

Dental-TriageBench: Benchmarking Multimodal Reasoning for Hierarchical Dental Triage

Ziyi He^{1*}, Yushi Feng^{1*}, Shuangyu Yang^{2,3}, Yinghao Zhu¹, Xichen Zhang⁵
Tai Pak Chuen Patrick^{2,3}, Lo Hei Yuet^{2,3}, Songying Wu^{2,3,4}, Weifa Yang^{2,3†}, Lequan Yu^{1†}

¹School of Computing and Data Science, The University of Hong Kong

²Faculty of Dentistry, The University of Hong Kong

³The Prince Philip Dental Hospital

⁴Li Ka Shing Faculty of Medicine, The University of Hong Kong

⁵The Hong Kong University of Science and Technology

Abstract

Dental triage is a safety-critical clinical routing task that requires integrating multimodal clinical information (e.g., patient complaints and radiographic evidence) to determine complete referral plans. We present DENTAL-TRIAGEBENCH, the first expert-annotated benchmark for reasoning-driven multimodal dental triage. Built from authentic outpatient workflows, it contains 246 de-identified cases annotated with expert-authored golden reasoning trajectories, together with hierarchical triage labels. We benchmark 21 proprietary, open-source, and medical-domain MLLMs against three junior dentists serving as the human baseline, and find a substantial human-model gap on fine-grained treatment-level triage. Further analyses show that accurate triage requires both complaint and OPG information, and that model errors concentrate on cases with multiple referral domains, where MLLMs tend to produce overly narrow referral sets and omission-heavy errors. DENTAL-TRIAGEBENCH provides a realistic testbed for developing multimodal clinical AI systems that are more clinically grounded, coverage-aware, and safer for downstream care.

1 Introduction

Dental triage is a safety-critical gatekeeping task at the entry point of care, where clinicians must decide which specialty pathway a patient should enter before definitive chairside tests are available. This decision relies on heterogeneous evidence available at intake. At intake, patients typically present with chief complaints that reflect their perceived symptoms and care needs, but these complaints may not fully localize the underlying problem or reveal concurrent conditions; orthopantomograms (OPGs), by contrast, expose structural findings across the teeth and surrounding maxillofacial region (Turosz et al., 2023; Wang et al., 2025b). Dental triage is there-

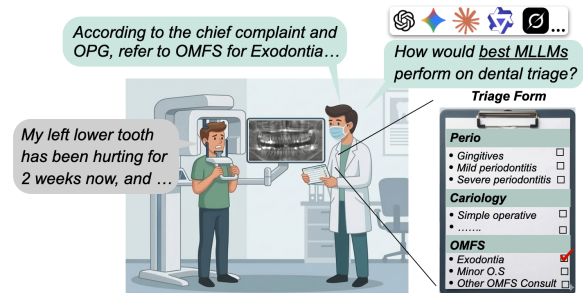


Figure 1: **Multimodal dental triage.** Given a patient’s multimodal information (e.g., chief complaint, orthopantomogram), the model predicts a multi-label specialty referral plan that reflects the patient’s treatment needs.

fore inherently multimodal, requiring clinicians to integrate textual complaints with radiographic evidence. It is not merely a perception problem, but a multimodal clinical decision-making task with direct implications for downstream care and patient safety.

Recent progress in dental AI has substantially improved panoramic radiograph understanding, with benchmarks and models increasingly targeting visual question answering, report generation, and image-grounded dialogue (Turosz et al., 2023; Hao et al., 2025; Meng et al., 2025; Huang et al., 2025; Balel et al., 2026). However, these settings primarily evaluate perception, description, or question answering, and stop short of the core challenge of real dental triage. In practice, triage requires clinicians to jointly reason over patient complaints and radiographic evidence, identify multiple concurrent conditions, and map them to a complete, hierarchical referral plan under clinical protocols. This makes the task fundamentally different from generating plausible image-grounded answers: it is multi-label, protocol-constrained, and safety-sensitive. The distinction is especially important because triage errors are asymmetric: over-referral increases workload, but under-referral leaves clin-

* Equal contribution. † Corresponding authors.

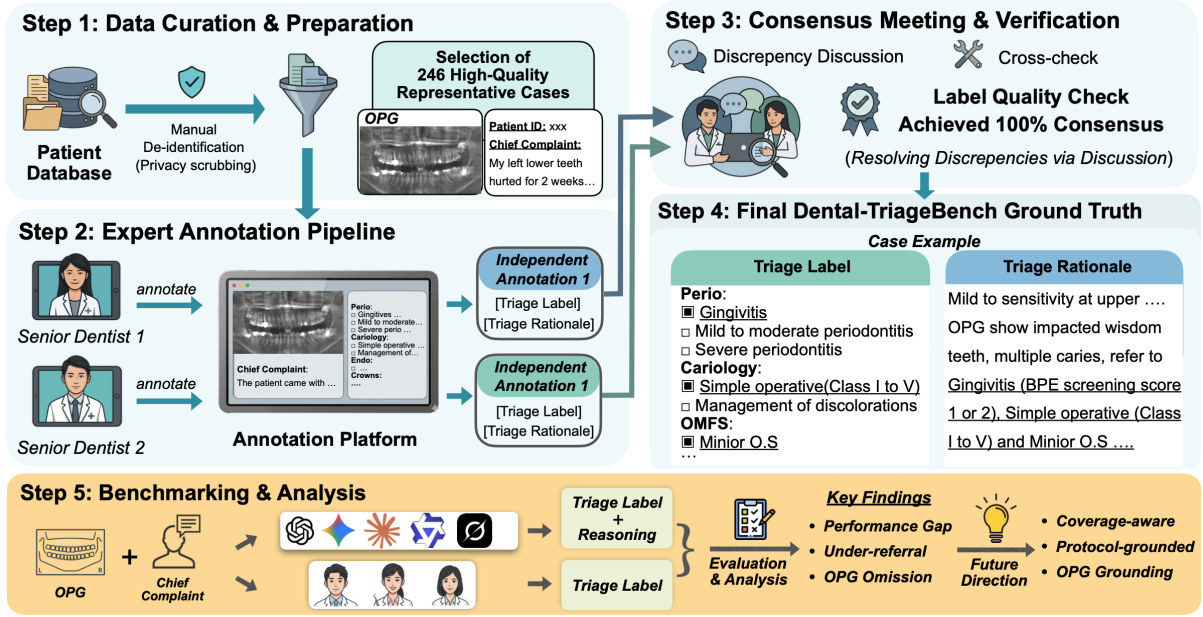


Figure 2: **Overview of DENTAL-TRIAGEBENCH.** Steps 1–4 illustrate the dataset construction pipeline, and Step 5 summarizes the main experimental settings and key findings.

ically relevant problems uncovered and can delay necessary care (Williams et al., 2024; Masannek et al., 2024; Lu et al., 2024). Existing evaluations therefore do not fully test whether multimodal models can deliver complete and safe referral coverage, rather than merely produce clinically plausible interpretations.

To address this gap, we introduce DENTAL-TRIAGEBENCH, the first expert-annotated benchmark for reasoning-driven multimodal dental triage. The benchmark is curated from authentic outpatient workflows where each case consists of an OPG and a patient chief complaint. Every case is annotated by two senior dentists with expert-authored golden reasoning trajectories, together with hierarchical multi-label triage labels spanning 8 coarse-grained domains and 22 fine-grained categories. We formulate dental triage as a hierarchical multimodal multi-label prediction task with rationale generation, and benchmark 21 proprietary, open-source, and medical-domain MLLMs under zero-shot prompting. Our results show that current models remain substantially below human performance and are far from reliable deployment in this safety-critical setting. Even the proprietary models like Claude-Opus-4.7 and GPT-5.5 trail junior dentists, particularly on fine-grained and recall-oriented metrics. We further find a pronounced gap between broad specialty-level routing and clinically precise treatment-level triage: while the best model reaches

0.302 Macro-F1 on the 22 fine-grained triage labels, its performance improves to 0.493 after aggregation to the 8 coarse-grained triage domains. This suggests that current MLLMs are better at identifying the general specialty pathway than at producing complete and treatment-specific referral plans.

Beyond aggregate scores, we use DENTAL-TRIAGEBENCH to study how and why current MLLMs fail on realistic multimodal triage. We find that errors concentrate on cases with multiple referral domains, where models tend to predict referral sets that are systematically narrower than both ground truth and human annotations. To better characterize these failures, we develop a clinician-grounded taxonomy covering radiographic misinterpretation, evidence omission, protocol misalignment, hallucination, and complaint neglect, and scale this analysis with a cross-family multi-judge majority vote validated against expert annotations. Together, these results suggest that the main bottleneck is not generating plausible dental language, but making complete, protocol-aligned, and coverage-aware triage decisions under realistic multimodal uncertainty.

In summary, our contributions are as follows:

- We introduce the first benchmark for reasoning-driven multimodal dental triage. DENTAL-TRIAGEBENCH pairs OPGs and chief complaints with expert-annotated hier-

archical triage labels and, crucially, expert-authored golden reasoning trajectories that capture clinically grounded referral logic.

- **We provide a rigorous evaluation of current MLLMs against human dentists.** We benchmark 21 proprietary, open-source, and medical-domain MLLMs on this task, and further evaluate three junior dentists on the same benchmark as a human baseline, showing that current models remain substantially below human performance, especially on fine-grained treatment-level triage.
- **We derive actionable insights for safer multimodal clinical AI.** Our analyses show that dental triage requires genuine cross-modal reasoning, and that current MLLMs systematically under-cover required referrals in complex multi-domain cases, highlighting the need for future systems that are more coverage-aware, protocol-aligned, and clinically safe.

2 Related Work

2.1 Medical MLLMs for Clinical Evaluation

Recent work on medical LLMs has moved beyond exam-style QA toward more realistic clinical evaluation (Zhang et al., 2026). Benchmarks such as ClinicBench, CliMedBench, and LLMEval-Med assess LLMs on authentic clinical scenarios and physician-validated tasks, while triage-specific studies examine emergency acuity assessment and decision support under real or benchmarked triage settings (Liu et al., 2024; Ouyang et al., 2024; Zhang et al., 2025; Williams et al., 2024; Masannek et al., 2024; Lu et al., 2024). In parallel, medical MLLMs and multimodal benchmarks have expanded image-grounded dialogue, visual question answering, report generation, and clinical visual reasoning across diverse medical modalities (Li et al., 2023; Moor et al., 2023; Zhang et al., 2024; Tu et al., 2024; Chen et al., 2024c,b; Zhou et al., 2025). These efforts are closely related to our setting, but they remain predominantly text-centric, diagnosis-centric, or organized around VQA and report generation rather than hierarchical multi-label referral planning with clinician-aligned failure analysis (Feng et al., 2026; Xiao et al., 2025; Feng et al., 2025).

2.2 Dental AI on panoramic radiographs

In dentistry, AI on panoramic radiographs has historically centered on task-specific computer vision problems such as tooth detection and numbering, restoration or state classification, and multi-disease diagnosis (Estai et al., 2022; Zhu et al., 2023). Recent reviews confirm rapid progress in dentomaxillofacial imaging, while also noting that most existing studies remain narrowly task-specific and only weakly connected to real clinical workflows (Turosz et al., 2023). More recent dental multimodal efforts begin to move toward broader foundation and instruction-following settings, including MMOral and OralGPT for panoramic-X-ray question answering, dialogue, and reporting, DentVLM for comprehensive multimodal dental diagnosis, DentVFM for generalist dental representation learning, OPGAgent for auditable OPG interpretation, OMNI-Dent for clinically guided explainable diagnosis, and recent hybrid LLM pipelines for panoramic report generation (Hao et al., 2025; Meng et al., 2025; Huang et al., 2025; Jang et al., 2026; Balel et al., 2026; Yu et al., 2026). However, existing dental benchmarks still largely formulate image interpretation as diagnosis, reporting, or question answering; they do not directly evaluate the outpatient triage problem in which complaints and OPG findings must be jointly mapped to complete, hierarchical, multi-domain referral decisions.

