Scientific Reports

https://doi.org/10.1038/s41598-026-57392-7

Article in Press

Benchmarking GPT-5, Gemini 2.5 Pro, Grok 4, and other LLMs on pediatric dentistry questions from a dental specialization exam

Received: 24 November 2025

Accepted: 5 June 2026

Published online: 11 June 2026

Cite this article as: Kayaci S.T., Ilhan H.O., Tassoker M. *et al.* Benchmarking GPT-5, Gemini 2.5 Pro, Grok 4, and other LLMs on pediatric dentistry questions from a dental specialization exam. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-57392-7

Sukriye Turkoglu Kayaci, Hamza Osman Ilhan, Melek Tassoker, Taha Cap & Helin Demir

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

Benchmarking GPT-5, Gemini 2.5 Pro, Grok 4, and Other LLMs on Pediatric Dentistry Questions from a Dental Specialization Exam

Sukriye Turkoglu Kayaci[1], Hamza Osman Ilhan[2], Melek Tassoker[3*], Taha Cap[4], and Helin Demir[5]

[1]Department of Pediatric Dentistry, Faculty Of Dentistry, University of Health Sciences, Istanbul, Turkiye
[2]Department of Computer Engineering, Yildiz Technical University, Istanbul, Türkiye
[3]Department Of Oral And Maxillofacial Radiology, Faculty Of Dentistry, Necmettin Erbakan University, Konya, Türkiye
[4]Department of Control and Automation Engineering, Yildiz Technical University, Istanbul, Türkiye
[5]Department of Pediatric Dentistry, Faculty Of Dentistry, University of Health Sciences, Istanbul, Turkiye
[*]mtassoker@erbakan.edu.tr

ABSTRACT

Artificial intelligence (AI), particularly large language models (LLMs), is an increasingly prominent tool in medical and dental education. Trained with deep learning and NLP techniques, these models interpret meaning, generate text, and manage complex information. They hold potential for practical educational applications, such as supporting exam preparation and personalized learning. Moreover, their performance in clinical case recognition suggests an emerging potential for use in diagnostic decision-support systems. This study aimed to evaluate the performance of state-of-the-art large language models (LLMs) on pediatric dentistry questions from the Dentistry Specialization Examination (DUS) in Türkiye, a high-stakes national exam for postgraduate training. A total of 119 pediatric dentistry questions from the past ten years of the DUS were compiled and presented to 11 recently developed LLMs (17 with reasoning mode activations), including GPT-5, Gemini 2.5 Pro, Grok-4, and DeepSeek R1. Each model's accuracy (%) and average response generation time (seconds) were calculated and compared. The Gemini 2.5 Pro (92.44%) demonstrated significantly higher mean scores compared to all other models except GPT-4 (78.15%), GPT-5 (90.76%), GPTOSS (78.99, 75.63%), and Grok-4 (88.24%). Similar patterns were also observed for GPT-5 and Grok-4. In contrast, Qwen-3 (49.58 - Reasoning, 54.62 - No Reasoning), and MedGemma (58.82%) exhibited notably lower accuracy rates across most comparisons. Overall, these findings highlight Gemini 2.5 Pro and GPT-5 as achieving the highest accuracy levels among the models, whereas Qwen-3, Qwen-3 (R), and MedGemma demonstrated the weakest performance. While models such as Gemma, LLaMA, and Mistral demonstrated faster response times (<1 sec), they exhibited relatively low accuracy. In contrast, reasoning-intensive modes (e.g., DeepSeek R1) improved accuracy but required excessively long generation times (up to 68 sec). LLM performance on pediatric dentistry questions was highly variable. Top models, notably Gemini 2.5 Pro (92.44%) and GPT-5 (90.76%), approached expert-level accuracy, while others (e.g., Qwen-3) performed poorly. A critical speed-accuracy trade-off was evident: reasoning modes improved scores but were impractically slow, whereas faster models had low accuracy. This variability necessitates careful validation before LLMs are used in high-stakes dental education or assessment.

ARTICLE IN PRESS

Introduction

Artificial intelligence (AI) has increasingly become a prominent tool in medical and dental education, with large language models (LLMs) attracting particular attention[1–3]. These models, trained on vast corpora using deep learning and natural language processing (NLP) techniques, are designed to interpret meaning, generate coherent text, and assist in managing complex information[4,5]. Beyond their technical capabilities, LLMs hold potential for practical applications such as supporting exam preparation, providing formative feedback, and facilitating personalized learning experiences[1,6,7]. Moreover, their demonstrated performance in recognizing clinical cases and engaging in diagnostic reasoning suggests an emerging potential for their use as diagnostic decision-support systems[8–10].

A notable implementation of LLMs is through conversational agents, or chatbots, which enable text- and voice-based dialogue and are reshaping how individuals interact with technology. These developments have prompted evaluation of LLMs not only as general-purpose communication tools but also within professional education and assessment contexts[11,12]. Previous studies have examined LLM performance on dental licensing and admissions exams[12–14]. A recent meta-analysis of eleven

ARTICLE IN PRESS

studies evaluated dental licensing examinations from eight countries, reporting integrated accuracy rates of 54% for GPT-3.5, 72% for GPT-4, and 56% for Bard. GPT-4 outperformed GPT-3.5 and Bard, passing more than half of the evaluated exams[13].

Recent advancements have introduced new generations of LLMs. In August 2025, OpenAI released GPT-5, its most advanced multimodal system to date[15,16]. Around the same period, Google DeepMind introduced Gemini 2.5 Pro, featuring enhanced reasoning and problem-solving capabilities[15]. In July 2025, xAI launched Grok-4, which integrates real-time information retrieval and advanced analytical functions[17]. Older LLM versions have already been tested in multiple dental domains, including endodontics[18], restorative dentistry[19], oral radiology[20], anatomy[21], oral surgery[22], oral pathology[6], prosthodontics[8], and orthodontics[9].

The Dentistry Specialization Education Entrance Examination (DUS) represents the competitive national examination for entry into postgraduate specialty training in dentistry. Administered by the Assessment, Selection and Placement Center (OSYM), the DUS assesses candidates' proficiency across biomedical sciences and clinical dental knowledge, with outcomes determining access to accredited postgraduate programs[23]. Given its comprehensive scope and high-stakes nature, the DUS serves as an important benchmark for evaluating whether LLMs can perform reliably in a national, discipline-specific context, combining knowledge-based items with clinical case scenarios that require diagnostic reasoning[24]. Crucially, the DUS is not limited to simple fact-recall. A significant portion of the examination is designed to assess higher-order cognitive skills, presenting candidates with complex clinical case scenarios. These questions require test-takers to engage in sophisticated diagnostic reasoning, etiologic identification, and clinical decision-making, effectively mirroring the diagnostic challenges faced in real-world pediatric dentistry practice. Therefore, evaluating LLM performance on this exam provides a direct proxy for their potential in diagnostic support.

