Scientific Reports

https://doi.org/10.1038/s41598-026-62380-y

Article in Press

Performance of large language models in a high-stakes dental assessment: evidence from the Turkish dentistry specialization examination

Received: 26 February 2026

Accepted: 10 July 2026

Published online: 13 July 2026

Cite this article as: Güçlü M., Yelbay M. & Karadağ İ. Performance of large language models in a high-stakes dental assessment: evidence from the Turkish dentistry specialization examination. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-62380-y

Merter Güçlü, Murtaza Yelbay & İlkim Karadağ

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

ARTICLE IN PRESS

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

ARTICLE IN PRESS

Performance of Large Language Models in a High-Stakes Dental Assessment: Evidence from the Turkish Dentistry Specialization Examination

Merter GÜÇLÜ*, Murtaza YELBAY*, İlkim KARADAĞ**

Yozgat Bozok University Faculty of Dentistry, Department of Periodontology, Yozgat, Turkey*

Ankara University Faculty of Dentistry, Department of Periodontology, Ankara, Turkey**

Corresponding author: Merter Güçlü

Email: merter.guclu@bozok.edu.tr

Abstract

Background: Large language models (LLMs) are increasingly used in dental education; however, their performance on high-stakes national examinations remains insufficiently explored. The Dentistry Specialization Examination (DUS) in Türkiye is a standardized gateway to postgraduate dental training and includes a dedicated periodontology component. No previous study has evaluated contemporary LLM accuracy on DUS periodontology questions or examined whether Bloom's Revised Taxonomy predicts model performance.

Methods: A total of 128 validated periodontology single-best-answer multiple-choice questions from publicly released DUS booklets (2012–2021) were analyzed. Each question was presented in Turkish to ChatGPT-5, Gemini 2.5 Pro, Claude Sonnet 4.5, and DeepSeek v3.2 using a standardized single-answer prompt. Responses were scored against the official key. Two periodontology specialists independently classified items according to Bloom's Revised Taxonomy and eight periodontal content domains. Accuracy differences were assessed using Cochran's Q test and Bonferroni-adjusted McNemar tests. The association between Bloom level and accuracy was examined using binary logistic regression.

ARTICLE IN PRESS

Results: Accuracy rates ranged from 80.47% to 91.41% across the four large language models. Incorrect responses were distributed across content domains, with a greater number observed in the non-surgical and surgical periodontal therapy domains. Regarding cognitive demand, inter-rater reliability for Bloom classification was high ($\kappa = 0.948$); however, Bloom level was not a statistically significant predictor of model accuracy ($p > 0.05$). Differences in performance across models were observed; a statistically significant difference was identified only between Gemini 2.5 Pro and DeepSeek v3.2 ($p = 0.0081$).

Conclusion: Contemporary LLMs showed high accuracy on released DUS periodontology single-best-answer multiple-choice items. While performance was strong in this structured, text-based examination context, the ability of these models to replicate higher-order clinical reasoning remains uncertain. Within the context of the analyzed DUS periodontology dataset, these findings contribute to ongoing discussions regarding the design and potential vulnerabilities of high-stakes single-best-answer MCQ-based assessments in the era of artificial intelligence. Consequently, it may be important for educators to understand the potential role of LLMs in students' preparation for assessments, suggesting that future research may explore modifications in test development to better capture student understanding.

Keywords: Dental education, Artificial intelligence, Large language models, Periodontology

Background

Artificial intelligence (AI) systems are increasingly capable of performing cognitive tasks traditionally associated with human reasoning, including information synthesis, pattern recognition, and structured problem solving (1). Recent advances in deep learning architectures

ARTICLE IN PRESS

have significantly expanded the application of AI in biomedical domains, enabling systems not only to process complex datasets but also to generate context-sensitive responses in educational and clinical settings (2). Among these developments, large language models (LLMs) such as ChatGPT, Gemini, and Claude have demonstrated remarkable proficiency in answering knowledge-based questions across multiple disciplines (3). As these models become integrated into digital learning environments, their influence extends beyond information retrieval to potentially reshaping assessment practices in health professions education (4,5).

The rapid improvement in LLM performance has raised critical questions regarding the robustness and validity of traditional high-stakes examinations. In dentistry, several studies have reported variable yet often high accuracy rates of AI-based language models when responding to standardized knowledge assessments (6,7). Specialty-specific analyses, including periodontology-focused evaluations, suggest that model performance may differ across domains, reflecting both the strengths and constraints of discipline-specific reasoning (7,8). At the same time, AI tools have been shown to support conceptual understanding in areas such as dental trauma and endodontics (9). While these findings primarily examine model performance, they simultaneously prompt a broader concern: if AI systems can reliably answer multiple-choice examination questions, what does this imply about the cognitive depth and discriminatory capacity of such assessments?

In Türkiye, the Dentistry Specialization Examination (DUS) functions as the national high-stakes gateway to postgraduate dental training (10). Structured as a multiple-choice examination covering both foundational sciences and eight clinical disciplines, DUS plays a decisive role in specialty placement and professional trajectory (11). Periodontology represents a core discipline within this structure, encompassing diagnostic reasoning, therapeutic planning, and management of

ARTICLE IN PRESS

periodontal and peri-implant diseases. Given its clinical centrality, performance in periodontology is expected to reflect meaningful disciplinary competence. Evaluating LLM accuracy within this context therefore offers not only insight into AI capability, but also an opportunity to examine how contemporary AI systems interact with the structural characteristics of a national licensing assessment. Periodontology was selected as it encompasses a broad range of foundational and clinical domains, including microbiology, histopathology, diagnostic reasoning, and both surgical and non-surgical treatment approaches, as well as oral implantology, thereby enabling a multidimensional evaluation of LLM performance within a single specialty. However, performance may vary across different dental disciplines, and the findings should therefore be interpreted within the context of periodontology.