3 Dental-TriageBench

We introduce DENTAL-TRIAGEBENCH, a clinically grounded multimodal benchmark for dental triage. Each case consists of a patient’s chief complaint and an orthopantomogram (OPG), with expert-annotated multi-label triage decisions and supporting clinical rationale.

3.1 Data Curation

All cases were retrospectively collected from historical clinical records of dental school clinics. To protect patient privacy, all textual notes and OPG images were manually de-identified before annotation, with personal identifiers removed from both modalities. After excluding low-quality or clinically inconclusive samples, the final benchmark contains 246 patient cases.

The original clinical triage decision in each historical outpatient record served as a workflow anchor rather than as unverified ground truth. Two licensed

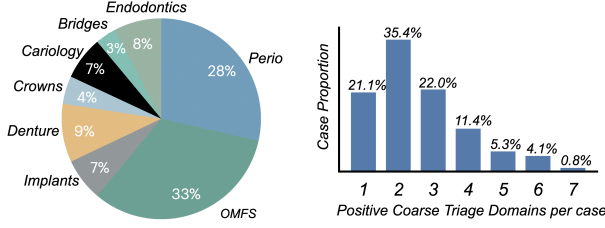


Figure 3: Data statistics of DENTAL-TRIAGEBENCH.

senior dentists independently reviewed each de-identified chief complaint–OPG pair, verified or corrected the structured labels, and authored the associated rationale. Their first-round labels were submitted independently in our annotation application before any discussion. Across fine-grained binary triage-label decisions, pre-adjudication agreement was strong (Cohen’s $\kappa = 0.83$); approximately 5% of cases required discussion to resolve label-level discrepancies. The dentists adjudicated these cases to produce the final consensus labels and rationales.

The triage taxonomy is clinically hierarchical, consisting of **8 coarse-grained triage domains** and **22 fine-grained triage labels** that cover major referral destinations and treatment types in routine dental triage. The full taxonomy is provided in Appendix E.

3.2 Benchmark Characteristics

As shown in the bar chart of Figure 3, the label distribution is strongly long-tailed. Common triage domains such as OMFS and Periodontology account for a large fraction of positive labels, whereas domains such as Bridges and Crowns appear much less frequently. We retain this naturally imbalanced distribution because it reflects authentic outpatient triage rather than a benchmark constructed for class balance. As a result, the task is more clinically realistic and more demanding: models must not only identify common referral patterns, but also preserve coverage over less frequent yet clinically relevant triage categories.

Second, DENTAL-TRIAGEBENCH preserves the complexity of real clinical triage rather than artificially balancing cases. First, the benchmark is inherently **multi-label**: At the coarse-grained domain level, each case is associated with **2.60 positive triage domains** on average, as shown in the pie chart in Figure 3, only 21.1% of patients require referral to a single domain, the remaining **78.9%**

involve multiple co-occurring conditions and multi-domain referral needs. This combination of long-tailed label prevalence and multi-label complexity makes DENTAL-TRIAGEBENCH a challenging testbed for current MLLMs.

3.3 Task Definition

We formulate dental triage as a **multimodal multi-label classification** task with rationale generation. Given a patient case $\mathcal{X} = (\mathcal{X}_{img}, \mathcal{X}_{text})$, where $\mathcal{X}_{img}$ is the OPG and $\mathcal{X}_{text}$ is the chief complaint, the model predicts $\mathcal{Y} = (\mathbf{y}, \mathcal{R})$.

Here, $\mathbf{y} \in \{0, 1\}^{|\mathcal{L}|}$ is a multi-hot label vector over the 22 fine-grained triage labels, and $\mathcal{R}$ is a natural-language rationale that explains the predicted triage decision. Because patients frequently present with multiple concurrent conditions, this task is inherently multi-label rather than single-label.

3.4 Evaluation

We evaluate structured triage prediction using *Macro-F1*, *Macro-Recall*, *Micro-F1*, and *Exact Match Ratio (EMR)*. These metrics capture complementary aspects of performance under severe class imbalance and multi-label prediction: class-balanced quality, overall label recovery, and exact case-level correctness. Additional details are provided in Appendix F.2.

4 Experiments

4.1 Experiment Setup

To evaluate the capability boundaries of current MLLMs on specialized clinical decision-making, we benchmark 21 representative multimodal models, including 9 proprietary models (Anthropic, 2025; Google, 2026a,b; OpenAI, 2025; xAI, 2025; Kimi, 2026; Alibaba, 2026; Anthropic, 2026; OpenAI, 2026b), 8 open-source general-purpose models (GLM-V Team, 2025; Wang et al., 2025a; Xiamo, 2025; Liu et al., 2026; Bai et al., 2025b,a; Li et al., 2024b,a), and 4 medical-domain models (Chen et al., 2024a; Team et al., 2025; Google, 2026c; Sellergren et al., 2025). All models are evaluated in a zero-shot setting to assess their inherent clinical alignment, and the full task instruction prompt is provided in Table 10.

For each case, models are asked to predict the exact triage decisions in the 22 *fine-grained triage label space* in a single run. Our primary evaluation is therefore conducted on this *fine-grained*

Model	Fine-grained (22 labels)				Coarse-grained (8 domains)			
	Macro-F1	Macro-Rec	Micro-F1	Exact-Match	Macro-F1	Macro-Rec	Micro-F1	Exact-Match
Proprietary MLLMs								
Claude-Sonnet-4.5-20250929	0.269	0.275	0.350	0.024	0.442	0.466	0.565	0.130
Claude-Opus-4.7	0.288	0.298	0.420	0.077	0.463	0.468	0.634	0.171
GPT-5.2	0.272	0.239	0.404	0.041	0.398	0.320	0.576	0.215
GPT-5.5	0.284	0.242	0.422	0.098	0.437	0.363	0.581	0.195
Gemini-3-Flash	0.302	0.310	0.459	0.049	<u>0.488</u>	0.492	0.656	<u>0.220</u>
Gemini-3-Pro	0.294	0.262	0.447	0.118	0.493	0.444	0.641	0.224
Grok-4	0.232	0.187	0.329	0.094	0.392	0.317	0.529	0.187
Kimi-K2.5	0.249	0.204	0.348	0.053	0.393	0.311	0.537	0.187
Qwen-VL-Max-Latest	0.222	0.206	0.283	0.037	0.398	0.319	0.517	0.183
Open-source MLLMs								
GLM-4.1V-9B-Thinking	0.072	0.050	0.086	0.016	0.133	0.089	0.182	0.069
InternVL3_5-8B	0.088	0.056	0.093	0.016	0.176	0.104	0.196	0.045
MiMo-VL-7B-RL	<u>0.125</u>	<u>0.093</u>	<u>0.175</u>	0.020	<u>0.266</u>	<u>0.184</u>	<u>0.386</u>	<u>0.146</u>
Ministral-3-14B-Instruct-2512	0.189	0.158	0.247	0.029	0.327	0.254	0.453	0.163
Qwen2.5-VL-7B-Instruct	0.101	0.077	0.148	<u>0.024</u>	0.191	0.135	0.301	0.102
Qwen3-VL-8B-Instruct	0.089	0.064	0.127	<u>0.024</u>	0.208	0.131	0.281	0.106
LLaVA-onevision-qwen2-7b	0.064	0.049	0.095	0.000	0.134	0.084	0.187	0.045
LLaVA-v1.6-mistral-7b-hf	0.014	0.008	0.019	0.004	0.032	0.017	0.058	0.008
Medical MLLMs								
HuatuoGPT-Vision-7B	0.074	0.058	0.101	0.004	0.121	0.072	0.188	0.061
Lingshu-7B	<u>0.076</u>	0.056	<u>0.115</u>	<u>0.012</u>	<u>0.172</u>	<u>0.108</u>	<u>0.277</u>	<u>0.094</u>
MedGemma-27b-it	0.121	0.094	0.173	0.029	0.266	0.194	0.357	0.130
MedGemma-1.5-4b-it	0.059	0.036	0.068	0.008	0.116	0.068	0.140	0.041
Human Baseline	0.402	0.431	0.525	0.083	0.562	0.613	0.727	0.272

Table 1: Benchmark results on DENTAL-TRIAGEBENCH across 22 fine-grained triage labels and 8 aggregated coarse-grained domains. We report Macro-F1, Macro-Recall, Micro-F1, and Exact-Match for 21 proprietary, open-source, and medical MLLMs, together with the human baseline. Within each model domain, the best results are in **bold** and the second-best results are underlined.

multi-label prediction task. To additionally assess whether models can still recover broader referral needs, we aggregate the 22 fine-grained label predictions into their corresponding 8 *coarse-grained triage domains* (e.g., mapping *Single-unit Crowns* to *Crowns*) and report performance at this higher level as well. Only after the senior-dentist consensus ground truth was finalized, we invited three junior dentists to independently complete the full benchmark under the same fine-grained label space. We report their average performance as the human baseline in the main paper and provide individual scores in Appendix G.3.

4.2 Main Results

We report the overall performance in Table 1, and we make three observations below.

Current MLLMs remain substantially below human performance on dental triage Overall, current MLLMs remain substantially below **junior dentists**, especially on the clinically more demanding fine-grained triage task. On the 22 *fine-grained triage labels*, the best-performing proprietary model, Gemini-3-Flash, achieves a Macro-F1 of 0.302 and a Micro-F1 of 0.459, while the junior dentists average reaches 0.402 and 0.525, respec-

tively. This gap persists across all four metrics, indicating that even the strongest frontier MLLMs still struggle to match human-level performance in precise treatment-level triage. The gap becomes even more notable when considering Macro-Recall, where junior dentists achieve 0.431, suggesting that human raters are better at covering the full set of required referrals in multi-label cases.

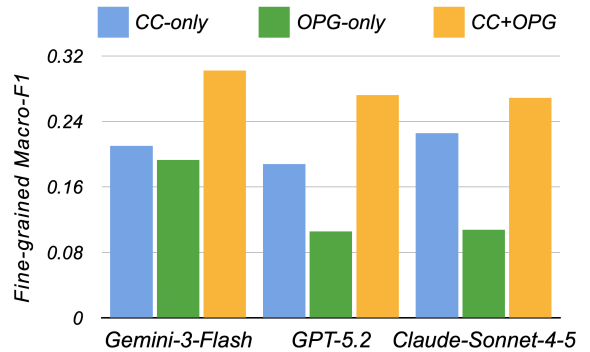


Figure 4: CC-only: only chief complaint; OPG-only: only OPG; CC+OPG: both chief complaint and OPG

Proprietary frontier models consistently outperform open-source models Gemini-3-Flash yields the strongest overall **fine-grained** performance, ranking first on Macro-F1, Macro-Recall, and Micro-F1, whereas Gemini-3-Pro achieves the

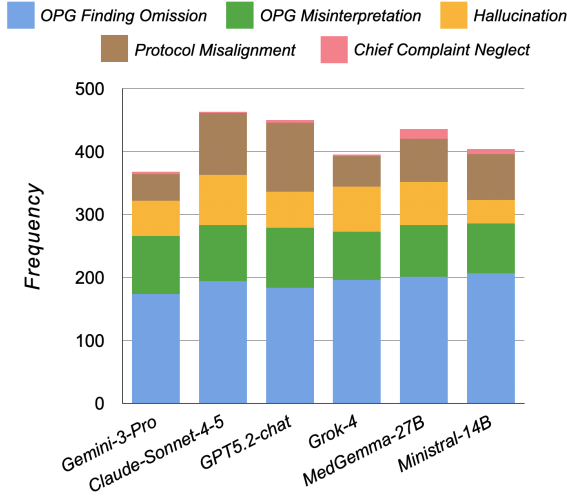


Figure 5: Distribution of five clinically grounded failure dimensions across six representative MLLMs, estimated by Gemini-3-Pro over all benchmark cases.

best Exact-Match score (0.118). In contrast, open-source and medical MLLMs lag far behind proprietary ones. The best open-source model, Ministral-3-14B-Instruct-2512, reaches only 0.189 Macro-F1 and 0.247 Micro-F1 on the fine-grained task, while the strongest medical model, MedGemma-27b-it, achieves 0.121 Macro-F1 and 0.173 Micro-F1. Notably, domain-specific medical MLLMs do not outperform strong general-purpose open models, indicating that current medical adaptation alone does not suffice for this highly specialized multimodal dental triage task.