Previous investigations have evaluated LLM performance on various dental licensing and admissions examinations. Studies examining the Integrated National Board Dental Examination (INBDE), Advanced Dental Admission Test (ADAT), and related assessments have shown promising but variable results, with models generally excelling on fact-recall questions while struggling on items requiring multi-step reasoning or numerical computation. Similar analyses in medical licensing contexts, such as the United States Medical Licensing Examination (USMLE)[12,25], also report strong performance in some domains but inconsistent results when questions are translated or reflect local curricular emphases[13,14]. Meta-analytic work reinforces these observations, highlighting both the rapid progress of newer model generations and their fragility across settings and question formats.

Despite these rapid technological advances, no comprehensive study has yet assessed the knowledge and performance of these new LLM generations specifically in pediatric dentistry. In Türkiye, the Dentistry Specialization Examination (DUS) serves as the national gateway to postgraduate specialty training in dentistry. The current study aims to evaluate the performance of 11 contemporary LLMs—including the most well-known state-of-the-art models such as GPT-5, Grok-4, DeepSeek R1, and Gemini 2.5 Pro—on pediatric dentistry questions drawn from the past ten years of DUS examinations.

The main contributions of this study can be summarized as follows:

- The accuracy of diverse state-of-the-art LLMs were evaluated on official DUS questions under controlled conditions.
- Specific content domains where LLMs demonstrate strong performance were identified as well as areas prone to systematic errors.
- The implications of these findings are discussed in relation to dental education, exam preparation strategies, and the responsible integration of AI tools into postgraduate training.

By grounding the evaluation in a nationally significant, high-stakes examination, this study provides critical evidence on both the readiness and limitations of LLMs for supporting postgraduate dental specialization preparation in Türkiye.

Materials and Methods

Dataset Information

The evaluation dataset consisted of 119 multiple-choice questions (MCQs) drawn from the Dentistry Specialization Examination (DUS). One canceled question was excluded. Questions were obtained from the official exam booklets published by OSYM[23]. All questions were translated from Turkish into English prior to submission to the AI models. Each item had five answer options (A–E), with a single correct response validated by the national examination authority (OSYM). Questions were transcribed from their original Microsoft Word 2019 version into a structured Excel 2021 file, including the stem, options, and official answer key. Image-based items were excluded to focus exclusively on text-based reasoning[13,14]. Questions were sampled from exams administered between 2015 and 2025, reflecting the standardized examination framework.

Responses from the chatbots were categorized as either "correct" or "incorrect" based on the official answer key provided by OSYM. The accuracy rate for each chatbot was calculated as the percentage of correct answers out of the total number of questions attempted[6,12].

ARTICLE IN PRESS

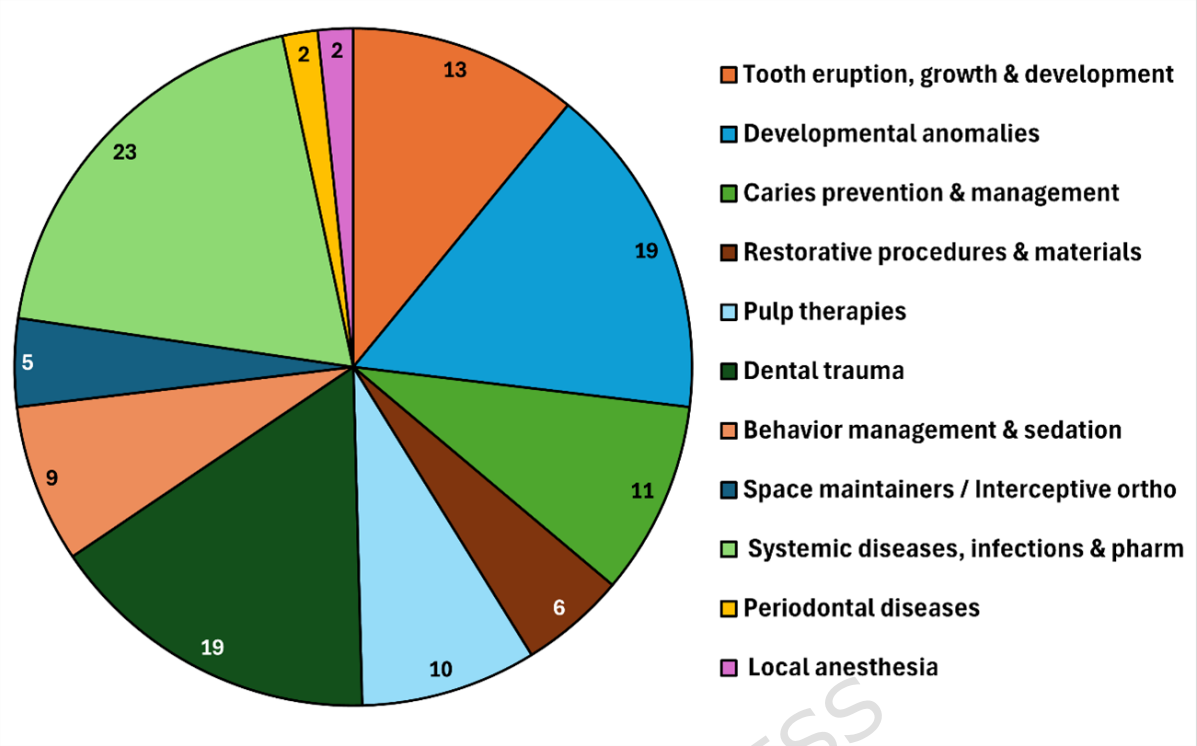


Figure 1. The distribution of the total number of questions according to 11 pediatric dentistiry topics

The questions were categorized into 11 pediatric dentistry topics as; Tooth eruption, growth and development, developmental anomalies, caries prevention and management, restorative procedures and materials, pulp therapies, dental trauma, behaviour management and sedation, space maintainers/interceptive orthodontics, systemic diseases, infections and pharm, periodontal diseases, local anesthesia[1,10,26,27].

