Evaluating Frontier Language Models on Clinician-Reviewed Dental Questions: A Reproducible Benchmark

Francisco Teixeira Barbosa
Foundation for Oral Rehabilitation
francisco@for.org

Daniel Robles Cantero
Universidad Europea Miguel de Cervantes
drobles@clinica.uemc.es

Aritza Brizuela Velasco
Universidad Europea Miguel de Cervantes
abrizuela@uemc.es

June 11, 2026

Abstract

Large language models (LLMs) are increasingly used to answer clinical questions, but general medical benchmarks do not directly test whether models can answer guideline-grounded dental questions that occur in periodontology, implant dentistry, oral-systemic medicine, pharmacology, and patient communication. We present a clinician-reviewed dental question-answering benchmark containing 30 open-ended questions across six clinical domains, each paired with explicit *must include* and *must avoid* rubric criteria. We evaluated eight contemporary LLMs available through OpenRouter on June 10–11, 2026, using single-trial, temperature-0 prompting and a rubric-based LLM judge. The leading deployment accuracies were GPT-5.2 at 96.7%, Claude Opus 4.8 and GPT-5.5 at 93.3%, and Gemini 3.1 Pro at 90.0%, but bootstrap confidence intervals over questions were wide and overlapped substantially. Claude Fable 5 answered only 25 of 30 questions because of refusals concentrated in oral-systemic topics; on answered questions it reached 96.0% accuracy. Two additional OpenAI-family judges showed only moderate agreement with the primary Claude Opus 4.8 judge (81.7–83.8% verdict agreement; Cohen's $\kappa = 0.506$–0.524), emphasizing that absolute accuracy levels are judge-dependent. The benchmark, transcripts, rubrics, and analysis code are released to support reproducible, specialty-specific evaluation of LLMs in dentistry.

1 Introduction

Medical question answering has become a central test case for large language models. MultiMedQA and Med-PaLM showed that LLMs can encode substantial clinical knowledge while also exposing gaps in factuality, reasoning, potential harm, and bias that matter for clinical deployment [1]. HealthBench extended open-ended health evaluation through physician-written rubric criteria, underscoring the value of explicit scoring rubrics over answer-only multiple-choice accuracy [2]. Biomedical QA benchmarks such as PubMedQA provide useful tests of biomedical reasoning over abstracts [3], but dentistry-specific clinical questions remain underrepresented in widely used benchmarks.

Dentistry is not a narrow variant of general medicine for evaluation purposes. Clinically useful dental answers often depend on specialty-specific consensus frameworks and guidelines, including the 2017 World Workshop classification of periodontal and peri-implant diseases [4, 5], the EFP

S3 treatment guideline for stages I–III periodontitis [6], AAOMS guidance on medication-related osteonecrosis of the jaw [7], and current infective-endocarditis prophylaxis statements [8]. Small omissions can change clinical meaning: for example, failing to distinguish periodontal stability endpoints from treatment thresholds, or giving jurisdiction-insensitive antibiotic prophylaxis advice, may produce a fluent but unsafe answer.

Recent dental LLM studies have evaluated models on dental multiple-choice questions [9], implantology scenarios [10], and open-ended periodontology examination answers [11]. Broader dental benchmark efforts include DentalBench [12] and the concurrent GlobalDentBench [13]. GlobalDentBench is substantially larger and multinational, whereas the present work contributes a smaller but fully public, auditable benchmark focused on clinically realistic, guideline-grounded dental questions with explicit scoring rubrics, answer transcripts, judge verdicts, and cross-judge agreement analysis.

2 Methods

2.1 Benchmark Dataset

The benchmark contains 30 open-ended questions across six domains: periodontal diagnosis, periodontal treatment, implants and peri-implantitis, oral-systemic medicine, pharmacology, and patient communication. Each domain contributes five questions. Question difficulty labels were assigned as basic (7 questions), intermediate (14 questions), or advanced (9 questions). Each question includes a rubric specifying required concepts and prohibited errors. An answer is scored correct only if it satisfies all required criteria and commits no prohibited error.

The rubrics were authored and reviewed by a periodontist and checked against guideline sources before the reported run. The validation log records verification against primary or authoritative sources for the highest-risk factual claims, including periodontal staging and grading, EFP treatment endpoints, diabetes-periodontitis effect sizes, MRONJ drug classes, infective endocarditis prophylaxis, and anticoagulant management. No patient records, radiographs, or protected health information were used.

