# MRAREBENCH: A MULTIMODAL RARE-DISEASE BENCHMARK FOR EVIDENCE–DIAGNOSIS CORRESPONDENCE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multimodal large language models (MLLMs) now approach specialist-level performance on chest radiograph interpretation, dermatology and pathology image classification, and medical visual question answering. These gains, however, come from common diseases, where the clinical narrative alone often narrows the differential. Rare diseases reverse these conditions. Each disease is individually uncommon and sparsely documented, prevalence gives little guidance across thousands of candidates, and the decisive findings are frequently visual and distributed across several images and modalities. Rare disease therefore offers a stringent test of whether a MLLM reaches a diagnosis from evidence or from prior familiarity. Current rare disease benchmarks describe patients in text and omit medical images, while multimodal medical benchmarks supply images but concentrate on common conditions. To address this limitation, we introduce MRareBench, the first benchmark for multimodal rare disease diagnosis. MRareBench comprises 908 evaluation items spanning multiple medical images and diverse imaging modalities, and pairs forward diagnosis with evidence verification under controlled disclosure of the diagnosis. Evaluation of 37 MLLMs shows that visual evidence improves rare disease diagnosis. A gap remains between evidence recognition and open diagnosis. MLLMs often identify relevant visual evidence after receiving the rare disease diagnosis but are less reliable in using multimodal evidence to identify the disease independently. The results indicate a difficulty in inferring rare disease diagnoses from multimodal evidence. MRareBench provides the measurement needed to advance multimodal diagnosis in rare-disease settings.

## 1 INTRODUCTION

Multimodal large language models (MLLMs) have rapidly advanced medical image understanding. On established tasks such as chest radiograph interpretation, dermatology and pathology image classification, and medical visual question answering, they now approach specialist-level performance Li et al. (2023); Saab et al. (2024); Sellergren et al. (2025). These gains, however, are concentrated on common and well-represented conditions, leaving unclear whether predictions rest on case-specific evidence or familiar diagnostic priors.

In medical diagnosis, a decision may be supported either by prior knowledge or by evidence observed in the current case. Such priors include disease prevalence, frequent symptom–disease associations, and high-frequency diagnostic patterns reinforced through training and data exposure. For common diseases, these priors can narrow the differential before the image is fully examined. Evidence-based diagnosis, however, requires the conclusion to be justified by the clinical and visual findings actually present in the patient rather than primarily by prior expectation. Whether MLLMs can sustain such evidence-based diagnosis when these priors are weak remains unresolved.

Rare disease provides a stringent setting for testing evidence-based diagnosis. Although rare diseases collectively affect more than 300 million people worldwide, each individual condition is uncommon Groft et al. (2021); Nguengang Wakap et al. (2020). Prevalence alone therefore provides little useful guidance for narrowing a differential that may contain thousands of plausible candidates, while the scarcity of documented cases limits the statistical regularities available for learning each



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

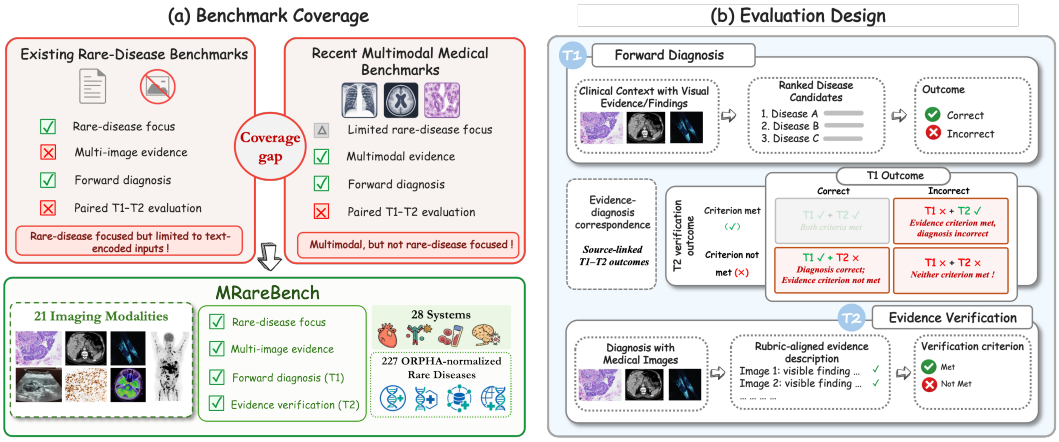



Figure 1: **MRareBench: Benchmark Coverage and Paired Evaluation. (a) Coverage.** MRareBench combines rare-disease focus, multi-image evidence, forward diagnosis (T1), and evidence verification (T2). **(b) Paired evaluation.** Source-linked T1 outcomes and case-aggregated T2 scores are cross-classified; correspondence is the rate of correct T1 LC Top-1 diagnoses among T2-positive model–case pairs.

condition Choi et al. (2017). At the same time, many rare diseases are distinguished by characteristic radiological, pathological, dermatological, or ophthalmological patterns, often distributed across multiple images and imaging modalities.

Table 1: **Benchmark Design Comparison.** MRareBench vs. existing benchmarks (✓ yes, △ partial, × no).

| Benchmark | Pub. | Rare | Multi-img. | Evid. ctrl. | Fwd/bwd |
|---|---|---|---|---|---|
| RareBench (Chen et al., 2024d) | KDD 2024 | ✓ | × | × | × |
| RareArena (Chen et al., 2026b) | Lancet Digit. Health 2026 | ✓ | × | △ | × |
| MedR-Bench (Qiu et al., 2025) | Nat. Commun. 2025 | △ | × | × | × |
| GMAI-MMBench (Chen et al., 2024c) | NeurIPS 2024 | △ | × | × | × |
| MedXpertQA (Zuo et al., 2025) | ICML 2025 | △ | △ | × | × |
| MedFrameQA (Yu et al., 2025) | 2025 | × | ✓ | × | × |
| MTBBench (Vasilev et al., 2026) | NeurIPS 2025 | × | ✓ | × | × |
| MRareBench | This paper | ✓ | ✓ | ✓ | ✓ |

Existing benchmarks leave two gaps that motivate the design in Figure 1. **(a)** *Coverage gap.* Rare disease benchmarks typically represent cases through textual descriptions and provide no medical images Chen et al. (2024d); Qiu et al. (2025). Multimodal medical benchmarks increasingly include multi-image cases but predominantly cover common or comparatively well-represented conditions Chen et al. (2024c); Zuo et al. (2025); Yu et al. (2025); Vasilev et al. (2026). As a result, multimodal rare disease diagnosis has remained largely unevaluated. **(b)** *Measurement gap.* Appending images does not resolve the limitations of conventional forward-only evaluation. Forward accuracy alone cannot reveal how much the images contribute to the diagnosis, or whether a model that can report strong supporting evidence can also identify the disease correctly. MRareBench addresses this gap through a controlled evaluation design that measures how evidence reporting changes when the diagnosis is revealed and whether strong evidence reporting is accompanied by correct open diagnosis. We refer to the latter relationship as *evidence–diagnosis correspondence*.

Hence, we introduce **MRareBench**, a benchmark for multimodal rare disease diagnosis and failure analysis spanning 227 rare diseases and a broad range of imaging modalities. MRareBench comprises two complementary evaluation tracks. The *forward-diagnosis* track (T1) requires a ranked differential diagnosis from the clinical context and images, and evaluates each case under the limited views to characterize the diagnostic contribution of visual evidence. The *evidence-verification* track (T2) evaluates rubric-aligned visual evidence reporting by comparing conditions in which the diagnosis is either supplied or withheld. Source-linked cases connect the two tracks, which enables us to perform the analysis of how diagnostic correctness relates to evidence reporting.

Across 37 MLLMs, our study shows that visual evidence contributes diagnostic signal beyond clinical context alone in multimodal rare disease diagnosis. A clear behavioral asymmetry is observed between evidence reporting and open disease identification. Supplying the rare disease diagnosis consistently improves evidence reporting across models. Strong evidence reporting, however, does not reliably produce a correct open diagnosis. The correct rare disease is often absent from the model differential when open diagnosis fails. The results therefore suggest a limitation in diagnostic hypothesis formation, in which the current models do not consistently convert multimodal evidence into the rare disease hypothesis. Our contributions are as follows:

- **Benchmark.** We introduce MRareBench for multimodal rare disease diagnosis. MRareBench contains 300 forward diagnosis cases and 608 evidence verification items. The benchmark spans 227 rare diseases, multi image evidence, and a broad range of imaging modalities.
- **Paired evaluation.** MRareBench connects open diagnosis with controlled evidence verification. The evaluation measures the diagnostic contribution of medical images, the effect of diagnosis disclosure on evidence reporting, and the relation between evidence reporting and correct open diagnosis.
- **Findings.** Across 37 MLLMs, visual evidence provides additional diagnostic signal in rare disease cases. Diagnosis disclosure consistently improves evidence reporting, but improved evidence reporting does not reliably produce correct open diagnosis. The results further suggest a limitation in the formation of appropriate diagnostic hypotheses from multimodal evidence.

## 2 Related Work

**Multimodal Medical Benchmarks.** Medical multimodal evaluation began with single-image visual question answering on radiology and pathology images Lau et al. (2018); He et al. (2021), and has since expanded along two axes: coverage, where large-scale suites span many departments, imaging modalities, and task formats under a single protocol Chen et al. (2024c), and difficulty, where expert-level benchmarks target specialist judgment rather than factual recall Zuo et al. (2025); Qiu et al. (2025). A parallel line replaces isolated questions with authentic case narratives derived from published case reports Wu et al. (2025), and more recent multi-image benchmarks require findings to be integrated across studies or along a longitudinal course Yu et al. (2025); Vasilev et al. (2026); Hu et al. (2023); Jin et al. (2024a). Two concurrent efforts address the evidence itself: MEDSYN evaluates multi-evidence synthesis over complex cases Chen et al. (2026a), and CURE varies the evidence supplied to a model to separate inference from retrieval Gu et al. (2026). A further body of work asks what a benchmark score is made of, since high accuracy can arise from spurious cues rather than the intended capability Geirhos et al. (2020). Visual question answering models exploit language priors and question–answer regularities, which motivated balanced protocols that force the answer onto the image Goyal et al. (2017); Dancette et al. (2021), and general-domain audits measure how much apparent multimodal skill survives once the image is withheld Chen et al. (2024b). Medicine raises the same concern: reported expert-level accuracy can partly reflect image-independent answering Jin et al. (2024b); Zafar et al. (2026), and overlap between benchmark items and training corpora inflates measured performance Xu et al. (2024).

**Rare-Disease Benchmarks.** Rare-disease evaluation operates under long-tailed disease distributions, phenotypes that overlap heavily across conditions, and few documented patients per condition. The dominant formulation is phenotype-centric: a case is represented as a structured list of clinical features, and a model ranks candidate conditions against that list Chen et al. (2024d); Kim et al. (2024). Later work keeps this input format and enriches the inference procedure through external knowledge sources and agentic pipelines that split diagnosis into retrieval and inference steps Chen et al. (2026b); Zhao et al. (2026). A layer of standardized clinical semantics underpins both: Orphanet identifiers fix disease identity across heterogeneous nomenclature, the Human Phenotype Ontology fixes feature identity, and Phenopacket-style schemas make individual cases interoperable Rath et al. (2012); Jacobsen et al. (2022). The evidence exposed to the model remains textual throughout.

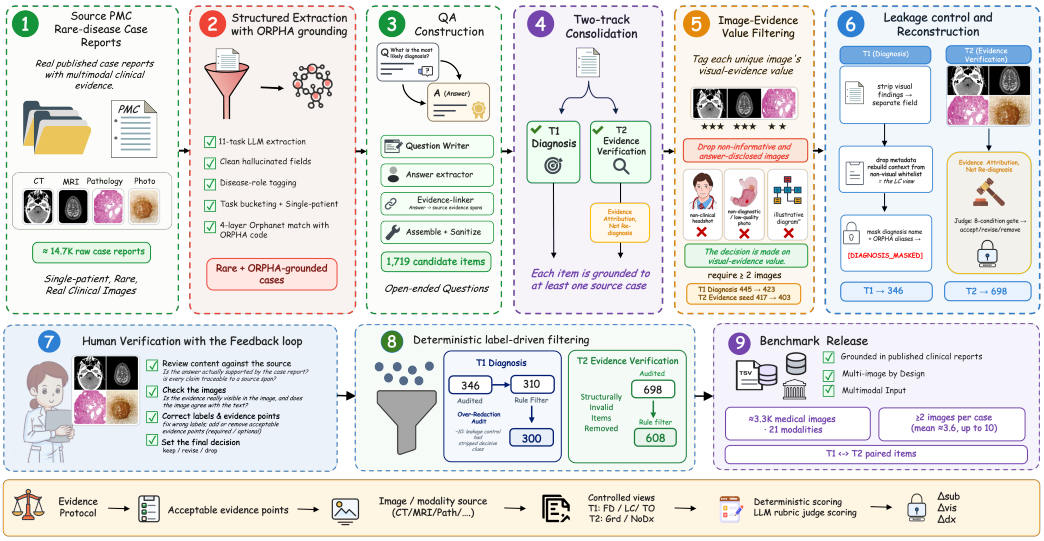



Figure 2: **MRareBench Construction Pipeline.** Nine stages turn published PMC rare-disease case reports into forward diagnosis and T2 evidence verification items, with ORPHA grounding, image-evidence filtering, leakage control, and human verification.

# 3 METHOD

## 3.1 BENCHMARK DESIGN

MRareBench comprises two tracks: forward diagnosis (T1), which evaluates whether a model can identify the disease, and evidence verification (T2), which evaluates whether it can ground a given diagnosis in the visible findings.

**Forward Diagnosis.** T1 is posed as open-set differential diagnosis. Formally, a case is $x_i = (c_i, \mathcal{I}_i)$ with clinical context $c_i$ and image set $\mathcal{I}_i = \{v_{i1}, \ldots, v_{im_i}\}$, and a model $f$ produces

$$\hat{y}_i = \big(\hat{y}_{i1}, \ldots, \hat{y}_{iK}\big) = f(c_i, \mathcal{I}_i), \qquad \hat{y}_{ik} \in \mathcal{D}, \quad K = 10, \tag{1}$$

where $\mathcal{D}$ is the unbounded space of disease names and the gold label $y_i \in \mathcal{D}$ is the Orphanet-normalized diagnosis. No candidate set is supplied, so the model must construct the differential itself. Open-set generation shifts the burden to evaluation. We therefore score a prediction as correct when it matches any Orphanet-registered alias of the gold diagnosis under normalized string matching,

$$\mathrm{match}(\hat{y}, y) = \max_{a \in \mathcal{A}(y)} \mathbb{1}\big[\pi\big(\sigma(\hat{y}), \sigma(a)\big)\big], \tag{2}$$

where $\mathcal{A}(y)$ is the alias set of $y$, $\sigma$ is clinical-string normalization, and $\pi$ is the string-matching predicate defined in the appendix. We report Recall@n over the ranked list.

**Controlled Evidence Views.** A case exposes three evidence channels: the clinical context $c_i$, the imaging-findings text $t_i$ that narrates what the medical images contain, and the images $\mathcal{I}_i$. Toggling the latter two yields three views,

$$\mathrm{FD} : (c_i, t_i, \mathcal{I}_i), \qquad \mathrm{LC} : (c_i, \mathcal{I}_i), \qquad \mathrm{TO} : (c_i), \tag{3}$$

under which a model is evaluated on the same case. The limited-context view (LC) is the primary image-grounded condition; the findings-disclosed view (FD) measures performance when the findings are also stated in text, and the text-only view (TO) is the image-withheld comparator to LC. Writing $R_{\mathrm{FD}}$, $R_{\mathrm{LC}}$, and $R_{\mathrm{TO}}$ for Recall@1 under each view, we report

$$\Delta_{\mathrm{sub}} = R_{\mathrm{FD}} - R_{\mathrm{LC}}, \qquad \Delta_{\mathrm{vis}} = R_{\mathrm{LC}} - R_{\mathrm{TO}}. \tag{4}$$

The gap $\Delta_{\mathrm{sub}}$ measures the change in diagnosis recall when the findings are additionally disclosed in text while the images remain available. The visual gap $\Delta_{\mathrm{vis}}$ measures the change when images are available: LC and TO share the same clinical context, with TO removing the images and the image reference in the response instruction.

**Evidence Verification.** T2 evaluates whether a model can ground a diagnosis in what the images show. Under the supplied-diagnosis condition (Grd), an item is presented as $(u_i, d_i)$, where $u_i$ is the evidence input and $d_i$ the target disease, and the model returns a free-text evidence statement

$$e_i = f(u_i, d_i), \tag{5}$$

which is scored against a rubric $P_i = \{p_{i1}, \ldots, p_{in_i}\}$ of required evidence points extracted from the source report, with a judge assigning binary hits $b_{ip} = \mathbb{1}[\, p \text{ is satisfied in } e_i \,]$. Supplying $d_i$ fixes the target hypothesis, so the resulting score isolates evidence reporting from disease identification.

**Diagnosis Conditions.** Two conditions differ only in whether the disease identity is given: $\mathrm{Grd}$ supplies the diagnosis alongside the images, and $\mathrm{NoDx}$ withholds the name while keeping the images, the question, and the response instruction fixed. A third condition, $\mathrm{Rec}$, supplies the diagnosis but withholds the images, measuring evidence reporting without visual input. For a fixed model, let $S_c^{(\ell)}$ be the mean item score under condition $c$ at tier $\ell \in \{\mathrm{L1}, \mathrm{full}\}$ on the Grd–NoDx common-valid items. The diagnosis-conditioning gap is

$$\Delta_{\mathrm{dx}}^{(\ell)} = S_{\mathrm{Grd}}^{(\ell)} - S_{\mathrm{NoDx}}^{(\ell)}, \tag{6}$$

Level 1 is the primary diagnosis-conditioning contrast; the full-tier contrast also includes integration and attribution credit.