The distribution of the total number of questions across these 11 topics is shown in Figure 1. To validate the dataset's relevance for diagnostic benchmarking, the questions were analyzed for content. This analysis confirmed that the items frequently move beyond simple knowledge retrieval, focusing instead on clinical decision-making and etiological interpretation. A representative example is provided in Figure 2, which presents a real-world dental trauma scenario requiring the selection of an appropriate treatment based on specific diagnostic findings (e.g., vital pulp, exposure size, time elapsed).

Models

In this study, a total of 17 experimental conditions were evaluated across 11 distinct large language model families (Mistral, Llama Maverick, Gemma, MedGemma, Qwen, DeepSeek, OpenAI GPT-OSS, GPT-4o, GPT5, Gemini 2.5 Pro and Grok), representing a diverse range of proprietary and open-weight systems[4,15–17]. To ensure the strict reproducibility of our benchmarking results, the exact API model identifiers, specific version snapshots (as of September 2025), and hyperparameter configurations for each model are comprehensively detailed in Table 1.

A standardized evaluation protocol was applied to maximize comparability. Specifically, the temperature parameter was set to 0.0 for all models where the API permitted, ensuring deterministic outputs. However, it should be noted that GPT-5 was evaluated at its fixed default temperature of 1.0 due to provider-imposed API restrictions active during the testing window; thus, its results reflect a probabilistic variance inherent to its default setting. Furthermore, in Table 1, models designated with a reasoning indicator (R) or "Reasoning Mode" refer to configurations where the model explicitly employs a Chain-of-Thought (CoT) process—either through a native "thinking" architecture (e.g., Gemini 2.5 Pro, DeepSeek R1) or via system prompts instructing the model to generate a rationale before the final answer. To prevent the truncation of these extensive reasoning traces, the maximum output token limit was adjusted accordingly for these specific runs.

Mistral Small

Mistral Small 3 is a 24-billion-parameter instruction-tuned, text-only model[36]. It is an open-weight release under the Apache 2.0 license and provides both base and instruction-tuned checkpoints. Peer-reviewed, exam-style evaluations in medicine have increasingly included Mistral family models, with reports showing competitive performance on expert-level multiple-choice

ARTICLE IN PRESS

77. A systemically healthy 9-year-old girl is brought in after falling and breaking her tooth at school approximately three hours earlier. Clinical examination reveals a complicated crown fracture of her maxillary right lateral incisor. Pulp exposure is 2 mm, the pulp is vital, and there is no tenderness to percussion or palpation. Radiographic examination reveals no pathology or fractures in the roots.

Which of the following is the most appropriate treatment for this patient?

A) Direct pulp capping

B) Indirect pulp capping

C) Partial pulpotomy

D) Root canal treatment

E) Tooth extraction

Figure 2. Representative DUS question requiring clinical decision-making

ARTICLE IN PRESS

Table 1. Detailed Model Specifications and Parameters

Model Name Test Condition	Exact API Identifier	Release Version Date	Temp.	Top_p	Max Output Tokens	Reasoning / System Setting
Gemini 2.5 Pro[28,29]	gemini-2.5-pro-001	Aug 2025	0.0	0.95	8192	Native Thinking Mode
GPT-5[30]	gpt-5-preview-2025-08-14	Aug 2025	1.0	1.0	4096	Default
Grok-4[31]	grok-4-latest	July 2025	0.0	0.9	4096	Default
GPT-OSS[32]	gpt-oss-120b-instruct	Aug 2025	0.0	1.0	2048	Standard Prompting
GPT-4o[33]	gpt-4o-2024-05-13	May 2024	0.0	1.0	4096	Default
GPT-OSS (R)[32]	gpt-oss-120b-instruct	April 2025	0.0	1.0	8192	CoT Prompting Enabled
DeepSeek R1 (R)[34]	deepseek-reasoner	Jan 2025	0.0	0.95	8192	Active Reasoning
DeepSeek R1[34]	deepseek-chat	Jan 2025	0.0	0.95	4096	Standard Mode
Llama 4 Maverick (R)[35]	llama-4-mav-17b-instruct	April 2025	0.0	0.9	8192	CoT Prompting Enabled
Llama 4 Maverick[35]	llama-4-mav-17b-instruct	April 2025	0.0	0.9	2048	Standard Prompting
Mistral Small[36]	mistral-small-2502	Feb 2025	0.0	0.7	2048	Standard Prompting
Mistral Small (R)[36]	mistral-small-2502	Feb 2025	0.0	0.7	8192	CoT Prompting Enabled
Gemma 3[37]	gemma-3-27b-it	March 2025	0.0	1.0	2048	Standard Prompting
Gemma 3 (R)[37]	gemma-3-27b-it	March 2025	0.0	1.0	8192	CoT Prompting Enabled
MedGemma[38]	medgemma-27b-v1	July 2025	0.0	1.0	2048	Domain Tuned
Qwen-3[39]	qwen-3-max-2025-05	May 2025	0.0	0.8	2048	Standard Mode
Qwen-3 (R)[39]	qwen-3-max-2025-05	May 2025	0.0	0.8	8192	Reasoning Budget Active

ARTICLE IN PRESS

critical-care questions and broader clinical QA, albeit typically below top proprietary systems[40].

Llama 4 Maverick

Llama 4 Maverick is a mixture-of-experts (MoE) model with 17 billion active parameters across 128 experts, instruction-tuned and multimodal. Tested under both default and reasoning-enabled modes across 119 dental multiple-choice questions. Llama 3.1 lineage demonstrated strong performance on clinical MCQs and diagnostic tasks, offering proximate evidence of the family's exam-like capabilities[35].

Gemma 3

A 27-billion-parameter instruction-tuned model supporting text and image inputs with a 128K token context window. Evaluated under default and reasoning-enabled conditions across 119 dental multiple-choice questions . It supports multimodal inputs (text and images), an extended 128K token context window, and provides strong baseline performance on reasoning, summarization, and multilingual tasks—achieved through distillation and post-training that leverages reinforcement learning (RL) from human, machine, and execution feedback[37,41].

MedGemma

Derived from Gemma-3, MedGemma is a 27B text-only model designed to enhance medical text comprehension. According to its 2025 technical report, MedGemma demonstrates advanced medical understanding and reasoning while maintaining the general capabilities of the Gemma 3 base models[37,38].

Qwen3

The flagship MoE model of the Qwen3 family (235-billion total parameters; $\tilde{2}$2-billion activated per token), integrating thinking and non-thinking modes with a user-adjustable reasoning budget. Post-training explicitly targets long chain-of-thought (CoT) via a four-stage regimen—long-CoT cold start, reasoning RL, thinking-mode fusion (with think vs. no think flags in the chat template), and general RL—after which increasing the thinking budget generally improves performance on reasoning centric benchmarks[39,42].