Coarse-grained domain-level routing is more clinically deployable than fine-grained triage prediction When predictions are aggregated from the 22 *fine-grained triage labels* into 8 *coarse-grained triage domains*, all models improve substantially, while the overall ranking pattern remains largely unchanged. For example, Gemini-3-Flash improves from 0.302 to 0.488 in Macro-F1 and from 0.459 to 0.656 in Micro-F1. This suggests that current MLLMs are better suited to broad specialty-level routing than to precise treatment-level triage, making coarse-grained routing a more realistic near-term deployment target. However, even in this setting, the best model still remains clearly below human performance.

Paired case-level bootstrapping (5,000 resamples) further confirms that the human-model gaps in fine- and coarse-grained Macro-/Micro-F1 are positive for all four strongest original frontier MLLMs, with every 95% confidence interval excluding zero.

Stratifying cases by involvement of a fine-grained label with fewer than 10 positives yields the same pattern both with and without tail labels: humans attain higher case-F1 and recall, whereas the best MLLM has higher omission risk. Full estimates are reported in Appendix G.1; these analyses support case-level and hierarchical conclusions, not definitive claims for each rare fine-grained label.

5 Further Analysis

5.1 Multimodality is essential for dental triage

In this section, we examine the necessity of multimodal input for dental triage. The fundamental premise of DENTAL-TRIAGEBENCH is that referral decisions must be made from both what the patient reports and what the radiograph reveals: The chief complaints encode symptom location and care intent, whereas OPGs expose structural and often asymptomatic findings that may independently trigger referral. We therefore evaluate three top-performing MLLMs under chief complaint-only, OPG-only, and multimodal settings. Figure 4 shows that removing either modality degrades performance for all 3 models, i.e., $F1_{\text{Fusion}} > \max(F1_{\text{Text}}, F1_{\text{Image}})$, confirming that neither input source alone is sufficient for reliable triage. We further provide a representative case in Appendix I.1 to illustrate what information is lost in each unimodal setting.

Crucially, the value of multimodal input becomes even clearer at the case level. As shown in Table 2, when measured at the case level, adding the missing modality yields case-level F1 gains greater than 0.1 for 13.0% of cases for Claude Sonnet 4.5, 17.5% for Gemini 3 Flash, and 34.1% for GPT-5.2; gains greater than 0.2 are observed for 11.0%, 10.2%, and 27.6% of cases, respectively. Taken together, these results confirm that our benchmark probes genuine cross-modal evidence integration rather than unimodal shortcut learning.

Model	$\Delta F1 > 0$	$\Delta F1 > 0.1$	$\Delta F1 > 0.2$
Claude Sonnet 4.5	17.1%	13.0%	11.0%
Gemini 3 Flash	26.0%	17.5%	10.2%
GPT-5.2	35.4%	34.1%	27.6%

Table 2: Percentage of cases with positive case-level F1 gain after adding the missing modality.

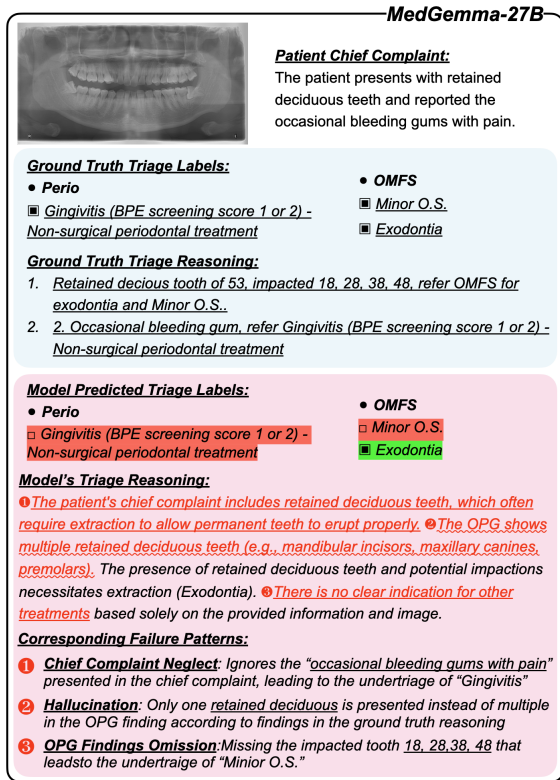


Figure 6: A failure taxonomy case study on MedGemma-27B, with correct triage label on **green**, and missing triage label on **red**.

5.2 Clinician-Grounded Failure Taxonomy

As shown in the main results, all evaluated MLLMs underperform human dentists on DENTAL-TRIAGEBENCH. However, aggregate metrics such as F1 and Exact Match do not reveal why these models fail or whether their errors follow clinically meaningful patterns. We therefore conduct a finer-grained failure analysis of model-generated rationales. Senior dentists first reviewed a subset of outputs and identified five recurring failure modes: (1) OPG Misinterpretation, (2) OPG Finding Omission, (3) Protocol Misalignment, (4) Hallucination, and (5) Chief Complaint Neglect. Detailed definitions are provided in Appendix H.1.

We scale the taxonomy-based analysis with three judge models from different families—Gemini-3-Pro, GPT-5.4, and GLM-5 (GLM-5-Team et al., 2026; OpenAI, 2026a)—and use their majority vote as the default label for each failure dimension. Each judge receives the clinician-defined criteria, model rationale, and expert rationale; the prompt is provided in Appendix J.3. Against senior-dentist labels on 100 sampled model-case outputs, the individual judges achieve 93.6%, 95.2%, and

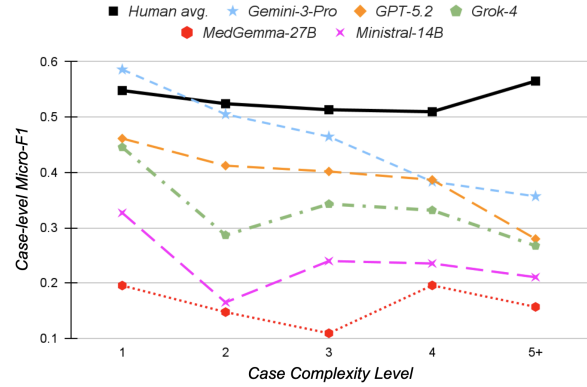


Figure 7: Comparison of case-level Micro-F1 between models and human baseline among different case complexity levels.

91.4% average accuracy, respectively, and the majority vote achieves 94.6%. Appendix H.2 reports dimension-level confusion matrices and discusses residual disagreement.

Figure 5 shows that model failures are highly non-uniform across dimensions. The most prevalent failure mode across all six models is **OPG Finding Omission**, which consistently occurs far more often than any other category, suggesting that current MLLMs more often fail to incorporate clinically relevant radiographic evidence than actively misread it. OPG Misinterpretation is also common, but substantially less frequent, indicating that omission rather than misperception is the dominant radiographic bottleneck on this task.

By contrast, Protocol Misalignment and Hallucination vary more substantially across models and thus account for much of the model-specific failure profile. For example, Claude-Sonnet-4-5 and GPT5.2 exhibit higher rates of protocol misalignment, whereas hallucination is more pronounced in Claude-Sonnet-4-5, Grok-4, and MedGemma-27B. In comparison, Chief Complaint Neglect is rare across all models, suggesting that the main challenge is not simply ignoring the complaint text, but failing to integrate radiographic findings with referral criteria into an appropriate triage decision. Representative case studies are shown in Figure 6 and Appendix I.2.

5.3 MLLMs Struggle with Complex Triage Cases

The failure taxonomy above suggests that MLLM errors are not random, but reflect a broader difficulty in turning clinical evidence into appropriate

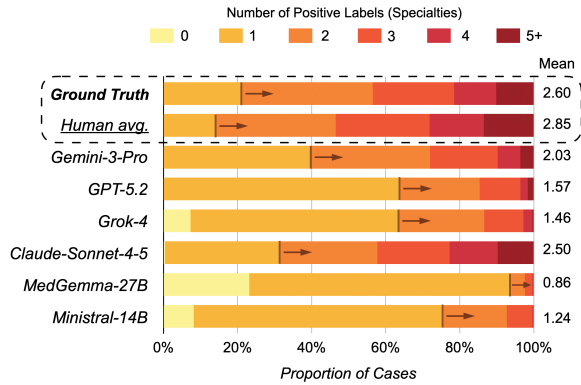


Figure 8: Distribution of the number of positive triage domains per case. The arrow marks the threshold for cases with at least two positive domains.

triage decisions. We next show that this limitation is most evident on cases requiring referral coverage across multiple triage domains, where *triage domains* denote the 8 coarse-grained specialty-level referral categories in our taxonomy and *case complexity* is defined as the number of such domains required by a case in the ground truth, i.e., how many specialty-level referral domains must be covered for that patient.

As case complexity increases, the performance gap between MLLMs and human dentists widens substantially. Figure 7 shows that human performance remains relatively stable, decreasing only from 0.548 on single-domain cases to 0.510 on cases with four positive domains, and recovering to 0.565 on the most complex cases with 5+ positive domains. In contrast, all evaluated MLLMs degrade much more sharply. For example, Gemini-3-Pro drops from 0.586 to 0.357 from single-domain to 5+ domain cases, while GPT-5.2 declines from 0.461 to 0.280. These results suggest that the main challenge in DENTAL-TRIAGEBENCH is not merely predicting individual labels, but maintaining adequate referral coverage for multi-problem cases. More broadly, current MLLMs remain substantially weaker than human dentists when triage requires simultaneous reasoning over multiple co-occurring referral needs.

5.4 Performance under Clinician-Rated Difficulty

Case complexity does not capture subtle radiographic evidence, cross-modal ambiguity, or borderline referral decisions. We therefore asked a senior dentist to rate all cases on three 1–5 ordinal dimensions: *OPG Evidence Difficulty*, *Cross-*

modal Clinical Ambiguity, and *Referral Decision Boundary Complexity*. A second senior dentist independently rated 50 random cases, yielding linear-weighted Cohen’s $\kappa = 0.67$. Across all three dimensions, Gemini-3-Pro is competitive with humans at difficulty level 1 but degrades substantially more as difficulty rises; for example, its Micro-F1 falls from 0.742 to 0.274 across OPG difficulty levels 1–5, whereas human performance changes from 0.637 to 0.513. Correlations with referral-domain count are 0.498, 0.482, and 0.778, respectively, showing that the first two dimensions capture factors beyond case complexity. Appendix G.2 gives definitions, rating distributions, reliability details, and all stratified results.

5.5 MLLMs systematically under-cover referral needs

To examine whether current MLLMs systematically under-cover referral needs, we next analyze the number of positive coarse-grained triage domains predicted per case. If MLLMs struggle to maintain adequate referral coverage, this should manifest as systematically narrower referral sets than those assigned in ground truth and by human dentists. Figure 8 shows exactly this pattern: ground truth and human annotations frequently include multiple positive triage domains for the same case, whereas most MLLMs predict substantially fewer. Nearly 80% of ground-truth and human-annotated cases involve at least two positive triage domains, and about half involve more than two, but such broad referral sets are much less common in model predictions. These results suggest that, as case complexity increases, current MLLMs tend to compress multi-domain referral needs into narrower predictions rather than preserve the breadth of coverage required for realistic triage.