**Hierarchical Required-Evidence Recall.** A model can hit rubric points without acquiring the evidence from the visual content. Once the diagnosis is supplied, its characteristic pattern can be described from memory. HierReqRecall withholds higher-level credit unless the required lower-level evidence is also reported: visible findings precede cross-image integration, which precedes diagnostic attribution. A gate $\gamma_{ip} \in \{0, 1\}$ requires all Level-1 points for Level-2 credit and all Level-1 and Level-2 points for Level-3 credit. Each required point $p \in P_i$ carries an evidence-bucket weight $w_p$, and the score over $N$ valid T2 items is

$$\mathrm{HierReqRecall} = \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{p \in P_i} w_p \, \gamma_{ip} \, b_{ip}}{\sum_{p \in P_i} w_p}, \tag{7}$$

which reduces to flat required-evidence recall when $w_p = 1$ and $\gamma_{ip} = 1$ for all $p$. More details are described in appendix.

**Evidence–Diagnosis Correspondence.** The two tracks share a set of source-linked cases, which lets us ask whether a model that reports the supporting evidence also identifies the disease on its own. For model $f$ and linked case $i$, let

$$h_{f,i} = \mathbb{1}\big[\hat{y}_{f,i,1} \text{ matches } y_i\big], \qquad z_{f,i} = \mathbb{1}\big[S_{f,i} \geq \tau\big], \tag{8}$$

where $\hat{y}_{f,i,1}$ is the top-ranked LC diagnosis from model $f$, and $S_{f,i}$ is the arithmetic mean of its valid T2 item scores linked to T1 case $i$, computed before thresholding. Let $\Omega$ denote the valid model–case pairs. The metric is defined as:

$$\widehat{\mathrm{Corr}}_{\mathrm{pooled}} = \frac{\sum_{(f,i) \in \Omega} h_{f,i} z_{f,i}}{\sum_{(f,i) \in \Omega} z_{f,i}}. \tag{9}$$

### 3.2 Benchmark Construction

**Source and Pipeline.** MRareBench draws on real PMC case reports published between 2005 and 2025, seeded by the PMCID list of (Wu et al., 2025). A nine-stage pipeline converts each heterogeneous report into an auditable structured record, summarized in Figure 2. Each record holds typed clinical evidence blocks, candidate disease mentions, image sets, figure links, and Orphanet-normalized diagnosis candidates, and preserves every PMCID, sentence identifier, figure identifier, and modality label.

**Masking and Channel Construction.** The pipeline separates the model-facing input from every text span that reveals the answer. Each report is split into four regions: a base region with non-revealing clinical context, a diagnosis region naming the disease and its Orphanet aliases, a visual region narrating the findings the images carry, and a metadata region covering the title, abstract, conclusion, and captions. Masking removes the diagnosis and metadata regions from every view, and the image-grounded view additionally drops the visual region. Constructing the image-grounded channel is the harder step. A source-grounded correction pass re-checks each item against the original report: it repairs figure misalignment, removes answer-revealing cues from the question stem, and relocates narrated findings into a separate findings-disclosed field while demographics, history, and laboratory context stay shared across views. Quality control is performed automatically and then reviewed by clinicians, who confirm that each item carries the visual information required to answer it and is solvable from the released input. Details are reported in Appendix D.

**Released Sets.** The two tracks contain 300 forward-diagnosis cases and 608 evidence-verification items, spanning 227 rare diseases and 21 imaging modalities. Table 1 compares design coverage against prior benchmarks.

## 4 Experiments and Analysis

### 4.1 Setup and Cohorts

We evaluate 37 MLLMs grouped by access into three cohorts: *closed-source* and *open-source* general models, and *medical* domain-specialized models. Table A4 lists every model with its reproduction identifier, and results are reported per cohort. Full per-view tables, with Recall@10 and MRR, appear in Tables A16 through A18. Figure 3 ranks all models by diagnosis accuracy under different conditions.

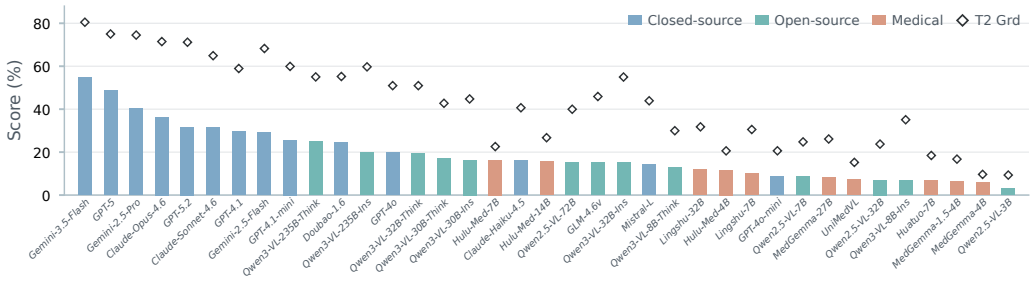



Figure 3: **Model Leaderboard.** 37 models ranked by T1 LC Recall@1, colored by access cohort. Hollow diamonds mark the same models' T2 Grd full-tier HierReqRecall, which is measured on a different scale.

### 4.2 Diagnostic Contribution of Visual Evidence

For a rare disease the clinical narrative rarely narrows the differential on its own, so what the images show has to carry the diagnosis. Withholding them costs every cohort Recall@1, and the strongest models lose the most, so this signal is available only to models able to read it. The limited-context and text-only views share the clinical context; the latter removes the images and the image reference in the instruction. Their contrast measures the observed contribution of image availability. The opposite contrast measures what a prose description of the findings adds while images remain available, and for the closed-source cohort it is about a third of the findings-disclosed score. A benchmark that states the imaging findings in text therefore measures something other than image-grounded diagnosis.

Table 2: **Quantitive Results across Models.** Ranked diagnosis (T1), evidence verification (T2), and their operational correspondence. **T1** is Recall@1 (%); **T2** is primary-judge full-tier HierReqRecall (%); **Corr** uses $\tau$=0.5.

| | T1 | | | | | T2 | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | **FD** | **LC** | $\Delta_{\text{sub}}$ | **TO** | $\Delta_{\text{vis}}$ | **Grd** | **NoDx** | $\Delta_{\text{dx}}^{\text{full}}$ | **Corr** |
| **Closed-source** | | | | | | | | | |
| GPT-4o-mini (Hurst et al., 2024) | 20.3 | 8.7 | 11.67 | 7.7 | 1.00 | 20.6 | 4.1 | 16.5 | 14.6 |
| GPT-4o (Hurst et al., 2024) | 37.7 | 19.7 | 18.00 | 17.3 | 2.33 | 51.0 | 19.1 | 31.9 | 19.4 |
| GPT-4.1-mini (OpenAI, 2025a) | 42.0 | 25.7 | 16.33 | 15.0 | 10.67 | 59.9 | 27.2 | 32.8 | 27.1 |
| GPT-4.1 (OpenAI, 2025a) | 43.3 | 29.7 | 13.67 | 18.3 | 11.33 | 58.9 | 29.1 | 29.9 | 31.9 |
| GPT-5 (Singh et al., 2025) | 55.0 | 48.7 | 6.33 | **27.7** | 21.00 | 75.0 | 42.4 | 32.6 | 51.4 |
| GPT-5.2 (OpenAI, 2025b) | 45.0 | 31.7 | 13.33 | 22.7 | 9.00 | 71.1 | 39.1 | 32.0 | 30.9 |
| Gemini-2.5-Flash (Comanici et al., 2025) | 43.7 | 29.0 | 14.67 | 13.0 | 16.00 | 68.3 | 41.9 | 26.3 | 31.4 |
| Gemini-2.5-Pro (Comanici et al., 2025) | 51.3 | 40.3 | 11.00 | 22.7 | 17.67 | 74.5 | 43.1 | 31.4 | 42.5 |
| Gemini-3.5-Flash (Google DeepMind, 2026) | **63.7** | **54.7** | 9.00 | 26.0 | 28.67 | 80.5 | 63.6 | 16.9 | 54.3 |
| Claude-Haiku-4.5 (Anthropic, 2025) | 30.3 | 16.0 | 14.33 | 13.0 | 3.00 | 40.7 | 11.3 | 29.4 | 23.5 |
| Claude-Sonnet-4.6 (Anthropic, 2026b) | 45.0 | 31.3 | 13.67 | 22.0 | 9.33 | 64.9 | 30.1 | 34.8 | 30.2 |
| Claude-Opus-4.6 (Anthropic, 2026a) | 51.3 | 36.3 | 15.00 | 22.7 | 13.67 | 71.5 | 38.5 | 33.0 | 37.6 |
| Doubao-Seed-1.6-Vision (ByteDance Seed, 2025) | 34.3 | 24.7 | 9.67 | 12.3 | 12.33 | 55.2 | 27.6 | 27.6 | 24.0 |
| Mistral-Large-2512 (Mistral AI, 2025) | 32.0 | 14.3 | 17.67 | 9.3 | 5.00 | 43.9 | 9.7 | 34.2 | 22.7 |
| **Open-source** | | | | | | | | | |
| Qwen3-VL-8B-Instruct (Bai et al., 2025a) | 14.3 | 6.7 | 7.67 | 5.0 | 1.67 | 35.1 | 15.3 | 19.8 | 7.7 |
| Qwen3-VL-8B-Thinking (Bai et al., 2025a) | 19.7 | 13.0 | 6.67 | 8.0 | 5.00 | 30.0 | 10.6 | 19.4 | 14.8 |
| Qwen3-VL-30B-Instruct (Bai et al., 2025a) | 27.0 | 16.3 | 10.67 | 8.3 | 8.00 | 44.8 | 19.9 | 24.9 | 16.1 |
| Qwen3-VL-30B-Thinking (Bai et al., 2025a) | 28.0 | 17.0 | 11.00 | 10.7 | 6.33 | 42.7 | 15.9 | 26.9 | 14.4 |
| Qwen3-VL-32B-Instruct (Bai et al., 2025a) | 24.3 | 15.0 | 9.33 | 8.0 | 7.00 | 55.0 | 26.5 | 28.4 | 18.0 |
| Qwen3-VL-32B-Thinking (Bai et al., 2025a) | 34.0 | 19.3 | 14.67 | 11.3 | 8.00 | 51.0 | 19.9 | 31.1 | 18.3 |
| Qwen3-VL-235B-Instruct (Bai et al., 2025a) | 28.7 | 20.0 | 8.67 | 9.7 | 10.33 | 59.7 | 26.6 | 33.1 | 17.1 |
| Qwen3-VL-235B-Thinking (Bai et al., 2025a) | 35.7 | 25.0 | 10.67 | 12.7 | 12.33 | 55.0 | 22.0 | 33.1 | 26.3 |
| Qwen2.5-VL-3B (Bai et al., 2025b) | 7.7 | 3.0 | 4.67 | 3.3 | - 0.33 | 9.3 | 5.1 | 4.3 | 0.0 |
| Qwen2.5-VL-7B (Bai et al., 2025b) | 12.3 | 8.7 | 3.67 | 5.0 | 3.67 | 24.8 | 7.1 | 17.7 | 7.7 |
| Qwen2.5-VL-32B (Bai et al., 2025b) | 19.7 | 7.0 | 12.67 | 5.3 | 1.67 | 23.7 | 8.4 | 15.4 | 14.3 |
| Qwen2.5-VL-72B (Bai et al., 2025b) | 26.0 | 15.3 | 10.67 | 10.0 | 5.33 | 40.0 | 14.4 | 25.6 | 17.9 |
| GLM-4.6v (Z.ai, 2025) | 30.7 | 15.3 | 15.33 | 9.0 | 6.33 | 45.9 | 21.9 | 24.0 | 14.3 |
| **Medical** | | | | | | | | | |
| Lingshu-7B (Xu et al., 2025) | 18.0 | 10.0 | 8.00 | 4.7 | 5.33 | 30.5 | 15.3 | 15.3 | 6.2 |
| Lingshu-32B (Xu et al., 2025) | 23.7 | 12.0 | 11.67 | 4.7 | 7.33 | 31.8 | 14.7 | 17.1 | 11.8 |
| MedGemma-4B (Sellergren et al., 2025) | 10.7 | 5.7 | 5.00 | 3.7 | 2.00 | 9.6 | 3.0 | 6.6 | 0.0 |
| MedGemma-27B (Sellergren et al., 2025) | 21.7 | 8.0 | 13.67 | 4.7 | 3.33 | 26.1 | 3.4 | 22.7 | 5.4 |
| Hulu-Med-4B (Jiang et al., 2025) | 19.7 | 11.7 | 8.00 | 6.3 | 5.33 | 20.6 | 8.3 | 12.3 | 5.6 |
| Hulu-Med-7B (Jiang et al., 2025) | 20.3 | 16.3 | 4.00 | 6.7 | 9.67 | 22.6 | 11.5 | 11.1 | 14.3 |
| Hulu-Med-14B (Jiang et al., 2025) | 26.0 | 15.7 | 10.33 | 7.3 | 8.33 | 26.7 | 12.5 | 14.2 | 18.2 |
| UniMedVL (Ning et al.) | 12.7 | 7.3 | 5.33 | 3.7 | 3.67 | 15.2 | 8.2 | 7.1 | 7.4 |
| MedGemma-1.5-4B (Sellergren et al., 2026) | 11.7 | 6.3 | 5.33 | 3.7 | 2.67 | 16.7 | 6.9 | 9.8 | 6.2 |
| HuatuoGPT-Vision-7B (Chen et al., 2024a) | 13.0 | 6.7 | 6.33 | 4.7 | 2.00 | 18.4 | 7.2 | 11.2 | 12.1 |

Table 3: **Findings-text gain by modality.**

| **Primary modality** | **General** | **Medical** |
|---|---|---|
| Pathology (micro) | 2.8 | 0.7 |
| X-ray | 11.7 | 10.0 |
| MRI | 12.2 | 10.8 |
| Other (pooled) | 15.0 | 10.9 |
| CT | 19.1 | 11.9 |

**Findings-Text Gain by Modality.** The findings-text gain varies sharply across modalities, as shown in Table 3. Pathology microscopy sits at one end, where describing the findings in prose adds almost nothing once the slide is available; computed tomography sits at the other, with the largest gain from disclosing the findings text. A smaller gap indicates less additional benefit from the description when images are already available. These differences motivate withholding imaging findings from the LC context: including them changes the information available for diagnosis, particularly for the radiological modalities with larger gains.

**Scale and Visual Evidence.** The contrast between the limited-context and text-only views measures how far a diagnosis rests on the images rather than on the clinical narrative, and it grows with model size: along the Qwen3-VL ladder it rises steadily from 8B to 235B. Smaller models lose little when the images are removed, which is to say their accuracy comes mostly from what the narrative

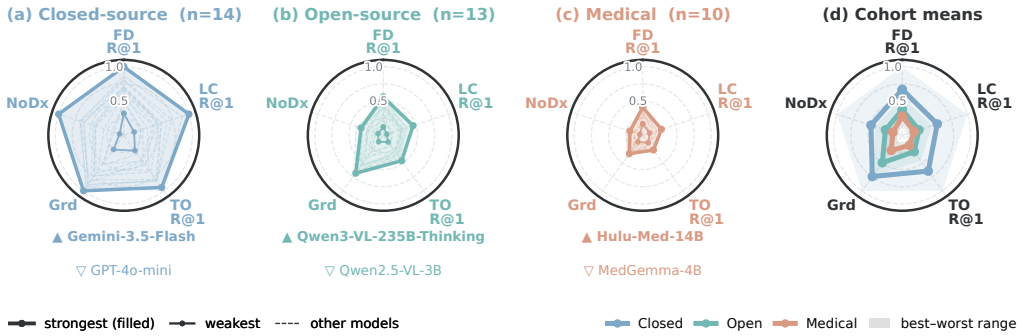



Figure 4: **Comparison of the Cohorts. (a–c)** Per-model profiles; triangles mark each cohort's strongest and weakest model. **(d)** Cohort means and best-worst ranges.

already suggests. Medical models do not follow this trend. At matched parameter count they draw less from the images than their general-purpose counterparts, with a fitted slope of $+3.3$ points per decade against $+9.2$ in Figure 5(a), and on evidence reporting they sit below open-general throughout, the gap widening from 6 to 10 points between 7B and 30B even though diagnosis accuracy is matched within $\pm 3$ points. Domain tuning appears to add disease knowledge rather than the ability to read evidence out of an image, a trade that favours conditions common enough for knowledge to narrow the differential. Figure 4 shows the same pattern distributionally.

Table 4: **Cohort averages across controlled T1 and T2 conditions.** T1 values are Recall@1 (%); T2 values are primary-judge full-tier HierReqRecall (%).

| | | T1 ranked diagnosis (R@1) | | | | | T2 verification (Hier) | |
|---|---|---|---|---|---|---|---|---|
| **Cohort** | **N** | **FD** | **LC** | **TO** | $\Delta_{\text{sub}}$ | $\Delta_{\text{vis}}$ | **Grd** | **NoDx** |
| Closed-source | 14 | 42.5 | 29.3 | 17.8 | 13.2 | 11.5 | 59.7 | 30.5 |
| Open-source | 13 | 23.7 | 14.0 | 8.2 | 9.7 | 5.8 | 39.8 | 16.4 |
| Medical | 10 | 17.7 | 10.0 | 5.0 | 7.8 | 5.0 | 21.8 | 9.1 |

### 4.3 Evidence Verification

When both the disease name and the images are given, several general models reach full-tier HierReqRecall in the 60–80 range, and the gating makes these scores demanding: credit for relating findings or attributing them to the disease is released only after the underlying visible finding has been stated. Withholding the images lowers reporting for all 37 models, so these scores reflect what the models read from the images rather than what they recall about the disease. Models can therefore describe the visual evidence of a rare disease once they are told which disease to look for.