2.2 Models and Execution

We evaluated eight models through the OpenRouter chat-completions API on June 10–11, 2026. The model roster is shown in Table 1. All candidate-answer calls used temperature 0 and a maximum completion budget of 12,000 tokens. Each model answered each question once. The run therefore produced 240 model-question rows. Latency was measured as wall-clock time for each candidate-answer request.

2.3 Rubric-Based Judging

The primary judge was Claude Opus 4.8, queried through the same OpenRouter client. The judge received the clinical question, the question-specific rubric, and the candidate answer, and returned a structured JSON verdict containing the number of required criteria satisfied, the total number of required criteria, any prohibited violations, a Boolean correctness verdict, and a brief explanation. This LLM-as-judge design follows the general direction of scalable model-assisted evaluation [14] and health-domain rubric evaluation [2], but we treated judge dependence as an empirical quantity rather than an assumption.

To measure judge dependence, all non-empty stored answers were independently re-scored with GPT-5.2 and GPT-5.5 as secondary judges. This check was motivated by reported self-recognition

Table 1: Model roster evaluated in the reported run. Model identifiers are the OpenRouter identifiers recorded by the benchmark harness.

Label	OpenRouter model identifier	Tier
Claude Fable 5	`anthropic/claude-fable-5`	Flagship
Claude Opus 4.8	`anthropic/claude-opus-4.8`	Flagship
GPT-5.5	`openai/gpt-5.5`	Flagship
GPT-5.2	`openai/gpt-5.2`	Flagship
Gemini 3.1 Pro	`google/gemini-3.1-pro-preview`	Flagship
Qwen3.7 Plus	`qwen/qwen3.7-plus`	Efficient
Llama 4 Maverick	`meta-llama/llama-4-maverick`	Open-weight
DeepSeek V3.2	`deepseek/deepseek-v3.2`	Open-weight

and self-preference biases in LLM-based evaluation [15, 16]. Because Claude Fable 5 refused five questions and produced no answer text for those rows, the paired judge-agreement analyses included 235 answered rows. We report raw verdict agreement and Cohen's kappa for each secondary judge pass.

2.4 Outcomes

The primary outcome was deployment accuracy: the proportion of all 30 questions scored correct for a model, counting refusals and empty answers as incorrect. We also report answer rate, accuracy among answered rows, and mean latency. For per-model uncertainty, we computed bootstrap 95% intervals over questions using 10,000 resamples in `src/analysis.py`, resetting Python's `random.Random(42)` per model. These intervals reflect the small question set and should not be interpreted as formal population-level confidence intervals over all possible dental questions.

3 Results

3.1 Overall Accuracy

Across all 240 model-question rows, primary-judge accuracy was 81.7%. Table 2 summarizes per-model accuracy, answer rate, accuracy on answered rows, and mean latency. GPT-5.2 achieved the highest deployment accuracy at 96.7% (29/30), followed by Claude Opus 4.8 and GPT-5.5 at 93.3% (28/30), Gemini 3.1 Pro at 90.0% (27/30), Qwen3.7 Plus at 83.3% (25/30), Claude Fable 5 at 80.0% (24/30), DeepSeek V3.2 at 70.0% (21/30), and Llama 4 Maverick at 46.7% (14/30).

Figure 1 shows the same ranking with bootstrap intervals. The top four models have heavily overlapping intervals, and this study should not be cited as evidence of a statistically resolved ordering among those frontier systems. The larger separation between the leading cluster and the open-weight models was more pronounced.

3.2 Domain Patterns

Figure 2 shows accuracy by clinical domain. Pharmacology was the clearest differentiator: Llama 4 Maverick scored 0% in this domain, while GPT-5.2, GPT-5.5, Claude Opus 4.8, Claude Fable 5, and DeepSeek V3.2 performed better. DeepSeek V3.2 was weakest in periodontal diagnosis and treatment. Qwen3.7 Plus showed strong performance for an efficient-tier model, with 83.3% deployment accuracy overall.

Table 2: Primary benchmark results. Accuracy counts refusals as incorrect. Confidence intervals are bootstrap intervals over 30 questions.