A key clinical question, however, is whether these narrower referral sets reflect fewer unnecessary referrals or more missed necessary ones. In dental triage, the latter is generally more concerning: while slight over-referral may increase downstream review burden, under-triage can delay appropriate specialty evaluation and leave clinically important needs unaddressed. Figure 9 shows that human raters occupy a region of slightly higher over-referral burden but markedly lower omission risk, whereas MLLMs incur substantially higher omission without a commensurate reduction in over-referral. This suggests that human dentists adopt

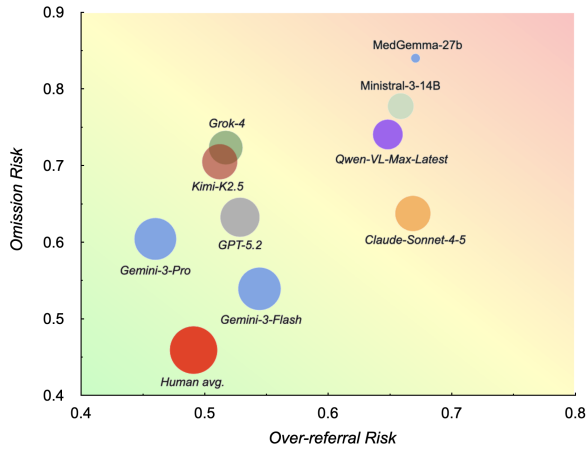


Figure 9: Trade-off between omission risk and over-referral burden. The lower-left region is preferred, indicating fewer missed necessary referrals and fewer unnecessary referrals. Bubble size denotes fine-grained Macro-F1 reported in Table 1.

a safer strategy by preserving broader referral coverage under uncertainty, whereas current MLLMs more often sacrifice necessary referrals without achieving a clear efficiency benefit. In this sense, the narrower referral sets produced by MLLMs reflect not a desirable conservative strategy, but a clinically riskier form of under-coverage.

5.6 Implications for Safer Triage Systems

Our analyses motivate four concrete baseline strategies. A *finding-first pipeline* should explicitly enumerate complaint-derived and radiographic evidence before referral prediction. A *protocol-guided checklist* should then inspect every coarse domain before mapping supported findings to fine-grained labels. Training, decoding, and evaluation should be *coverage-aware*, accounting for omission risk, referral-set coverage, and label-cardinality calibration in addition to F1. Finally, systems should use *uncertainty-aware deferral* for multi-domain cases, subtle OPG findings, or uncertain evidence-to-protocol mappings. These strategies directly target the observed omission-heavy errors and prioritize decision support with human oversight rather than autonomous triage.

6 Conclusion

We introduce DENTAL-TRIAGEBENCH, the first expert-annotated benchmark for reasoning-driven multimodal dental triage from chief complaints and OPGs, and show across 21 MLLMs that current systems remain substantially below human performance, especially on fine-grained referral planning,

with clinically concerning omission-heavy under-coverage on complex multi-domain cases. These findings point to several directions for future work: developing coverage-aware triage models that preserve referral completeness under multi-problem uncertainty; improving OPG grounding through dental-specific multimodal pretraining, structured radiographic supervision, or finding-first pipelines; aligning model decisions more tightly with real triage protocols; and advancing uncertainty-aware clinical AI through abstention, selective deferral, and safety-oriented evaluation. We hope DENTAL-TRIAGEBENCH will support the development of multimodal dental AI systems that are more clinically grounded, protocol-aligned, coverage-aware, and safe for downstream care.

7 Limitations

DENTAL-TRIAGEBENCH is modest in size and comes from one clinical setting. The consistent imaging workflow, annotation protocol, and triage convention enable a controlled paired comparison on the same cases, but results should not be interpreted as external validation across institutions, populations, devices, or regional protocols; multi-center evaluation is an important next step. Each of the eight primary coarse referral domains has at least 20 positive cases, but some fine-grained labels are rare. We therefore make case-level and hierarchical claims rather than definitive per-label claims for every tail category. The benchmark uses OPG-complaint pairs and does not capture the full evidence available in chairside diagnosis. Finally, although the rationale analysis uses a cross-family majority vote validated against senior dentists, automatic failure labels remain imperfect instance-level approximations, particularly for protocol misalignment and subtle hallucination.

8 Ethical Considerations

Dental triage is a safety-critical task, and this work should not be interpreted as evidence for autonomous clinical deployment. Instead, DENTAL-TRIAGEBENCH is intended to expose the current limitations of MLLMs under realistic multimodal triage conditions. As our results show, even the strongest models remain below human performance and exhibit omission-heavy errors that may be clinically unsafe, particularly in complex cross-domain cases. All benchmark cases were retrospectively collected from authentic clin-

ical workflows and manually de-identified prior to annotation, with personal identifiers removed from both text and OPG images; all labels were reviewed by licensed senior dentists. Because the benchmark is derived from real clinical data, its use should remain subject to appropriate privacy, institutional, and data-governance requirements. More broadly, benchmark performance should not be conflated with clinical competence, since real-world triage depends on additional examination findings and contextual information beyond chief complaints and OPGs. We therefore position DENTAL-TRIAGEBENCH as a research benchmark for developing safer and more clinically grounded decision-support systems, rather than as a substitute for qualified dental professionals.

Acknowledgments

This work was supported in part by the Research Grants Council of Hong Kong through grants 27206123, 17200125, C5055-24G, and T45-401/22-N; the Hong Kong Innovation and Technology Fund through grant ITS/544/24FP; and the University of Hong Kong Seed Fund for Collaborative Research.

References

- Alibaba. 2026. [Openai-compatible vision api for qwen-vl models](#). Official developer documentation for qwen-vl-max / qwen-vl-max-latest.
- Anthropic. 2025. [Introducing claude sonnet 4.5](#). Official announcement.
- Anthropic. 2026. [Introducing claude opus 4.7](#). Official announcement.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. [Qwen3-vl technical report](#). *Preprint*, arXiv:2511.21631.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. [Qwen2.5-vl technical report](#). *Preprint*, arXiv:2502.13923.
- Yunus Balel, Kaan Sağtaş, Fatih Teke, and Mehmet Ali Kurt. 2026. A novel hybrid large language model approach for reporting panoramic radiographs and performance comparison with current large language models. *Journal of imaging informatics in medicine*.
- Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. 2024a. [Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale](#). *Preprint*, arXiv:2406.19280.
- Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyang Huang, Yanzhou Su, Benyou Wang, Shaoting Zhang, Bin Fu, Jianfei Cai, Bohan Zhuang, Eric J. Seibel, Junjun He, and Yu Qiao. 2024b. [Gmailmmbench: A comprehensive multimodal evaluation benchmark towards general medical AI](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Andrew Johnston, Alaa Youssef, Louis Blankemeier, Christian Bluethgen, Stephan Altmayer, and 1 others. 2024c. A vision-language foundation model to enhance efficiency of chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*.
- Mohamed Estai, Marc Tennant, Dieter Gebauer, Andrew Brostek, Janardhan Vignarajan, Maryam Mehdizadeh, and Sajib Saha. 2022. Deep learning for automated detection and numbering of permanent teeth on panoramic images. *Dentomaxillofacial Radiology*, 51(2):20210296.
- Yushi Feng, Tsai Hor Chan, Guosheng Yin, and Lequan Yu. 2025. [Democratizing large language model-based graph data augmentation via latent knowledge graphs](#). *Neural Networks*, 191:107777.
- Yushi Feng, Junye Du, Yingying Hong, Qifan Wang, and Lequan Yu. 2026. Pass: Probabilistic agentic super-net sampling for interpretable and adaptive chest x-ray reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI-26.
- GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, and 168 others. 2026. [Glm-5: from vibe coding to agentic engineering](#). *Preprint*, arXiv:2602.15763.
- GLM-V Team. 2025. Glm-4.1v-thinking and glm-4.5v: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*.
- Google. 2025. [Gemini 2 flash | gemini api | google ai for developers](#). Official developer documentation.
- Google. 2026a. [Gemini 3 flash preview | gemini api | google ai for developers](#). Official developer documentation.
- Google. 2026b. [Gemini 3 pro preview | gemini api | google ai for developers](#). Official developer documentation.

- Google. 2026c. [google/medgemma-1.5-4b-it](#). Official Hugging Face model card.
- Jing Hao, Yuxuan Fan, Yanpeng Sun, Kaixin Guo, Lin Lizhuo, Jinrong Yang, Qiyong Hemis Ai, Lun M Wong, Hao Tang, and Kuo Feng Hung. 2025. [Towards better dental AI: A multimodal benchmark and instruction dataset for panoramic x-ray analysis](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Xinrui Huang, Fan Xiao, Dongming He, Anqi Gao, Dandan Li, Xiaofan Zhang, Shaoting Zhang, and Xudong Wang. 2025. Towards generalist intelligence in dentistry: Vision foundation models for oral and maxillofacial radiology. *arXiv preprint arXiv:2510.14532*.
- Leeje Jang, Yao-Yi Chiang, Angela M Hastings, Patimaporn Pungchanchaikul, Martha B Lucas, Emily C Schultz, Jeffrey P Louie, Mohamed Estai, Wen-Chen Wang, Ryan HL Ip, and 1 others. 2026. Omni-dent: Towards an accessible and explainable ai framework for automated dental diagnosis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 415–424.
- Kimi. 2026. Kimi k2.5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*.
- Bo Li, Hao Zhang, Kaichen Zhang, Dong Guo, Yuanhan Zhang, Renrui Zhang, Feng Li, Ziwei Liu, and Chunyuan Li. 2024a. [Llava-next: What else influences visual instruction tuning beyond data?](#)
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024b. [Llava-onevision: Easy visual task transfer](#). *arXiv preprint arXiv:2408.03326*.
- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. [Llava-med: Training a large language-and-vision assistant for biomedicine in one day](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, and 101 others. 2026. [Ministral 3](#). *Preprint*, arXiv:2601.08584.
- Fenglin Liu, Zheng Li, Hongjian Zhou, Qingyu Yin, Jingfeng Yang, Xianfeng Tang, Chen Luo, Ming Zeng, Haoming Jiang, Yifan Gao, Priyanka Nigam, Sreyashi Nag, Bing Yin, Yining Hua, Xuan Zhou, Omid Rohanian, Anshul Thakur, Lei A. Clifton, and David A. Clifton. 2024. [Large language models are poor clinical decision-makers: A comprehensive benchmark](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 13696–13710. Association for Computational Linguistics.
- Meng Lu, Brandon Ho, Dennis Ren, and Xuan Wang. 2024. [Triageagent: Towards better multi-agents collaborations for large language model-based clinical triage](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, volume EMNLP 2024 of *Findings of ACL*, pages 5747–5764. Association for Computational Linguistics.
- Lars Masannek, Linea Schmidt, Antonia Seifert, Tristan Kölsche, Niklas Huntemann, Robin Jansen, Mohammed Mehsin, Michael Bernhard, Sven G Meuth, Lennert Böhm, and 1 others. 2024. Triage performance across large language models, chatgpt, and untrained doctors in emergency medicine: comparative study. *Journal of medical Internet research*, 26:e53297.
- Zijie Meng, Jin Hao, Xiwei Dai, Yang Feng, Jiaxiang Liu, Bin Feng, Huikai Wu, Xiaotang Gai, Hengchuan Zhu, Tianxiang Hu, and 1 others. 2025. [Dentvllm: A multimodal vision-language model for comprehensive dental diagnosis and enhanced clinical practice](#). *arXiv preprint arXiv:2509.23344*.
- Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. [Medflamingo: a multimodal medical few-shot learner](#). In *Machine Learning for Health, ML4H@NeurIPS 2023, 10 December 2023, New Orleans, Louisiana, USA*, volume 225 of *Proceedings of Machine Learning Research*, pages 353–367. PMLR.
- OpenAI. 2025. [Introducing gpt-5.2](#). Official announcement.
- OpenAI. 2026a. [Introducing gpt-5.4](#). Official announcement.
- OpenAI. 2026b. [Introducing gpt-5.5](#). Official announcement.
- Zetian Ouyang, Yishuai Qiu, Linlin Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2024. [Climedbench: A large-scale chinese benchmark for evaluating medical large language models in clinical scenarios](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 8428–8438. Association for Computational Linguistics.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2025. [Medgemma technical report](#). *Preprint*, arXiv:2507.05201.
- LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui

- Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. 2025. [Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning](#). *Preprint*, arXiv:2506.07044.
- Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, and 1 others. 2024. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138.
- Natalia Turosz, Kamila Chęcińska, Maciej Chęciński, Anita Brzozowska, Zuzanna Nowak, and Maciej Sikora. 2023. Applications of artificial intelligence in the analysis of dental panoramic radiographs: an overview of systematic reviews. *Dentomaxillofacial Radiology*, 52(7):20230284.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025a. [InternV3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency](#). *Preprint*, arXiv:2508.18265.
- Yin-Chih Chelsea Wang, Tsao-Lun Chen, Shankeeth Vinayahalingam, Tai-Hsien Wu, Chu Wei Chang, Hsuan Hao Chang, Hung-Jen Wei, Mu-Hsiung Chen, Ching-Chang Ko, David Anssari Moin, and 1 others. 2025b. Artificial intelligence to assess dental findings from panoramic radiographs—a multinational study. *arXiv preprint arXiv:2502.10277*.
- Christopher YK Williams, Travis Zack, Brenda Y Miao, Madhumita Sushil, Michelle Wang, Aaron E Kornblith, and Atul J Butte. 2024. Use of a large language model to assess clinical acuity of adults in the emergency department. *JAMA network open*, 7(5):e248895.
- xAI. 2025. [Grok 4](#). Official announcement.
- Ziyan Xiao, Ruiyang Zhang, Yushi Feng, Lingting Zhu, Liang Peng, and Lequan Yu. 2025. [A dynamic agent framework for large language model reasoning for medical and visual question answering](#). In *IEEE/CVF International Conference on Computer Vision, ICCV 2025 - Workshops, Honolulu, HI, USA, October 19-20, 2025*, pages 1154–1163. IEEE.
- Xiaomi. 2025. MIMO-VL technical report. *arXiv preprint arXiv:2506.03569*.
- Zhaolin Yu, Litao Yang, Ben Babicka, Ming Hu, Jing Hao, Anthony Huang, James Huang, Yueming Jin, Jiasong Wu, and Zongyuan Ge. 2026. [Opgagent: An agent for auditable dental panoramic x-ray interpretation](#). *Preprint*, arXiv:2603.00462.
- Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, and 1 others. 2024. A generalist vision–language foundation model for diverse biomedical tasks. *Nature medicine*, 30(11):3129–3141.
- Ming Zhang, Yujiong Shen, Zelin Li, Huayu Sha, Binze Hu, Yuhui Wang, Chenhao Huang, Shichun Liu, Jingqi Tong, Changhao Jiang, Mingxu Chai, Zhiheng Xi, Shihan Dou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025. [Llmeval-med: A real-world clinical benchmark for medical llms with physician validation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 4888–4914. Association for Computational Linguistics.
- Ruiyang Zhang, Yushi Feng, Liang Peng, and Lequan Yu. 2026. [ProActIS: Boosting LLMs’ proactive information seeking in dynamic clinical workflows with a lightweight exam proposer](#). In *Findings of the Association for Computational Linguistics: EMNLP 2026*. Association for Computational Linguistics.
- Tianhong Zhou, Yin Xu, Yingtao Zhu, Chuxi Xiao, Haiyang Bian, Lei Wei, and Xuegong Zhang. 2025. [DrVD-bench: Do vision-language models reason like human doctors in medical image diagnosis?](#) In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Junhua Zhu, Zhi Chen, Jing Zhao, Yueyuan Yu, Xiaojuan Li, Kangjian Shi, Fan Zhang, Feifei Yu, Keying Shi, Zhe Sun, and 1 others. 2023. Artificial intelligence in the diagnosis of dental diseases on panoramic radiographs: a preliminary study. *BMC Oral Health*, 23(1):358.

A AI Usage Declaration

We declare that the textual content of this paper was originally drafted by the authors. We utilized Gemini-3-Pro solely for proofreading and grammatical error correction. In addition, some illustrative figures were generated with assistance from Nano-Banana. All authors have carefully reviewed the polished text and figures, and take full responsibility for the final content.

B Data Availability

Because DENTAL-TRIAGEBENCH is built from real clinical cases, the dataset cannot be released for unrestricted public access. To support responsible research use, we plan to release the benchmark under a controlled access mechanism: researchers will be able to obtain the data through an application process subject to institutional approval and data use conditions.

C Ethics Review and Data Governance

Medical records were retrospectively retrieved from a university-affiliated dental teaching hospital, with all patient data anonymized prior to annotation and analysis. The study protocol was approved by the relevant institutional ethics review board.

D Human Rater Recruitment and Compensation

The three junior dentists who participated in our human baseline evaluation were recruited through voluntary participation. All raters were licensed junior dentists with relevant clinical training, and none were recruited through crowdsourcing platforms. Each rater was compensated for their time and professional effort at a rate aligned with reasonable local professional wages, which we considered appropriate for the participants' expertise and local compensation standards.

E Triage Label Taxonomy

Our benchmark adopts a **hierarchical triage taxonomy** adapted from authentic hospital outpatient triage forms and further refined by senior dental experts. The complete taxonomy is illustrated in Figure 10. The taxonomy is defined at two levels:

- **Coarse-grained triage domains (8 categories): domain-level routing.** The 8 coarse-grained triage domains (e.g., Periodontology, Endodontics, and OMFS) correspond to the

major clinical departments involved in dental triage. This level answers the high-level question: *Which specialty should evaluate or manage this patient?* In our further analyses, these coarse-grained triage domains are used to study referral breadth, case complexity, and omission risk.

- **Fine-grained triage labels (22 categories): treatment-oriented complexity stratification.** Within each coarse-grained triage domain, the 22 fine-grained triage labels capture clinically meaningful distinctions in referral type and treatment complexity. In academic dental hospitals, such distinctions are essential for assigning patients to the appropriate level of provider (e.g., undergraduate students versus postgraduate specialists) and for allocating suitable treatment time and resources. The primary benchmark prediction task is therefore defined at this fine-grained level.

For instance, within each coarse-grained triage domain of *Endodontics*, our taxonomy distinguishes between the fine-grained triage labels *Endodontics (Anterior/Premolars)* and *Molar Endodontics*. From a purely symptomatic perspective, both may require root canal treatment. However, from a triage and resource allocation perspective, an anterior tooth often involves a simpler canal system that can be handled by a junior practitioner within a standard appointment slot, whereas a molar typically presents a more complex canal anatomy, requiring a more experienced endodontist, specialized equipment such as an operating microscope, and substantially more chair time.

Distinguishing Triage from Diagnosis It is important to clarify that the 22 fine-grained triage labels are **not definitive pathological diagnoses**. In dentistry, definitive diagnosis (e.g., irreversible pulpitis) often requires additional clinical examinations, such as cold testing or electric pulp testing, which cannot be performed from an OPG alone. Instead, these labels represent **treatment-oriented triage categories** that support downstream referral and resource planning. This formulation keeps the benchmark aligned with the real clinical scope of triage: making the safest and most efficient referral decision from limited multimodal evidence (OPG + chief complaint), rather than forcing the model to infer a definitive diagnosis beyond the available information.

F Output Parsing and Evaluation Metrics

F.1 Output Parsing and Structured Output Extraction

To ensure consistent evaluation across heterogeneous MLLMs, we standardized all model outputs into a unified JSON schema before computing metrics. Each evaluated model was instructed to directly generate its prediction in a predefined JSON format containing the structured triage decision fields required for evaluation. The target schema specifies the full set of triage keys and enables deterministic extraction of fine-grained labels and, when available, the accompanying rationale.

Because some models occasionally produced malformed JSON or omitted required keys, we adopted a two-stage parsing pipeline. First, each model was given up to three attempts to produce a valid JSON output under the same task instruction, with retries triggered only when the output could not be parsed as valid JSON or failed schema validation due to missing required fields. If the model still failed after three attempts, we applied an external reformatting step using Gemini-2.0-Flash (Google, 2025). Importantly, this step was used only to convert the model’s original response into the target schema and did not perform a new clinical prediction.

The reformatting model received the raw model output and was instructed to preserve the original semantic content while reorganizing it into the predefined JSON structure. It was not provided with ground-truth labels and was not allowed to infer additional findings, add unsupported referrals, or revise the clinical meaning of the original answer. In other words, this fallback mechanism was strictly schema-preserving rather than prediction-correcting. The prompt used for this reformatting step is provided in Table 11.

After structured parsing, the predicted fine-grained triage labels were deterministically mapped into binary label vectors for evaluation. Coarse-grained domain predictions were then obtained by aggregating the corresponding fine-grained labels according to the benchmark taxonomy.

F.2 Evaluation Metrics

We evaluate structured triage prediction under both the **fine-grained 22-label space** and the **coarse-grained 8-domain space**. In both settings, each case is treated as a multi-label prediction problem.

Let $\mathcal{L}$ denote the label set, where $|\mathcal{L}| = 22$ for the fine-grained task and $|\mathcal{L}| = 8$ for the coarse-grained task. For each case $i \in \{1, \dots, N\}$, the ground-truth and predicted label vectors are denoted by $\mathbf{y}_i, \hat{\mathbf{y}}_i \in \{0, 1\}^{|\mathcal{L}|}$.

We report four complementary metrics: **Macro-F1**, **Micro-F1**, **Macro-Recall**, and **Exact Match Ratio (EMR)**. These metrics together capture class-balanced discrimination, overall label recovery, sensitivity to omission, and strict case-level correctness under imbalanced multi-label prediction.

Macro-F1 For each label $\ell \in \mathcal{L}$, we compute its precision P_ℓ , recall R_ℓ , and F1 score:

$$P_\ell = \frac{TP_\ell}{TP_\ell + FP_\ell}, \quad R_\ell = \frac{TP_\ell}{TP_\ell + FN_\ell},$$
$$F1_\ell = \frac{2P_\ell R_\ell}{P_\ell + R_\ell}.$$

Macro-F1 is then defined as the unweighted average across labels:

$$\text{Macro-F1} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} F1_\ell.$$

Macro-F1 measures how well a model performs *across all referral categories equally*, regardless of label frequency. In our setting, this is important because DENTAL-TRIAGEBENCH exhibits a pronounced long-tail distribution, and clinically less frequent referral categories (e.g., Bridges or Implants) remain important despite having fewer samples.

Micro-F1 Micro-F1 aggregates true positives, false positives, and false negatives over all labels before computing the F1 score:

$$P_{\text{micro}} = \frac{\sum_\ell TP_\ell}{\sum_\ell (TP_\ell + FP_\ell)},$$
$$R_{\text{micro}} = \frac{\sum_\ell TP_\ell}{\sum_\ell (TP_\ell + FN_\ell)},$$
$$\text{Micro-F1} = \frac{2P_{\text{micro}}R_{\text{micro}}}{P_{\text{micro}} + R_{\text{micro}}}.$$

Micro-F1 reflects *overall label-level prediction quality* across the benchmark, giving more weight to frequent categories. In the dental triage setting, it captures how reliably a model recovers the dominant referral needs over all patient-label decisions, and is therefore useful as a population-level summary of structured triage performance.

Macro-Recall Macro-Recall is defined as the un-weighted average of per-label recall:

$$\text{Macro-Recall} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} R_{\ell}.$$

We report Macro-Recall separately because recall is clinically especially important in triage. Missing a required referral leaves clinically relevant needs uncovered and may delay downstream care, whereas modest over-referral is often less harmful. Macro-Recall therefore directly reflects a model’s ability to *cover the full range of necessary domain referrals*, including less frequent categories.

Exact Match Ratio (EMR) Exact Match Ratio measures the proportion of cases for which the predicted label set exactly matches the ground-truth label set:

$$\text{EMR} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\hat{\mathbf{y}}_i = \mathbf{y}_i),$$

where $\mathbb{I}(\cdot)$ is the indicator function. EMR is a strict case-level metric: a prediction is counted as correct only if *all* required labels are recovered and no extra labels are added. In our task, this metric is clinically meaningful because dental triage is inherently multi-label and incomplete referral sets can be unsafe even when the main domain is correctly identified. EMR therefore captures whether a model can produce a *fully correct and complete triage decision* for an individual patient.