**Effect of Supplying the Diagnosis.** Withholding the disease name, while holding the images, question, instruction, and rubric fixed, lowers Level-1 visible-finding recall for every one of the 37 models, by 24.7 points on average; the secondary judge gives a smaller effect in the same direction for every model described in App. J. We read Level-1 rather than the full tier here, since the higher tiers credit attribution to the disease and would be trivially favoured by supplying its name. The findings recovered under the supplied diagnosis were visible in the images all along, so what the name provides is not evidence but a target: models report the findings once they know which disease to look for, and largely miss them when they do not. In rare disease nothing else supplies that target, since neither prevalence nor the clinical narrative points at the condition in advance.

### 4.4 Evidence–Diagnosis Correspondence

Among the model–case pairs whose evidence reporting clears the threshold, only 26.3% are also diagnosed correctly when the disease name is withheld, as Figure 5(b) shows. The matched pairs

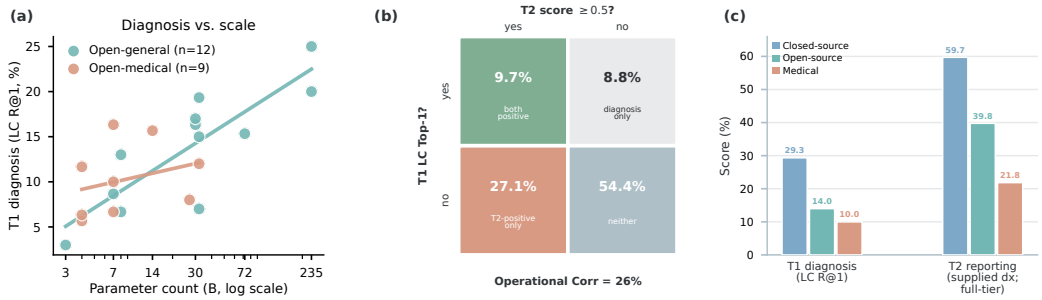



Figure 5: **Scale, Cohorts, and Evidence–Diagnosis Correspondence. (a)** Open-general and open-medical models at matched parameter count, with per-cohort log-linear fits over the 21 models with a published parameter count. **(b)** Diagnosis and evidence-reporting outcomes on the same cases, pooled over models. **(c)** Cohort means for diagnosis accuracy and evidence reporting.

that fail the threshold reach $14.0\%$, so reporting the evidence nearly doubles the chance of naming the disease and still leaves three quarters of those cases undiagnosed. Reporting what the images show is therefore related to identifying the disease, but far from sufficient for it. The rate is stable across the reporting choices, moving by under 3 points as the tier, judge, or threshold varies, and never rising above $39.5\%$ under any setting in Appendix Table A15.

**Where the Correct Disease Lands.** The failures are not ranking errors. Table A19 labels diagnoses that miss at the top rank by hit position: roughly $73\%$ omit the correct rare disease from the ten-candidate differential altogether, and only $27\%$ place it at ranks 2–10. The differential is otherwise populated with rare-disease candidates in Table A12, so the models are searching the right space and not retrieving the right hypothesis from it. For a common condition the answer would ordinarily enter the differential on prior likelihood alone; here nothing puts it there but the evidence, and the evidence does not.

## 5 Conclusion

We introduced MRareBench, the first benchmark for systematically evaluating multimodal rare disease diagnosis in MLLMs. Our evaluation of 37 closed-source, open-source, and medical MLLMs shows that visual evidence provides diagnostic information beyond clinical context alone, but current models remain substantially limited in converting this evidence into the correct rare disease diagnosis. Supplying the diagnosis markedly improves evidence reporting, whereas withholding the diagnosis consistently reduces the visual findings models recover. Strong evidence reporting also does not reliably translate into correct open diagnosis: even when relevant evidence is recognized, the correct disease is frequently absent from the generated differential rather than merely being ranked incorrectly. These findings point to limitations in diagnostic hypothesis formation in current MLLMs. Improving multimodal rare disease diagnosis therefore requires better mechanisms for translating multimodal evidence into appropriate diagnostic hypotheses. MRareBench provides a controlled evaluation framework and a basis for developing more evidence-grounded multimodal diagnostic systems.

## 6 Limitations

MRareBench is constructed from published rare-disease case reports, whose selection and modality distribution may not represent routine clinical populations. Its results therefore characterize model behavior on this benchmark and should not be interpreted as general clinical competence.

## 7 AI USE STATEMENT

Generative AI tools were used for language improving and editing. AI-assisted content was reviewed by the authors. The model-assisted benchmark construction and evaluation process was overseen and manually checked by the authors, with detailed protocols and audit results provided in the appendix.

## 8 ETHICS STATEMENT

MRareBench is built from open-access case reports in PubMed Central and does not introduce newly collected patient data. We did not attempt to identify any patient; potentially identifying information was removed or obscured, including identifiable regions in images, and clinicians participated in benchmark curation and manual verification. The benchmark evaluates model behavior on these cases and does not establish general clinical competence.

## 9 REPRODUCIBILITY STATEMENT

Data and code for reproducing the evaluation are available here. The MRareBench evaluator is integrated into the official repository. The appendix provides additional details on benchmark construction, model configurations, prompts, scoring, and evaluation results.

## REFERENCES


Anthropic. Introducing claude haiku 4.5, October 2025. URL https://www.anthropic.com/news/claude-haiku-4-5. Blog post.

Anthropic. Introducing Claude Opus 4.6, 2 2026a. URL https://www.anthropic.com/news/claude-opus-4-6. Blog post, February 5, 2026.

Anthropic. Introducing claude Sonnet 4.6, 2 2026b. URL https://www.anthropic.com/news/claude-sonnet-4-6. Blog post, February 17, 2026.

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, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo 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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

ByteDance Seed. Technical introduction to the Seed1.6 model series, 2025. URL https://seed.bytedance.com/en/seed1_6. Doubao-Seed-1.6-Vision (Volcengine Ark model id doubao-seed-1.6-vision-250815).

Boqi Chen, Xudong Liu, Jiachuan Peng, Marianne Frey-Marti, Kyle Lam, Bang Zheng, Lin Li, and Jianing Qiu. Medsyn: Benchmarking multi-evidence synthesis in complex clinical cases for multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23631–23655, 2026a.

Haichao Chen, Zhengyun Zhao, Songchi Zhou, Shikai Hu, Jinyuan Wang, Ye Jin, Xianghong Jin, Yih Chung Tham, Xiaofei Wang, Weizhi Ma, et al. Rarearena: a comprehensive benchmark

dataset unveiling the potential of large language models in rare disease diagnosis. *The Lancet Digital Health*, 8(2), 2026b.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, et al. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 7346–7370, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024b.

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37: 94327–94427, 2024c.

Xuanzhong Chen, Xiaohao Mao, Qihan Guo, Lun Wang, Shuyang Zhang, and Ting Chen. Rarebench: can llms serve as rare diseases specialists? In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 4850–4861, 2024d.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 787–795, 2017.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord. Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1574–1583, 2021.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Google DeepMind. Gemini 3.5 flash model card, 5 2026. URL https://deepmind.google/models/model-cards/gemini-3-5-flash/. Model card, published May 19, 2026.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Stephen C Groft, Manuel Posada, and Domenica Taruscio. Progress, challenges and global approaches to rare diseases. *Acta paediatrica*, 110(10):2711–2716, 2021.

Yannian Gu, Zhongzhen Huang, Linjie Mu, Xizhuo Zhang, Shaoting Zhang, and Xiaofan Zhang. Cure: A multimodal benchmark for clinical understanding and retrieval evaluation. *arXiv preprint arXiv:2603.19274*, 2026.

Xuehai He, Zhuo Cai, Wenlan Wei, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Towards visual question answering on pathology images. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 708–718, 2021.

Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada, Ronald M Summers, and Yingying Zhu. Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4156–4165, 2023.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Julius OB Jacobsen, Michael Baudis, Gareth S Baynam, Jacques S Beckmann, Sergi Beltran, Orion J Buske, Tiffany J Callahan, Christopher G Chute, Mélanie Courtot, Daniel Danis, et al. The ga4gh phenopacket schema defines a computable representation of clinical data. *Nature biotechnology*, 40(6):817–820, 2022.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

Congyun Jin, Ming Zhang, Weixiao Ma, Yujiao Li, Yingbo Wang, Yabo Jia, Yuliang Du, Tao Sun, Haowen Wang, Cong Fan, et al. Rjua-meddqa: A multimodal benchmark for medical document question answering and clinical reasoning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5218–5229, 2024a.

Qiao Jin, Fangyuan Chen, Yiliang Zhou, Ziyang Xu, Justin M Cheung, Robert Chen, Ronald M Summers, Justin F Rousseau, Peiyun Ni, Marc J Landsman, et al. Hidden flaws behind expert-level accuracy of multimodal gpt-4 vision in medicine. *NPJ Digital Medicine*, 7(1):190, 2024b.

Junyoung Kim, Kai Wang, Chunhua Weng, and Cong Liu. Assessing the utility of large language models for phenotype-driven gene prioritization in the diagnosis of rare genetic disease. *The American Journal of Human Genetics*, 111(10):2190–2202, 2024.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251, 2018.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564, 2023.

Mistral AI. Introducing Mistral 3, 12 2025. URL https://mistral.ai/news/mistral-3/. Blog post, December 2, 2025. Mistral Large 3 (version 2512); model card: https://docs.mistral.ai/models/model-cards/mistral-large-3-25-12.

Stéphanie Nguengang Wakap, Deborah M Lambert, Annie Olry, Charlotte Rodwell, Charlotte Gueydan, Valérie Lanneau, Daniel Murphy, Yann Le Cam, and Ana Rath. Estimating cumulative point prevalence of rare diseases: analysis of the orphanet database. *European journal of human genetics*, 28(2):165–173, 2020.

Junzhi Ning, Wei Li, Cheng Tang, Jiashi Lin, Chenglong Ma, Chaoyang Zhang, Jiyao Liu, Ying Chen, Shujian Gao, Yuandong Pu, et al. Unimedvl: Unifying medical multimodal understanding and generation through observation-knowledge-analysis. In *Forty-third International Conference on Machine Learning*.

OpenAI. Introducing GPT-4.1 in the API, 4 2025a. URL https://openai.com/index/gpt-4-1/. Blog post, April 14, 2025. Covers GPT-4.1, GPT-4.1 mini, and GPT-4.1 nano.

OpenAI. Update to GPT-5 system card: GPT-5.2, 12 2025b. URL https://openai.com/index/gpt-5-system-card-update-gpt-5-2/. System card update, December 11, 2025. Covers gpt-5.2-instant and gpt-5.2-thinking.

OpenAI. Introducing GPT-5.4, 3 2026a. URL https://openai.com/index/introducing-gpt-5-4/. Blog post, March 5, 2026. Covers GPT-5.4 and GPT-5.4 Pro.

OpenAI. Introducing GPT-5.4 mini and nano, 3 2026b. URL https://openai.com/index/introducing-gpt-5-4-mini-and-nano/. Blog post, March 17, 2026. Covers GPT-5.4 mini and GPT-5.4 nano.

Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Yanjie Fan, Weike Zhao, Zhuoxia Chen, Hongfei Gu, Chuanjin Peng, Ya Zhang, Yanfeng Wang, et al. Quantifying the reasoning abilities of llms on clinical cases. *Nature Communications*, 16(1):9799, 2025.

Ana Rath, Annie Olry, Ferdinand Dhombres, Maja Miličić Brandt, Bruno Urbero, and Segolene Ayme. Representation of rare diseases in health information systems: the orphanet approach to serve a wide range of end users. *Human mutation*, 33(5):803–808, 2012.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, et al. Medgemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Kiril Vasilev, Alexandre Misrahi, Eeshaan Jain, Phil F Cheng, Petros Liakopoulos, Olivier Michielin, Michael Moor, and Charlotte Bunne. Mtbbench: A multimodal sequential clinical decision-making benchmark in oncology. *Advances in Neural Information Processing Systems*, 38, 2026.

Kevin Wu, Eric Wu, Rahul Thapa, Kevin Wei, Angela Zhang, Arvind Suresh, Jacqueline J Tao, Min Woo Sun, Alejandro Lozano, and James Zou. Medcasereasoning: Evaluating and learning diagnostic reasoning from clinical case reports. *arXiv preprint arXiv:2505.11733*, 2025.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Cheng Xu, Shuhao Guan, Derek Greene, M Kechadi, et al. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*, 2024.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

Suhao Yu, Haojin Wang, Juncheng Wu, Luyang Luo, Jingshen Wang, Cihang Xie, Pranav Rajpurkar, Carl Yang, Yang Yang, Kang Wang, et al. Medframeqa: A multi-image medical vqa benchmark for clinical reasoning. *arXiv preprint arXiv:2505.16964*, 2025.

Anas Zafar, Leema Krishna Murali, and Ashish Vashist. Beyond accuracy: Evaluating visual grounding in multimodal medical reasoning. In *Catch, Adapt, and Operate: Monitoring ML Models Under Drift Workshop*, 2026.

Z.ai. GLM-4.6V: Open source multimodal models with native tool use, December 2025. URL https://z.ai/blog/glm-4.6v. Blog post.

Weike Zhao, Chaoyi Wu, Yanjie Fan, Pengcheng Qiu, Xiaoman Zhang, Yuze Sun, Xiao Zhou, Shuju Zhang, Yu Peng, Yanfeng Wang, et al. An agentic system for rare disease diagnosis with traceable reasoning. *Nature*, 651(8106):775–784, 2026.

Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. In *International Conference on Machine Learning*, pp. 80961–80990. PMLR, 2025.

# Appendix

Table A1: **MRareBench Core Composition.**

| Set | Items | Primary evaluation | Role |
|---|---|---|---|
| T1 Diagnosis | 300 | Recall@1/3/5/10, MRR, MR | Limited-Context rare-disease differential diagnosis |
| T2 Evidence Verification | 608 | Open-ended tiered required-evidence recall (HierReqRecall) | Verify visible evidence supporting the established diagnosis |
| T2 paired with T1 by PMCID | 402 | Paired T1–T2 analysis | Released mapping to 289 T1 cases; the pre-removal provenance mapping is 412 T2 rows to 297 T1 cases |
| T2 not paired with T1 | 206 | T2-only analysis | Expand visual-evidence coverage beyond the T1 paired subset |
| **Total filtered task instances** | **908** | Summary row | **T1 + T2; paired rows are subsets, not additional items** |

## A Dataset Dimensions

MRareBench contains at least two images per item (mean 3.52 for T1 and 3.70 for T2), 9.66 acceptable evidence points per T2 item on average.

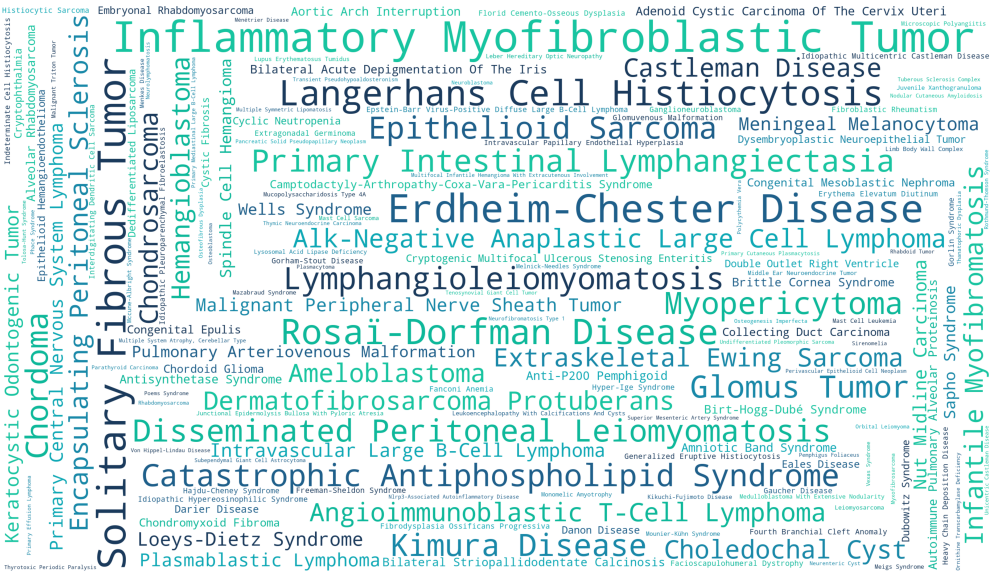



Figure A1: **Visualisation of Rare-Disease Names.**

## B Rare-Disease Hypothesis Density

RareRatio@10 is the share of normalized top-10 candidates labelled as rare diseases. Across models, LC RareRatio@10 ranges from 72% to 14%; removing the image lowers rare-hypothesis density for 36 of the 37 models, with GPT-4o as the exception.

## C Redaction Validity

Redaction validity is assessed separately from memorization: residual finding text and removal of legitimate non-visual context can bias the evidence-channel contrasts.