Cognitive complexity constitutes a fundamental dimension of examination validity. Bloom's Revised Taxonomy provides a hierarchical framework ranging from lower-order processes such as remembering and understanding to higher-order skills including analysis and evaluation. Prior educational research indicates that many health science multiple-choice examinations disproportionately emphasize lower-order cognitive levels, potentially favoring pattern recognition over integrative reasoning (12). Emerging commentary has suggested that the rise of LLMs necessitates a reassessment of assessment design, particularly in relation to cognitive demand and construct validity (13). However, neither international dental licensing examinations nor the Turkish DUS have been systematically analyzed in terms of Bloom-level cognitive distribution, and no prior investigation has examined whether cognitive demand predicts LLM performance within a periodontology-focused high-stakes assessment.

ARTICLE IN PRESS

Accordingly, the present study was designed not merely as a comparison of language models, but as an examination-focused evaluation of a national specialty examination in the era of artificial intelligence. We assessed the performance of four contemporary LLMs (ChatGPT-5, Gemini 2.5 Pro, Claude Sonnet 4.5, and DeepSeek v3.2) on periodontology questions from the Turkish Dentistry Specialization Examination (DUS). Beyond overall accuracy comparisons, we examined whether Bloom's Revised Taxonomy level predicted model performance and characterized domain-specific error patterns across periodontal subfields. By integrating cognitive stratification and subdomain analysis, this study aims to contribute to the ongoing discourse on the potential vulnerabilities and future design of high-stakes dental assessments in an AI-augmented educational landscape, within the context of the analyzed DUS periodontology dataset. These findings suggest that LLMs may have the potential to function as preparatory tools for high-stakes examinations, thereby indicating a potential need for future research exploring modifications in test development to more accurately capture true student understanding.

ARTICLE IN PRESS

Methods

Study design

This cross-sectional analytical study was designed as an examination-focused evaluation of a national high-stakes examination under AI-based response conditions. Rather than serving solely as a model comparison, the study framework treated large language models (LLMs) as standardized external respondents to examine the structural and cognitive characteristics of periodontology items within the Turkish Dentistry Specialization Examination (DUS). Because the study relied exclusively on publicly available examination materials and did not involve human participants, patient data, or identifiable information, ethical approval was not required. All

ARTICLE IN PRESS

analyses were conducted on archived, publicly released content to ensure transparency and reproducibility. Importantly, the objective of this study was not to optimize or benchmark individual model performance, but to evaluate how contemporary LLMs interact with the structural and cognitive properties of a high-stakes examination under standardized conditions. The primary endpoint of this study was the overall accuracy of each LLM on the DUS periodontology item corpus. Secondary and exploratory analyses included the association between Bloom's cognitive level and model accuracy, and the descriptive mapping of error distribution across periodontal content domains. Accuracy comparisons across models were treated as confirmatory analyses, whereas Bloom-level and domain-based analyses were considered exploratory in nature.

Data source and item selection

A total of 130 periodontology single-best-answer multiple-choice questions administered in official DUS examinations between 2012 and 2021 were obtained from publicly accessible booklets released by the Assessment, Selection and Placement Center (ÖSYM). These examination years were included because they represent the complete set of DUS question booklets officially released online; subsequent examinations are not publicly available and therefore could not be analyzed. Two annulled items were excluded, resulting in a final dataset of 128 valid single-best-answer questions. The use of an entire released corpus minimized selection bias and enabled comprehensive content-level analysis of the periodontology component during the specified period.

ARTICLE IN PRESS

Content-domain Mapping

To evaluate the structural distribution of the examination, all 128 items were classified into eight predefined periodontology content domains based on DUS curricular objectives and standard specialty training frameworks: biological foundations of periodontology, periodontal pathology, gingival diseases, periodontal diseases, periodontal diagnosis and clinical evaluation, non-surgical periodontal therapy, surgical periodontal therapy, and oral implantology. Two board-certified periodontology specialists independently reviewed each item and assigned domain labels according to the primary knowledge and decision-making construct assessed in the question stem and answer options. Disagreements were resolved through consensus discussion. This classification enabled domain-level descriptive analysis and facilitated mapping of model errors across disciplinary areas, allowing exploration of potential variation in performance across content domains.