DeepSeek R1

A reinforcement-learning-based reasoning model trained to support chain-of-thought (CoT) reasoning. Beyond medical reasoning tasks such as USMLE-style items and clinical case analyses, recent peer reviewed research has also assessed DeepSeek in a dental context, showing competitive accuracy on multiple-choice endodontics questions derived from the Dentistry Specialization Entrance Exam (DUS)[34].

OpenAI GPT-OSS

An open-weight MoE reasoning model with 120-billion total parameters, released under the Apache 2.0 license[32]. The model was designed as a publicly accessible counterpart to OpenAI's proprietary reasoning models, with the aim of supporting reproducible academic research and fostering broader community development of reasoning-oriented LLMs. Independent evaluations on medical QA datasets indicate that GPT-OSS performs competitively with leading closed-weight systems, though peer reviewed studies have not yet investigated its application in dental education or specialty examinations, leaving its reliability in this domain untested[43].

GPT-5

Provider-hosted model with text-only and multimodal capabilities[30]. Under a unified evaluation protocol, GPT-5 demonstrated state-of-the-art performance surpassing GPT-4o and even licensed human experts across multiple medical QA datasets such as MedQA, MMLU-Medical, USMLE self-assessment, and visual tasks like MedXpertQA and VQA-RAD[15,16].

Gemini 2.5 Pro

Positioned as Google's most advanced reasoning-capable model, Gemini 2.5 Pro leverages a sparse MoE architecture with native multimodal support and a context window of up to one million tokens[28,29]. It delivers state-of-the-art performance on benchmarks assessing coding, math, reasoning, and long context comprehension, such as AIME 2025, GPQA and various factual grounding tasks. In the medical domain, Gemini 2.5 Pro was evaluated using MRCGP-style primary care MCQs, achieving an impressive 95% accuracy—far exceeding the average clinician peer score of 73%[44]. At the time of data collection (March 2025), Gemini 2.5 Pro was accessed as an experimental preview release, and performance characteristics may evolve in the General Availability (GA) version.

Grok 4

Developed as a provider-hosted system with tool-integrated reasoning, Grok 4 includes native tool use, real-time search, and structured output capabilities; xAI has not publicly disclosed its parameter count[31,45].

ARTICLE IN PRESS

GPT-4o

GPT-4o is the most recent multimodal model in the GPT-4 family, released by OpenAI in 2024. It is accessible only as a provider-hosted system, and specific details such as parameter count or internal architecture have not been publicly disclosed[33,46].

Experimental Protocol

All models were evaluated under a standardized zero-shot setting[47]. Each question was presented in plain text with five options and accompanied by an identical system prompt:

"You are a board-certified dental expert. Answer the following multiple-choice question by returning ONLY the letter of the best option (A, B, C, D, or E). Do not output any explanation."

The system prompt explicitly instructed models to return 'ONLY the letter'. Despite this, post-processing (as described above) was applied to handle verbose outputs. No in-context examples, fine-tuning, or retrieval augmentation were applied. Each model produced exactly one output per question. To maximize reproducibility, temperature was set to 0.0 for all models where the API permitted. However, due to provider-imposed constraints, GPT-5 was evaluated at its fixed default temperature of 1.0. Consequently, while inference for most models was strictly deterministic, GPT-5 results reflect its default probabilistic behavior. Reasoning features were enabled for models that natively support them, creating two distinct evaluation tiers: standard and reasoning-enhanced.

Data Parsing and Scoring Logic

To ensure consistent scoring, model outputs were processed using a standardized parsing script. A regular expression was employed to extract the first valid option letter (A-E) from the response text. Formats such as '(A)', 'Option A', or 'A)' were normalized to 'A'. Any response that did not contain a clear selection, including empty strings, formatting errors, or policy-based refusals (e.g., 'I cannot answer medical questions'), was recorded as 'incorrect'. No points were awarded for partial answers.

Evaluation Metrics

Accuracy was calculated as the proportion of correct answers relative to the 119-item dataset. This measure directly reflected a model's ability to align with the official key provided by OSYM. Average response time was recorded as wall-clock API latency recorded during inference, reported in seconds. Both accuracy and timing were reported separately for default and reasoning-enabled conditions to assess potential trade-offs between reasoning effort and computational overhead.

Computational Environment

All evaluations were conducted using official API access. Inference latency was automatically captured within the request loop. No external optimizations, caching, or prompt engineering were applied. Fine-tuning, retrieval-augmented generation, or third-party frameworks were not employed. API requests were initiated from Istanbul, Türkiye, using a standard commercial fiber connection. Requests were processed sequentially (no concurrency) to avoid rate-limit penalties. Streaming was disabled to capture the full response generation time. Consequently, reported times reflect the total 'wall-clock' latency observed in this specific environment.

Statistical Analysis

All analyses were conducted using IBM SPSS Statistics for Windows (version 25; IBM Corp., Armonk, NY, USA). Descriptive statistics (mean, standard deviation, minimum, maximum) were reported along with skewness and kurtosis values to examine the normality of the distributions[48]. Differences between models were assessed using the Friedman test, and when significant results were obtained, pairwise comparisons were performed using the Wilcoxon signed-rank test with Bonferroni correction[49]. A significance level of $p < 0.05$ was adopted for all statistical tests.

Experimental Results

Table 2 presents the descriptive statistics of the measurements obtained from 17 LLMs. The mean values of the models ranged from 0.496 (Qwen-3 (R)) to 0.924 (Gemini2.5Pro). The relatively low standard deviations indicate that the accuracy rates were largely clustered within similar ranges.