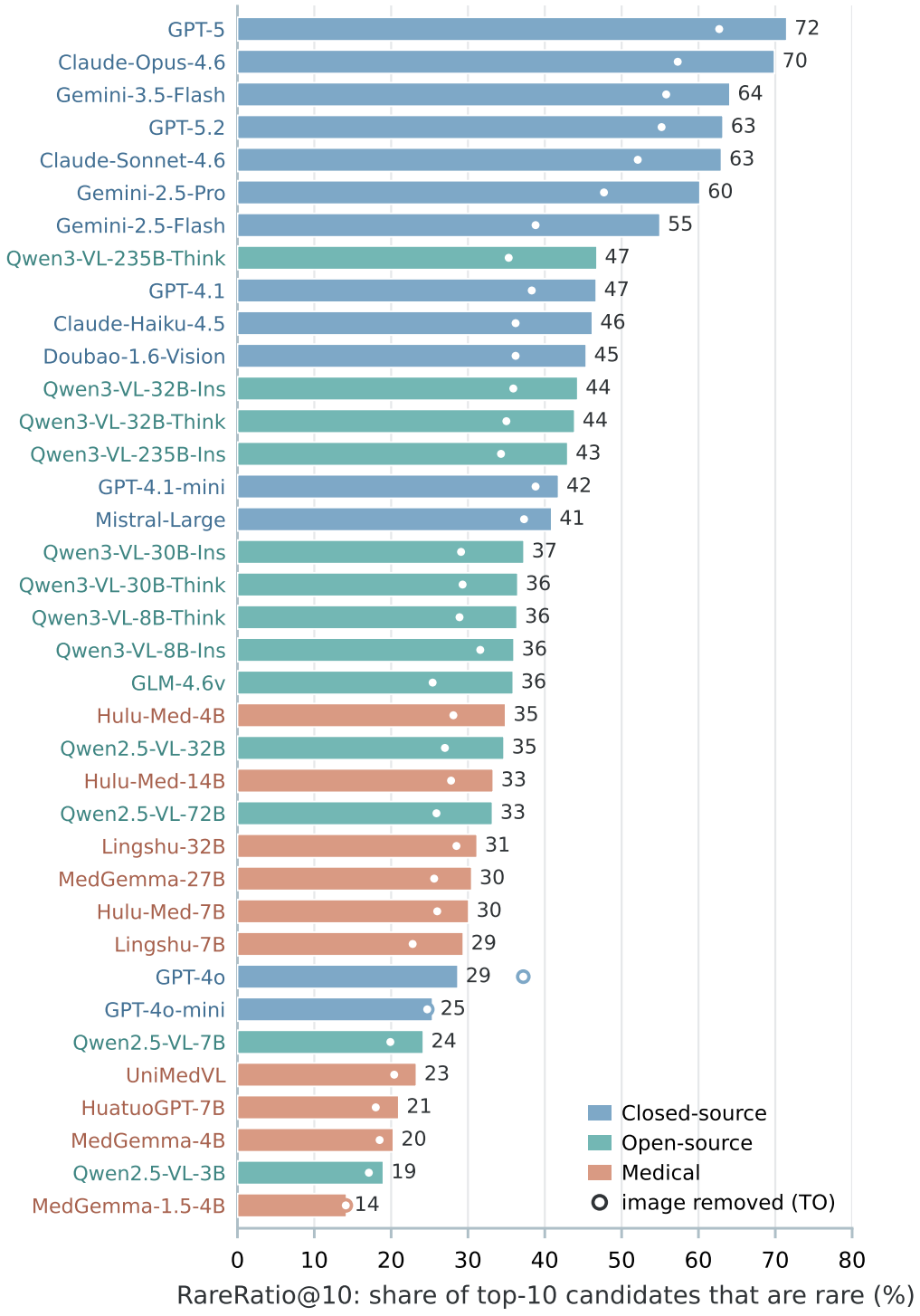



Figure A2: **Per-Model Rare-Disease Hypothesis Density.** Bars show LC RareRatio@10; hollow markers show the image-removed TO value. Colors denote access cohort.

**Construction.** LC is reconstructed from six whitelisted non-visual field types; stripped image-derived findings are retained separately and the image set is unchanged. A stripped finding reappears verbatim in the released LC context in 1 of 300 cases (0.3%).

Table A2: **Construction and Filtering of the Released Benchmark.**

| | T1 Diagnosis | T2 Evidence Verification |
|---|---|---|
| Audited candidate pool | 346 | 665 |
| *Removed by rule* | | |
| Item not medically valid | – | 21 |
| Answer not traceable to the source article | – | 4 |
| Text or image support for the answer too weak | – | 17 |
| Disease not sufficiently rare | – | 10 |
| Text alone directly answers the question | 11 | – |
| Image contributes nothing to the answer | 0 | – |
| Text states the diagnosis outright | 1 | – |
| No imaging-findings text, so FD cannot be built | 13 | – |
| Marked unrepairable at review | 3 | – |
| Images rejected by content-safety screening | 8 | – |
| *Removed at structural revalidation* | | |
| Remaining context insufficient after redaction | 10 | – |
| Invalid figure-slot mapping | – | 4 |
| Not reconstructible in the evaluation set | – | 1 |
| **Released benchmark** | **300** | **608** |

Table A3: **Independent Clinical Re-Adjudication of the Diagnosis Set.**

| **Field and label distribution** ($N$=120) | **Agree** | $\kappa$ |
|---|---|---|
| Item validity: 110 valid, 9 ambiguous, 1 specialist, **0 wrong** | 40/40 | 1.00 |
| Provenance: 118 supported, 1 unclear, 1 insufficient | 40/40 | 1.00 |
| Image–text match: 110 yes, 10 partial, **0 mismatch** | 40/40 | 1.00 |
| Leakage disposition: 52 pass, 68 fixable, **0 remove** | 39/40 | 0.95 |
| Final action: 22 accept, 97 revise, 1 escalate, **0 exclude** | 39/40 | 0.92 |

## D Item Construction and Quality Control

Both tracks reach their released size through a deterministic, label-driven filter whose rules are the source code that produced the delivered files. Table A2 accounts for every removal. A source-grounded, model-output-blind judge assigns each candidate a structured verdict with per-field evidence; the filter then removes an item if and only if a rule fires on those labels. The governing principle is to remove unusable items rather than difficult ones, and to remove rather than rewrite. Items whose narrative carries strong clinical clues are therefore kept: symptoms, laboratory values, and history are what a rare-disease case report legitimately contains, and stripping them would destroy clinical realism. Only two conditions are treated as disqualifying leakage, a text that states the diagnosis outright and a text from which the diagnosis follows directly, and both are removed to zero.

**Independent Re-Adjudication.** Two clinically trained reviewers independently re-adjudicated a stratified sample of the T1 pool after the released set had been fixed, so their review validates the pipeline. Table A3 reports the results.

## E Source-Memorization Audit by Publication Year

Source-level pretraining exposure cannot be excluded. We therefore audit publication-year signatures in deterministic LC outputs from 27 general-purpose models over all 300 cases (2005–2025), defining $\Delta_{\text{mem}} = \text{R@1}(< 2023) - \text{R@1}(\geq 2023)$. Publication year has negligible point-biserial correlation with correctness (mean $\bar{r} = +0.02$; all $|r| \leq 0.09$; none significant at $p < 0.05$). Newer-report gain is not significantly correlated with model strength ($r = +0.34$, $p > 0.05$), and year-subset rankings correlate with the overall ranking at 0.93–0.98 for LC/FD and at least 0.83 for TO. Tables A8 and A9 give the summary and per-model results. These are proxy audits: they

Table A4: **Evaluated Models and Reproduction Identifiers.**

| Model | Identifier |
|---|---|
| *Closed-source* | |
| GPT-4o-mini (Hurst et al., 2024) | `gpt-4o-mini`† |
| GPT-4o (Hurst et al., 2024) | `gpt-4o`† |
| GPT-4.1-mini (OpenAI, 2025a) | `gpt-4.1-mini`† |
| GPT-4.1 (OpenAI, 2025a) | `gpt-4.1`† |
| GPT-5 (Singh et al., 2025) | `gpt-5-2025-08-07`† |
| GPT-5.2 (OpenAI, 2025b) | `gpt-5.2`† |
| Gemini-2.5-Flash (Comanici et al., 2025) | `gemini-2.5-flash`† |
| Gemini-2.5-Pro (Comanici et al., 2025) | `gemini-2.5-pro-nothinking`† |
| Gemini-3.5-Flash (Google DeepMind, 2026) | `gemini-3.5-flash`† |
| Claude-Haiku-4.5 (Anthropic, 2025) | `claude-haiku-4-5-20251001`† |
| Claude-Sonnet-4.6 (Anthropic, 2026b) | `claude-sonnet-4-6`† |
| Claude-Opus-4.6 (Anthropic, 2026a) | `claude-opus-4-6`† |
| Doubao-Seed-1.6-Vision (ByteDance Seed, 2025) | `doubao-seed-1-6-vision-250815`† |
| Mistral-Large-2512 (Mistral AI, 2025) | `mistralai/mistral-large-2512`† |
| *Open-source* | |
| Qwen3-VL-8B-Instruct (Bai et al., 2025a) | `qwen3-vl-8b-instruct`† |
| Qwen3-VL-8B-Thinking (Bai et al., 2025a) | `qwen3-vl-8b-thinking`† |
| Qwen3-VL-30B-Instruct (Bai et al., 2025a) | `qwen3-vl-30b-a3b-instruct`† |
| Qwen3-VL-30B-Thinking (Bai et al., 2025a) | `qwen3-vl-30b-a3b-thinking`† |
| Qwen3-VL-32B-Instruct (Bai et al., 2025a) | `qwen3-vl-32b-instruct`† |
| Qwen3-VL-32B-Thinking (Bai et al., 2025a) | `qwen3-vl-32b-thinking`† |
| Qwen3-VL-235B-Instruct (Bai et al., 2025a) | `qwen3-vl-235b-a22b-instruct`† |
| Qwen3-VL-235B-Thinking (Bai et al., 2025a) | `qwen3-vl-235b-a22b-thinking`† |

| Model | Identifier |
|---|---|
| *Open-source (cont.)* | |
| Qwen2.5-VL-3B (Bai et al., 2025b) | `Qwen/Qwen2.5-VL-3B-Instruct` |
| Qwen2.5-VL-7B (Bai et al., 2025b) | `Qwen/Qwen2.5-VL-7B-Instruct` |
| Qwen2.5-VL-32B (Bai et al., 2025b) | `Qwen/Qwen2.5-VL-32B-Instruct` |
| Qwen2.5-VL-72B (Bai et al., 2025b) | `Qwen/Qwen2.5-VL-72B-Instruct` |
| GLM-4.6v (Z.ai, 2025) | `glm-4.6v`† |
| *Medical* | |
| Lingshu-7B (Xu et al., 2025) | `lingshu-medical-mllm/Lingshu-7B` |
| Lingshu-32B (Xu et al., 2025) | `lingshu-medical-mllm/Lingshu-32B` |
| MedGemma-4B (Sellergren et al., 2025) | `google/medgemma-4b-it` |
| MedGemma-27B (Sellergren et al., 2025) | `google/medgemma-27b-it` |
| MedGemma-1.5-4B (Sellergren et al., 2026) | `google/medgemma-1.5-4b-it` |
| Hulu-Med-4B (Jiang et al., 2025) | `ZJU-AI4H/Hulu-Med-4B` |
| Hulu-Med-7B (Jiang et al., 2025) | `ZJU-AI4H/Hulu-Med-7B` |
| Hulu-Med-14B (Jiang et al., 2025) | `ZJU-AI4H/Hulu-Med-14B` |
| UniMedVL (Ning et al.) | `uni-medical/UniMedVL` |
| HuatuoGPT-Vision-7B (Chen et al., 2024a) | `FreedomIntelligence/HuatuoGPT-Vision-7B` |
| *Auxiliary (construction & judging)* | |
| Question generation | `gpt-5.4`† (OpenAI, 2026a) |
| Judge (primary) | `gpt-5.4-mini`† (OpenAI, 2026b) |
| Judge (secondary) | `deepseek-v4-pro`† (Xu et al., 2026) |

Table A5: **Evidence Reporting by Task Type.**

| Task type | N | Grd ReqR | Grd Hier |
|---|---|---|---|
| Cross-image relation | 204 | 57.0 | 37.5 |
| Modality contribution | 188 | 64.8 | 47.8 |
| Evidence verification | 153 | 59.7 | 40.3 |
| Diagnostic grounding | 13 | 59.0 | 38.0 |
| Others (long tail) | 50 | 67.1 | 49.5 |

Table A6: **Diagnosis Performance by Imaging Modality.**

| Cohort | Modality | N | FD | LC | TO | $\Delta_{\mathbf{vis}}$ | $\Delta_{\mathbf{sub}}$ |
|---|---|---|---|---|---|---|---|
| General | Pathology (micro) | 103 | 23.1 | 20.3 | 13.2 | 7.1 | 2.8 |
| | CT | 77 | 44.1 | 25.1 | 11.9 | 13.1 | 19.1 |
| | MRI | 61 | 33.1 | 20.9 | 11.8 | 9.1 | 12.2 |
| | X-ray | 13 | 32.8 | 21.1 | 9.7 | 11.4 | 11.7 |
| | Ultrasound | 12 | 51.9 | 27.8 | 26.2 | 1.5 | 24.1 |
| | Other (pooled) | 34 | 35.0 | 19.9 | 15.1 | 4.8 | 15.0 |
| Medical | Pathology (micro) | 103 | 12.0 | 11.4 | 6.5 | 4.9 | 0.7 |
| | CT | 77 | 20.6 | 8.7 | 3.5 | 5.2 | 11.9 |
| | MRI | 61 | 21.1 | 10.3 | 4.4 | 5.9 | 10.8 |
| | X-ray | 13 | 18.5 | 8.5 | 3.1 | 5.4 | 10.0 |
| | Ultrasound | 12 | 30.8 | 15.8 | 8.3 | 7.5 | 15.0 |
| | Other (pooled) | 34 | 17.4 | 6.5 | 4.4 | 2.1 | 10.9 |

Table A7: **Additional Evaluation Dimensions.**

| Track | Dimension | What it examines |
|---|---|---|
| T1 diagnosis | Parseability and first hit | Usable output; first ORPHA-normalized hit rank. |
| T1 diagnosis | Judge Top-1 check | Synonym-sensitive rank-1 correctness (sanity check on R@1). |
| T1 diagnosis | Rare-candidate labels | Rare-disease density of the top-10 differential. |
| T2 verification | Required evidence coverage | Coverage of each required visual evidence point. |
| T2 verification | Tiered required recall | Coverage gated by evidence-hierarchy prerequisites (the headline T2 score). |

do not rule out source-level exposure and cannot inspect closed training data, and the small residual year-to-year variation is confounded with the disease-difficulty distribution, which we do not

control. Limited-Context T1 performance shows no measurable dependence on publication year, no recency-memorization signature, and a model ranking that is stable across publication years.

Table A8: **Expected and Observed Signatures of Source Memorization.**

| **Prediction if memorization inflates scores** | **Observed** ($N$=27) | **Holds?** |
|---|---|---|
| Per-case accuracy tracks publication year (continuous, all 300 cases) | $\bar{r}$=+0.02; all $\lvert r\rvert\leq$0.09; 0/27 sig. ($p<0.05$) | No |
| Newer-report gain grows with model recency/strength | corr. with overall R@1 =+0.34 (n.s.); top models flat | No |
| Publication year reshuffles the model ranking | Spearman $\rho$=0.93–0.98 (LC/FD), $\geq$ 0.83 (TO) | No |

## F T1 Judge Diagnostic Definitions

The T1 judge Top-1 check is an audit field rather than a headline benchmark score. For each case, the released evaluator gives the judge the diagnosis, accepted synonyms or abbreviations, and the model's parsed ranked differential, and the judge acts as a strict attending-physician evaluator that answers YES only when the criterion is completely and explicitly satisfied. Top-1 is a binary synonym-sensitive correctness check,

$$\mathrm{Top1}_i = \mathbb{1}\{\hat{y}_{i,1} \in G_i\}, \tag{10}$$

scored 1 only when the first ranked diagnosis is the diagnosis or an accepted synonym. Tables A16–A18 report its per-view mean as a percentage,

$$\mathrm{Top1} = \frac{100}{N}\sum_i \mathrm{Top1}_i, \tag{11}$$

paired across the two judges (GPT-5.4-mini/DeepSeek-V4-Pro); a YES/NO judge line converts to 1/0, and a missing or non-matching line defaults to 0.

## G Metric Formula Details

The evaluator stores aggregate scores and per-row traces. Table A7 indexes the additional dimensions whose formulas follow. Recall and correspondence formulas use proportions in $[0, 1]$; tables display percentages and score differences in percentage points. Thus $\tau = 0.5$ corresponds to a score of $50\%$.