Taken together, these four metrics provide complementary views of performance. Macro-F1 and Macro-Recall emphasize balanced performance across the full triage taxonomy, including rare but clinically meaningful categories; Micro-F1 summarizes overall label-level accuracy across the dataset; and EMR evaluates whether the model can recover the complete referral set for each case without omission or spurious additions. Reporting all four is therefore important for assessing both *predictive quality* and *clinical safety* in hierarchical multi-label dental triage.

G Additional Results

G.1 Bootstrap and Tail-Label Robustness

We assess stability with 5,000 paired case-level bootstrap resamples over all 246 cases. For every reported metric and model, the 95% confidence interval for the human-minus-model gap remains

above zero (Table 4), showing that the human advantage is not an artifact of a small set of sampled cases.

We additionally define a tail fine-grained label as one with fewer than 10 positive occurrences and stratify cases by whether at least one such label is present. Table 5 shows that the qualitative pattern holds in both strata: junior dentists have higher case-F1 and case-recall, while the best-performing MLLM has greater omission risk. Thus, rare labels remain clinically relevant parts of the action space, but the main case-level conclusion is not driven solely by them.

G.2 Clinician-Rated Difficulty Analysis

A senior dentist rated all 246 cases from 1 (easiest) to 5 (hardest) along three dimensions: **OPG Evidence Difficulty**, the difficulty of detecting subtle, overlapping, or ambiguous triage-relevant radiographic findings; **Cross-modal Clinical Ambiguity**, the difficulty of reconciling complaint and OPG evidence when symptoms are non-localized, inconsistent, or confounded; and **Referral Decision Boundary Complexity**, the difficulty of assigning labels under borderline thresholds, protocol rules, or multiple plausible pathways. A second senior dentist independently rated 50 random cases. Linear-weighted Cohen’s κ across the ratings was 0.67. The first and second dentists’ mean scores were 2.80/2.93 for OPG evidence, 2.66/2.58 for cross-modal ambiguity, and 2.68/2.73 for referral boundaries.

The Pearson correlations between each difficulty score and case complexity (the number of positive coarse domains) are 0.498 for OPG Evidence Difficulty, 0.482 for Cross-modal Clinical Ambiguity, and 0.778 for Referral Decision Boundary Complexity. The moderate first two correlations indicate that the explicit ratings capture clinical factors not reducible to referral-set cardinality; the stronger third correlation is expected because multi-domain cases often involve more treatment thresholds and pathways.

G.3 Individual Human Baseline Results

To ensure a fair comparison between human raters and MLLMs, we provided the three junior dentists with the same task instructions used for model evaluation. The full instruction prompt is shown in Table 10.

Models	Perio	Cariology	Endo	Crowns	Bridges	Denture	Implants	OMFS
Proprietary MLLMs								
Claude-Sonnet-4.5-20250929	0.598	0.340	<u>0.446</u>	0.368	0.130	0.571	<u>0.259</u>	0.821
Gemini-3-Flash	<u>0.652</u>	<u>0.443</u>	0.337	0.375	<u>0.222</u>	0.723	0.237	0.913
Gemini-3-Pro	0.735	0.368	0.444	0.413	0.235	<u>0.610</u>	0.333	0.807
GPT-5.2	0.494	0.378	0.338	0.255	0.069	0.571	0.226	<u>0.850</u>
Grok-4	0.561	0.452	0.354	0.250	0.125	0.577	0.120	0.695
Kimi-K2.5	0.435	0.436	0.396	0.245	0.148	0.482	0.207	0.797
Qwen-VL-Max-Latest	0.449	0.410	0.453	0.311	0.154	0.469	0.203	0.734
Open-source MLLMs								
GLM-4.1V-9B-Thinking	0.133	0.039	<u>0.349</u>	0.053	0.000	0.154	0.080	0.259
InternVL3_5-8B	0.258	0.218	0.197	0.065	0.080	0.310	0.136	0.143
MiMo-VL-7B-RL	0.288	<u>0.338</u>	0.297	0.114	0.074	<u>0.329</u>	0.078	<u>0.612</u>
Ministral-3-14B-Instruct-2512	0.376	0.319	0.385	0.195	0.067	0.420	<u>0.179</u>	0.679
Qwen2.5-VL-7B-Instruct	0.274	0.039	0.320	0.098	<u>0.077</u>	0.154	0.078	0.489
Qwen3-VL-8B-Instruct	0.161	0.358	0.143	0.000	0.080	0.286	0.185	0.448
LLaVA-onevision-qwen2-7b	<u>0.306</u>	0.215	0.111	<u>0.140</u>	0.000	0.000	0.120	0.181
LLaVA-v1.6-mistral-7b-hf	0.169	0.044	0.000	0.000	0.000	0.000	0.044	0.000
Medical MLLMs								
HuatuoGPT-Vision-7B	0.317	<u>0.073</u>	0.070	0.056	0.000	<u>0.222</u>	0.039	0.191
Lingshu-7B	0.133	<u>0.043</u>	0.136	<u>0.146</u>	0.080	0.235	0.082	<u>0.523</u>
MedGemma-1.5-4b-it	0.104	0.070	<u>0.179</u>	0.167	0.000	0.065	0.143	0.205
MedGemma-27b-it	<u>0.265</u>	0.361	0.421	0.143	<u>0.071</u>	0.209	<u>0.122</u>	0.537

Table 3: Class-wise F1 scores on the 8 coarse-grained domains. Within each model domain, the best results are in **bold** and the second-best results are underlined.

Comparison	Fine Macro-F1 gap	Fine Micro-F1 gap	Coarse Macro-F1 gap	Coarse Micro-F1 gap
Human – Claude-Sonnet-4.5	0.133 [0.091, 0.173]	0.175 [0.138, 0.211]	0.120 [0.081, 0.160]	0.162 [0.130, 0.194]
Human – Gemini-3-Flash	0.100 [0.059, 0.140]	0.066 [0.035, 0.096]	0.074 [0.028, 0.120]	0.071 [0.040, 0.102]
Human – Gemini-3-Pro	0.108 [0.066, 0.149]	0.079 [0.042, 0.115]	0.068 [0.018, 0.120]	0.086 [0.052, 0.120]
Human – GPT-5.2	0.130 [0.089, 0.172]	0.121 [0.086, 0.157]	0.164 [0.117, 0.209]	0.151 [0.114, 0.187]

Table 4: Human–model performance gaps with paired case-level 95% bootstrap confidence intervals from 5,000 resamples.

Case subset	N	Avg. GT labels	Human Case-F1	Best MLLM Case-F1	Human Case-Recall	Best MLLM Case-Recall	Omission risk (Human / MLLM) ↓
All cases	246	2.85	0.523	0.484	0.606	0.521	0.444 / 0.536
With ≥1 tail fine-grained label	24	3.67	0.563	0.476	0.620	0.506	0.432 / 0.580
Without tail fine-grained labels	222	2.77	0.519	0.485	0.604	0.523	0.445 / 0.529

Table 5: Case-level performance stratified by involvement of low-frequency fine-grained labels. “Best MLLM” selects the best model for each reported case-level metric.

Difficulty dimension	Model	Level 1	Level 2	Level 3	Level 4	Level 5
OPG evidence (9.3/30.5/31.3/25.6/3.3%)	Human avg.	0.637	0.514	0.522	0.518	0.513
	Gemini-3-Pro	0.742	0.536	0.453	0.393	0.274
	GPT-5.2	0.618	0.391	0.370	0.387	0.367
	Grok-4	0.442	0.315	0.366	0.293	0.269
	MedGemma-27B	0.170	0.098	0.187	0.185	0.125
Cross-modal ambiguity (16.3/34.6/23.2/17.5/8.5%)	Human avg.	0.628	0.522	0.526	0.485	0.471
	Gemini-3-Pro	0.663	0.520	0.455	0.403	0.248
	GPT-5.2	0.516	0.398	0.371	0.393	0.327
	Grok-4	0.531	0.334	0.261	0.289	0.272
	MedGemma-27B	0.217	0.179	0.133	0.145	0.030
Referral boundary (13.8/32.1/29.3/20.7/4.1%)	Human avg.	0.606	0.528	0.528	0.478	0.559
	Gemini-3-Pro	0.654	0.521	0.430	0.443	0.224
	GPT-5.2	0.479	0.396	0.393	0.389	0.360
	Grok-4	0.574	0.244	0.340	0.322	0.288
	MedGemma-27B	0.201	0.131	0.159	0.168	0.110

Table 6: Micro-F1 stratified by clinician-rated difficulty. Parenthesized values give the percentage of cases at levels 1–5 for each dimension.

To complement the averaged human baseline reported in the main paper, we provide the individual performance of the three junior dentists in Table 7.

Although minor variation exists across raters, their overall performance is relatively consistent at both the fine-grained and coarse-grained levels. In par-

ticular, all three dentists substantially outperform current MLLMs on the fine-grained triage task, while showing similar strength on coarse-grained specialty routing. These results suggest that the human baseline is not driven by a single unusually strong or weak annotator. We therefore report the average across the three junior dentists in the main paper as a more stable and representative estimate of human performance.

G.4 Class-wise performance across 8 coarse-grained domains

Table 3 reports class-wise F1 scores on the 8 coarse-grained domains. A consistent pattern across model families is that performance is much stronger on frequent and visually salient categories such as *OMFS* and *Perio*, but substantially weaker on lower-support categories such as *Bridges*, *Implants*, and *Crowns*. This class-wise disparity reflects the pronounced label imbalance of DENTAL-TRIAGEBENCH and provides additional context for the gap between *Micro-F1* and *Macro-F1* in Table 1: micro-averaging is driven more by high-frequency categories, while macro-averaging weights rare classes equally and is therefore more sensitive to failures on the long tail.

H Failure Taxonomy and LLM-as-Judge Details

H.1 Definitions of the Five Failure Dimensions

To move beyond aggregate set-prediction metrics, we characterize model errors using five clinician-grounded failure dimensions. These dimensions are designed to identify the *root cause* of a model’s failure by comparing the model-generated rationale against the expert-annotated ground-truth rationale and the original patient complaint. Importantly, the taxonomy is applied primarily at the level of **reasoning**, rather than inferred solely from final label discrepancies. A model may therefore exhibit a failure dimension even when its final predicted triage labels are partially or fully correct, if its rationale reveals flawed clinical justification.

The five dimensions are treated as **strictly mutually exclusive at the level of each specific clinical finding**. In other words, for any individual symptom, lesion, or anatomical issue, we assign only one underlying failure type according to a strict hierarchy. The attribution proceeds in the following order: (1) **Hallucination**, (2) **Chief Complaint Neglect**, (3) **OPG Finding Omission**, (4) **OPG Mis-**

interpretation, and (5) **Protocol Misalignment**. This ordering ensures that downstream errors are attributed to the earliest underlying failure. For example, if the model misjudges the severity of a lesion on OPG, any subsequent missing referral caused by that misjudgment is attributed to **OPG Misinterpretation** rather than **Protocol Misalignment**. We define the five failure dimensions as follows.

D1: Chief Complaint Neglect. This dimension captures failures in which the model’s rationale completely ignores the core symptom, concern, or treatment request expressed in the patient’s chief complaint. It is triggered only when the text complaint contains clinically relevant information that should contribute to triage, but the model fails to acknowledge it at all in its reasoning. This category does not apply when the complaint is mentioned but interpreted imperfectly; rather, it is reserved for complete neglect of the salient textual signal.

D2: OPG Finding Omission. This dimension captures failures in which the model completely fails to mention an important radiographic finding that is explicitly identified in the expert ground-truth rationale. Typical examples include missing visible caries, bone loss, impacted teeth, retained roots, or other clinically relevant OPG abnormalities. Crucially, D2 is assigned only when the model omits the finding entirely. If the model does mention the correct finding or anatomical region but misjudges its severity or pathological nature, the error is instead categorized as **D3: OPG Misinterpretation**.

D3: OPG Misinterpretation. This dimension captures failures in which the model identifies the relevant lesion, anatomical region, or radiographic abnormality, but misinterprets its clinical severity, extent, or pathological depth relative to the expert rationale. Examples include describing deep caries as superficial, severe periodontal destruction as mild disease, or a non-restorable tooth as restorable. D3 is intended to capture incorrect visual-clinical grounding rather than total omission. Under the hierarchy, once D3 is assigned for a finding, downstream triage mistakes caused by this initial misjudgment are not separately counted as protocol errors.