Model	Accuracy (95% CI)	Answer rate	Accuracy on answered	Mean latency
GPT-5.2	96.7% [90.0, 100]	100%	96.7%	14.8 s
Claude Opus 4.8	93.3% [83.3, 100]	100%	93.3%	12.1 s
GPT-5.5	93.3% [83.3, 100]	100%	93.3%	20.1 s
Gemini 3.1 Pro	90.0% [76.7, 100]	100%	90.0%	20.5 s
Qwen3.7 Plus	83.3% [70.0, 96.7]	100%	83.3%	40.8 s
Claude Fable 5	80.0% [66.7, 93.3]	83.3%	96.0%	16.5 s
DeepSeek V3.2	70.0% [53.3, 86.7]	100%	70.0%	34.1 s
Llama 4 Maverick	46.7% [30.0, 63.3]	100%	46.7%	25.7 s

3.3 Clinical Error Analysis

To interpret errors beyond aggregate scores, we performed a post-hoc clinical review of the 44 model-question rows marked incorrect by the primary judge. This audit did not alter the headline metrics in Table 2, which remain the stored primary-judge verdicts. Excluding five refusals and five primary-judge internal-consistency candidates, 34 rows represented clear clinical answer errors. The clear clinical errors clustered into periodontal treatment endpoints and protocols (10 rows), pharmacology safety and guideline nuance (8 rows), peri-implant evidence overstatement or planning omissions (7 rows), periodontal diagnostic thresholds (6 rows), and patient-communication omissions (3 rows). A full row-level audit is included in the repository.

The clearest clinical pattern was not absence of dental terminology. Many failed answers were fluent and detailed, but missed operational thresholds or overstated evidence beyond supported endpoints. For example, several models correctly linked keratinized mucosa around implants to plaque control, inflammation, recession, and brushing comfort, but then overstated the evidence by implying proven marginal-bone or survival benefit. Similarly, pharmacology failures often concerned safety-critical details: clindamycin persisted as a prophylaxis alternative despite the 2021 AHA change, and some answers advised dentists to alter DOAC dosing without consultation for routine extraction.

The five primary-judge internal-consistency candidates were treated as adjudication flags rather than post-hoc corrections. Retaining them preserves reproducibility of the stored primary-judge endpoint, while showing why future versions should include prespecified manual adjudication for judge-disagreement and logically inconsistent verdicts.

3.4 Refusals and Empty Answers

Claude Fable 5 refused five of 30 questions. The refused topics clustered in oral-systemic evidence and peri-implant soft-tissue topics: diabetes and periodontitis, pregnancy and periodontal treatment, smoking and periodontitis, periodontitis and Alzheimer's disease, and supracrestal tissue attachment around implants. This produced an 83.3% answer rate and an 80.0% deployment accuracy, despite 96.0% accuracy among answered questions. Stored per-row provenance shows that the same refusal behavior reproduced across both Amazon Bedrock and Anthropic first-party serving. For clinical deployment, this distinction matters: a model that refuses clinically relevant questions may be safer than one that hallucinates, but it is not equivalent to a system that gives a correct, calibrated answer.