**T1 Ranking and Candidate-Label Metrics.** The prompt requests ten candidates. Let $\hat{Y}_i = (\hat{y}_{i,1}, \ldots, \hat{y}_{i,K_i})$ be the parsed ranked differential used for evaluation, where $K_i \leq 10$, and let $G_i$ be the set containing the ORPHA-normalized diagnosis and accepted aliases. The first matched rank is

$$r_i = \min\{j \in \{1, \ldots, K_i\} : \mathrm{match}(\hat{y}_{i,j}, y_i) = 1\}, \tag{12}$$

with $r_i = \infty$ when no candidate matches, including an empty parsed list. The evaluator stores parse failures and hit positions as

$$\mathrm{ParseFail} = \frac{1}{N}\sum_i \mathbb{1}\{|\hat{Y}_i| = 0\}, \tag{13}$$

$$\mathrm{HitRate} = \frac{1}{N}\sum_i \mathbb{1}\{r_i < \infty\}. \tag{14}$$

Table A9: **Per-Model Publication-Year Audit.**

| Model | Overall | $<$2023 | $\geq$2023 | $\Delta_{\mathrm{mem}}$ | $r$(yr) |
|---|---|---|---|---|---|
| **Closed-source** | | | | | |
| GPT-4o-mini | 8.7 | 7.7 | 11.0 | $-3.3$ | $+0.06$ |
| GPT-4o | 19.7 | 17.7 | 24.2 | $-6.5$ | $+0.08$ |
| GPT-4.1-mini | 25.7 | 23.9 | 29.7 | $-5.7$ | $+0.01$ |
| GPT-4.1 | 29.7 | 30.6 | 27.5 | $+3.1$ | $+0.02$ |
| GPT-5 | 48.7 | 49.3 | 47.3 | $+2.0$ | $+0.01$ |
| GPT-5.2 | 31.7 | 29.2 | 37.4 | $-8.2$ | $+0.04$ |
| Gemini-2.5-Flash | 29.0 | 29.7 | 27.5 | $+2.2$ | $-0.01$ |
| Gemini-2.5-Pro-NoThink | 40.3 | 41.6 | 37.4 | $+4.3$ | $-0.04$ |
| Gemini-3.5-Flash | 54.7 | 54.1 | 56.0 | $-2.0$ | $+0.07$ |
| Claude-Haiku-4.5 | 16.0 | 15.3 | 17.6 | $-2.3$ | $-0.03$ |
| Claude-Sonnet-4.6 | 31.3 | 29.2 | 36.3 | $-7.1$ | $+0.04$ |
| Claude-Opus-4.6 | 36.3 | 37.3 | 34.1 | $+3.3$ | $-0.02$ |
| Doubao-Seed-1.6-Vision | 24.7 | 22.5 | 29.7 | $-7.2$ | $+0.07$ |
| Mistral-Large-2512 | 14.3 | 14.4 | 14.3 | $+0.1$ | $+0.02$ |
| **Open-source** | | | | | |
| Qwen3-VL-8B-Instruct | 6.7 | 4.8 | 11.0 | $-6.2$ | $+0.09$ |
| Qwen3-VL-8B-Thinking | 13.0 | 12.0 | 15.4 | $-3.4$ | $+0.02$ |
| Qwen3-VL-30B-Instruct | 16.3 | 14.8 | 19.8 | $-4.9$ | $-0.02$ |
| Qwen3-VL-30B-Thinking | 17.0 | 14.8 | 22.0 | $-7.1$ | $-0.01$ |
| Qwen3-VL-32B-Instruct | 15.0 | 15.3 | 14.3 | $+1.0$ | $-0.01$ |
| Qwen3-VL-32B-Thinking | 19.3 | 18.2 | 22.0 | $-3.8$ | $+0.01$ |
| Qwen3-VL-235B-Instruct | 20.0 | 17.2 | 26.4 | $-9.1$ | $+0.04$ |
| Qwen3-VL-235B-Thinking | 25.0 | 22.5 | 30.8 | $-8.3$ | $+0.04$ |
| Qwen2.5-VL-3B | 3.0 | 2.4 | 4.4 | $-2.0$ | $+0.03$ |
| Qwen2.5-VL-7B | 8.7 | 8.1 | 9.9 | $-1.8$ | $-0.01$ |
| Qwen2.5-VL-32B | 7.0 | 5.3 | 11.0 | $-5.7$ | $+0.04$ |
| Qwen2.5-VL-72B | 15.3 | 12.4 | 22.0 | $-9.5$ | $+0.06$ |
| GLM-4.6v | 15.3 | 13.9 | 18.7 | $-4.8$ | $+0.06$ |
| **Medical** | | | | | |
| Lingshu-7B | 10.0 | 9.6 | 11.0 | $-1.4$ | $+0.01$ |
| Lingshu-32B | 12.0 | 11.0 | 14.3 | $-3.3$ | $+0.06$ |
| MedGemma-4B | 5.7 | 5.7 | 5.5 | $+0.2$ | $-0.03$ |
| MedGemma-27B | 8.0 | 5.7 | 13.2 | $-7.4$ | $+0.08$ |
| Hulu-Med-4B | 11.7 | 13.4 | 7.7 | $+5.7$ | $-0.06$ |
| Hulu-Med-7B | 16.3 | 17.7 | 13.2 | $+4.5$ | $-0.09$ |
| Hulu-Med-14B | 15.7 | 15.3 | 16.5 | $-1.2$ | $-0.03$ |
| UniMedVL | 7.3 | 5.7 | 11.0 | $-5.2$ | $+0.02$ |
| MedGemma-1.5-4B | 6.3 | 6.2 | 6.6 | $-0.4$ | $+0.01$ |
| HuatuoGPT-Vision-7B | 6.7 | 7.2 | 5.5 | $+1.7$ | $+0.01$ |

Let $\mathcal{H} = \{i : r_i < \infty\}$ denote cases with a matched candidate. Standard ranked-diagnosis summaries are

$$\mathrm{R@}k = \frac{1}{N}\sum_i \mathbb{1}\{r_i \leq k\}, \tag{15}$$

$$\mathrm{MRR} = \frac{1}{N}\sum_{i\in\mathcal{H}} \frac{1}{r_i}, \tag{16}$$

$$\mathrm{MR} = \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} r_i, \qquad |\mathcal{H}| > 0. \tag{17}$$

MR is conditional on a hit within the evaluated differential; it is undefined and reported as N/A if $\mathcal{H}$ is empty. MRR assigns zero contribution to cases without a hit.

**Candidate-Label Counts.** Each generated candidate receives a label $\ell(c) \in \{\text{rare}, \text{nonrare}, \text{generic}, \text{unknown}\}$. In addition to RareRatio@10, the evaluator records label-specific counts for $a \in \{\text{rare}, \text{unknown}, \text{generic}\}$:

$$\text{Count}_a@10 = \frac{1}{N} \sum_i C_{ia}, \tag{18}$$

$$C_{ia} = \sum_{j=1}^{K_i} \mathbb{1}\{\ell(\hat{y}_{i,j}) = a\}. \tag{19}$$

**T2 Required-Evidence Recall.** For T2 open-ended evidence verification, each case has required evidence points $P_i$, and the judge returns required hits $b_{ip}$. The headline flat recall, reported alongside the tiered score in Table A14, is

$$\text{ReqRecall} = \frac{1}{N} \sum_i \frac{1}{|P_i|} \sum_{p \in P_i} b_{ip}. \tag{20}$$

## H Evidence Hierarchy for T2

T2 scores rubric-aligned evidence reporting. Level 1 covers visible observations, Level 2 cross-image integration, and Level 3 diagnostic attribution; later credit is gated by earlier visual evidence. Table A10 gives the weights, Table A13 the matched Grd/NoDx recalls, and Table A14 the Grd/Rec image contrast.

**Gate Computation.** Partition the required points of item $i$ into $P_i^{(1)}$, $P_i^{(2)}$, and $P_i^{(3)}$ according to the evidence buckets in Table A10. Define $a_i = \prod_{p \in P_i^{(1)}} b_{ip}$ and $q_i = \prod_{p \in P_i^{(2)}} b_{ip}$. The binary gates are

$$\gamma_{ip} = \begin{cases} 1, & p \in P_i^{(1)}, \\ a_i, & p \in P_i^{(2)}, \\ a_i q_i, & p \in P_i^{(3)}. \end{cases} \tag{21}$$

An empty product equals one: a missing prerequisite level does not close the gate. A failed gate removes credit from the numerator while the required point remains in the weighted denominator. Only present required points enter that denominator; optional points do not contribute.

**Per-Level Scores and Matched Means.** Level-1 recall is the unweighted mean of hits in $P_i^{(1)}$. More generally, raw level-$k$ recall is $|P_i^{(k)}|^{-1} \sum_{p \in P_i^{(k)}} b_{ip}$ when that level is present; an absent level has no per-level score and is omitted from its mean. Full-tier recall uses the weighted, gated formula in the main text. For a fixed model, write $s_j^{c,\ell}$ for a T2 item score and $\mathcal{V}$ for the items with valid outputs in both Grd and NoDx. Then

$$S_c^{(\ell)} = \frac{1}{|\mathcal{V}|} \sum_{j \in \mathcal{V}} s_j^{c,\ell}, \qquad \Delta_{\text{dx}}^{(\ell)} = \frac{1}{|\mathcal{V}|} \sum_{j \in \mathcal{V}} \big(s_j^{\text{Grd},\ell} - s_j^{\text{NoDx},\ell}\big). \tag{22}$$

A valid output is non-empty and contains no `Failed to obtain answer` marker (case-insensitive). The primary Grd–NoDx comparison has 22,317 valid model–item pairs, with 579–608 items per model and defined L1 and full-tier scores throughout. Grd–Rec uses its own common-valid intersection, comprising 22,316 pairs. The reported 24.7-point L1 gap is the unweighted mean of the 37 model-level gaps.

## I Robustness of T2 Scoring to Judge Choice

The primary judge is `gpt-5.4-mini`; `deepseek-v4-pro` (Xu et al., 2026) re-scores the same frozen predictions without model re-inference. Table A11 reports rank agreement. Item-level agreement is weaker: the item-weighted Cohen's $\kappa$ is $0.35$ under Grd and $0.32$ under Rec. The secondary

Table A10: **Evidence Hierarchy and Scoring Weights.**

| Evidence bucket | Level | Weight |
|---|---|---|
| Modality / imaging type | 1 | 0.50 |
| Macroscopic pattern or region | 1 | 1.00 |
| Microstructural pattern | 1 | 1.20 |
| Other visible finding | 1 | 1.00 |
| Cross-image integration | 2 | 2.00 |
| Diagnostic attribution | 3 | 0.75 |

judge also scores $15.4$ points below the primary under Grd. Judge choice therefore preserves the model ranking while shifting the absolute scale. Table A15 gives the full sensitivity analysis. There, $N_{11}$ and $N_{01}$ count diagnose-and-verify and verify-only pairs, **Pooled** is $N_{11}/(N_{11}+N_{01})$, and **Equal-M** averages the per-model rate instead.

**Case Aggregation and Correspondence.** The released mapping links 402 T2 items to 289 T1 cases. For model $f$ and case $i$, let $\mathcal{J}_{f,i}$ contain linked T2 items in the Grd–NoDx common-valid intersection with a defined score under the selected condition, tier, and judge. Fixing these choices, let $s_{f,j}$ be the item score. For non-empty $\mathcal{J}_{f,i}$,

$$S_{f,i} = \frac{1}{|\mathcal{J}_{f,i}|} \sum_{j \in \mathcal{J}_{f,i}} s_{f,j}, \qquad z_{f,i} = \mathbb{1}[S_{f,i} \geq \tau]. \tag{23}$$

Let $h_{f,i}$ indicate a correct T1 LC Top-1 diagnosis and $\Omega$ contain the resulting valid model–case pairs. The pooled empirical correspondence is

$$\widehat{\mathrm{Corr}}_{\mathrm{pooled}} = \frac{\sum_{(f,i)\in\Omega} h_{f,i} z_{f,i}}{\sum_{(f,i)\in\Omega} z_{f,i}}. \tag{24}$$

The headline uses Grd, full-tier scores, the primary judge, and $\tau = 0.5$: among 10,629 valid pairs, 3,911 are T2-positive and 1,028 are also diagnosis-positive, giving $26.3\%$. Equal-M instead averages the defined per-model conditional rates, giving $19.3\%$ for this setting. A rate with no T2-positive denominator is undefined and omitted from Equal-M. Cases without valid linked scores are excluded. Secondary-judge NoDx results cover 35 models. Confidence intervals use 10,000 bootstrap replicates over model and T1-case clusters. Correspondence describes co-occurring task outcomes; it does not establish a shared internal evidence representation.

## J Robustness of Judge Choice

The main table (Table 2) reports the descriptive full-tier gap $\Delta_{\mathrm{dx}}^{\mathrm{full}}$ under the primary judge. Table A13 reports the primary Level-1 contrast on the same Grd–NoDx common-valid rows. The matched mean $\Delta_{\mathrm{dx}}^{\mathrm{L1}}$ is $24.7$ points, and all 37 model-level contrasts are positive.

## K Prompt and Rubric Templates

These are the prompts used in the evaluation, reproduced verbatim. Braces mark the fields filled in for each item. Where the model is shown images, they are attached ahead of the text below.

### Prompts sent to the evaluated models

**T1 LC Ranked-Differential Prompt.** The baseline diagnosis view shows the images but removes the sentences that narrate what the images contain. The model receives the case context, then the question, then the output instruction.

**T1 LC assembled text message**

```
{context}

Question: {question}

You are an expert physician. Based on the clinical presentation and the
medical images, list the 10 most likely diagnoses, ranked from MOST likely
(1) to LEAST likely (10). Output ONLY a numbered list, one standard clinical
disease name per line (e.g. "1. Disease name"). Do not include explanations,
reasoning, or extra text.
```

The findings-disclosed view (FD) puts the imaging-findings text back into the context and changes nothing else. The text-only view (TO) keeps the LC context but attaches no image, and drops the reference to images from the instruction. The three views share the clinical context; their differences are the two evidence channels and the image reference in the TO instruction.

**T1 TO assembled text message (no image attached)**

```
{context}

Question: {question}

You are an expert physician. Based on the clinical presentation, list the
10 most likely diagnoses, ranked from MOST likely (1) to LEAST likely (10).
Output ONLY a numbered list, one standard clinical disease name per line
(e.g. "1. Disease name"). Do not include explanations, reasoning, or extra
text.
```

**How the Three T2 Conditions Differ.** All three conditions share the same question and the same instruction. Grd opens the context with the sentence `Established diagnosis: {disease}.` NoDx removes that one sentence and changes nothing else; the images stay attached. We checked all 608 items: the sentence is present under Grd, and no disease name survives anywhere in the NoDx input. The instruction still tells the model that a diagnosis is established, so NoDx withholds the disease name. Rec keeps the Grd text but attaches no images.

### Prompts sent to the judge and labeling model

**Rare-Candidate Ontology Prompt.** Each distinct candidate name from the ranked differentials is classified on its own, without the case context. These labels give RareRatio@10.

**Rare-Candidate Ontology Prompt (RareRatio@10 labelling)**

```
You are a strict rare-disease ontology assistant. Classify ONE diagnosis
candidate from a ranked differential diagnosis list.

Candidate diagnosis: {candidate}

Return whether this candidate is a specific rare disease entity. Do not judge
whether it is correct for a patient. Do not use the case context. Classify
broad or non-specific terms such as "infection", "tumor", "autoimmune
disease", "genetic disorder", or "syndrome" as generic unless they name a
specific disease entity.

Reply with valid JSON only, exactly using this schema:
{
  "candidate": "...",
  "label": "rare | nonrare | generic | unknown",
  "is_rare_disease": true,
  "is_specific_disease": true,
  "orpha_code_if_known": "ORPHA:xxxx or null"
}
Use label "unknown" if you cannot judge reliably.
```

**T2 Required-Evidence Rubric Prompt.** A per-item binary rubric over that row's own acceptable-evidence points: coverage of each required and optional point, plus a contradiction check on each required point. Defaulting to NO makes required-evidence recall stricter than free-form plausibility.

**T2 Required-Evidence Rubric Prompt (binary, point-wise)**

```
You are an extremely strict attending-physician evaluator of a model's
description of the VISIBLE evidence in medical images. For every YES/NO
judgment default to NO; answer YES only when the criterion is clearly met
(not vaguely, not by merely repeating the question or the diagnosis name).

Established diagnosis (given to the model; NOT what you are scoring): {diagnosis}

Reference rationale (ground truth of what the images show):
{rationale}

MODEL ANSWER TO EVALUATE:
{answer}

REQUIRED evidence points ({n_req}):
{required points, numbered}

OPTIONAL evidence points ({n_opt}):
{optional points, numbered}

Output ONLY YES/NO, one per line, in EXACTLY this order:
- Lines 1..{n_req} = COVERAGE of each REQUIRED point (in order): does the
answer explicitly and correctly state this point? (YES/NO)
- Next {n_opt} line(s) = COVERAGE of each OPTIONAL point (in order): same
question. (YES/NO)
- Final {n_req} line(s) = CONTRADICTION of each REQUIRED point (in order):
does the answer state something that DIRECTLY CONTRADICTS this point (claims
it is absent, or claims the opposite finding)? Answer YES only for a clear
contradiction, otherwise NO.

Reply with EXACTLY {2*n_req + n_opt} lines, each only YES or NO, no numbering,
no other text.
```

## L Supplementary Evidence-Channel Figures

Figure A5 plots image contribution against diagnosis performance in panel (a), and T1 diagnosis against supplied-diagnosis T2 evidence reporting by access cohort in panel (b).

Figure A3 resolves the correspondence tuple to the model–case level and shows the two off-diagonal outcomes, T2-positive only and diagnosis-positive only. Table A15 gives the integer counts, thresholds, weighting rules, and clustered intervals. Figure A6 plots aggregate T1 and T2 scores per model.

## M Per-Model Correspondence Rate and Hit-Position

Table A19 places two distinct summaries side by side: T1 LC hit-position categories over the released diagnosis set, and the prespecified operational correspondence rate on the effective source-linked subset. The hit-position columns split non-Top-1 T1 cases by output position: **Retr.** is gold outside the top 10 and **Commit.** is gold at ranks 2–10, with **miss**$=100-$R@10. The correspondence columns are generated from the frozen common-valid outputs with no additional model inference. Per-model effective $N$ ranges from 272 to 289, and the pooled row reproduces the 10,629-pair, 26.3% headline result.