D4: Protocol Misalignment. This dimension captures failures in which the model correctly understands the relevant pathology or clinical find-

Rater	Fine-grained (22 labels)				Coarse-grained (8 specialties)			
	Macro-F1	Macro-Rec	Micro-F1	Exact-Match	Macro-F1	Macro-Rec	Micro-F1	Exact-Match
Junior Dentist 1	0.394	0.440	0.511	0.049	0.567	0.647	0.696	0.203
Junior Dentist 2	0.409	0.439	0.520	0.077	0.547	0.604	0.730	0.260
Junior Dentist 3	0.402	0.413	0.546	0.122	0.571	0.586	0.754	0.354
Average	0.402	0.431	0.525	0.083	0.562	0.613	0.727	0.272

Table 7: Individual performance of the three junior dentists on DENTAL-TRIAGEBENCH. We report the same evaluation metrics as in the main paper.

ing, but still maps it to the wrong triage label under the clinical referral rules. D4 is therefore the most restrictive category: it is assigned only when there is no preceding error of complaint neglect, OPG omission, or OPG misinterpretation for that specific finding. In other words, the pathology is perceived correctly, but the final referral decision does not follow the intended clinical triage protocol. Examples include assigning cariology when the rationale correctly describes caries extending to the pulp, or failing to route a correctly identified impacted wisdom tooth to the appropriate OMFS category. The full expert-authored Clinical Triage Protocol can be found at the end of the Table 12.

D5: Hallucination. This dimension captures failures in which the model fabricates symptoms, anatomical structures, lesions, or radiographic findings that are unsupported by either the patient complaint or the expert ground-truth rationale. Hallucination is triggered solely from the generated reasoning text and is independent of whether the final triage labels happen to be correct. For example, if the model invents an impacted tooth, a lesion, or a symptom not present in the case, the output is marked as D5 even if the predicted referral set coincidentally overlaps with the ground truth.

Attribution principle. The purpose of this taxonomy is not to count every observable downstream mistake independently, but to identify the earliest clinically meaningful source of failure in the model’s reasoning process. This design helps distinguish perception failures from decision-rule failures and prevents the same underlying mistake from being counted multiple times across dimensions. As a result, the taxonomy provides a more interpretable account of how multimodal LLMs fail in dental triage. And the detailed prompt of scoring guideline to the LLM Judge is demonstrated in Table 12.

Judge	Validation outputs	Avg. accuracy
Gemini-3-Pro	100	93.6%
GPT-5.4	100	95.2%
GLM-5	100	91.4%
Majority vote	100	94.6%

Table 8: Judge agreement with senior-dentist failure annotations.

H.2 Human–Judge Agreement

We expanded expert validation to 100 sampled model–case outputs. A senior dentist independently assigned binary labels for all five failure dimensions using the same criteria as the judges. Table 8 compares three judge families and their majority vote against these human labels. Cross-family voting reduces dependence on any single vendor family; it does not, however, guarantee perfect instance-level attribution.

Table 9 reports per-dimension confusion counts for the majority vote. Rows are majority-vote predictions and columns are human labels: M0/H0 is a shared negative, M0/H1 a false negative, M1/H0 a false positive, and M1/H1 a shared positive. Protocol Misalignment has 91% accuracy because judges tend to apply written protocol criteria more literally and flag deviations that dentists may regard as acceptable. Hallucination has 92% accuracy because dentists are more sensitive to subtle unsupported extrapolations, whereas judges more often reserve the label for explicit fabrication. We therefore use these labels to estimate aggregate failure patterns and avoid treating every automatic instance label as definitive.

I Additional Case Studies

I.1 A Representative Case on Modality Ablation

We provide a representative modality-ablation case for Gemini-3-Flash in Figure 11. This case illustrates a clear division of labor between the two

Failure dimension	M0/H0	M0/H1	M1/H0	M1/H1	Accuracy
D1 Chief Complaint Neglect	97	0	1	2	99%
D2 OPG Finding Omission	10	3	1	86	96%
D3 OPG Misinterpretation	63	2	3	32	95%
D4 Protocol Misalignment	69	2	7	22	91%
D5 Hallucination	72	6	2	20	92%

Table 9: Per-dimension confusion counts between the majority-vote judge (M) and senior-dentist annotations (H) on 100 outputs.

modalities. The chief complaint, “*retained deciduous teeth and occasional bleeding gums with pain*”, is sufficient to suggest gingival inflammation and some form of extraction need, allowing the text-only model to correctly predict *Gingivitis* and *Exodontia*. However, it does not specify the number, position, or surgical complexity of the affected teeth. As a result, under the complaint-only setting, the model fails to activate *Minor O.S.*, thereby under-calling the full OMFS treatment plan. By contrast, when the OPG is available, the model can directly identify the retained tooth and multiple impacted third molars, which provide the critical visual evidence for surgical complexity and support the correct joint prediction of both *Exodontia* and *Minor O.S.*. This example highlights why dental triage is intrinsically multimodal: text can often capture symptoms and broad treatment intent, whereas radiographic evidence is necessary to determine the full procedural scope and referral complexity.

I.2 A Failure Case Study on Gemini-3-Pro

We provide an additional failure case study for Gemini-3-Pro in Figure 12. In this case, the model misinterprets both the status of tooth 36 and the severity of tooth 27 on the OPG. These OPG misinterpretation errors lead it to assign incorrect triage labels, namely *Molar Endodontics* and *Crowns*, resulting in false positives and thus over-triage. Although the model correctly recognizes the impaction of tooth 48, it fails to translate this finding into the appropriate OMFS referral for *Exodontia*, leading to under-triage.

J Prompt Template Details

J.1 Prompt Template for Model Inference

Table 10 illustrates the task instruction to all models for 0-shot inference.

J.2 Prompt Template for Output Formatting

We present the output formatting prompt in Table 11 to another LLM when the model output is

inconsistent with the standardized JSON output format

J.3 Prompt Template for LLM Judge

Table 12 displays the scoring guideline for LLM judge used in the Section 5.2.



Periodontology • <i>Gingivitis (BPE screening score 1 or 2) - Non-surgical periodontal treatment</i> • <i>Mild to moderate periodontitis (BPE screening score 3) - Non-surgical periodontal treatment</i> • <i>Severe periodontitis (BPE screening score 4) - Complex periodontal treatment</i>
Cariology • <i>Simple operative (Class I to V)</i> • <i>Management of discolorations (e.g. Veneers and/or vital bleaching)</i>
Endodontics • <i>Endodontics (Anterior and/or premolars)</i> • <i>Molar endodontics</i>
Crowns • <i>Single-unit Crowns (Anterior)</i> • <i>Crowns (Posterior; incl. Onlays)</i>
Bridges • <i>Resin bonded bridges (Anterior; <= 2 missing teeth)</i> • <i>Resin bonded bridges (Posterior; <= 2 missing teeth)</i> • <i>Conventional bridges (<= 4 units)</i>
Denture • <i>Removable partial dentures (Simple; bound or free-end saddle)</i> • <i>Extensive removable partial denture (Acrylic base; mucosally supported)</i> • <i>Overdenture (incl. complete or partial)</i> • <i>Complete dentures</i> • <i>Combination case (Edentulous in one arch only)</i>
Implants • <i>Implants (Single tooth)</i> • <i>Implants (Multiple teeth)</i>
OMFS • <i>Exodontia</i> • <i>Minor O.S</i> • <i>Other OMSF Consult</i>

Figure 10: **Hierarchical taxonomy of the dental triage label space in our benchmark.** The taxonomy is organized into two levels: 8 coarse-grained triage domains (e.g., Periodontology, Endodontics, and OMFS) and 22 fine-grained triage categories, which specify treatment-oriented referral categories within each coarse-grained triage domain.



Patient Chief Complaint:

The patient presents with retained deciduous teeth and occasional bleeding gums with pain

- ☒ True Positive (Correct Triage)
- ☐ False Negative (Under-triage)

Predicted Triage Output:

• Perio

☒ Gingivitis (BPE screening score 1 or 2) -
Non-surgical periodontal treatment

• OMFS

☒ Exodontia
☒ Minor O.S.

Triage Reasoning:

The OPG shows impacted 18, 28, 38, 48 and retained deciduous teeth which require Exodontia and Minor O.S.. And the patient's complaint of bleeding gums and pain suggest gingivitis, requiring non-surgical periodontal treatment.

Chief Complaint+OPG



Predicted Triage Output:

• Perio

☒ Gingivitis (BPE screening score 1 or 2) -
Non-surgical periodontal treatment

• OMFS

☒ Exodontia
☐ Minor O.S.

Triage Reasoning:

The patient reports bleeding gums and pain, indicating gingival inflammation requiring non-surgical periodontal treatment. The presence of retained deciduous teeth in an adult requires exodontia to address the pain and allow for potential orthodontic or restorative management of the permanent dentition.

Chief Complaint-only



Miss findings of the impacted teeth 18, 28, 38, 48 that required Minor O.S. shown on the OPG

Predicted Triage Output:

• Perio

☐ Gingivitis (BPE screening score 1 or 2) -
Non-surgical periodontal treatment

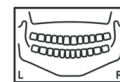
• OMFS

☒ Exodontia
☒ Minor O.S.

Triage Reasoning:

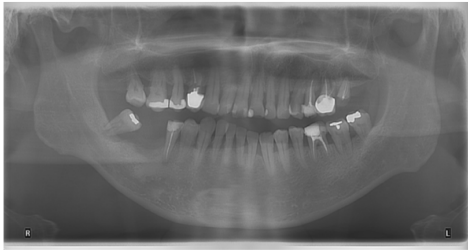
The OPG reveals impacted third molars in all four quadrants. Specifically, the mandibular third molars (38 and 48) are horizontally or mesio-angularly impacted, which often necessitates surgical extraction (Minor O.S./Exodontia) to prevent pericoronitis or damage to adjacent second molars.

OPG-only



Miss information of the bleeding gums with pain situation reported by the patient in the chief complaint

Figure 11: A representative modality-ablation case on Gemini-3-Flash. The chief complaint alone supports *Gingivitis* and *Exodontia*, but misses the surgical complexity required for *Minor O.S.*. With access to the OPG, the model correctly recovers the full OMFS referral plan.



Patient Chief Complaint:

The patient presents with gum disease and reported fractured teeth on the upper left.

Ground Truth Triage Labels:

• **Perio**

■ Severe periodontitis (BPE screening score 4) - Complex periodontal treatment

• **OMFS**

■ Exodontia

Ground Truth Triage Reasoning:

1. Gum disease, OPG show bone resorption, 36 furcation involved, refer for Severe periodontitis (BPE screening score 4) - Complex periodontal treatment
2. 27 fractured tooth, retained root, 46, 47 missing, 48 mesially angulated with adjacent bone resorption, all of which refer OMFS for exodontia

Model Predicted Triage Labels:

• **Perio**

■ Severe periodontitis (BPE screening score 4) - Complex periodontal treatment

• **Crowns**

■ Crowns (Posterior; incl. Onlays)

• **OMFS**

□ Exodontia

• **Endo**

■ Molar endodontics

Model's Triage Reasoning:

The patient presents with gum disease and fractured upper left teeth. The OPG reveals generalized horizontal bone loss ① Tooth 36 exhibits a periapical radiolucency indicating a need for molar endodontics. The upper left molars (26, 27) are root-filled and heavily restored; the reported fracture in this area necessitates assessment for posterior crowns to restore function. ② Tooth 48 is impacted.

Corresponding Failure Patterns:

- ① **OPG Misinterpretation:** Misinterpreted the status of 36 and the severity of 27.
- ② **Protocol Misalignment:** Correct identification of impacted 48, but fails to map this finding to any OMFS label.