The normality of the distributions was assessed using skewness and kurtosis statistics. Following the ± 1.5 thresholds recommended by Tabachnick and Fidell[48], it was observed that the distributions of some LLMs deviated from normality. In particular, the skewness values of Gemini2.5Pro, GPT5, and Grok4 were critically negative, while their kurtosis values were markedly high, indicating a pronounced peakedness of the distributions. Moreover, several other models, including GPT4o, GPTOSS, GPTOSS (R), Llama4M (R), Qwen3, DeepSeekR1, and Qwen3 (R), also exceeded the ± 1.5 threshold in skewness

ARTICLE IN PRESS

Table 2. Descriptive statistics of the measurements obtained from the LLMs

LLM (R – Reasoning Mode)	Accuracy (Over 119 MCQs)	Number of Correct Answers	SD	Skewness	Kurtosis
Gemini2.5Pro	0.924	110	0.266	-3.251	8.716
GPT5	0.908	108	0.266	-3.251	8.716
Grok4	0.882	105	0.324	-2.404	3.843
GPTOSS	0.79	94	0.409	-1.442	0.079
GPT4o	0.782	93	0.415	-1.38	-0.098
GPTOSS (R)	0.756	90	0.431	-1.209	-0.547
DeepSeekR1 (R)	0.748	89	0.436	-1.156	-0.674
DeepSeekR1	0.723	86	0.45	-1.008	-1.002
Llama 4 Maverick (R)	0.706	84	0.458	-0.915	-1.182
Llama 4 Maverick	0.697	83	0.461	-0.871	-1.263
Mistral Small	0.655	78	0.477	-0.663	-1.588
Mistral Small (R)	0.655	78	0.477	-0.663	-1.588
Gemma3	0.613	73	0.489	-0.472	-1.808
Gemma3 (R)	0.613	73	0.489	-0.472	-1.808
MedGemma	0.588	70	0.494	-0.363	-1.9
Qwen3	0.546	65	0.5	-0.188	-1.999
Qwen3 (R)	0.496	59	0.502	0.017	-2.034

and/or kurtosis coefficients. Based on these findings, non-parametric tests were employed for the comparisons across artificial intelligence models.

The Friedman test was conducted to compare the artificial intelligence tools, and the findings are presented in Table 3. The results of the Friedman test indicated significant differences in accuracy rates among the artificial intelligence models, $\chi^2(16) = 214.171$, $p < .001$. The effect size, calculated as Kendall's $W = .112$, suggested that these differences corresponded to a small effect.

Post-hoc analyses using the Wilcoxon signed-rank test with Bonferroni correction revealed that certain models clearly outperformed others. The Gemini 2.5 Pro model demonstrated significantly higher mean scores compared to all other models except GPT-4, GPT-5, GPTOSS, and Grok-4. Similar patterns were also observed for GPT-5 and Grok-4.

In contrast, Qwen-3 (R), Qwen-3, and MedGemma exhibited notably lower accuracy rates across most comparisons. Overall, these findings highlight Gemini 2.5 Pro and GPT-5 as achieving the highest accuracy levels among the models, whereas Qwen-3, Qwen-3 (R), and MedGemma demonstrated the weakest performance. Topic-based best and worst performances are in Table 4. Notably, this high performance extended to topics heavily reliant on case recognition and diagnostic reasoning. As shown in Table 4, top-performing models like Gemini 2.5 Pro and GPT-5 achieved high accuracy in complex clinical areas such as 'Pulp Therapies' (90.0%) and 'Dental Trauma' (84.2%). This suggests their capability is not limited to factual recall but extends to interpreting clinical scenarios.

Response times also varied considerably as shown in Table 5. Reasoning-enabled (R) models (e.g., GPT-5.0, GPT-4.0 Reasoning, Gemma-3 Reasoning) tended to require longer times to generate responses. The fastest models were Gemma 3, Mistral Small, and LLaMA 4 Maverick, while the slowest were Qwen-3 (R) and DeepSeek R1 (R). This highlights a trade-off between accuracy and time, with higher accuracy often requiring longer processing.

Discussion

In this study, 11 recently state of the art developed LLMs (17 models with reasoning mode activations) were evaluated on pediatric dentistry questions from the DUS, and their accuracy rates and average response generation times were compared. The findings revealed considerable variation in performance among the models. The highest accuracy rates were achieved by Gemini 2.5 Pro (92.44%), GPT-5 (90.76%), and Grok-4 (88.24%). The performance of these three models approached expert-level accuracy. Moderate accuracy levels were observed in DeepSeek R1 (Standard and High Reasoning modes, 72.26 - 74.78%), OpenAI GPT-OSS (Standard 78.99%, High Reasoning 75,63%), and GPT-4o (78.15%). While these models achieved reliable performance within the 74–79% range, they fell short compared to the top-tier models. In contrast, lower-performing models included Qwen3 (Standard 54.62%, High Reasoning 49.57%), MedGemma (58.82%), Gemma 3 (61.34%), Mistral Small (65.54%), and LLaMA 4 Maverick (69.74–70.58%). Their accuracy rates, ranging between 52% and 71%, were not clinically adequate. Gemini 2.5 Pro's superior performance may be explained by its training on larger, multimodal corpora and fine-grained instruction tuning, which enhance its ability to manage domain-specific terminology and multi-step reasoning.

ARTICLE IN PRESS

Table 3. Comparison of accuracy rates of LLMs

LLMs (R - Reasoning mode)	Accuracy	SD	Mean rank	Chi-square	p	Kendall W	Difference
Gemini 2.5 Pro (1)	0.924	0.266	10.807				1 > 2;3;7;9-17
GPT-5 (5)	0.908	0.266	10.807				5 > 7;9-17
Grok-4 (8)	0.882	0.324	10.45				8 > 9-16
GPTOSS (6)	0.79	0.409	9.664				6 > 11;14;15
GPT-4o (4)	0.782	0.415	9.592				4 > 11;14;15
GPTOSS (R)(7)	0.756	0.431	9.378				7 > 11;14;15
DeepSeek (R)(17)	0.748	0.436	9.307				–
DeepSeek-R1 (16)	0.723	0.45	9.092				–
Llama 4 Maverick (R) (10)	0.706	0.458	8.95	214.171	$< .001$	0.112	10 > 15
Llama 4 Maverick (9)	0.697	0.461	8.878				9 > 15
Mistral Small (12)	0.655	0.477	8.521				–
Mistral Small (R) (13)	0.655	0.477	8.521				–
Gemma-3 (2)	0.613	0.489	8.164				4;5;6;8 > 2
Gemma-3 (R)(3)	0.613	0.489	8.164				4;5;6;8 > 3
MedGemma (11)	0.588	0.494	7.95				17 > 11
Qwen-3 (14)	0.546	0.5	7.592				15;16 > 14
Qwen-3 (R) (15)	0.496	0.502	7.164				16;17 > 15