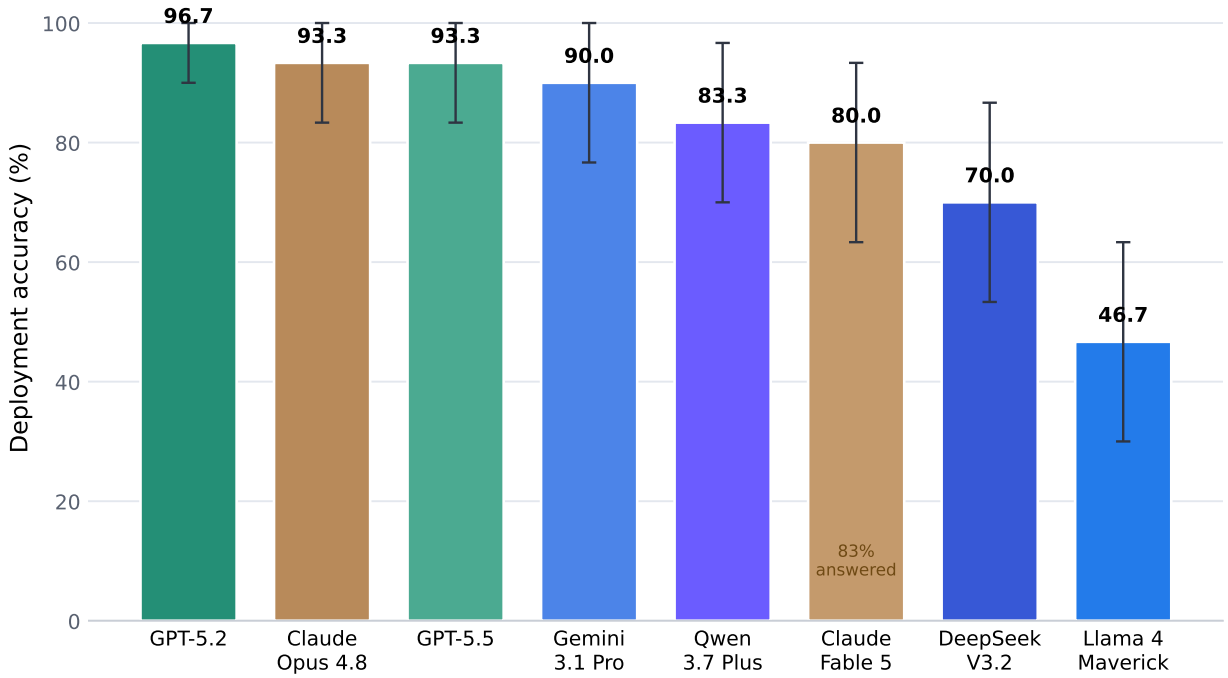

Figure 1: Deployment accuracy by model. Error bars show bootstrap 95% intervals over questions. Refusals are counted as incorrect.

3.5 Judge Agreement

The GPT-5.2 secondary judge agreed with the Claude Opus 4.8 primary judge on 81.7% of paired answered rows, with Cohen's $\kappa = 0.506$. The GPT-5.5 secondary judge agreed on 83.8% of paired answered rows, with $\kappa = 0.524$. Both OpenAI judges were stricter than the primary judge overall. For example, GPT-5.5 scored its own GPT-5.5 answers at 76.7%, whereas the Claude Opus 4.8 primary judge scored the same GPT-5.5 answers at 93.3%. This pattern argues against a simple same-family favoritism explanation and reinforces that absolute accuracy should be interpreted as judge-dependent.

4 Discussion

This benchmark suggests that contemporary frontier LLMs can answer many guideline-grounded dental questions, but it also shows why specialty-specific evaluation remains necessary. The highest-scoring systems performed well on the small question set, yet their confidence intervals overlap, their errors occur in clinically meaningful places, and the judge-agreement analysis shows only moderate inter-judge reliability. Ranking frontier models from a 30-question benchmark would therefore be overconfident. The more defensible conclusion is that a leading cluster performed substantially better than weaker open-weight systems on this particular dental task set, while refusal behavior and domain-specific failure modes changed the deployment interpretation.

Open-ended dental QA differs from multiple-choice dental examinations. Multiple-choice benchmarks are easier to scale and compare, but they can hide clinically important omissions. In this benchmark, a candidate answer can be fluent and plausible yet still fail if it omits a required ac-