Table A11: **Agreement between the Two Evidence Judges.**

| Condition | Cohort | #M | $r$ | $\rho$ |
|---|---|---|---|---|
| Grd | Closed | 14 | 0.97 | 0.97 |
| Grd | Open | 13 | 0.97 | 0.95 |
| Grd | Medical | 10 | 0.90 | 0.95 |
| Rec | Closed | 14 | 0.93 | 0.93 |
| Rec | Open | 13 | 0.98 | 0.91 |
| Rec | Medical | 10 | 0.71 | 0.76 |
| Grd | All | 37 | 0.97 | 0.98 |
| Rec | All | 37 | 0.93 | 0.92 |

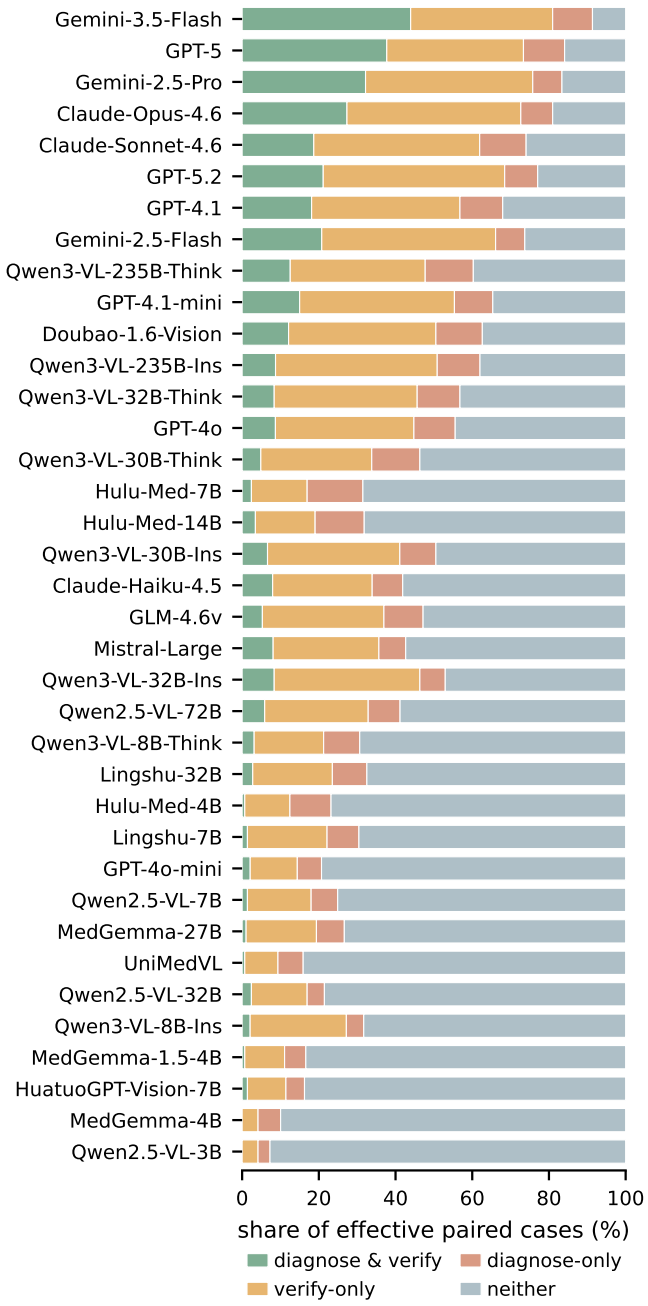



Figure A3: **Per-Model Breakdown of the Cross-Task Outcome** 2×2, sorted by T1 LC accuracy. Figure 5(b) pools them.

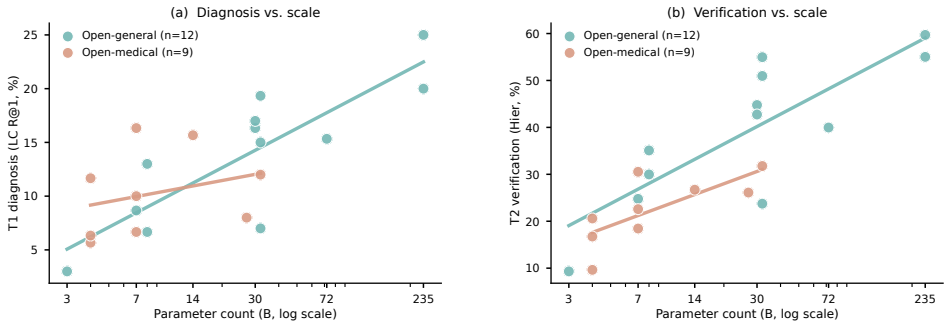



Figure A4: **Comparison of Open-General and Open-Medical Models at Matched Parameter Count** (log scale), with per-cohort log-linear fits. Both panels cover the same 21 models with a published parameter count. (a) T1 LC R@1. (b) T2 HierReqRecall.

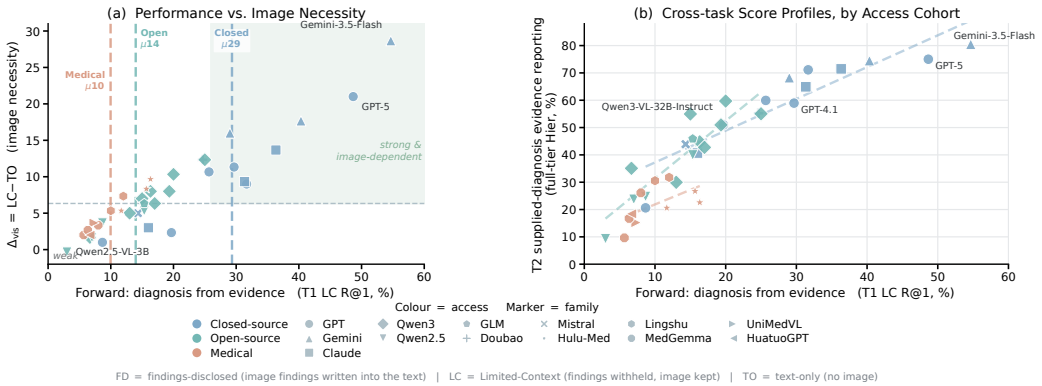



Figure A5: **Two Per-Model Scatter Views Plotted against Diagnosis Accuracy.** Both panels share the T1 LC Recall@1 axis. (a) $\Delta_{\text{vis}}$=LC−TO. (b) Matched Grd full-tier HierReqRecall by access cohort. No equality diagonal is drawn: T1 and T2 are different tasks.

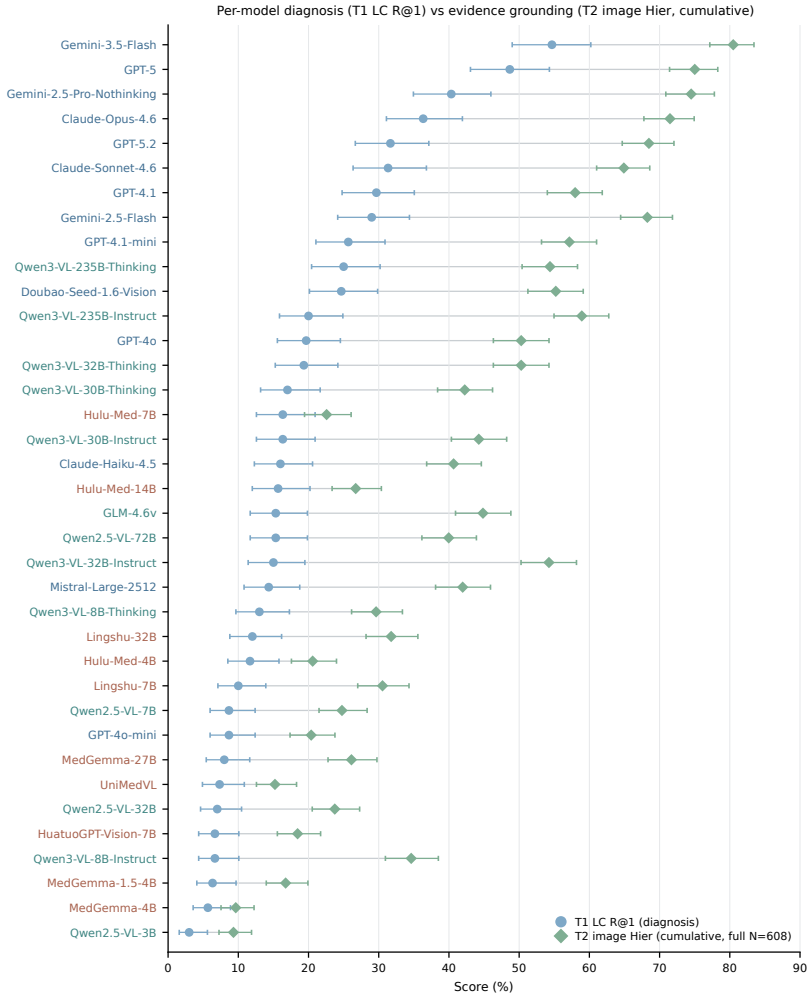



Figure A6: **Per-Model Diagnosis and Supplied-Diagnosis Evidence Reporting.**

Table A12: **Rare-Disease Hypothesis Density.**

| Model | FD RareRatio | LC RareRatio | TO RareRatio | $\Delta_{sub}$ | $\Delta_{vis}$ |
|---|---|---|---|---|---|
| **Closed-source** | | | | | |
| GPT-4o-mini | 33.4 | 25.4 | 24.7 | 8.0 | 0.7 |
| GPT-4o | 49.0 | 28.7 | 37.2 | 20.3 | -8.5 |
| GPT-4.1-mini | 46.6 | 41.8 | 38.8 | 4.8 | 3.0 |
| GPT-4.1 | 49.6 | 46.7 | 38.3 | 2.9 | 8.4 |
| GPT-5 | 72.1 | 71.5 | 62.7 | 0.6 | 8.8 |
| GPT-5.2 | 67.0 | 63.2 | 55.2 | 3.8 | 8.0 |
| Gemini-2.5-Flash | 55.1 | 55.0 | 38.8 | 0.1 | 16.2 |
| Gemini-2.5-Pro-NoThink | 60.9 | 60.2 | 47.7 | 0.7 | 12.5 |
| Gemini-3.5-Flash | 63.8 | 64.1 | 55.8 | -0.3 | 8.3 |
| Claude-Haiku-4.5 | 51.4 | 46.2 | 36.2 | 5.1 | 10.0 |
| Claude-Sonnet-4.6 | 64.5 | 63.0 | 52.1 | 1.6 | 10.9 |
| Claude-Opus-4.6 | 70.1 | 69.9 | 57.3 | 0.2 | 12.5 |
| Doubao-Seed-1.6-Vision | 47.0 | 45.4 | 36.2 | 1.6 | 9.2 |
| Mistral-Large-2512 | 47.1 | 40.9 | 37.3 | 6.3 | 3.6 |
| **Open-source** | | | | | |
| Qwen3-VL-8B-Instruct | 40.0 | 36.0 | 31.6 | 3.9 | 4.4 |
| Qwen3-VL-8B-Thinking | 38.8 | 36.4 | 28.9 | 2.4 | 7.5 |
| Qwen3-VL-30B-Instruct | 40.6 | 37.3 | 29.1 | 3.3 | 8.2 |
| Qwen3-VL-30B-Thinking | 41.1 | 36.5 | 29.3 | 4.6 | 7.2 |
| Qwen3-VL-32B-Instruct | 47.5 | 44.3 | 35.9 | 3.3 | 8.4 |
| Qwen3-VL-32B-Thinking | 46.5 | 43.9 | 35.0 | 2.7 | 8.9 |
| Qwen3-VL-235B-Instruct | 45.4 | 43.0 | 34.3 | 2.4 | 8.7 |
| Qwen3-VL-235B-Thinking | 49.8 | 46.8 | 35.3 | 3.0 | 11.5 |
| Qwen2.5-VL-3B | 25.2 | 19.0 | 17.1 | 6.1 | 1.9 |
| Qwen2.5-VL-7B | 29.7 | 24.2 | 19.9 | 5.5 | 4.3 |
| Qwen2.5-VL-32B | 39.3 | 34.7 | 27.0 | 4.6 | 7.7 |
| Qwen2.5-VL-72B | 37.9 | 33.2 | 25.9 | 4.7 | 7.2 |
| GLM-4.6v | 41.2 | 35.9 | 25.4 | 5.2 | 10.5 |
| **Medical** | | | | | |
| Lingshu-7B | 30.6 | 29.4 | 22.8 | 1.2 | 6.6 |
| Lingshu-32B | 34.7 | 31.2 | 28.5 | 3.5 | 2.7 |
| MedGemma-4B | 23.5 | 20.3 | 18.5 | 3.2 | 1.8 |
| MedGemma-27B | 33.0 | 30.5 | 25.6 | 2.5 | 4.8 |
| Hulu-Med-4B | 36.5 | 34.9 | 28.1 | 1.6 | 6.8 |
| Hulu-Med-7B | 32.3 | 30.1 | 26.0 | 2.2 | 4.0 |
| Hulu-Med-14B | 33.6 | 33.3 | 27.8 | 0.3 | 5.6 |
| UniMedVL | 30.2 | 23.3 | 20.4 | 6.9 | 2.9 |
| MedGemma-1.5-4B | 17.2 | 14.2 | 14.1 | 2.9 | 0.1 |
| HuatuoGPT-Vision-7B | 26.8 | 21.0 | 18.0 | 5.8 | 3.0 |

## N Ranked Differential-Diagnosis Evaluation

Each model emits a top-10 ranked differential over the 300-item diagnosis set under each input condition (FD/LC/TO). The primary metric is deterministic. Tables A16–A18 report, per condition, Recall@1/@10, MRR, and mean first-hit rank (MR), with the LLM-judge **Top-1** check beside R@1 as a synonym-sensitive validity check on the same rank-1 event; R@1 and Top-1 track each other closely and both judges reproduce run-to-run. All three tables cover the 37 models of Table 2, and rows follow its source-type, family, and scale/mode order.

Table A13: **Evidence Recall with and without Diagnosis Disclosure.** L1–L3 are raw per-level recalls; Hier is the weighted, gated full-tier score. Values are percentages; the L1 gap is in percentage points. $N$ counts common-valid items.

| | | **Grd** (diagnosis + image) | | | | **NoDx** (name withheld + image) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **N** | **L1** | **L2** | **L3** | **Hier** | **L1** | **L2** | **L3** | **Hier** | $\Delta^{\mathbf{L1}}_{\mathbf{dx}}$ |
| **Closed-source** | | | | | | | | | | |
| GPT-4o-mini | 601 | 27.0 | 65.6 | 80.2 | 20.6 | 6.2 | 23.2 | 28.6 | 4.1 | 20.8 |
| GPT-4o | 600 | 59.4 | 83.4 | 88.0 | 51.0 | 24.9 | 38.7 | 32.2 | 19.1 | 34.5 |
| GPT-4.1-mini | 579 | 68.6 | 89.4 | 93.4 | 59.9 | 36.2 | 55.1 | 46.1 | 27.2 | 32.4 |
| GPT-4.1 | 598 | 66.7 | 85.5 | 89.1 | 58.9 | 35.7 | 47.2 | 38.8 | 29.1 | 31.1 |
| GPT-5 | 608 | 82.1 | 92.0 | 95.7 | 75.0 | 50.7 | 59.5 | 49.5 | 42.4 | 31.3 |
| GPT-5.2 | 584 | 78.9 | 92.9 | 95.5 | 71.1 | 47.7 | 62.7 | 56.7 | 39.1 | 31.1 |
| Gemini-2.5-Flash | 608 | 76.8 | 90.7 | 92.1 | 68.3 | 51.8 | 61.7 | 44.7 | 41.9 | 25.0 |
| Gemini-2.5-Pro-Nothinking | 608 | 81.7 | 93.3 | 93.8 | 74.5 | 52.1 | 64.9 | 46.9 | 43.1 | 29.6 |
| Gemini-3.5-Flash | 608 | 86.6 | 95.3 | 91.6 | 80.5 | 71.7 | 81.5 | 67.4 | 63.6 | 14.9 |
| Claude-Haiku-4.5 | 608 | 49.8 | 72.2 | 78.5 | 40.7 | 15.9 | 22.1 | 17.9 | 11.3 | 33.8 |
| Claude-Sonnet-4.6 | 608 | 72.8 | 87.5 | 89.6 | 64.9 | 38.1 | 45.5 | 32.1 | 30.1 | 34.7 |
| Claude-Opus-4.6 | 608 | 78.9 | 92.0 | 90.6 | 71.5 | 47.1 | 55.3 | 36.0 | 38.5 | 31.8 |
| Doubao-Seed-1.6-Vision | 608 | 65.0 | 86.5 | 89.3 | 55.2 | 35.9 | 53.0 | 39.1 | 27.6 | 29.1 |
| Mistral-Large-2512 | 581 | 52.4 | 77.5 | 85.7 | 43.9 | 13.6 | 22.5 | 18.9 | 9.7 | 38.8 |
| **Open-source** | | | | | | | | | | |
| Qwen3-VL-8B-Instruct | 600 | 45.0 | 74.0 | 84.0 | 35.1 | 22.1 | 36.5 | 33.2 | 15.3 | 22.9 |
| Qwen3-VL-8B-Thinking | 601 | 38.3 | 67.9 | 73.5 | 30.0 | 15.0 | 25.8 | 16.5 | 10.6 | 23.2 |
| Qwen3-VL-30B-Instruct | 601 | 54.5 | 80.6 | 85.0 | 44.8 | 26.5 | 39.4 | 31.4 | 19.9 | 28.1 |
| Qwen3-VL-30B-Thinking | 601 | 52.6 | 76.6 | 83.7 | 42.7 | 21.6 | 36.0 | 25.1 | 15.9 | 31.0 |
| Qwen3-VL-32B-Instruct | 600 | 65.3 | 85.8 | 89.5 | 55.0 | 35.3 | 45.3 | 31.0 | 26.5 | 30.0 |
| Qwen3-VL-32B-Thinking | 600 | 60.5 | 83.0 | 85.2 | 51.0 | 27.1 | 45.8 | 34.8 | 19.9 | 33.4 |
| Qwen3-VL-235B-Instruct | 600 | 68.7 | 85.8 | 89.2 | 59.7 | 34.8 | 47.5 | 35.7 | 26.6 | 33.9 |
| Qwen3-VL-235B-Thinking | 601 | 64.0 | 85.4 | 91.3 | 55.0 | 29.7 | 45.2 | 34.9 | 22.0 | 34.3 |
| Qwen2.5-VL-3B | 608 | 13.7 | 38.2 | 61.5 | 9.3 | 7.7 | 13.5 | 13.0 | 5.1 | 6.0 |
| Qwen2.5-VL-7B | 608 | 32.3 | 62.2 | 78.3 | 24.8 | 10.7 | 17.1 | 17.1 | 7.1 | 21.6 |
| Qwen2.5-VL-32B | 608 | 30.7 | 60.2 | 75.3 | 23.7 | 11.9 | 22.0 | 14.0 | 8.4 | 18.8 |
| Qwen2.5-VL-72B | 608 | 50.0 | 79.3 | 85.5 | 40.0 | 20.9 | 36.1 | 30.1 | 14.4 | 29.0 |
| GLM-4.6v | 594 | 55.5 | 82.3 | 87.2 | 45.9 | 29.1 | 47.9 | 41.6 | 21.9 | 26.4 |
| **Medical** | | | | | | | | | | |
| Lingshu-7B | 608 | 39.8 | 63.0 | 84.4 | 30.5 | 22.1 | 27.2 | 24.0 | 15.3 | 17.7 |
| Lingshu-32B | 608 | 41.9 | 71.1 | 88.3 | 31.8 | 21.4 | 31.3 | 26.0 | 14.7 | 20.6 |
| MedGemma-4B | 608 | 14.4 | 43.4 | 66.4 | 9.6 | 4.9 | 12.7 | 20.2 | 3.0 | 9.5 |
| MedGemma-27B | 608 | 33.7 | 58.4 | 75.7 | 26.1 | 5.2 | 10.0 | 9.2 | 3.4 | 28.6 |
| Hulu-Med-4B | 608 | 28.2 | 59.0 | 80.9 | 20.6 | 12.6 | 23.3 | 24.2 | 8.3 | 15.6 |
| Hulu-Med-7B | 608 | 30.7 | 60.0 | 81.9 | 22.6 | 17.3 | 26.6 | 26.8 | 11.5 | 13.4 |
| Hulu-Med-14B | 608 | 35.4 | 66.0 | 85.2 | 26.7 | 18.5 | 29.9 | 30.3 | 12.5 | 16.9 |
| UniMedVL | 608 | 21.2 | 44.9 | 66.8 | 15.2 | 12.2 | 16.7 | 19.4 | 8.2 | 9.0 |
| MedGemma-1.5-4B | 608 | 23.8 | 47.7 | 69.7 | 16.7 | 11.8 | 12.3 | 6.2 | 6.9 | 11.9 |
| HuatuoGPT-Vision-7B | 608 | 25.8 | 46.7 | 73.8 | 18.4 | 12.7 | 11.8 | 4.1 | 7.2 | 13.1 |