Figure 12: **Additional failure case study of Gemini-3-Pro.** The model misinterprets the status of tooth 36 and overestimates the severity of tooth 27 on the OPG, leading to incorrect positive predictions for *Molar Endodontics* and *Crowns*. Although it correctly identifies the impaction of tooth 48, it fails to map this finding to the appropriate OMFS referral for *Exodontia*. This example illustrates how OPG misinterpretation can simultaneously produce both over-triage and under-triage errors within the same case.

Prompt for task instruction for model inference

You are a triage expert and dentist, you are given a patient's chief complaint and one OPG image, and a list of triage/treatment labels, you need to triage the patient to the most appropriate triage/treatment labels and provide your reasoning for the triage (where the label is 1). The reasoning should be a string that explains your triage decision based on the chief complaint and OPG image findings. Please give your response exactly in the following JSON format where 0 stands for no, 1 stands for yes:

```
```json
{
 "triage_output": {
 "Perio": {
 "Gingivitis (BPE screening score 1 or 2) - Non-surgical periodontal treatment": 0,
 "Mild to moderate periodontitis (BPE screening score 3) - Non-surgical periodontal treatment": 0,
 "Severe periodontitis (BPE screening score 4) - Complex periodontal treatment": 0
 },
 "Cariology": {
 "Simple operative (Class I to V)": 0,
 "Management of discolorations (e.g. Veneers and/or vital bleaching)": 0
 },
 "Endo": {
 "Endodontics (Anterior and/or premolars)": 0,
 "Molar endodontics": 0
 },
 "Crowns": {
 "Single-unit Crowns (Anterior)": 0,
 "Crowns (Posterior; incl. Onlays)": 0
 },
 "Bridges": {
 "Resin bonded bridges (Anterior; <= 2 missing teeth)": 0,
 "Resin bonded bridges (Posterior; <= 2 missing teeth)": 0,
 "Conventional bridges (<= 4 units)": 0
 },
 "Denture": {
 "Removable partial dentures (Simple; bound or free-end saddle)": 0,
 "Extensive removable partial denture (Acrylic base; mucosally supported)": 0,
 "Overdenture (incl. complete or partial)": 0,
 "Complete dentures": 0,
 "Combination case (Edentulous in one arch only)": 0
 },
 "Implants": {
 "Implants (Single tooth)": 0,
 "Implants (Multiple teeth)": 0
 },
 "OMFS": {
 "Exodontia": 0,
 "Minor O.S.": 0,
 "Other OMFS Consult": 0
 }
 },
 "reasoning": "Your concise analysis here (max 200 words)"
}
```

And pay attention to the following:

1. Other OMFS Consult includes the following items: Tumors, cysts, TMJ, fractures, mucosal lesions.
2. Bear in mind that there could be multiple triage/treatment labels that are relevant to the patient, meaning that the patient has multiple conditions or needs multiple treatments.
3. Your output should be a json object that is strictly formatted to the above json format, please do not add any other text or comments to the output or add any new keys to the json object.

The chief complaint: {chief\_complaint}

And the OPG image is:

Table 10: The task instruction prompt for model inference

### Prompt for task instruction for model inference

You are an expert in JSON formatting, you are given an invalid JSON object or free text from another model, you need to format its answer to a valid JSON object. The format should strictly follow this:

```
{{
 "triage_output": {{
 "Perio": {{
 "Gingivitis (BPE screening score 1 or 2) - Non-surgical periodontal treatment": 0,
 "Mild to moderate periodontitis (BPE screening score 3) - Non-surgical periodontal treatment": 0,
 "Severe periodontitis (BPE screening score 4) - Complex periodontal treatment": 0
 }},
 "Cariology": {{
 "Simple operative (Class I to V)": 0,
 "Management of discolorations (e.g. Veneers and/or vital bleaching)": 0
 }},
 "Endo": {{
 "Endodontics (Anterior and/or premolars)": 0,
 "Molar endodontics": 0
 }},
 "Crowns": {{
 "Single-unit Crowns (Anterior)": 0,
 "Crowns (Posterior; incl. Onlays)": 0
 }},
 "Bridges": {{
 "Resin bonded bridges (Anterior; <= 2 missing teeth)": 0,
 "Resin bonded bridges (Posterior; <= 2 missing teeth)": 0,
 "Conventional bridges (<= 4 units)": 0
 }},
 "Denture": {{
 "Removable partial dentures (Simple; bound or free-end saddle)": 0,
 "Extensive removable partial denture (Acrylic base; mucosally supported)": 0,
 "Overdenture (incl. complete or partial)": 0,
 "Complete dentures": 0,
 "Combination case (Edentulous in one arch only)": 0
 }},
 "Implants": {{
 "Implants (Single tooth)": 0,
 "Implants (Multiple teeth)": 0
 }},
 "OMFS": {{
 "Exodontia": 0,
 "Minor O.S.": 0,
 "Other OMFS Consult": 0
 }}
 }},
 "reasoning": "model's analysis (in string format, not dictionary or any other format)"
}}
```

Pay attention:

1. The original model output might not contain all the keys in the json object, for those keys that are not present in the original model output, you should set the value to 0.
2. Any other free text that is not related to the key-value pairs, you should convert them to the reasoning field if that makes sense.
3. The original input to the model is: 'You are a triage expert. Analyze the patient's chief complaint, medical history, and OPG image. Output the triage decision strictly in JSON format.'

Now the original model output is: {model\_output}

Table 11: The prompt for standardizing model output

## Prompt for llm as judge for failure analysis

### # Role

You are an expert Chief Dentist and a rigorous AI evaluator. Your task is to perform a Failure Analysis on an AI model's performance in a Multimodal Dental Triage task.

### # Input Data

1. [Patient Input]:
  - Chief Complaint: {chief\_complaint}
2. [Label Discrepancies] (Crucial for Evaluation):
  - Missing Labels (False Negatives - model failed to predict): {missing\_labels}
  - Extra Labels (False Positives - model wrongly predicted): {extra\_labels}
3. [Ground Truth (GT)]:
  - GT Reasoning: {gt\_reasoning}
4. [Model]:
  - Model Reasoning: {model\_reasoning}

### # Core Evaluation Philosophy (Doctor's Workflow)

**\*\*CRITICAL:\*\*** You must evaluate the model primarily based on its **\*\*[Model Reasoning]\*\*** compared to the **\*\*[GT Reasoning]\*\***. Do NOT use label discrepancies as the sole trigger for finding errors.

- Example: If the model fabricates a tooth in its reasoning, it is a Hallucination (D5) EVEN IF it accidentally guessed the correct final triage label.
- Example: If the model misjudges a tooth's severity (D3), any downstream missing restorative labels (like missing a Crown or Implant) are considered a CASCADING consequence of D3. Do NOT tag D4 for the missing restoration.

### # Task Instruction: FIND THE ROOT CAUSE

Evaluate the model on the following 5 STRICTLY MUTUALLY EXCLUSIVE error dimensions. For each specific clinical issue/tooth mentioned in the GT or Model reasoning, assign it to ONLY ONE underlying root cause based on this **\*\*Strict Hierarchical Rule\*\***:

### **\*\*Hierarchy of Evaluation (Evaluate Top to Bottom for each finding):\*\***

1. **\*\*Did the model fabricate something? -> D5.\*\*** (e.g., mentions an impacted lower wisdom tooth that isn't there, or invents a symptom).
2. **\*\*Did the model completely ignore the core text symptom? -> D1.\*\***
3. **\*\*Did the model completely fail to mention an important OPG finding? -> D2.\*\*** (e.g., GT mentions generalized bone loss or caries on tooth 11, but the model reasoning never mentions it).
4. **\*\*Did the model mention the correct finding but misjudge its physical severity/nature? -> D3.\*\*** (e.g., GT says "caries into pulp", but model says "superficial caries"; GT says "severe periodontitis", but model says "mild").
5. **\*\*Did the model perceive EVERYTHING correctly about a lesion (NO D1, D2, or D3 errors), but simply assigned the wrong final Triage label? -> ONLY THEN tag D4.\*\***

### # The 5 Error Dimensions Definitions:

#### **\*\*[Text] D1: Chief Complaint Neglect\*\***

- The model's reasoning completely ignores the core symptoms/requests in the [Patient Input].

#### **\*\*[OPG] D2: OPG Finding Omission\*\***

- The model completely fails to mention important visual findings explicitly listed in the [GT Reasoning].
- **\*Rule:** If the GT mentions a finding and the model reasoning does not mention it at all, it is D2. If the model mentions it but gets the severity wrong, it is D3, NOT D2.

#### **\*\*[OPG] D3: OPG Misinterpretation\*\***

- The model notices the correct lesion/area but misjudges its severity or pathological depth compared to the GT.
- **\*Rule:** If D3 occurs, DO NOT tag D4 for any missing downstream labels (like missing an extraction or crown) caused by this initial misjudgment.

#### **\*\*[Triage] D4: Protocol Misalignment\*\***

- **\*\*Strict Prerequisite:\*\*** The model's reasoning perfectly matches the GT reasoning regarding the anatomical/pathological condition (No D1, D2, or D3 errors for this specific finding).

- The error is PURELY a failure to map a correctly understood pathology to the correct Triage label based on clinical rules. Refer to the [Appendix: Clinical Triage SOPs (Reference Only)] for more details.

**\*\*[Hallucination] D5: Hallucination\*\***

- The model fabricates anatomical structures, lesions, or symptoms in its reasoning that do not exist in the GT or text.  
 - **\*Rule:\*** Must be triggered based on the reasoning text, regardless of whether the final triage labels are right or wrong.

**# Output Format**

Return ONLY a valid JSON object. Do not include markdown formatting.

For the "score", output strictly 1 or 0.

For the "reason", explicitly and briefly state the reasoning of marking this dimension as 1. If the score is 0 for this dimension, leave reason as "N/A".

```
{
 "D1_Neglect": {"score": 0, "reason": "Explanation / N/A"},
 "D2_Omission": {"score": 0, "reason": "Explanation / N/A"},
 "D3_Misinterpretation": {"score": 0, "reason": "Explanation / N/A"},
 "D4_Protocol_Confusion": {"score": 0, "reason": "Explanation / N/A"},
 "D5_Hallucination": {"score": 0, "reason": "Explanation / N/A"}
}
```

**# [Appendix: Clinical Triage Protocol]**

**\*\*1. Perio (Periodontal)\*\***

- IF Bone loss  $\leq$  1/3 root -> Mild; 1/3-1/2 -> Moderate;  $>$  1/2 -> Severe.  
 - IF Calculus present -> Require Perio treatment.  
 - IF Root furcation involved -> Require Complex periodontal treatment.

**\*\*2. Cariology vs. Endo\*\***

- IF Caries extends to pulp -> MUST select Endo (NOT Cariology/Simple Operative).  
 - IF Tooth is non-restorable -> Refer Exodontia.  
 - IF Discoloration BUT vital pulp & no caries/Endo -> Consider Veneers/bleaching.  
 - IF Anterior/Premolar AND Molar need RCT -> Select both Endo categories.

**\*\*3. Restorations (Crowns / Bridges / Implants)\*\***

- IF Posterior tooth needs RCT -> MUST select Crowns (Posterior/Onlays).  
 - IF Anterior tooth needs RCT -> NO Crown usually (unless gross structure missing or necrosis discoloration).  
 - IF Non-wisdom tooth extracted -> MUST consider replacement (Bridges, Dentures, or Implants).

**\*\*4. OMFS (Oral Surgery) Routing Rules\*\***

- Exodontia (Simple): No incision/bone removal. E.g., Maxillary vertical wisdom teeth; Erupted mandibular wisdom without adjacent 2nd molar blocking it; Retained roots.  
 - Minor O.S. (Complex): Requires incision/flap/bone removal. E.g., ALL impacted wisdom teeth; Non-impacted lower wisdom IF adjacent 2nd molar is present.  
 - Rule Exception: IF simultaneous impacted lower (Minor O.S.) + non-impacted upper (Exodontia) -> Can select Minor O.S. only (done in one session), OR both.  
 - Other OMFS Consult: Tumors, cysts, TMJ, fractures, mucosal lesions.

Table 12: The prompt for LLM-as-Judge for failure analysis