Large Language Models and Testing Environment

Four contemporary large language models were evaluated: ChatGPT-5 (OpenAI; accessed via the official ChatGPT web interface with a Plus subscription account; no further version specification or snapshot identifier was available within the interface at the time of access), Gemini 2.5 Pro (Google DeepMind; accessed via the official Gemini web interface with a paid Pro subscription), Claude Sonnet 4.5 (Anthropic; accessed via the official Claude web interface with a paid subscription), and DeepSeek v3.2 (DeepSeek; accessed via the official DeepSeek web interface, free of charge). These models were selected based on availability, contemporary relevance, and documented use in biomedical or educational contexts. All models were accessed via their official web interfaces on 15 December 2025 using Google Chrome on a desktop computer, without API

ARTICLE IN PRESS

access. Default system settings were maintained; no parameter tuning (e.g., temperature modification), retrieval augmentation, external browsing, or plugin-based assistance was enabled. By restricting models to their base parametric knowledge without external tools, the study design aimed to approximate examination-like closed-book conditions and to isolate internal language-model reasoning capacity. Although deterministic settings were not explicitly enforced, all models were evaluated under default configurations to reflect real-world usage conditions consistent with authentic examination settings. This approach was chosen to reflect authentic high-stakes examination conditions, where only a single response is permitted per question. While some studies have used repeated querying strategies, single-response protocols have also been adopted to simulate real examination conditions, consistent with prior LLM evaluation approaches.

Administration Protocol

Each of the 128 questions was presented individually in its original Turkish format. To standardize response conditions, the following uniform instruction was used: “I will ask you questions from the Clinical Sciences Test – Department of Periodontology, previously administered in the Turkish Dentistry Specialization Examination (DUS). The questions will be single-best-answer multiple-choice questions. Please provide only the correct answer option (A, B, C, D, or E) without any explanation or commentary.” Each question was entered in a new browser session with cleared history and no prior conversational context to eliminate potential carry-over effects. No clarification prompts, iterative refinements, or multi-turn interactions were permitted. Responses were recorded verbatim. This strict single-turn protocol was implemented to replicate the static, time-independent response structure of multiple-choice examinations. This standardized prompting approach was adopted to ensure comparability across models; however, different prompting strategies may influence model performance and should be explored in future research.

ARTICLE IN PRESS

All questions and prompts were presented in Turkish to reflect the original examination format; however, language-related differences in model performance cannot be excluded and should be considered in future research.

Scoring Procedure

Model responses were compared against the official ÖSYM answer key. Outputs were coded dichotomously as correct (1) or incorrect (0). Responses containing multiple selections, ambiguous outputs, or non-standard formats were coded as incorrect to preserve alignment with high-stakes single-best-answer scoring conventions.

Cognitive-demand Stratification (Bloom's Revised Taxonomy)

To evaluate the cognitive structure of the examination, each item was independently classified into one of six levels of Anderson and Krathwohl's revised Bloom's taxonomy: remember, understand, apply, analyze, evaluate, and create. Two periodontology specialists performed independent classification. Inter-rater agreement was assessed using Cohen's kappa (κ), with $\kappa > 0.80$ interpreted as high agreement. Bloom-level classification was incorporated to examine whether model accuracy varied systematically according to cognitive demand, thereby providing insight into whether performance reflected lower-order pattern recognition or robustness across hierarchical cognitive levels. All inferential analyses involving Bloom categories were conducted separately for each rater's classification to assess consistency of findings across independent cognitive labeling. Bloom classifications were not averaged; all analyses were performed separately for each rater to preserve independent evaluation and assess consistency across raters. No composite or averaged Bloom score was used in inferential analyses.

ARTICLE IN PRESS

Error Distribution Analysis

Beyond overall accuracy, all incorrect responses were mapped to their respective content domains. This allowed descriptive identification of domain-level error clustering across models. By linking incorrect responses to curricular constructs, this analysis aimed to explore whether certain disciplinary areas (e.g., procedural therapy vs. foundational knowledge) were disproportionately susceptible to AI misclassification.

Statistical analysis

Given the dichotomous outcome structure and repeated-measures design (each item answered by four models), overall accuracy differences were assessed using Cochran's Q test. When significant, pairwise McNemar tests with Bonferroni correction (adjusted $\alpha = 0.0083$) were performed. To examine the predictive value of cognitive demand, binary logistic regression models were conducted separately for each LLM and each rater's Bloom classification, treating Bloom level as an ordinal predictor of answer correctness. Model fit was evaluated using McFadden and Nagelkerke pseudo-R^2 statistics. Statistical significance was set at $p < 0.05$ unless otherwise adjusted. Descriptive statistics and cross-tabulations were used to characterize distributional patterns. All statistical analyses were performed using IBM SPSS Statistics for Windows, version 23.0 (IBM Corp., Armonk, NY, USA).

ARTICLE IN PRESS

ARTICLE IN PRESS

Results

Distribution of Periodontology Question Domains

A total of 128 valid periodontology questions were included in the analysis. All items were answered by the four large language models and independently classified according to Bloom's Revised Taxonomy by two raters. The complete dataset is provided in the supplementary material. Based on the eight-domain classification, the questions were distributed as follows: periodontal pathology (n = 23), gingival diseases (n = 19), biological foundations of periodontology (n = 18), surgical periodontal therapy (n = 18), non-surgical periodontal therapy (n = 17), periodontal diseases (n = 16), periodontal diagnosis and clinical evaluation (n = 12), and oral implantology (n = 5). The majority of items were concentrated in disease-related and treatment-oriented domains, whereas diagnostic and implantology domains were comparatively less represented within this dataset.