Table A14: **Evidence Recall with and without Medical Images.** $N$ counts Grd–Rec common-valid items. ReqR is flat required-evidence recall and Hier is the full-tier score (%). $\Delta_{\mathrm{img|SupDx}}$ is the Grd minus Rec mean full-tier score on these items (pp).

| | | Grd | | Rec | | |
|---|---|---|---|---|---|---|
| **Model** | **N** | **ReqR** | **Hier** | **ReqR** | **Hier** | $\boldsymbol{\Delta_{\mathrm{img\|SupDx}}}$ |
| **Closed-source** | | | | | | |
| GPT-4o-mini | 601 | 44.3 | 20.6 | 16.4 | 3.0 | 17.6 |
| GPT-4o | 600 | 69.2 | 51.0 | 27.2 | 9.9 | 41.1 |
| GPT-4.1-mini | 580 | 77.2 | 59.9 | 19.6 | 10.2 | 49.7 |
| GPT-4.1 | 598 | 74.6 | 58.9 | 36.8 | 17.2 | 41.8 |
| GPT-5 | 608 | 86.5 | 75.0 | 29.2 | 12.9 | 62.1 |
| GPT-5.2 | 585 | 84.7 | 71.2 | 13.8 | 7.4 | 63.8 |
| Gemini-2.5-Flash | 608 | 82.3 | 68.3 | 36.1 | 15.7 | 52.5 |
| Gemini-2.5-Pro-Nothinking | 608 | 86.2 | 74.5 | 41.1 | 20.9 | 53.6 |
| Gemini-3.5-Flash | 608 | 89.3 | 80.5 | 42.5 | 23.5 | 57.0 |
| Claude-Haiku-4.5 | 608 | 59.4 | 40.7 | 31.1 | 13.2 | 27.5 |
| Claude-Sonnet-4.6 | 608 | 78.8 | 64.9 | 36.7 | 19.2 | 45.7 |
| Claude-Opus-4.6 | 608 | 83.5 | 71.5 | 38.7 | 19.2 | 52.2 |
| Doubao-Seed-1.6-Vision | 608 | 73.5 | 55.2 | 30.6 | 12.7 | 42.5 |
| Mistral-Large-2512 | 581 | 63.3 | 43.9 | 30.9 | 13.9 | 30.0 |
| **Open-source** | | | | | | |
| Qwen3-VL-8B-Instruct | 600 | 57.6 | 35.1 | 33.6 | 14.0 | 21.1 |
| Qwen3-VL-8B-Thinking | 601 | 50.5 | 30.0 | 22.8 | 10.0 | 20.0 |
| Qwen3-VL-30B-Instruct | 601 | 65.2 | 44.8 | 38.5 | 17.1 | 27.7 |
| Qwen3-VL-30B-Thinking | 601 | 62.8 | 42.7 | 30.5 | 13.6 | 29.2 |
| Qwen3-VL-32B-Instruct | 600 | 73.8 | 55.0 | 38.2 | 17.4 | 37.5 |
| Qwen3-VL-32B-Thinking | 600 | 69.4 | 51.0 | 36.4 | 17.4 | 33.5 |
| Qwen3-VL-235B-Instruct | 600 | 75.8 | 59.7 | 37.8 | 17.3 | 42.4 |
| Qwen3-VL-235B-Thinking | 598 | 72.9 | 55.1 | 39.1 | 19.7 | 35.4 |
| Qwen2.5-VL-3B | 608 | 27.2 | 9.3 | 14.5 | 2.7 | 6.6 |
| Qwen2.5-VL-7B | 608 | 46.4 | 24.8 | 24.2 | 6.7 | 18.0 |
| Qwen2.5-VL-32B | 608 | 44.5 | 23.7 | 24.1 | 5.8 | 17.9 |
| Qwen2.5-VL-72B | 608 | 62.1 | 40.0 | 34.3 | 13.8 | 26.2 |
| GLM-4.6v | 594 | 66.3 | 45.9 | 35.5 | 15.2 | 30.7 |
| **Medical** | | | | | | |
| Lingshu-7B | 608 | 52.2 | 30.5 | 26.1 | 7.5 | 23.0 |
| Lingshu-32B | 608 | 55.9 | 31.8 | 30.0 | 9.9 | 21.9 |
| MedGemma-4B | 608 | 29.5 | 9.6 | 15.5 | 1.7 | 7.9 |
| MedGemma-27B | 608 | 46.0 | 26.1 | 26.8 | 8.5 | 17.6 |
| Hulu-Med-4B | 608 | 43.6 | 20.6 | 26.2 | 7.4 | 13.2 |
| Hulu-Med-7B | 608 | 45.5 | 22.6 | 24.4 | 9.9 | 12.7 |
| Hulu-Med-14B | 608 | 50.2 | 26.7 | 26.6 | 7.3 | 19.4 |
| UniMedVL | 608 | 34.1 | 15.2 | 19.6 | 5.8 | 9.4 |
| MedGemma-1.5-4B | 608 | 36.8 | 16.7 | 16.4 | 2.5 | 14.2 |
| HuatuoGPT-Vision-7B | 608 | 38.7 | 18.4 | 26.1 | 6.7 | 11.7 |

Table A15: **Sensitivity of Evidence–Diagnosis Correspondence.**

| Cond. | Tier | Judge | $\tau$ | M | $N_{11}$ | $N_{01}$ | Pooled % [95% CI] | Equal-M % [95% CI] |
|---|---|---|---|---|---|---|---|---|
| Grd | L1 | Secondary | 0.4 | 37 | 1073 | 3493 | 23.5 [17.5, 29.4] | 19.5 [14.7, 24.8] |
| Grd | L1 | Secondary | 0.5 | 37 | 1009 | 3308 | 23.4 [17.3, 29.3] | 19.2 [14.4, 24.4] |
| Grd | L1 | Secondary | 0.6 | 37 | 703 | 2119 | 24.9 [17.9, 31.5] | 18.3 [13.1, 23.7] |
| Grd | L1 | Primary | 0.4 | 37 | 1446 | 4622 | 23.8 [18.3, 29.4] | 19.8 [15.2, 24.9] |
| Grd | L1 | Primary | 0.5 | 37 | 1383 | 4408 | 23.9 [18.3, 29.6] | 19.6 [14.9, 24.7] |
| Grd | L1 | Primary | 0.6 | 37 | 1136 | 3499 | 24.5 [18.3, 30.7] | 18.9 [14.0, 24.1] |
| Grd | full-tier | Secondary | 0.4 | 37 | 652 | 1854 | 26.0 [18.4, 33.0] | 17.4 [12.2, 23.1] |
| Grd | full-tier | Secondary | 0.5 | 37 | 542 | 1448 | 27.2 [19.1, 34.5] | 16.8 [11.3, 23.0] |
| Grd | full-tier | Secondary | 0.6 | 37 | 451 | 1176 | 27.7 [19.4, 35.0] | 16.1 [10.5, 22.5] |
| Grd | full-tier | Primary | 0.4 | 37 | 1123 | 3312 | 25.3 [19.0, 31.6] | 19.3 [14.4, 24.6] |
| Grd | full-tier | Primary | 0.5 | 37 | 1028 | 2883 | 26.3 [19.6, 32.7] | 19.3 [14.3, 24.9] |
| Grd | full-tier | Primary | 0.6 | 37 | 919 | 2511 | 26.8 [20.0, 33.3] | 19.5 [14.4, 25.1] |
| NoDx | L1 | Secondary | 0.4 | 35 | 675 | 1723 | 28.1 [20.3, 35.9] | 21.8 [15.8, 28.2] |
| NoDx | L1 | Secondary | 0.5 | 35 | 638 | 1598 | 28.5 [20.6, 36.5] | 21.9 [15.9, 28.5] |
| NoDx | L1 | Secondary | 0.6 | 35 | 393 | 824 | 32.3 [21.9, 41.5] | 21.2 [14.1, 29.1] |
| NoDx | L1 | Primary | 0.4 | 37 | 859 | 2245 | 27.7 [20.1, 35.1] | 21.0 [15.4, 27.1] |
| NoDx | L1 | Primary | 0.5 | 37 | 809 | 2080 | 28.0 [20.1, 35.7] | 20.7 [15.0, 26.8] |
| NoDx | L1 | Primary | 0.6 | 37 | 604 | 1438 | 29.6 [20.7, 38.0] | 20.3 [14.2, 27.2] |
| NoDx | full-tier | Secondary | 0.4 | 35 | 374 | 702 | 34.8 [23.6, 44.4] | 22.0 [14.4, 30.9] |
| NoDx | full-tier | Secondary | 0.5 | 35 | 296 | 478 | 38.2 [26.3, 48.2] | 23.5 [14.5, 34.2] |
| NoDx | full-tier | Secondary | 0.6 | 35 | 252 | 386 | 39.5 [27.2, 49.9] | 23.6 [13.9, 35.6] |
| NoDx | full-tier | Primary | 0.4 | 37 | 619 | 1283 | 32.5 [23.2, 41.4] | 22.8 [15.8, 30.9] |
| NoDx | full-tier | Primary | 0.5 | 37 | 524 | 1031 | 33.7 [23.8, 42.9] | 22.8 [15.6, 31.1] |
| NoDx | full-tier | Primary | 0.6 | 37 | 428 | 851 | 33.5 [22.5, 43.3] | 21.8 [14.5, 30.8] |

Table A16: **Per-Model Ranked Diagnosis with Findings Disclosed.**

| Model | R@1 | Top-1 (G/D) | R@10 | MRR | MR↓ |
|---|---|---|---|---|---|
| **Closed-source** | | | | | |
| GPT-4o-mini | 20.3 | 12.3/9.7 | 46.0 | .275 | 3.04 |
| GPT-4o | 37.7 | 34.0/26.0 | 60.3 | .444 | 2.39 |
| GPT-4.1-mini | 42.0 | 37.3/29.7 | 63.7 | .483 | 2.23 |
| GPT-4.1 | 43.3 | 39.3/29.7 | 62.7 | .483 | 2.28 |
| GPT-5 | 55.0 | 56.7/50.3 | 75.0 | .620 | 1.71 |
| GPT-5.2 | 45.0 | 43.0/38.7 | 73.3 | .541 | 2.16 |
| Gemini-2.5-Flash | 43.7 | 41.0/29.7 | 68.7 | .515 | 2.05 |
| Gemini-2.5-Pro-Nothinking | 51.3 | 46.0/35.0 | 76.3 | .597 | 1.99 |
| Gemini-3.5-Flash | 63.7 | 58.7/48.0 | 85.0 | .709 | 1.72 |
| Claude-Haiku-4.5 | 30.3 | 24.7/20.3 | 54.0 | .378 | 2.44 |
| Claude-Sonnet-4.6 | 45.0 | 43.0/35.7 | 74.3 | .542 | 2.22 |
| Claude-Opus-4.6 | 51.3 | 46.3/39.0 | 79.3 | .606 | 1.98 |
| Doubao-Seed-1.6-Vision | 34.3 | 30.0/22.7 | 59.0 | .425 | 2.24 |
| Mistral-Large-2512 | 32.0 | 29.7/20.3 | 56.3 | .397 | 2.38 |
| **Open-source** | | | | | |
| Qwen3-VL-8B-Instruct | 14.3 | 9.7/8.0 | 38.0 | .214 | 3.04 |
| Qwen3-VL-8B-Thinking | 19.7 | 15.3/12.0 | 43.3 | .271 | 2.76 |
| Qwen3-VL-30B-Instruct | 27.0 | 20.0/16.0 | 51.0 | .335 | 2.75 |
| Qwen3-VL-30B-Thinking | 28.0 | 20.0/16.0 | 51.0 | .355 | 2.34 |
| Qwen3-VL-32B-Instruct | 24.3 | 20.7/14.7 | 50.0 | .316 | 2.81 |
| Qwen3-VL-32B-Thinking | 34.0 | 27.3/20.0 | 55.7 | .405 | 2.34 |
| Qwen3-VL-235B-Instruct | 28.7 | 23.0/16.7 | 52.3 | .352 | 2.79 |
| Qwen3-VL-235B-Thinking | 35.7 | 28.7/20.3 | 62.7 | .430 | 2.68 |
| Qwen2.5-VL-3B | 7.7 | 3.0/2.3 | 31.3 | .140 | 3.87 |
| Qwen2.5-VL-7B | 12.3 | 7.3/5.3 | 32.7 | .187 | 2.88 |
| Qwen2.5-VL-32B | 19.7 | 15.7/12.7 | 43.7 | .264 | 3.00 |
| Qwen2.5-VL-72B | 26.0 | 18.7/14.0 | 49.7 | .336 | 2.39 |
| GLM-4.6v | 30.7 | 26.0/19.0 | 53.7 | .380 | 2.52 |
| **Medical** | | | | | |
| Lingshu-7B | 18.0 | 12.0/9.7 | 37.0 | .241 | 2.52 |
| Lingshu-32B | 23.7 | 18.0/14.0 | 45.0 | .302 | 2.59 |
| MedGemma-4B | 10.7 | 6.3/4.0 | 32.3 | .168 | 3.42 |
| MedGemma-27B | 21.7 | 16.7/12.7 | 44.0 | .277 | 2.88 |
| Hulu-Med-4B | 19.7 | 11.0/9.0 | 40.3 | .261 | 2.69 |
| Hulu-Med-7B | 20.3 | 12.0/9.7 | 44.7 | .276 | 2.85 |
| Hulu-Med-14B | 26.0 | 17.7/13.7 | 47.3 | .320 | 2.55 |
| UniMedVL | 12.7 | 7.7/6.0 | 32.0 | .182 | 3.16 |
| MedGemma-1.5-4B | 11.7 | 5.3/4.3 | 34.7 | .185 | 3.20 |
| HuatuoGPT-Vision-7B | 13.0 | 7.3/5.0 | 35.0 | .196 | 2.94 |

