 (95.2)	1 (4.8)		
	Gemini 2.0 Flash	17 (81.0)	4 (19.0)		
	Gemini 2.5 Flash	16 (76.2)	5 (23.8)		
	Gemini 2.5 Pro	20 (95.2)	1 (4.8)		
	Qwen 2.5-Max	16 (76.2)	5 (23.8)		
C-MCQs	ChatGPT-4o mini	115 (72.3)	44 (27.7)	71.940	<0.001
	Copilot	139 (87.4)	20 (12.6)		
	DeepSeek-V3	132 (83.0)	27 (17.0)		
	GPT-4o	146 (91.8)	13 (8.2)		
	Gemini 2.0 Flash	133 (83.6)	26 (16.4)		
	Gemini 2.5 Flash	152 (95.6)	7 (4.4)		
	Gemini 2.5 Pro	153 (96.2)	6 (3.8)		
	Qwen 2.5-Max	119 (74.8)	40 (25.2)		
Total Questions	ChatGPT-4o mini	130 (72.2)	50 (27.8)	74.047	<0.001
	Copilot	156 (86.7)	24 (13.3)		
	DeepSeek-V3	149 (82.8)	31 (17.2)		
	GPT-4o	166 (92.2)	14 (7.8)		
	Gemini 2.0 Flash	150 (83.3)	30 (16.7)		
	Gemini 2.5 Flash	168 (93.3)	12 (6.7)		
	Gemini 2.5 Pro	173 (96.1)	7 (3.9)		
	Qwen 2.5-Max	135 (75.0)	45 (25.0)		

When stratified by format, LLMs demonstrated broadly consistent accuracy across MCQs and C-MCQs. The chi-square test showed no significant effect of question type on performance ($\chi^2(1, N = 1440) = 1.42$, $p = .233$, $\Phi = .031$), indicating that format alone did not systematically influence results (Table 3). Nevertheless, model-level differences were evident: Gemini 2.5 Pro (96.1%) and Gemini 2.5 Flash (93.3%) consistently outperformed lower-performing systems such as ChatGPT-4o mini (72.2%) and Qwen 2.5-Max (75.0). These findings highlight that architecture and training strategies, rather than item format, are the primary determinants of LLM accuracy.

Discussion

To our knowledge, this is the first study to systematically benchmark multiple LLMs using official periodontology questions from a national specialty examination. The results demonstrated substantial variability in accuracy depending on the model, subject domain, and question format. These findings are in line with prior research in medicine and dentistry, which consistently reported that AI performance fluctuates across item types and knowledge areas (1,9,13).

Among the subject domains, implantology yielded the lowest accuracy. This is likely attributable to the numerical and procedural demands of the field—such as implant torque thresholds, grafting protocols, and healing timelines—that exceed the current reasoning and quantitative capacities of LLMs (12, 13). In contrast, systemic disease–related questions achieved the highest accuracy, reflecting the models' relative strength in fact-based, guideline-driven domains where knowledge recall predominates (2, 10). Similar patterns have been observed in endodontics and oral surgery, where LLMs perform well in foundational knowledge tasks but underperform in items requiring procedural reasoning and sequential decision-making (8, 14).

Furthermore, C-MCQs accentuated performance gaps between models compared to standard MCQs. This reinforces the well-documented limitations of LLMs in handling multi-step reasoning, logical consistency, and integrative judgment (2, 5, 13). Prior studies similarly noted that while LLMs are effective in simulating recall-based responses, they struggle with questions requiring structured reasoning and logical chaining—competencies that are central to advanced clinical decision-making in periodontology (15, 16).

The topic-level analysis revealed pronounced variability in model accuracy, confirming that LLM performance is highly context dependent. Gemini 2.5 Pro consistently outperformed all other systems, achieving 100% accuracy in six of nine domains and maintaining high accuracy (87.5–89.3%) in the remainder. In contrast, ChatGPT-4o mini and Qwen 2.5-Max displayed markedly lower accuracies in critical categories such as Periodontium and Periodontal Treatment, underscoring the heterogeneity of model reliability across subject areas. These discrepancies align with prior findings in medical and dental education, where domain-specific challenges frequently expose the contextual and reasoning limitations of LLMs (2, 5, 17).