ARTICLE IN PRESS

Table 4. Topic-Based Best and Worst Performances of LLMs

Subspeciality	Best Performing Model(s)	Accuracy (%)	Worst Performing Model(s)	Accuracy (%)
Tooth eruption, growth & development (13 Qs)	Gemini 2.5 Pro / GPT 5 (12/13)	92.3%	Mistral Small (R) (3/13)	23.1%
Developmental anomalies (19 Qs)	Gemini 2.5 Pro / GPT 5 (17/19)	89.5%	Qwen3 (R) (8/19)	42.1%
Caries prevention & management (11 Qs)	GPT 5 / Gemini 2.5 Pro / Grok 4 (10/11)	90.9%	Qwen3 (R) (4/11)	36.4%
Restorative procedures & materials (6 Qs)	GPT 5 / Gemini 2.5 Pro / Grok 4 / GPT-OSS / GPT-4o (6/6)	100%	Qwen3 (R) / LLaMA (R) / Mistral (R) (2–3/6)	33.3–50.0%
Pulp therapies (10 Qs)	GPT 5 / Gemini 2.5 Pro (9/10)	90.0%	Mistral (R) (1/10)	10.0%
Dental trauma (19 Qs)	GPT 5 (16/19)	84.2%	Qwen3 (R) (5/19)	26.3%
Behavior management & sedation (9 Qs)	Gemini 2.5 Pro (9/9)	100%	Mistral Small (5/9)	55.6%
Space maintainers / Interceptive ortho (5 Qs)	GPT-OSS (R) / GPT 5 / Gemini 2.5 Pro / Grok 4 (4/5)	80.0%	Qwen3 / Gemma / MedGemma (2/5)	40.0%
Systemic diseases, infections & pharm. (23 Qs)	GPT 5 / Gemini 2.5 Pro (19/23)	82.6%	Qwen3 (9/23)	39.1%
Periodontal diseases (2 Qs)	Most models (2/2)	100%	GPT-OSS (R)(1/2)	50.0%
Local anesthesia (2 Qs)	GPT 5 / Gemini 2.5 Pro / Grok 4 / GPT-OSS / GPT-4o (2/2)	100%	Mistral Small (0/2)	0%

ARTICLE IN PRESS

Table 5. Observed Wall-Clock Latency of LLMs (sec)

LLM (R – Reasoning Mode)	Average Generation Time (sec)
GPT-4o	1.2
GPT-5	1.18
Gemini 2.5 Pro	9.05
Grok 4	21.71
Mistral Small	0.49
DeepSeek R1	28
LLaMA 4 Maverick	0.57
Gemma 3	0.34
Qwen3	11.2
MedGemma	2.1
OpenAI GPT-OSS	3.3
DeepSeek R1 (R)	36
Mistral Small (R)	2.40
LLaMA 4 Maverick ((R)	3.4
Gemma 3 (R)	0.70
Qwen3 (R)	68.24
OpenAI GPT-OSS (R)	12.56

Its 'thinking model' design also reduces hallucination and improves consistency, providing an advantage over GPT-5 and Grok-4[12,22,25].

Beyond overall accuracy rates, topic-based analyses provided further insights into the domain-specific strengths and weaknesses of the models. Subgroup analyses by topic revealed that the performance of LLMs was not homogeneous. In restorative procedures, periodontal diseases, and local anesthesia, several models achieved 100% accuracy, suggesting that AI models tend to perform better in areas dominated by knowledge-based and standardized approaches. On the other hand, models such as Qwen3, Mistral, and MedGemma demonstrated weaker performance with lower accuracy rates, particularly in caries management, trauma, and pulp therapies, which require more complex clinical reasoning. These findings indicate that LLMs have topic-specific strengths and weaknesses, and model selection for clinical and educational applications should take these differences into account.

Restorative Procedures & Materials, Periodontal Diseases, and Behavior Management models demonstrated perfect accuracy rates, while they also showed promising results in critical diagnostic areas such as case recognition and clinical decision-making. A key implication of this study is the strong potential for state-of-the-art LLMs to function as diagnostic decision-support systems in dentistry. This is evidenced by their high performance not just on knowledge-based questions, but specifically in areas requiring case recognition and diagnostic reasoning. For instance, top models achieved perfect (100%) accuracy in 'Restorative Procedures & Materials' and 'Behavior Management' , and over 90% in 'Caries Prevention & Management' and 'Pulp Therapies'. The ability to accurately navigate these complex, vignette-based questions—which form the core of clinical practice—suggests a robust capability for diagnostic support. However, the weaker performance of models like Qwen3 and MedGemma in these same areas indicates that not all LLMs possess this diagnostic acumen, making model-specific validation essential.

In terms of response times, Gemma 3 (0.34 s), LLaMA 4 (0.57 s), and Mistral Small (0.49 s) were the fastest models but exhibited relatively low accuracy. GPT-5 (1.18 s), GPT-4o (1.2 s), and Gemini 2.5 Pro (9 s) fell within the medium speed category while maintaining high accuracy, thus demonstrating a balanced performance. Models such as Qwen3 High Reasoning (68 s) and DeepSeek R1 High (36 s) showed improved accuracy but at the cost of excessively long response times, which limits their practical utility. These findings highlight a clear trade-off between accuracy and speed. High reasoning modes increased accuracy but substantially prolonged response times, reducing usability in real-time educational or clinical contexts[14,50]. Therefore, GPT-5 and Gemini 2.5 Pro emerged as the most balanced models, while Grok-4 also achieved high accuracy but with relatively longer response times. Although Gemma and LLaMA models offered speed advantages (<1 sec), they were inadequate in terms of accuracy, and reasoning-focused models like DeepSeek did not achieve acceptable response times for practical applications.

Recent evidence has demonstrated that Gemini 2.5 Pro achieves remarkable performance in radiology-related tasks. For instance, in MRI interpretation, ChatGPT-4o and Gemini 2.5 Pro both reached 100% accuracy in identifying the imaging plane and 98.46% in determining contrast-enhancement status[51]. Moreover, MRI sequence classification accuracy was reported as 93.1% for Gemini 2.5 Pro, compared with 97.7% for ChatGPT-4o and only 73.1% for Claude 4 Opus. These results highlight

ARTICLE IN PRESS

Gemini 2.5 Pro's strong capability in handling complex domain-specific diagnostic tasks, further supporting its high accuracy in pediatric dentistry questions observed in our study.