Table A17: **Per-Model Ranked Diagnosis with Images and Limited Context.**

| Model | R@1 | Top-1 (G/D) | R@10 | MRR | MR↓ |
|---|---|---|---|---|---|
| **Closed-source** | | | | | |
| GPT-4o-mini | 8.7 | 3.7/2.3 | 29.7 | .139 | 3.93 |
| GPT-4o | 19.7 | 16.0/13.0 | 34.7 | .238 | 2.72 |
| GPT-4.1-mini | 25.7 | 20.3/16.7 | 48.7 | .323 | 2.71 |
| GPT-4.1 | 29.7 | 27.3/20.7 | 52.7 | .356 | 2.68 |
| GPT-5 | 48.7 | 49.3/44.3 | 71.0 | .558 | 1.97 |
| GPT-5.2 | 31.7 | 30.0/25.3 | 60.0 | .400 | 2.73 |
| Gemini-2.5-Flash | 29.0 | 25.7/18.0 | 56.3 | .369 | 2.73 |
| Gemini-2.5-Pro-Nothinking | 40.3 | 39.0/27.7 | 68.3 | .494 | 2.20 |
| Gemini-3.5-Flash | 54.7 | 50.3/42.7 | 82.0 | .633 | 2.00 |
| Claude-Haiku-4.5 | 16.0 | 13.0/10.7 | 34.7 | .210 | 3.16 |
| Claude-Sonnet-4.6 | 31.3 | 29.0/24.3 | 60.7 | .395 | 2.74 |
| Claude-Opus-4.6 | 36.3 | 34.0/26.7 | 64.7 | .446 | 2.59 |
| Doubao-Seed-1.6-Vision | 24.7 | 22.0/17.0 | 47.3 | .318 | 2.49 |
| Mistral-Large-2512 | 14.3 | 13.0/9.0 | 41.3 | .222 | 3.24 |
| **Open-source** | | | | | |
| Qwen3-VL-8B-Instruct | 6.7 | 3.7/3.3 | 29.0 | .124 | 4.09 |
| Qwen3-VL-8B-Thinking | 13.0 | 9.3/6.7 | 31.3 | .181 | 3.06 |
| Qwen3-VL-30B-Instruct | 16.3 | 11.3/9.0 | 39.3 | .222 | 3.43 |
| Qwen3-VL-30B-Thinking | 17.0 | 11.7/8.0 | 38.3 | .232 | 2.99 |
| Qwen3-VL-32B-Instruct | 15.0 | 14.0/10.3 | 35.3 | .207 | 3.13 |
| Qwen3-VL-32B-Thinking | 19.3 | 15.7/12.7 | 43.3 | .262 | 3.02 |
| Qwen3-VL-235B-Instruct | 20.0 | 15.3/10.7 | 41.7 | .259 | 2.91 |
| Qwen3-VL-235B-Thinking | 25.0 | 21.0/15.7 | 50.3 | .327 | 2.74 |
| Qwen2.5-VL-3B | 3.0 | 1.0/0.7 | 14.7 | .062 | 3.82 |
| Qwen2.5-VL-7B | 8.7 | 4.3/3.7 | 25.3 | .126 | 3.82 |
| Qwen2.5-VL-32B | 7.0 | 5.7/3.3 | 30.3 | .125 | 4.35 |
| Qwen2.5-VL-72B | 15.3 | 10.7/8.0 | 36.3 | .206 | 3.57 |
| GLM-4.6v | 15.3 | 12.0/9.3 | 40.7 | .222 | 3.35 |
| **Medical** | | | | | |
| Lingshu-7B | 10.0 | 3.7/2.7 | 28.3 | .157 | 3.02 |
| Lingshu-32B | 12.0 | 8.7/6.7 | 33.0 | .182 | 3.15 |
| MedGemma-4B | 5.7 | 2.0/1.3 | 22.0 | .101 | 3.79 |
| MedGemma-27B | 8.0 | 5.3/4.3 | 31.0 | .149 | 3.65 |
| Hulu-Med-4B | 11.7 | 5.3/4.7 | 30.3 | .172 | 3.07 |
| Hulu-Med-7B | 16.3 | 10.7/9.3 | 36.0 | .219 | 3.01 |
| Hulu-Med-14B | 15.7 | 11.0/9.3 | 36.0 | .221 | 2.81 |
| UniMedVL | 7.3 | 6.3/2.7 | 24.7 | .124 | 3.35 |
| MedGemma-1.5-4B | 6.3 | 3.0/1.7 | 26.0 | .116 | 3.94 |
| HuatuoGPT-Vision-7B | 6.7 | 3.7/2.7 | 24.7 | .116 | 3.68 |

Table A18: **Per-Model Ranked Diagnosis from Clinical Text Only.**

| Model | R@1 | Top-1 (G/D) | R@10 | MRR | MR↓ |
|---|---|---|---|---|---|
| **Closed-source** | | | | | |
| GPT-4o-mini | 7.7 | 4.0/2.7 | 24.3 | .118 | 4.01 |
| GPT-4o | 17.3 | 13.0/10.3 | 39.3 | .233 | 3.23 |
| GPT-4.1-mini | 15.0 | 14.0/10.7 | 36.3 | .208 | 3.29 |
| GPT-4.1 | 18.3 | 15.7/12.0 | 37.0 | .235 | 2.92 |
| GPT-5 | 27.7 | 26.7/21.3 | 51.0 | .337 | 2.73 |
| GPT-5.2 | 22.7 | 19.7/16.7 | 47.3 | .292 | 3.08 |
| Gemini-2.5-Flash | 13.0 | 10.7/8.0 | 36.7 | .201 | 3.24 |
| Gemini-2.5-Pro-Nothinking | 22.7 | 18.0/14.3 | 50.7 | .302 | 3.14 |
| Gemini-3.5-Flash | 26.0 | 21.7/18.7 | 53.0 | .337 | 2.96 |
| Claude-Haiku-4.5 | 13.0 | 9.7/8.7 | 29.0 | .164 | 3.74 |
| Claude-Sonnet-4.6 | 22.0 | 17.7/15.3 | 47.0 | .290 | 3.04 |
| Claude-Opus-4.6 | 22.7 | 20.7/16.0 | 50.0 | .305 | 2.99 |
| Doubao-Seed-1.6-Vision | 12.3 | 8.0/5.7 | 30.3 | .174 | 3.34 |
| Mistral-Large-2512 | 9.3 | 7.0/4.3 | 34.0 | .162 | 3.59 |
| **Open-source** | | | | | |
| Qwen3-VL-8B-Instruct | 5.0 | 3.3/2.0 | 23.3 | .097 | 4.26 |
| Qwen3-VL-8B-Thinking | 8.0 | 4.7/4.3 | 27.3 | .130 | 3.96 |
| Qwen3-VL-30B-Instruct | 8.3 | 5.0/3.7 | 28.7 | .143 | 3.47 |
| Qwen3-VL-30B-Thinking | 10.7 | 7.7/5.3 | 28.0 | .149 | 3.67 |
| Qwen3-VL-32B-Instruct | 8.0 | 5.7/5.3 | 27.7 | .134 | 3.72 |
| Qwen3-VL-32B-Thinking | 11.3 | 7.3/6.3 | 28.7 | .166 | 2.94 |
| Qwen3-VL-235B-Instruct | 9.7 | 6.0/4.3 | 32.0 | .160 | 3.61 |
| Qwen3-VL-235B-Thinking | 12.7 | 9.3/7.3 | 34.3 | .190 | 3.33 |
| Qwen2.5-VL-3B | 3.3 | 1.0/1.0 | 12.7 | .055 | 4.13 |
| Qwen2.5-VL-7B | 5.0 | 1.7/1.3 | 18.7 | .088 | 3.75 |
| Qwen2.5-VL-32B | 5.3 | 2.3/2.0 | 21.7 | .099 | 3.97 |
| Qwen2.5-VL-72B | 10.0 | 5.7/4.7 | 26.7 | .146 | 3.48 |
| GLM-4.6v | 9.0 | 6.3/5.0 | 32.0 | .155 | 3.62 |
| **Medical** | | | | | |
| Lingshu-7B | 4.7 | 2.0/1.3 | 18.3 | .084 | 3.98 |
| Lingshu-32B | 4.7 | 2.0/1.7 | 24.0 | .101 | 4.03 |
| MedGemma-4B | 3.7 | 1.7/0.7 | 21.3 | .083 | 4.59 |
| MedGemma-27B | 4.7 | 3.7/2.0 | 21.7 | .095 | 3.92 |
| Hulu-Med-4B | 6.3 | 3.3/3.3 | 24.7 | .119 | 3.57 |
| Hulu-Med-7B | 6.7 | 3.7/3.3 | 25.0 | .116 | 3.92 |
| Hulu-Med-14B | 7.3 | 4.7/4.0 | 22.3 | .117 | 3.43 |
| UniMedVL | 3.7 | 1.7/1.0 | 17.0 | .070 | 4.45 |
| MedGemma-1.5-4B | 3.7 | 1.0/0.7 | 18.7 | .074 | 4.23 |
| HuatuoGPT-Vision-7B | 4.7 | 2.0/1.3 | 13.3 | .070 | 3.73 |

Table A19: **Per-Model Diagnosis Outcomes and Correspondence.** Retrieval failures omit the gold diagnosis from the top 10; commitment failures rank it from second to tenth.

| Model | LC R@1 | miss | Retr. | Commit. | $N_{eff}$ | Corr |
|---|---|---|---|---|---|---|
| **Closed-source** | | | | | | |
| Gemini-3.5-Flash | 54.7 | 18.0 | 39.7 | 60.3 | 289 | 54.3 |
| GPT-5 | 48.7 | 29.0 | 56.5 | 43.5 | 289 | 51.4 |
| Gemini-2.5-Pro-NoThink | 40.3 | 31.7 | 53.1 | 46.9 | 289 | 42.5 |
| Claude-Opus-4.6 | 36.3 | 35.3 | 55.5 | 44.5 | 289 | 37.6 |
| GPT-5.2 | 31.7 | 40.0 | 58.5 | 41.5 | 279 | 30.9 |
| Claude-Sonnet-4.6 | 31.3 | 39.3 | 57.3 | 42.7 | 289 | 30.2 |
| GPT-4.1 | 29.7 | 47.3 | 67.3 | 32.7 | 287 | 31.9 |
| Gemini-2.5-Flash | 29.0 | 43.7 | 61.5 | 38.5 | 289 | 31.4 |
| GPT-4.1-mini | 25.7 | 51.3 | 69.1 | 30.9 | 280 | 27.1 |
| Doubao-Seed-1.6-Vision | 24.7 | 52.7 | 69.9 | 30.1 | 289 | 24.0 |
| GPT-4o | 19.7 | 65.3 | 81.3 | 18.7 | 288 | 19.4 |
| Claude-Haiku-4.5 | 16.0 | 65.3 | 77.8 | 22.2 | 289 | 23.5 |
| Mistral-Large-2512 | 14.3 | 58.7 | 68.5 | 31.5 | 272 | 22.7 |
| GPT-4o-mini | 8.7 | 70.3 | 77.0 | 23.0 | 285 | 14.6 |
| **Open-source** | | | | | | |
| Qwen3-VL-235B-Thinking | 25.0 | 49.7 | 66.2 | 33.8 | 287 | 26.3 |
| Qwen3-VL-235B-Instruct | 20.0 | 58.3 | 72.9 | 27.1 | 287 | 17.1 |
| Qwen3-VL-32B-Thinking | 19.3 | 56.7 | 70.2 | 29.8 | 287 | 18.3 |
| Qwen3-VL-30B-Thinking | 17.0 | 61.7 | 74.3 | 25.7 | 287 | 14.4 |
| Qwen3-VL-30B-Instruct | 16.3 | 60.7 | 72.5 | 27.5 | 287 | 16.1 |
| GLM-4.6v | 15.3 | 59.3 | 70.1 | 29.9 | 284 | 14.3 |
| Qwen2.5-VL-72B | 15.3 | 64.7 | 76.4 | 23.6 | 289 | 17.9 |
| Qwen3-VL-32B-Instruct | 15.0 | 64.7 | 76.1 | 23.9 | 287 | 18.0 |
| Qwen3-VL-8B-Thinking | 13.0 | 68.7 | 78.9 | 21.1 | 287 | 14.8 |
| Qwen2.5-VL-7B | 8.7 | 74.7 | 81.8 | 18.2 | 289 | 7.7 |
| Qwen2.5-VL-32B | 7.0 | 69.7 | 74.9 | 25.1 | 289 | 14.3 |
| Qwen3-VL-8B-Instruct | 6.7 | 71.0 | 76.1 | 23.9 | 287 | 7.7 |
| Qwen2.5-VL-3B | 3.0 | 85.3 | 88.0 | 12.0 | 289 | 0.0 |
| **Medical** | | | | | | |
| Hulu-Med-7B | 16.3 | 64.0 | 76.5 | 23.5 | 289 | 14.3 |
| Hulu-Med-14B | 15.7 | 64.0 | 75.9 | 24.1 | 289 | 18.2 |
| Lingshu-32B | 12.0 | 67.0 | 76.1 | 23.9 | 289 | 11.8 |
| Hulu-Med-4B | 11.7 | 69.7 | 78.9 | 21.1 | 289 | 5.6 |
| Lingshu-7B | 10.0 | 71.7 | 79.6 | 20.4 | 289 | 6.2 |
| MedGemma-27B | 8.0 | 69.0 | 75.0 | 25.0 | 289 | 5.4 |
| UniMedVL | 7.3 | 75.3 | 81.3 | 18.7 | 289 | 7.4 |
| HuatuoGPT-Vision-7B | 6.7 | 75.3 | 80.7 | 19.3 | 289 | 12.1 |
| MedGemma-1.5-4B | 6.3 | 74.0 | 79.0 | 21.0 | 289 | 6.2 |
| MedGemma-4B | 5.7 | 78.0 | 82.7 | 17.3 | 289 | 0.0 |
| *Pooled* | 18.7 | 59.5 | 73.2 | 26.8 | 10,629 | 26.3 |

# O Qualitative Case Examples

## T1 Example

### Exact LC model input

**Patient Demographics.**
- The patient is a 53-year-old woman.

**Symptoms and Signs.**
- The patient presented with dyspnea, palpitations, and fatigue over a 6-month period.

**Medical History.**
- Past history included hypothyroidism, asthma, hyperlipidemia, and migraine.

**Laboratory Results.**
- Electrocardiogram showed sinus rhythm with first degree atrioventricular block.
- Initial laboratory investigations revealed mild anemia and mildly elevated inflammatory markers.
- White blood cell count, renal function, and liver function were not significantly abnormal.
- Blood cultures remained sterile.
- Extended autoimmune and infective screening was unrevealing.

**Clinical Timeline.**
- The patient underwent cardiac surgery for severe aortic valve dysfunction, and tissue obtained at surgery was later used for definitive diagnosis.

**Question.** Based on the patient's brief clinical history and the provided histopathology image, what diagnoses should be considered for this patient?

**Instruction:** You are an expert physician. Based on the clinical presentation and the medical images, list the 10 most likely diagnoses, ranked from MOST likely (1) to LEAST likely (10). Output ONLY a numbered list, one standard clinical disease name per line (e.g. "1. Disease name"). Do not include explanations, reasoning, or extra text.

**Model-visible image**

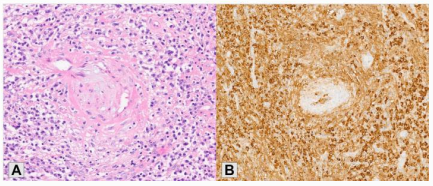



Image 1: histopathology

The histopathology image shown above is attached under FD and LC; TO attaches no medical image.

**Imaging-findings text (the FD-only channel).** Histopathologic examination of the surgical specimen showed a dense lymphoplasmacytic inflammatory infiltrate. FD prepends this sentence to the context above; LC and TO withhold it.

### T1 Variants

| Input channel | FD | LC | TO |
|---|---|---|---|
| Clinical context | Present | Present | Present |
| Imaging-findings text | Present | Withheld | Withheld |
| Medical images | Present | Present | Withheld |

### Answer

IgG4-related disease with myocardial involvement (IgG4-related myocardial mass causing severe aortic incompetence).

Figure A7: **T1 Example.** A leakage-controlled ranked-diagnosis case.

## T2 Example

**Model-visible images**



Image 1

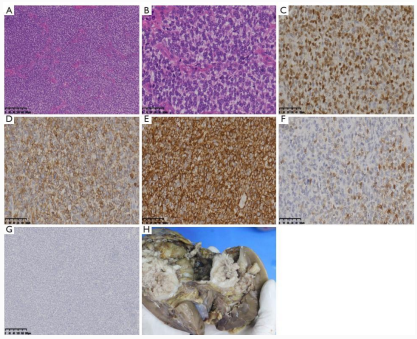


Image 2

**Exact evidence-verification input.**

**Established diagnosis.** Clear cell sarcoma of the kidney. *Relevant images: Image 1 and Image 2.*

**Question.** Using Image 1 and Image 2, describe the visible evidence in each image that supports the established diagnosis, including the main pattern seen in each image and how the two images relate to the same disease process.

**Instruction:** You are an expert physician. The diagnosis is already established. The images are labelled Image 1, Image 2, . . . For EACH image, refer to it explicitly by its label (e.g. "Image 1:") and state the main visible finding/pattern, then explain how the findings across the images fit together to support the established diagnosis. Be specific and describe only what is visible.

**T2 Variants**

| Input channel | Grd | NoDx | Rec |
|---|---|---|---|
| Established diagnosis | Present | Withheld | Present |
| Medical images | Present | Present | Withheld |

**Answer**

**Image 1** shows a large heterogeneous mass centered in the kidney, with both solid and lower-attenuation or necrotic-appearing components. **Image 2** shows the corresponding resected renal tumor together with highly cellular microscopic tumor tissue. Taken together, the images connect the renal mass with its gross and microscopic tumor appearance, which is compatible with the established diagnosis.

Figure A8: **T2 Example.** An open-ended evidence-verification case.