The variability between domains appears to reflect their distinct cognitive and procedural demands. Knowledge-driven categories such as Periodontal Pathology and Systemic Diseases primarily involve factual recall and adherence to established guidelines—tasks that LLMs are comparatively adept at handling (1). By contrast, procedural domains such as Implantology and Periodontal Treatment require the integration of clinical protocols, quantitative parameters (e.g., bone levels, implant success thresholds), and multi-step decision-making. These areas represent persistent weaknesses of current LLMs, which often falter when numerical reasoning and sequential logic are required (5, 9). Notably, these topic-specific discrepancies mirror earlier observations in related fields. Notably, these topic-specific discrepancies mirror earlier observations in related fields. Prior work has consistently shown that LLMs perform well on knowledge-based exam items but struggle with multi-step clinical reasoning. For instance, Kung et al. (18) reported that ChatGPT achieved higher accuracy on recall-oriented USMLE items but showed limitations in sequential reasoning, while Puleio et al. (19) similarly emphasized weaknesses in clinical decision-making tasks within dentistry. Our findings extend these patterns to the Turkish DUS setting, indicating that the distinction between "knowledge-based" and "procedural/decision-based" domains is consistent across both languages and specialties. From an educational perspective, these results suggest that LLMs can reliably assist students in consolidating foundational periodontal knowledge, particularly for topics emphasizing definitions, classifications, and systemic interactions. However, their role in procedural or treatment-planning contexts should remain limited, as errors in implantology or periodontal therapy may foster misconceptions and unwarranted confidence. From a clinical standpoint, the findings highlight the need for strict expert oversight: while advanced systems such as Gemini 2.5 Pro and GPT-4o show potential as adjunctive decision-support tools, lower-performing models (e.g., Qwen 2.5-Max, ChatGPT-4o mini) cannot be safely relied upon in high-stakes clinical decision-making, where patient safety is paramount.

The longitudinal analysis demonstrated that LLM performance remained remarkably stable across the 2012–2024 examination period. Accuracy rates ranged from 75% to 100% depending on the model and year, with Gemini 2.5 Pro consistently outperforming all others by maintaining >90% accuracy annually and achieving perfect scores in selected years. Although minor fluctuations were observed in models such as Copilot, Qwen 2.5-Max, and ChatGPT-4o mini, chi-square analysis confirmed these differences were not statistically significant. The small effect size reinforces the interpretation that model performance has been largely stable over time. This temporal stability indicates that the reliability of LLMs is primarily determined by intrinsic design factors—including architecture, training corpus scale, and fine-tuning strategies—rather than annual variations in examination content. Similar patterns have been reported across domains. For example, Kung et al. (18) showed that ChatGPT's accuracy in USMLE-style exams remained consistent across testing blocks despite variations in question focus. In dentistry, studies evaluating endodontic and periodontal content have likewise demonstrated year-to-year reproducibility of AI performance, underscoring the durability of these systems across exam cycles (1, 2). Interestingly, the fluctuations observed in weaker models (e.g., Qwen 2.5-Max, Copilot) mirror earlier findings where smaller or less specialized systems were more sensitive to question framing and dataset variation (10). This suggests that stability is not evenly distributed: while advanced models such as Gemini 2.5 Pro and GPT-4o exhibit robust performance regardless of exam year, lighter-weight systems remain vulnerable to contextual variability. From an educational perspective, these results indicate that LLMs can provide dependable

support for exam preparation across different years, offering students reproducible accuracy independent of the testing cycle. From a clinical perspective, the sustained stability of robust models underscores their potential as reliable adjuncts in longitudinal decision support, where consistency is critical for patient management. By contrast, weaker models demand cautious integration, as year-to-year fluctuations may undermine trustworthiness in real-world clinical contexts.

When stratified by question format, the models demonstrated broadly consistent accuracy across standard MCQs and C-MCQs. The chi-square test revealed no significant effect of question type, indicating that format alone did not systematically influence accuracy. However, closer inspection at the model level uncovered substantial disparities. Gemini 2.5 Pro and Gemini 2.5 Flash consistently outperformed weaker systems such as ChatGPT-4o mini and Qwen 2.5-Max. These results suggest that while LLMs as a group may appear resilient to differences in format, their effectiveness is ultimately determined by model architecture, training strategies, and fine-tuning approaches.

From an educational perspective, the absence of format-driven decline contrasts with prior research, where C-MCQs were shown to reduce accuracy due to their demand for multi-step reasoning and logical consistency (2, 13). This discrepancy in the present study may be partly attributable to the small number of C-MCQs, which limits the statistical power to detect subtle differences. Nevertheless, the consistent superiority of Gemini 2.5 Pro highlights that advanced models are better equipped for reasoning-intensive tasks such as integrating diagnostic cues, weighing prognostic indicators, and balancing treatment alternatives—skills that are fundamental in periodontal practice (1).

By contrast, weaker models frequently defaulted to oversimplified reasoning or produced partially correct but misleading outputs, a limitation also reported in implantology and radiology application (12). Clinically, this underscores the importance of selective adoption: while robust models may serve as valuable adjuncts in training complex decision-making, reliance on weaker systems without expert oversight risks undermining both educational outcomes and patient care.

This study has several limitations. First, only text-based DUS periodontology questions were included, while visual and radiographic items—essential for diagnosis—were excluded, limiting clinical applicability. Second, the number of C-MCQs was relatively small, which may have reduced statistical power to detect format-related differences. Third, all evaluations were performed via web-based interfaces, which may differ from API-based environments in prompt handling and memory management. Fourth, rapid model updates limit reproducibility, as performance benchmarks can quickly become outdated. Finally, the study did not incorporate human control groups which could provide a direct benchmark against LLM performance. Although outside the scope of the present design, future studies incorporating human–LLM comparative analyses would offer stronger interpretability and educational relevance.