Recent studies have shown that GPT-5 consistently outperforms baseline models, including GPT-4o, across text-based and multimodal medical benchmarks, in line with our findings[16]. Notably, GPT-5 achieved significant gains in reasoning and comprehension, surpassing GPT-4o and even exceeding human expert performance, underscoring a major advancement in LLM capabilities[16,51]. Recent work in ophthalmology has shown that GPT-5 models achieved near-perfect accuracy and superior rationale quality on board-style multiple-choice questions, clearly surpassing GPT-4o and earlier reasoning models, which further supports our findings in pediatric dentistry[52]. Contrarily, in brain tumor MRI tasks[53], models GPT-5 family were reported to achieve only moderate accuracy, with GPT-5 reaching 43.71% and GPT-5-mini performing similarly at 44.19%, indicating that their performance remains insufficient for clinical application[54]. In mammography tasks, although GPT-5 demonstrated improved performance compared to earlier GPT versions, it still lagged behind human experts and domain-specific fine-tuned models[55]. Consequently, it has been reported as unsuitable for direct use in high-risk clinical applications, with sensitivity and specificity values of 63.5% and 52.3%, respectively, compared to 86.9% and 88.9% achieved by human specialists[56–58].

Introduced on July 9, 2025, by Elon Musk under xAI, Grok-4 was presented as a next-generation artificial intelligence system, claimed to possess problem-solving abilities surpassing the doctoral level across all academic domains[45]. Research conducted with Grok-4 in the literature is extremely limited. Supporting our observations, a recent study in cardiology highlighted Grok's superiority in management-related tasks, outperforming other LLMs in areas such as congenital heart disease and heart failure, while GPT-4 showed slightly better overall performance and Gemini performed best in electrophysiology, though without statistical significance[13]. Overall, these studies show encouraging yet mixed results across different LLMs and highlight the need for more systematic research, particularly on Grok-4 where evidence is still very limited.

Future Work and Limitations

This study is limited by its reliance on a single dataset of pediatric dentistry questions from the DUS over a ten-year period, which may restrict the generalizability of the findings to other examinations. Moreover, as large language models are rapidly evolving, the results reflect only the performance of the most recent versions available at the time of data collection and may not be representative of future iterations. Another potential limitation is data contamination, as some exam-related content may have been included in the training data of the models, potentially influencing their performance[6,22]. Additionally, the lack of uniformity in temperature settings poses a limitation to direct comparability. Specifically, GPT-5 could not be adjusted to a temperature of 0.0 due to API restrictions, meaning its high performance must be interpreted with the caveat that it operated with higher variance (temperature=1.0) compared to the strictly deterministic settings of other models. Furthermore, latency measurements were derived from a single-pass evaluation. Due to the lack of repeated trials (n=3-5), we could not calculate stable measures of dispersion such as Interquartile Range (IQR). Thus, the reported times should be interpreted as indicative of the user experience under specific conditions rather than a robust benchmark of model compute speed. Finally, response time measurements reported in this study represent wall-clock API latency. These values are subject to external variables such as network conditions, server load, and rate limiting, and thus should be interpreted as a measure of real-world user wait time rather than intrinsic model generation speed.

Conclusions

This benchmarking study showed that large language models display variable performance on pediatric dentistry questions from a dental specialization exam. Gemini 2.5 Pro (92.44%), GPT-5 (90.76%), and Grok-4 (88.24%) achieved the highest accuracy, clearly outperforming GPT-4o and smaller models. While reasoning-focused settings improved accuracy for some models, they required substantially longer generation times, indicating a trade-off between speed and reliability[12,22,25]. Overall, the newest LLMs approach expert-level performance, but variability across models and slower response in reasoning-intensive modes highlight the need for further refinement before widespread use in high-stakes assessment or education[6,14].

Funding

No external funding was obtained for this study.

Competing interests

The authors declare no conflicts of interest.

ARTICLE IN PRESS

Data availability

The datasets generated and/or analysed during the current study are available in the https://www.osym.gov.tr/TR,15070/dus-cikmis-sorular.html accessed on 25.11.2025.

Author contributions statement

STK: concept and study design, statistical analysis, preparation of the figures, wrote the main manuscript text. HOI: analysis of the data, wrote the main manuscript text. MT and HD: study design, wrote the main manuscript text. TC: Analysis of the data. All authors reviewed and approved the final version of the manuscript.

References

1. Fang, Q. *et al.* Artificial intelligence (ai)-driven dental education: Exploring the role of chatbots in a clinical learning environment. *The J. Prosthet. Dent.* (2024).
2. Uribe, S. E. *et al.* Artificial intelligence chatbots and large language models in dental education: Worldwide survey of educators. *Eur. J. Dental Educ.* **28**, 865–876 (2024).
3. Gordon, M. *et al.* A scoping review of artificial intelligence in medical education: Beme guide no. 84. *Med. Teach.* **46**, 446–470 (2024).
4. Brown, T. *et al.* Language models are few-shot learners. *Adv. neural information processing systems* **33**, 1877–1901 (2020).
5. Russell, S. & Norvig, P. *Artificial Intelligence: A Modern Approach (4th Edition)* (Pearson, 2020).
6. Sallam, M. Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare*, vol. 11, 887 (MDPI, 2023).
7. Al-Moghrabi, D., Arqub, S. A., Maroulakos, M. P., Pandis, N. & Fleming, P. S. Can chatgpt identify predatory biomedical and dental journals? a cross-sectional content analysis. *J. Dent.* **142**, 104840 (2024).
8. Najeeb, M. & Islam, S. Artificial intelligence (ai) in restorative dentistry: current trends and future prospects. *BMC Oral Heal.* **25**, 592 (2025).
9. Bichu, Y. M. *et al.* Applications of artificial intelligence and machine learning in orthodontics: a scoping review. *Prog. orthodontics* **22**, 18 (2021).
10. Albano, D. *et al.* Artificial intelligence for radiographic imaging detection of caries lesions: a systematic review. *BMC Oral Heal.* **24**, 274 (2024).
11. Alkhaaldi, S. M. *et al.* Medical student experiences and perceptions of chatgpt and artificial intelligence: cross-sectional study. *JMIR Med. Educ.* **9**, e51302 (2023).
12. Kung, T. H. *et al.* Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. *PLoS digital health* **2**, e0000198 (2023).
13. Liu, M. *et al.* Large language models in dental licensing examinations: systematic review and meta-analysis. *Int. Dental J.* **75**, 213–222 (2025).
14. Kim, W., Kim, B. C. & Yeom, H.-G. Performance of large language models on the korean dental licensing examination: a comparative study. *Int. Dental J.* **75**, 176–184 (2025).
15. AlSaad, R. *et al.* Multimodal large language models in health care: applications, challenges, and future outlook. *J. medical Internet research* **26**, e59505 (2024).
16. Wang, S., Hu, M., Li, Q., Safari, M. & Yang, X. Capabilities of gpt-5 on multimodal medical reasoning. *arXiv preprint arXiv:2508.08224* (2025).
17. Comanici, G. *et al.* Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).
18. Aminoshariae, A., Kulild, J. & Nagendrababu, V. Artificial intelligence in endodontics: current applications and future directions. *J. endodontics* **47**, 1352–1357 (2021).
19. Revilla-León, M. *et al.* Artificial intelligence applications in restorative dentistry: A systematic review. *The J. prosthetic dentistry* **128**, 867–875 (2022).