ARTICLE IN PRESS

AI Performance on the DUS Periodontology Corpus

Accuracy rates across the four large language models ranged from 80.47% to 91.41%, with Gemini 2.5 Pro (117/128, 91.41%), ChatGPT-5 (115/128, 89.84%), Claude Sonnet 4.5 (107/128, 83.59%), and DeepSeek v3.2 (103/128, 80.47%). Detailed numbers, percentage distributions, and 95% confidence intervals (CIs) are presented in Table 1. Cochran's Q test demonstrated a statistically significant difference in overall performance among the four models ($Q = 11.731$, $p = 0.008$). Pairwise comparisons were subsequently conducted using McNemar tests with Bonferroni adjustment (adjusted significance threshold: $p \leq 0.0083$). A statistically significant difference was observed only between Gemini 2.5 Pro and DeepSeek v3.2 ($p = 0.0081$), indicating higher odds of

ARTICLE IN PRESS

correct responses for Gemini relative to DeepSeek under the paired comparison framework. No other pairwise comparisons met the adjusted significance threshold.

Table 1. Number and percentage of correct and incorrect responses across the four large language models. ChatGPT-5: Plus subscription, official web interface; Gemini 2.5 Pro: paid Pro subscription, official web interface; Claude Sonnet 4.5: paid subscription, official web interface; DeepSeek v3.2: free of charge, official web interface. All models accessed on 15 December 2025.

	ChatGPT	Gemini	Claude	DeepSeek
Correct	115 (89.84%) [95% CI: 83.3–94.5]	117 (91.41%) [95% CI: 85.1–95.6]	107 (83.59%) [95% CI: 76.0–89.6]	103 (80.47%) [95% CI: 72.6–86.9]
Incorrect	13 (10.16%)	11 (8.59%)	21 (16.41%)	25 (19.53%)

Topic-Specific Error Distribution Across Periodontal Subdomains

Incorrect responses were distributed across periodontal content domains for all four large language models. ChatGPT-5 produced a total of 13 errors, most frequently within non-surgical periodontal therapy (n = 3) and biological foundations of periodontology (n = 3), followed by surgical periodontal therapy (n = 3), periodontal diagnosis and clinical evaluation (n = 2), and periodontal diseases (n = 2). Gemini 2.5 Pro generated 11 errors, primarily within surgical periodontal therapy (n = 3) and non-surgical periodontal therapy (n = 3), with fewer errors in biological foundations (n = 2), periodontal pathology (n = 1), periodontal diseases (n = 1), and periodontal diagnosis and clinical evaluation (n = 1). Claude Sonnet 4.5 demonstrated 21 errors, distributed mainly across non-surgical periodontal therapy (n = 5) and biological foundations (n = 5), followed by periodontal pathology (n = 3) and periodontal diagnosis and clinical evaluation (n = 4), with a smaller proportion in surgical periodontal therapy (n = 3) and gingival diseases (n = 1). DeepSeek v3.2 exhibited 25 errors, most commonly in surgical periodontal therapy (n = 7) and non-surgical

ARTICLE IN PRESS

periodontal therapy (n = 6), with additional errors observed in biological foundations (n = 3), periodontal diagnosis and clinical evaluation (n = 3), periodontal pathology (n = 2), gingival diseases (n = 2), oral implantology (n = 1), and periodontal diseases (n = 1). Fewer incorrect responses were observed in gingival diseases and oral implantology domains across models within this dataset. The detailed quantitative breakdown of model-specific errors across all curricular subdomains is provided in Table 2.

Table 2. Distribution of incorrect responses across periodontal subdomains for the four large language models.

Periodontal Subdomain	ChatGPT	Gemini	Claude	DeepSeek	Total Errors per Domain
Biological foundations of periodontology	3	2	5	3	13
Periodontal pathology	0	1	3	2	6
Gingival diseases	0	0	1	2	3
Periodontal diseases	2	1	0	1	4
Periodontal diagnosis and clinical evaluation	2	1	4	3	10
Non-surgical periodontal therapy	3	3	5	6	17
Surgical periodontal therapy	3	3	3	7	16
Oral implantology	0	0	0	1	1
Total Errors per Model	**13**	**11**	**21**	**25**	**70**

ARTICLE IN PRESS

ARTICLE IN PRESS

Performance Variability and System-Level Differences

Inter-rater reliability for Bloom classification was high ($\kappa = 0.948$, $p < 0.001$), indicating strong agreement between raters. Pairwise model comparisons confirmed a statistically significant difference only between Gemini 2.5 Pro and DeepSeek v3.2 ($p = 0.0081$), as detailed above. Although several unadjusted comparisons yielded p values below 0.05, none remained statistically significant after Bonferroni correction except for the Gemini–DeepSeek comparison. Crosstabulations and detailed p values are presented in Tables 3 and 4.

Table 3. Crosstabs for pairwise comparisons of correct and incorrect answers among the four large language models.

Model Pair	Both Correct	Only First Correct	Only Second Correct	Both Incorrect
ChatGPT vs Gemini	110	5	7	6
ChatGPT vs Claude	102	13	5	8
ChatGPT vs DeepSeek	95	20	8	5
Gemini vs Claude	103	14	4	7
Gemini vs DeepSeek	96	21	7	4
Claude vs DeepSeek	90	17	13	8

Table 4. Pairwise McNemar test p values for comparisons among the four large language models. *$p \leq 0.0083$ after Bonferroni correction.