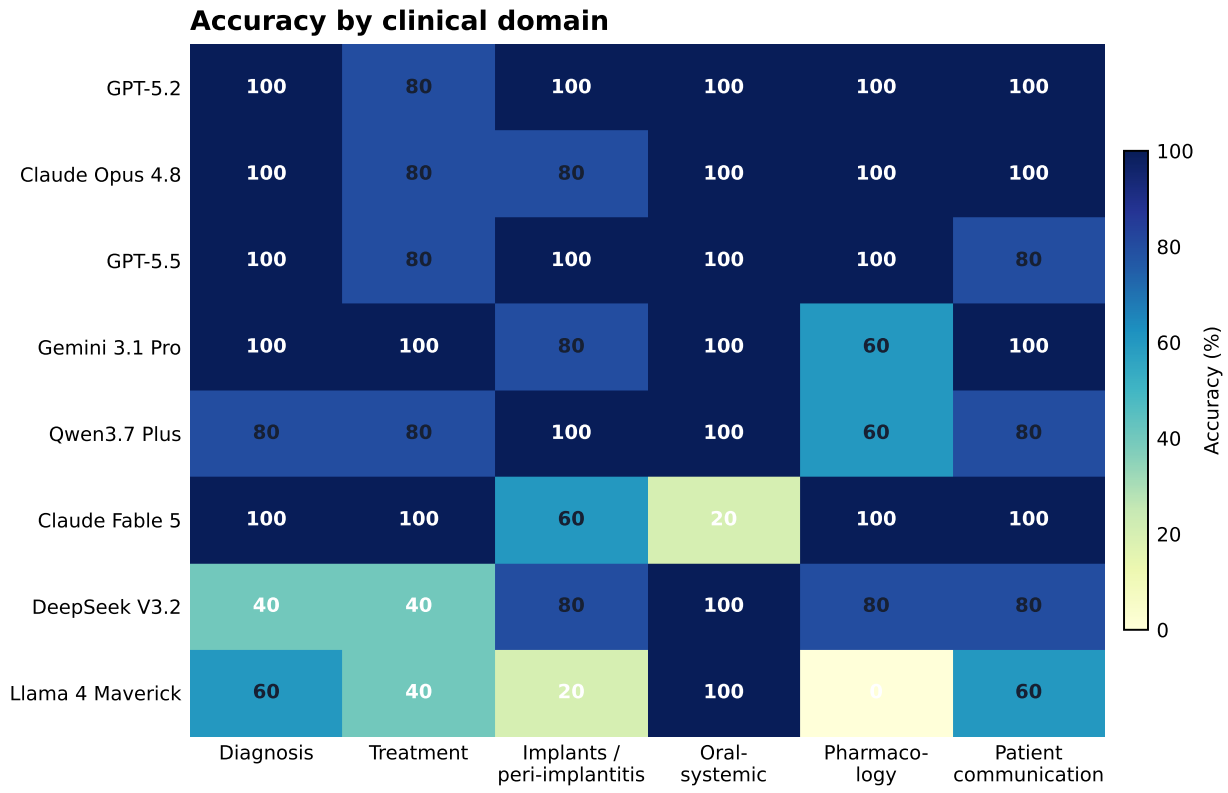

Figure 2: Accuracy by model and clinical domain. Each cell contains five questions.

tion threshold, fails to mention irreversibility of established bone loss in patient communication, or presents a jurisdiction-specific prophylaxis guideline as universal. The rubric-based design is intended to expose those omissions.

At the same time, the method remains preliminary. The question set is small, the rubrics were reviewed by one periodontist, only one candidate-answer trial was used, and LLM-as-judge scoring is not a substitute for independent human adjudication. The secondary-judge passes and internal-consistency audit quantify this problem rather than solve it. Future versions should expand the dataset, include multiple calibrated dental experts, use repeated trials, prespecify manual adjudication for judge-disagreement and internally inconsistent rows, and separate knowledge recall, clinical reasoning, patient communication, and potential-harm scoring.

5 Limitations

First, the benchmark contains only 30 questions, with five per domain. The bootstrap intervals are consequently wide, and small accuracy differences should be treated as noise. Second, the final eight-model run used a single trial at temperature 0. Earlier five-model end-to-end runs showed at most a one-question per-model shift, but full eight-model run-to-run stability was not measured. Third, the primary ground truth is a single-clinician, guideline-checked rubric set rather than a multi-expert consensus panel. Fourth, all primary verdicts rely on an LLM judge, secondary judges showed only moderate agreement with the primary judge, and the post-hoc audit identified five internally inconsistent primary-judge verdicts. Fifth, the results reflect model availability, model

versions, and OpenRouter routing on June 10–11, 2026. Model identifiers and behavior may drift. Finally, the benchmark evaluates text answers to simulated questions; it does not validate clinical deployment, patient outcomes, chairside decision support, or regulatory safety.

6 Data and Code Availability

The benchmark code, dataset, raw JSONL results, transcripts, judge verdicts, plotting scripts, and reproducible analysis script are available at:

https://github.com/Tuminha/llm-evaluation-for-dentistry

The repository commit containing the reported 240-row primary results file, both 235-row secondary judge files, and `src/analysis.py` is:

`416104585625b211732bd7355636fda8d625075f`

A post-hoc row-level clinical error audit is provided in `results/clinical_error_analysis.md`.

7 Ethics Statement

This study used clinician-authored benchmark questions and guideline-derived rubrics. It did not use patient records, radiographs, clinical images, or protected health information. The results should not be interpreted as validating any model for autonomous diagnosis, prescribing, treatment planning, or patient-specific clinical decision-making.

8 Author Contributions

Francisco Teixeira Barbosa designed the benchmark, implemented the evaluation harness, ran the analyses, and drafted the manuscript. Daniel Robles Cantero and Aritza Brizuela Velasco contributed clinical and scientific review and manuscript-development support.

9 Acknowledgments

The authors acknowledge the Foundation for Oral Rehabilitation for support during preparation of this benchmark and manuscript.

References

[1] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620:172–180, 2023. doi:10.1038/s41586-023-06291-2.