ARTICLE IN PRESS

ARTICLE IN PRESS

20. Hung, K., Montalvao, C., Tanaka, R., Kawai, T. & Bornstein, M. M. The use and performance of artificial intelligence applications in dental and maxillofacial radiology: A systematic review. *Dentomaxillofacial Radiol.* **49**, 20190107 (2020).

21. Lazarus, M. D., Truong, M., Douglas, P. & Selwyn, N. Artificial intelligence and clinical anatomical education: promises and perils. *Anat. Sci. Educ.* **17**, 249–262 (2024).

22. Schwendicke, F. a., Samek, W. & Krois, J. Artificial intelligence in dentistry: chances and challenges. *J. dental research* **99**, 769–774 (2020).

23. T.C. Ölçme, S. v. Y. M. O. Diş hekimliğinde uzmanlık eğitimi giriş sınavı (dus) (2025).

24. Sismanoglu, S. & Capan, B. S. Performance of artificial intelligence on turkish dental specialization exam: can chatgpt-4.0 and gemini advanced achieve comparable results to humans? *BMC Med. Educ.* **25**, 214 (2025).

25. Tangadulrat, P., Sono, S., Tangtrakulwanich, B. *et al.* Using chatgpt for clinical practice and medical education: cross-sectional survey of medical students' and physicians' perceptions. *JMIR medical education* **9**, e50658 (2023).

26. Alotaibi, S. & Deligianni, E. Ai in oral medicine: is the future already here? a literature review. *Br. Dental J.* **237**, 765–770 (2024).

27. Çoban, D., Yaşa, Y., Aktaş, A. & İlhan, H. O. Detection of jaw lesions on panoramic radiographs using deep learning method. *J. Imaging Informatics Medicine* 1–16 (2025).

28. DeepMind, G. Gemini 2.5 pro: Our most intelligent ai model (2025).

29. Team, G. D. Gemini 2.5: A multimodal mixture-of-experts for complex reasoning. *Google Deep. Tech. Rep.* (2025).

30. Singh, A. *et al.* Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267* (2025).

31. xAI. Grok 4 model card. *xAI Tech. Rep.* (2025).

32. Agarwal, S. *et al.* gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925* (2025).

33. Hurst, A. *et al.* Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

34. Guo, D. *et al.* Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature* **645**, 633–638 (2025).

35. Yin, S. *et al.* A survey on multimodal large language models. *Natl. Sci. Rev.* **11**, nwae403 (2024).

36. AI, M. Mistral small 3: Efficiency at scale (2025).

37. Team, G. *et al.* Gemma 3 technical report. *arXiv preprint arXiv:2503.19786* (2025).

38. Sellergren, A. *et al.* Medgemma technical report. *arXiv preprint arXiv:2507.05201* (2025).

39. Yang, A. *et al.* Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).

40. Qin, L. *et al.* A survey of multilingual large language models. *Patterns* **6** (2025).

41. Google. Google ai for developers. gemma 3 model overview (2025).

42. Aydin, O., Karaarslan, E., Erenay, F. S. & Bacanin, N. Generative ai in academic writing: A comparison of deepseek, qwen, chatgpt, gemini, llama, mistral, and gemma. *arXiv preprint arXiv:2503.04765* (2025).

43. Leon, M. Gpt-5 and open-weight large language models: Advances in reasoning, transparency, and control. *Inf. Syst.* 102620 (2025).

44. Armitage, R. Performance of leading large language models in may 2025 in membership of the royal college of general practitioners-style examination questions: a cross-sectional analysis. *arXiv preprint arXiv:2506.02987* (2025).

45. Adinath, D. & IS, S. Chatgpt-5 and grok-4: A comparative analysis of advancements in nlp. *Available at SSRN 5547358* (2025).

46. Shahriar, S. *et al.* Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency. *Appl. Sci.* **14**, 7782 (2024).

47. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. & Iwasawa, Y. Large language models are zero-shot reasoners. *Adv. neural information processing systems* **35**, 22199–22213 (2022).

48. Tabachnick, B., Fidell, L. & Ullman, J. *Using multivariate statistics* (Pearson, 2007).

49. Field, A. *Discovering statistics using IBM SPSS statistics* (Sage publications limited, 2024).

50. Corbella, S., Srinivas, S. & Cabitza, F. Applications of deep learning in dentistry. *Oral Surgery, Oral Medicine, Oral Pathol. Oral Radiol.* **132**, 225–238 (2021).

ARTICLE IN PRESS

ARTICLE IN PRESS

51. Salbas, A. & Buyuktoka, R. E. Performance of large language models in recognizing brain mri sequences: A comparative analysis of chatgpt-4o, claude 4 opus, and gemini 2.5 pro. *Diagnostics* **15**, 1919 (2025).

52. Antaki, F. *et al.* Performance of gpt-5 frontier models in ophthalmology question answering. *arXiv preprint arXiv:2508.09956* (2025).

53. Safari, M. *et al.* Performance of gpt-5 in brain tumor mri reasoning. *arXiv preprint arXiv:2508.10865* (2025).

54. Ozenbas, C. *et al.* Chatgpt-4o's performance in brain tumor diagnosis and mri findings: A comparative analysis with radiologists. *Acad. Radiol.* (2025).

55. Pesapane, F. *et al.* A preliminary investigation into the potential, pitfalls, and limitations of large language models for mammography interpretation. *Discov. Oncol.* **16**, 233 (2025).

56. Li, Q. *et al.* Is chatgpt-5 ready for mammogram vqa? *arXiv preprint arXiv:2508.11628* (2025).

57. Shen, Y. *et al.* Multi-modal large language models in radiology: principles, applications, and potential. *Abdom. Radiol.* **50**, 2745–2757 (2025).

58. Miaojiao, S. *et al.* Using a large language model for breast imaging reporting and data system classification and malignancy prediction to enhance breast ultrasound diagnosis: Retrospective study. *JMIR Med. Informatics* **13**, e70924 (2025).

ARTICLE IN PRESS