Model Pair	p value
ChatGPT vs Gemini	0.774
ChatGPT vs Claude	0.096
ChatGPT vs DeepSeek	0.036

ARTICLE IN PRESS

Gemini vs Claude	0.031
Gemini vs DeepSeek	0.008*
Claude vs DeepSeek	0.585

Bloom's Revised Taxonomy and Model Performance

Binary logistic regression analyses were conducted to examine whether Bloom's cognitive level predicted response accuracy. Bloom scores (1–6) were treated as an ordinal predictor, and answer correctness (0/1) was specified as the dependent variable. Model fit indices were low for both raters (McFadden $R^2 = 0.0015$; Nagelkerke $R^2 = 0.033$ and 0.034 for Rater 1 and Rater 2, respectively), indicating limited explanatory contribution of Bloom level within these models. No statistically significant association was observed between Bloom level and response accuracy for any of the four large language models (all $p > 0.05$). Regression coefficients were small in magnitude and consistent across raters. These findings indicate that, within this dataset, Bloom's cognitive level was not a statistically significant predictor of model accuracy. Full regression outputs, including B coefficients, standard errors, and 95% confidence intervals for both raters, are presented in Table 5 to ensure transparency.

ARTICLE IN PRESS

Table 5. Binary logistic regression analysis of model accuracy across Bloom's cognitive levels for both raters.

Prediction	B	Standard Error	Wald Statistic	*p*	Odds Ratio	95% CI Lower	95% CI Upper
Rater 1 - ChatGPT	-0.047	0.243	0.038	0.846	0.954	0.593	1.535
Rater 1 - Gemini	0.075	0.288	0.067	0.795	1.078	0.613	1.895

ARTICLE IN PRESS

Rater 1 - Claude	0.269	0.253	1.129	0.288	1.308	0.797	2.149
Rater 1 - DeepSeek	0.070	0.200	0.121	0.728	1.072	0.725	1.586
Rater 2 - ChatGPT	-0.016	0.241	0.004	0.948	0.984	0.614	1.579
Rater 2 - Gemini	0.104	0.286	0.133	0.716	1.110	0.633	1.945
Rater 2 - Claude	0.297	0.251	1.408	0.235	1.346	0.824	2.201
Rater 2 - DeepSeek	0.103	0.198	0.269	0.604	1.108	0.752	1.634

Discussion

In this study, four contemporary large language models showed high overall accuracy on periodontology questions from the Turkish Dentistry Specialization Examination (DUS), with correct response rates ranging from 80.47% to 91.41%. Rather than exhibiting wide variability, the models showed a high degree of convergence in performance; a statistically significant difference was observed only between the highest- and lowest-performing systems ($p = 0.0081$). Furthermore, cognitive demand operationalized via Bloom's Revised Taxonomy ($\kappa = 0.948$) was not identified as a statistically significant predictor of response accuracy ($p > 0.05$). Collectively, these findings suggest that contemporary, high-capacity LLMs can perform strongly on structured, text-based single-best-answer multiple-choice specialty examination items and warrant further consideration regarding the interaction between artificial intelligence systems and such structured examination formats. For contextual reference, the accuracy rates observed in this study (80.47–91.41%) substantially exceed chance-level performance on five-option single-best-answer items (20%), and are broadly consistent with accuracy rates reported for earlier-generation models on DUS periodontology questions — for example, ChatGPT-4.0 achieved 95% and Gemini Advanced achieved 90% correct response rates on DUS periodontology items in a prior study (10).

ARTICLE IN PRESS

Importantly, these findings should not be interpreted as a competitive benchmarking of model performance, but rather as an examination of how contemporary LLMs interact with the structural and cognitive characteristics of a high-stakes assessment context.

These results are consistent with a growing body of literature examining LLM performance in dental and medical licensing examinations. Torres-Zegarra et al. reported that GPT-4 and Bing achieved high accuracies (86.7% and 82.2%) on the Peruvian National Licensing Medical Examination, outperforming human examinees and demonstrating relatively stable performance across disciplines, although reduced accuracy was observed for items requiring country-specific contextual knowledge (14). Similarly, Chau et al. found that GPT-4 achieved 72.1% overall accuracy on INBDE and ORE questions, with periodontics-specific performance ranging from 86% to 94% (15). Dashti et al. demonstrated that GPT-4 significantly exceeded GPT-3.5 in INBDE, ADAT, and DAT assessments, particularly for knowledge-based items (16), while Tussie et al. reported that GPT-4 outperformed other contemporary models in INBDE question sets (17). Together, these studies indicate that LLMs tend to perform strongly on text-based, knowledge-oriented examination formats.

ARTICLE IN PRESS

However, prior research also suggests that performance declines in contexts requiring multi-step diagnostic reasoning, treatment planning, or complex clinical interpretation (18). Multimodal investigations have further shown reduced accuracy when image interpretation is required; for example, ChatGPT-4o and Claude 3.5 demonstrated substantially lower performance in mammographic BI-RADS classification tasks (19). A recent DUS-periodontology study similarly reported that GPT-4 performed better on foundational items than on clinically complex questions (20). Similarly, earlier-generation models demonstrated limited accuracy on DUS prosthodontics

ARTICLE IN PRESS

and pediatric dentistry questions, further supporting the notion that performance varies across specialty domains and model generations (21,22,23). More recent comparative studies further support these patterns. Erdal et al. reported that ChatGPT-5, ChatGPT-4o, and Gemini 2.5 Pro achieved high overall accuracy on DUS prosthodontics questions under independent querying (86%, 83%, and 85%, respectively), while sequential querying led to a significant decline in knowledge-question accuracy for GPT models but not for Gemini 2.5 Pro, suggesting model-specific sensitivity to contextual carryover effects (24). Similarly, Geduk et al. demonstrated high text-based accuracy across seven LLMs on DUS oral and maxillofacial radiology questions, while all models showed substantially reduced performance on image-based items, further underscoring the domain-specific limitations of current LLMs in visually demanding tasks (25). Within this broader context, the present findings align with existing evidence indicating that LLM performance is stronger in structured, text-based assessments than in clinically integrative or visually driven tasks.