Conclusion

This study is the first to systematically evaluate LLMs on official DUS periodontology questions. Results showed significant variability by topic and format, with Gemini 2.5 Pro achieving the highest accuracy. Implantology and C-MCQs highlighted persistent weaknesses, reflecting current limitations in multi-step reasoning and quantitative tasks. Overall performance was stable across years, indicating that LLMs are best viewed as supportive—not standalone—tools in education and clinical decision-making. Their optimal role is to complement expert oversight, reinforce factual knowledge, and aid exam preparation, while avoiding over-reliance. With advances in multimodal integration, multilingual benchmarking, and reasoning, LLMs may evolve into reliable partners in periodontal education and practice.

Declarations

Funding

This study received no funding.

Conflicts of interest / Competing interests

The authors declare that they have no financial or personal relationships with any of the companies that were developing the large language models evaluated in this study (OpenAI, Google DeepMind, Microsoft, Alibaba Cloud, or DeepSeek). The study was conducted independently for academic and scientific purposes only.

Ethics approval

Not applicable.

Availability of data and material

All data used in this study are publicly available and were obtained from the official Turkish Dental Specialization Examination (DUS) sources.

Authors' contributions

F.K. conceived the research topic. Y.E. collected the data. F.K. and Y.E. co-wrote the manuscript and approved the final version.

Acknowledgements

This study was not supported by any grants, funds, or institutional projects. The authors would like to thank the Usak University Faculty of Dentistry for its academic environment supporting AI-related research.

References

1. Kamath P, Kamath P, Saldanha SJR, et al. A brief exploration of artificial intelligence in dental healthcare: a narrative review. F1000Res. 2024;13:37 DOI:10.12688/f1000research.140481.2
2. El-Hakim M, Anthonappa R and Fawzy A. Artificial intelligence in dental education: a scoping review of applications, challenges, and gaps. Dent J. 2025;13:384. DOI:10.3390/dj13090384
3. Patil R and Gudivada V. A review of current trends, techniques, and challenges in large language models (LLMs). Appl Sci. 2024;14:2074. DOI:10.3390/app14052074
4. Ahmad P, Asif JA, Alam MK et al. A bibliometric analysis of Periodontology 2000. Periodontol 2000. 2020;82:286–97. DOI:10.1111/prd.12328
5. You W, Hao A, Li S, et al. Deep learning-based dental plaque detection on primary teeth: a comparison with clinical assessments. *BMC Oral Health*. 2020;20:141. DOI:10.1186/s12903-020-01114-6
6. Lin CC, Sun JS, Chang CH, et al. Performance of artificial intelligence chatbots in national dental licensing examination. J Dent Sci. 2025 DOI:10.1016/j.jds.2025.05.012
7. Ölçme, Seçme ve Yerleştirme Merkezi (ÖSYM). DUS çıkmış sorular. Available from: http://www.osym.gov.tr/TR,15070/dus-cikmis-sorular.html (Accessed April 21, 2025).
8. Sismanoglu S and 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;25:214 DOI:10.1186/s12909-024-06389-9
9. Naik S, Al-Kheraif AA and Vellappally S. Artificial intelligence in dentistry: assessing the informational quality of YouTube videos. PLoS One. 2025;20:e0316635. DOI:10.1371/journal.pone.0316635
10. Louca C, Tonni I, Leung A,et al. Artificial intelligence: friend or foe in the assessment of dental students?. J Dent. 2025;156:105676. DOI:10.1016/j.jdent.2025.105676
11. El-Hakim M, Anthonappa R and Fawzy A. Artificial intelligence in dental education: a scoping review of applications, challenges, and gaps. Dent J. 2025;13:384. DOI:10.3390/dj13090384
12. Bahadir HS, Keskin NB, Çakmak EŞK, et al. Comparison of diagnoses made by dentistry students and by artificial intelligence dentists. J Dent Educ. 2025;89:1165-73. DOI:10.1002/jdd.13810
13. Prakash K and Prakash R. An artificial intelligence-based dental semantic search engine as a reliable tool for dental students and educators. J Dent Educ. 2024;88:1257–66. DOI:10.1002/jdd.13560
14. Qutieshat A, Al Rusheidi A, Al Ghammari S et al. Comparative analysis of diagnostic accuracy in endodontic assessments: dental students vs. artificial intelligence. Diagnosis (Berl). 2024;11:259–65. DOI:10.1515/dx-2024-0034
15. Ardestani M, Kamalloo E and Rafiei D. LongRecall: a structured approach for robust recall evaluation in long-form text. arXiv preprint arXiv:2508.15085. 2025. DOI:10.48550/arXiv.2508.15085
16. Wang Y, Cui H and Kleinberg J. Microstructures and accuracy of graph recall by large language models. Adv Neural Inf Process Syst. 2024;37:65154-82. DOI:10.48550/arXiv.2402.11821
17. Lucas HC, Upperman JS and Robinson JR. A systematic review of large language models and their implications in medical education. Med Educ. 2024;58:1276–85. DOI:10.1111/medu.15402
18. 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:e0000198. DOI:10.1371/journal.pdig.0000198
19. Puleio F, Lo Giudice G, Bellocchio AM et al. Clinical, research, and educational applications of ChatGPT in dentistry: a narrative review. Appl Sci. 2024;14:10802. DOI:10.3390/app142310802