[2] R. K. Arora, J. Wei, R. Soskin Hicks, P. Bowman, J. Quinonero-Candela, F. Tsimpourlas, M. Sharman, M. Shah, A. Vallone, A. Beutel, J. Heidecke, and K. Singhal. HealthBench: Evaluating large language models towards improved human health. arXiv:2505.08775, 2025.

[3] Q. Jin, B. Dhingra, Z. Liu, W. Cohen, and X. Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of EMNLP-IJCNLP*, pages 2567–2577, 2019. doi:10.18653/v1/D19-1259.

[4] M. S. Tonetti, H. Greenwell, and K. S. Kornman. Staging and grading of periodontitis: Framework and proposal of a new classification and case definition. *Journal of Periodontology*, 89(Suppl 1):S159–S172, 2018. doi:10.1002/JPER.18-0006.

[5] P. N. Papapanou, M. Sanz, N. Buduneli, T. Dietrich, M. Feres, D. H. Fine, T. F. Flemmig, R. Garcia, W. V. Giannobile, F. Graziani, et al. Periodontitis: Consensus report of workgroup 2 of the 2017 World Workshop on the Classification of Periodontal and Peri-Implant Diseases and Conditions. *Journal of Clinical Periodontology*, 45(Suppl 20):S162–S170, 2018. doi:10.1111/jcpe.12946.

[6] M. Sanz, D. Herrera, M. Kebschull, I. Chapple, S. Jepsen, T. Berglundh, A. Sculean, M. Tonetti, and EFP Workshop Participants. Treatment of stage I–III periodontitis: The EFP S3 level clinical practice guideline. *Journal of Clinical Periodontology*, 47(Suppl 22):4–60, 2020. doi:10.1111/jcpe.13290.

[7] S. L. Ruggiero, T. B. Dodson, T. Aghaloo, R. Carlson, B. B. Ward, and D. Kademani. American Association of Oral and Maxillofacial Surgeons' position paper on medication-related osteonecrosis of the jaws–2022 update. *Journal of Oral and Maxillofacial Surgery*, 80(5):920–943, 2022. doi:10.1016/j.joms.2022.02.008.

[8] W. R. Wilson, M. Gewitz, P. B. Lockhart, A. F. Bolger, D. C. DeSimone, D. S. Kazi, D. J. Couper, A. Beaton, C. Kilmartin, J. M. Miro, et al. Prevention of viridans group streptococcal infective endocarditis: A scientific statement from the American Heart Association. *Circulation*, 143(20):e963–e978, 2021. doi:10.1161/CIR.0000000000000969.

[9] H. C. Nguyen, H. P. Dang, T. L. Nguyen, V. Hoang, and V. A. Nguyen. Accuracy of latest large language models in answering multiple choice questions in dentistry: A comparative study. *PLOS ONE*, 20(1):e0317423, 2025. doi:10.1371/journal.pone.0317423.

[10] X. Wu, G. Cai, B. Guo, L. Ma, S. Shao, J. Yu, Y. Zheng, L. Wang, and F. Yang. A multi-dimensional performance evaluation of large language models in dental implantology: Comparison of ChatGPT, DeepSeek, Grok, Gemini and Qwen across diverse clinical scenarios. *BMC Oral Health*, 25:1272, 2025. doi:10.1186/s12903-025-06619-6.

[11] S. Ramlogan, V. Raman, and S. Ramlogan. A pilot study of the performance of Chat GPT and other large language models on a written final year periodontology exam. *BMC Medical Education*, 25:727, 2025. doi:10.1186/s12909-025-07195-7.

[12] H. Zhu, Y. Xu, Y. Li, Z. Meng, and Z. Liu. DentalBench: Benchmarking and advancing LLMs capability for bilingual dentistry understanding. arXiv:2508.20416, 2025.

[13] J. Zhao, J. Liang, Z. Cai, J. Zhang, Z. Wen, S. Deng, W. Yi, C. Luo, H. Zhang, J. Chen, et al. GlobalDentBench: A multinational benchmark for evaluating LLM clinical reasoning in dentistry with expert calibration. arXiv:2605.24636, 2026.

[14] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *NeurIPS 2023 Datasets and Benchmarks Track*, 2023. arXiv:2306.05685.

[15] A. Panickssery, S. R. Bowman, and S. Feng. LLM evaluators recognize and favor their own generations. arXiv:2404.13076, 2024.

[16] K. Wataoka, T. Takahashi, and R. Ri. Self-preference bias in LLM-as-a-judge. arXiv:2410.21819, 2024.