Beyond descriptive accuracy comparisons, this study offers additional perspective through cognitive and domain-level stratification. Unlike some earlier reports suggesting reduced performance for more complex items, the present logistic regression analysis did not identify a statistically significant association between Bloom level and model accuracy. While this may indicate relatively stable performance across Bloom categories within the analyzed items, it should be interpreted cautiously. The distribution of cognitive levels within the released DUS dataset, as well as the inherent limitations of Bloom's taxonomy in capturing clinical reasoning complexity, may influence this observation. Therefore, the absence of statistical significance should not be interpreted as evidence that cognitive demand has no effect on AI performance in broader contexts. Furthermore, it is important to acknowledge the conceptual mismatch between Bloom's

ARTICLE IN PRESS

Taxonomy and LLM functioning. While Bloom's levels are designed to categorize human cognitive processes, LLMs operate through complex statistical pattern recognition. Consequently, the lack of association between Bloom level and model performance may partly reflect the limitations of applying a human-centric cognitive framework to non-biological intelligence systems. Detailed regression coefficients and confidence intervals are presented in Table 5 to support the transparency of these findings. The low R^2 values obtained further indicate that Bloom level alone accounts for minimal variance in model accuracy, suggesting that other factors — such as item content, domain-specific knowledge representation, or linguistic features of the questions — may contribute more substantially to performance variation.

Domain-level error mapping further illustrated that incorrect responses were not uniformly distributed across periodontal subfields. Across models, a greater proportion of errors occurred in non-surgical and surgical periodontal therapy domains, whereas fewer errors were observed in gingival diseases and oral implantology. Although these findings were descriptive and not subjected to formal inferential testing, they suggest that domains requiring procedural integration and therapeutic decision-making may present relatively greater challenges for language models. Variability across models was present but modest. Although Gemini 2.5 Pro demonstrated statistically higher odds of correct responses compared with DeepSeek, no other pairwise differences reached the adjusted significance threshold. This pattern may reflect differences in model architecture, training data composition, multilingual exposure, and domain representation, as emphasized in recent reviews of LLM development (26,27,28). Furthermore, the observed performance differences among the models should be interpreted with caution, as the proprietary nature of these systems limits access to their specific training datasets and update mechanisms. Consequently, performance variations cannot be definitively attributed to specific model

ARTICLE IN PRESS

architectures or domain-specific exposure, a challenge often referred to as the limited transparency (black-box nature) of proprietary LLMs. At the same time, the relatively narrow performance range observed in this study suggests a degree of convergence among contemporary high-capacity LLMs in handling structured, text-based examination content.

From an educational and assessment perspective, these findings raise important considerations. LLMs are increasingly used for self-directed learning, knowledge clarification, and formative feedback in health professions education (2,3,4,5). High performance on single-best-answer multiple-choice specialty examination items suggests that such systems may function as supplementary preparatory tools, particularly for foundational knowledge domains. However, as noted in prior commentary (13), strong performance on recognition-based single-best-answer MCQs does not necessarily equate to clinical competence. If high-stakes examinations predominantly assess lower-order cognitive processes, increasingly capable AI systems may achieve high accuracy without demonstrating integrative clinical reasoning.

Accordingly, within the context of the analyzed DUS periodontology dataset, the present findings contribute to ongoing discussions regarding the potential vulnerabilities of single-best-answer MCQ-based specialty assessments in the era of artificial intelligence. While this study does not directly evaluate examination validity, the ability of LLMs to achieve high accuracy under controlled conditions suggests that, within this dataset, examination formats emphasizing factual recall and structured recognition may warrant further investigation regarding their susceptibility to AI replication, particularly compared with formats requiring contextualized, case-based reasoning. Future research may explore whether incorporating higher-order cognitive tasks, clinical vignettes, and multimodal components into examination designs could contribute to

ARTICLE IN PRESS

preserving discriminatory capacity in the era of advancing AI capabilities. However, this should not be interpreted as evidence of inadequacy in current examination systems, but rather as an evolving interaction between assessment formats and advancing AI capabilities.

This study has several limitations. The analysis focused on a single-best-answer MCQ corpus from 2012–2021 without multimodal elements, providing a baseline snapshot as of December 2025. Furthermore, the high overall accuracy rates (>80%) across models may have introduced a 'ceiling effect,' potentially limiting the statistical power to detect more nuanced performance differences across cognitive levels. While we utilized a single-run protocol to replicate authentic 'real-time' examination conditions, we acknowledge that this design choice has implications for the reliability and reproducibility of the reported results. LLM outputs are known to vary across sessions due to inherent stochasticity in sampling mechanisms, and the accuracy values reported in this study should therefore be interpreted as reflecting a single evaluation instance rather than stable, reproducible performance estimates. We further clarify that repeated-run designs or deterministic settings would be necessary to assess the stability of model responses under controlled conditions. Additionally, the conceptual mismatch between Bloom's human-centric taxonomy and the statistical pattern recognition of LLMs should be considered when interpreting the lack of significant associations. Moreover, although domain-level classification was based on established DUS curricular structures defined by ÖSYM, alternative grouping strategies and more detailed statistical comparisons across subdomains were not performed and may be explored in future studies. Furthermore, Bloom's taxonomy levels were treated as ordinal predictors in the regression analyses; however, equal interval spacing between categories was assumed rather than empirically validated, which represents an additional methodological consideration. Additionally, the models evaluated in this study were accessed under different subscription conditions — three via paid

ARTICLE IN PRESS

accounts and one free of charge — which may have influenced output quality and should be considered when interpreting cross-model comparisons.

Building on these findings, future research should extend this work by integrating case-based tasks, sequential decision-making scenarios, and multimodal inputs to evaluate higher-order reasoning performance. Comparative analyses across additional dental specialties and formal psychometric examination of item characteristics in relation to AI performance may further clarify the interaction between assessment design and artificial intelligence capability. Longitudinal monitoring of LLM performance across model updates will also be important to determine whether observed gains reflect deeper reasoning improvements or enhanced pattern recognition within large-scale training corpora.

ARTICLE IN PRESS

Conclusion

Contemporary large language models showed high overall accuracy on the analyzed DUS periodontology single-best-answer multiple-choice questions, exhibiting a strong convergence in their performance capabilities. Cognitive demand, based on Bloom's taxonomy, was not identified as a statistically significant predictor of model accuracy within this specific dataset. These findings suggest that while current LLMs perform robustly in structured, text-based examination formats, their high success rates should be interpreted with caution when considering the structural characteristics of traditional single-best-answer multiple-choice assessments. Although LLMs may serve as supplementary tools for examination preparation, future studies may explore whether assessments could incorporate complex, case-based, and multimodal tasks to better capture authentic higher-order clinical reasoning in dental education.

ARTICLE IN PRESS

Declarations

Ethics approval and consent to participate

This study was based exclusively on the analysis of publicly available examination materials. It did not involve human participants, patients, clinical data, or identifiable personal information. Therefore, ethics committee approval and informed consent to participate were not required.

Consent for publication

Not applicable.

Availability of data and materials

The dataset generated and/or analyzed during the current study (including question classifications and large language model responses) is provided as supplementary material.

ARTICLE IN PRESS

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions

MG: Conceptualization, methodology, supervision, writing – original draft preparation.

IK: Formal analysis, data curation, visualization, writing – review and editing.

ARTICLE IN PRESS

MY: Data curation, investigation, validation, writing – review and editing.

All authors read and approved the final manuscript.

Acknowledgements

The authors thank the Assessment, Selection, and Placement Center (ÖSYM) for publicly providing the DUS examination booklets used in this study.

References

1. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. J Med Syst. 2023;4;47(1):33. doi:10.1007/s10916-023-01925-4. PMID:36869927; PMCID:PMC9985086.
2. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;9;2(2):e0000198. doi:10.1371/journal.pdig.0000198. PMID:36812645. PMCID:PMC9931230.
3. Eggmann F, Weiger R, Zitzmann NU, Blatz MB. Implications of large language models such as ChatGPT for dental medicine. J Esthet Restor Dent. 2023;35(7):1098-1102. doi:10.1111/jerd.13046. PMID: 37017291.
4. Danesh A, Pazouki H, Danesh K, Danesh F, Danesh A. The performance of artificial intelligence language models in board-style dental knowledge assessment: A preliminary study on ChatGPT. J Am Dent Assoc. 2023;154(11):970-974. doi:10.1016/j.adaj.2023.07.016. PMID: 37676187.

ARTICLE IN PRESS

5. Ali K, Barhom N, Tamimi F, Duggal M. ChatGPT-A double-edged sword for healthcare education? Implications for assessments of dental students. Eur J Dent Educ. 2024;28(1):206-211. doi:10.1111/eje.12937. PMID:37550893.
6. Suárez A, Díaz-Flores García V, Algar J, Gómez Sánchez M, Llorente de Pedro M, Freire Y. Unveiling the ChatGPT phenomenon: Evaluating the consistency and accuracy of endodontic question answers. Int Endod J. 2024;57(1):108-113. doi:10.1111/iej.13985. PMID: 37814369.
7. Danesh A, Pazouki H, Danesh F, Danesh A, Vardar-Sengul S. Artificial intelligence in dental education: ChatGPT's performance on the periodontic in-service examination. J Periodontol. 2024;95(7):682-687. doi:10.1002/JPER.23-0514. PMID:38197146.
8. Ozden I, Gokyar M, Ozden ME, Sazak Ovecoglu H. Assessment of artificial intelligence applications in responding to dental trauma. Dent Traumatol. 2024;40(6):722-729. doi:10.1111/edt.12965. PMID:38742754.
9. Mohammad-Rahimi H, Motamedian SR, Pirayesh Z, Haiat A, Zahedrozegar S, Mahmoudinia E, Rohban MH, Krois J, Lee JH, Schwendicke F. Deep learning in periodontology and oral implantology: A scoping review. J Periodontal Res. 2022;57(5):942-951. doi:10.1111/jre.13037. PMID:35856183.
10. 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;10;25(1):214. doi:10.1186/s12909-024-06389-9. PMID:39930399. PMCID: PMC11809121.
11. Culhaoglu AK, Kilicarslan MA, Deniz KZ. Evaluation of dental specialty entrance examination on approach and preferences for different levels of dental education and post graduation. J Dent Fac Atatürk Uni. 2021;31:420-7.

ARTICLE IN PRESS

ARTICLE IN PRESS

12. Zaidi NLB, Grob KL, Monrad SM, Kurtz JB, Tai A, Ahmed AZ, Gruppen LD, Santen SA. Pushing Critical Thinking Skills With Multiple-Choice Questions: Does Bloom's Taxonomy Work? Acad Med. 2018;93(6):856-859. doi:10.1097/ACM.0000000000002087. PMID:29215375.
13. Gonsalves, C. On ChatGPT: what promise remains for multiple choice assessment?. Journal of Learning Development in Higher Education, 27. doi:10.47408/jldhe.vi27.1009.
14. Torres-Zegarra BC, Rios-Garcia W, Ñaña-Cordova AM, Arteaga-Cisneros KF, Chalco XCB, Ordoñez MAB, Rios CJG, Godoy CAR, Quezada KLTP, Gutierrez-Arratia JD, Flores-Cohaila JA. Performance of ChatGPT, Bard, Claude, and Bing on the Peruvian National Licensing Medical Examination: a cross-sectional study. J Educ Eval Health Prof. 2023;20:30. doi:10.3352/jeehp.2023.20.30. PMID: 37981579. PMCID: PMC11009012.
15. Chau RCW, Thu KM, Yu OY, Hsung RT, Lo ECM, Lam WYH. Performance of Generative Artificial Intelligence in Dental Licensing Examinations. Int Dent J. 2024;74(3):616-621. doi:10.1016/j.identj.2023.12.007. PMID:38242810. PMCID:PMC11123518.
16. Dashti M, Ghasemi S, Ghadimi N, Hefzi D, Karimian A, Zare N, Fahimipour A, Khurshid Z, Chafjiri MM, Ghaedsharaf S. Performance of ChatGPT 3.5 and 4 on U.S. dental examinations: the INBDE, ADAT, and DAT. Imaging Sci Dent. 2024;54(3):271-275. doi:10.5624/isd.20240037. PMID:39371301. PMCID:PMC11450412.
17. Tussie, C., Starosta, A. Comparing the dental knowledge of large language models. *Br Dent J* (2024). https://doi.org/10.1038/s41415-024-8015-2
18. Tosun B, Yilmaz ZS. Comparison of artificial intelligence systems in answering prosthodontics questions from the dental specialty exam in Turkey. J Dent Sci.

ARTICLE IN PRESS

2025;20(3):1454-1459. doi:10.1016/j.jds.2025.01.025. PMID:40654425. PMCID:PMC12254736.

19. Karahan, B. N., Emekli, E., & Altın, M. A. (2025). Artificial Intelligence-Based Chatbots' Ability to Interpret Mammography Images: A Comparison of Chat-GPT 4o and Claude 3.5. European Journal of Therapeutics, 31(1),28–34. https://doi.org/10.58600/eurjther2599
20. Kandemir M, Sarıbaş EE. Comparative performance of large language models in answering periodontology questions from the Turkish Dental Specialty Examination: a cross-sectional study on accuracy and coverage. BMC Oral Health. 2025;19;25(1):1804. doi:10.1186/s12903-025-07241-2. PMID: 41257711. PMCID:PMC12629045.
21. Yüceer Çetiner E. 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. EJOMS. 2025;4(3):59-65.
22. Aşık A, Kuru E. Diş hekimliğinde uzmanlık eğitim giriş sınavında sorulan çocuk diş hekimliği sorularına ChatGPT'nin verdiği cevapların analizi: Kesitsel araştırma. Turkiye Klinikleri J Dental Sci. 2025;31(3):401-6.
23. Bilgin Avşar D, Ertan AA. Diş hekimliğinde uzmanlık sınavında sorulan protetik diş tedavisi sorularının ChatGPT-3.5 ve Gemini tarafından cevaplanma performanslarının karşılaştırmalı olarak incelenmesi: Kesitsel araştırma. Turkiye Klinikleri J Dental Sci. 2024;30(4):668-73.
24. Erdal SG, Çavuşculu Güdül AB, Köroğlu A. The effectiveness of large language models in dental specialty questions: a comparative study in the field of prosthodontics. BMC Med Educ. 2026;26:455. doi:10.1186/s12909-026-08808-5.

ARTICLE IN PRESS

ARTICLE IN PRESS

25. Geduk G, Hasırcı UC, Dumanlı Kusay D, Çayır Aras R, Çapar İ, Altın E, Şeker Ç. A comparative analysis of the performance of large language models in the dentistry specialty examination. Sci Rep. 2026;16:6739. doi:10.1038/s41598-026-37800-8.
26. Han, S., Wang, M., Zhang, J., Li, D., & Duan, J. (2024). A Review of Large Language Models: Fundamental Architectures, Key Technological Evolutions, Interdisciplinary Technologies Integration, Optimization and Compression Techniques, Applications, and Challenges. Electronics, 13(24), 5040. doi:10.3390/electronics13245040
27. Singhal, K., Azizi, S., Tu, T. et al. Large language models encode clinical knowledge. Nature 2023;620:172–180. doi:10.1038/s41586-023-06291-2.
28. Raza, M., Jahangir, Z., Riaz, M.B. et al. Industrial applications of large language models. Sci Rep 2025;15,13755. doi:10.1038/s41598-025-98483-1

ARTICLE IN PRESS