Medical thinking with images

Zonghai Yao [1,2] *, **Benlu Wang** [3] *, **Yifan Zhang** [1,4], **Junda Wang** [1,2], **Iris Xia** [3], **Zhipeng Tang** [2], **Shuo Han** [4], **Feiyun Ouyang** [1,4], **Zhichao Yang** [2], **Arman Cohan** [3], **Hong Yu** [1,2,4],
[1]Center for Healthcare Organization and Implementation Research, VA Bedford Health Care
[2]Manning College of Information and Computer Sciences, UMass Amherst, MA, USA
[3]Department of Computer Science, Yale University, CT, USA
[4]Miner School of Computer and Information Sciences, UMass Lowell, MA, USA
zonghaiyao@umass.edu, benlu.wang@yale.edu

Abstract

Large language models and vision-language models score high on many medical QA benchmarks; however, real-world clinical reasoning remains challenging because cases often involve multiple images and require cross-view fusion. We present **MedThinkVQA**, a benchmark that asks models to *think with multiple images*: read each image, merge evidence across views, and pick a diagnosis with stepwise supervision. We make three parts explicit: multi-image questions, expert-annotated stepwise supervision, and beyond-accuracy evaluation. Only MedThinkVQA combines all these parts in one expert-annotated benchmark. The dataset has 8,481 cases in total, with 751 test cases, and on average 6.51 images per case; it is expert-annotated and, at this level, larger and more image-dense than prior work (earlier maxima ≤ 1.43 images per case). On the test set, GPT-5 achieves 57.39% accuracy, approximately 15 percentage points below the strongest result on the most challenging prior benchmark of a similar kind, while other strong models are lower (Qwen2.5-VL-32B: 39.54%, MedGemma-27B: 37.55%, InternVL3.5-38B: 43.14%). Giving *expert* findings and summaries brings clear gains, but using models' *self-generated* ones brings small or negative gains. Step-level evaluation shows where models stumble: errors center on image reading and cross-view integration in both decisive and non-decisive steps ($> 70\%$); when a step is decisive for the final choice, reasoning slips become more common (32.26%), while scenario and pure-knowledge slips are relatively rare ($< 10\%$). These patterns isolate and quantify the core obstacle: *extracting and integrating cross-image evidence*, rather than language-only inference. Code and example data are available at Here.

1 Introduction

Medical QA has advanced fast with large language models (LLMs) and vision-language models (VLMs). Scores on exam-style datasets are high, and many tasks now appear to be saturated [Jin et al., 2021, Pal et al., 2022, Jin et al., 2019]. But the everyday diagnosis is not a single question and answer. As shown in Fig. 1 (left), clinicians review the clinical scenario and interpret several images, then proceed through the steps (e.g., Differential Diagnosis) before a diagnostic determination. Therefore, we need a benchmark that tests and evaluates the process on multi-image cases.

MedThinkVQA sets a clear three-step flow (Fig. 1, middle). Step 1 is *per-image findings*: detect and explain key signs on each image. The output consists of brief finding sentences for each image. Step 2 is a *case-level imaging summary*: merge signs across views into one summary. Step 3 is

*Equal contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

differential-diagnosis reasoning: give image-grounded eliminations for distractors and then pick one option. All intermediate information is sourced from the peer-reviewed Eurorad [2] repository and is expert-written and pedagogically designed. These expert-annotated intermediate evidences are not a verbatim chain-of-thought target. However, they serve as high-quality references that are essential for evaluating the *think with multiple images (TwI)* steps and for designing faithful training or judging signals. After Step 1–3, we add a *Medical Education Case Discussion* task. This is a long-form document generation task. It mirrors real practice: clinicians explain the background, clinical and imaging perspectives, clinical significance, outcomes, and take-home points. In medicine, this is a core mode of education and knowledge sharing, and it is an important skill. Prior benchmarks usually stop at the final diagnosis, but we include this post-diagnosis ability in evaluation.

Answer accuracy is useful, but it hides where the model fails. We add *stepwise automatic metrics* for Steps 1–2. We follow recent radiology–report generation work Yu et al. [2023], Ostmeier et al. [2024] and use ROUGE as a base, then add RadCliQ to probe whether the model captures fine clinical details in per-image findings and in the case-level summary. For Step 3 (differential diagnosis), we evaluate the *step-by-step* reasoning. Referencing all expert-annotated key information provided by the MedThinkVQA dataset, LLM-judges and expert evaluators check each step for *clinical correctness* and for *clinical criticality*. When a step is wrong, we attach an *error-type tag* from four buckets: clinical-scenario misunderstanding, missing image evidence, knowledge gap, or flawed reasoning. Our judging is reliable: on step-level labels, human–human agreement reaches Cohen's $\kappa = 0.82$, while human–LLM-judge agreement ranges from $\kappa = 0.70$ to $\kappa = 0.84$. For the Case Discussion, we grade both content and value. We use a structured six-part format (Background, Clinical Perspective, Imaging Perspective, Clinical Significance, Outcome, Take-Home Notes). For each part, LLM-judges and experts run a sentence-by-sentence clinical-correct check against the reference discussion. We then score education values on a 0–2 scale for each dimension, totaling 10 points. Thus, the data keep the common 5-option MCQ benchmarking format, and we add a fine-grained, diagnosing-style framework that goes *beyond accuracy* to show where models succeed or fail (see Fig. 1, right).

Table 1 places MedThinkVQA among recent multimodal medical QA datasets [Hu et al., 2024, Ye et al., 2024, Zuo et al., 2025]. Our cases are expert-annotated and include *clinical scenarios*, *per-image findings*, *case imaging summaries*, and *teaching notes*. There are 8,481 cases, with 751 for testing, and an average of 6.51 images per case. Prior expert-level benchmarks use far fewer images per case (max $\leq$1.43). Therefore, our setup emphasizes cross-view fusion, rather than single-view recognition. We map diagnoses to ICD-10 and include Orphanet-aligned rare conditions. We maintain expert distractors, apply confusion-aware pruning, and eliminate text-only solvable items, ensuring each question remains image-dependent and challenging. We also balance simple surface biases during sampling (for example, avoiding patterns like "the longest option is the correct one"), so shortcuts do not inflate scores.

On our test split, GPT-5 achieves the highest 57.39% accuracy, while other strong models are lower (Qwen2.5-VL-72B: 49.18%, MedGemma-27B: 42.02%, InternVL3.5-38B: 43.14%). This is $\sim 15\%$ points below the strongest result on the hardest prior benchmark of a similar kind, indicating the additional difficulty introduced by considering multiple images. Giving *expert* findings and summaries raises accuracy. Replacing them with models' *self-generated* ones gives small gains or hurts. The bottleneck is reading each image well and fusing evidence across images in the think-with-images steps, which our step-level analysis supports: across 202 labeled steps ($\sim$46% Critical), 44 steps carry non-empty error tags; among these error-bearing steps, 77.27% reflect image-understanding issues and 22.73% reflect reasoning, with medical knowledge (9.09%) and scenario setup (4.55%) much rarer; within error-bearing *Critical* steps (31/93), the share of reasoning rises to 32.26% while image understanding remains high at 70.97% (scenario and knowledge near 6–10%).

Contributions. (1) A benchmark for *multi-image* diagnostic reasoning with expert supervision at three steps. (2) A *beyond-accuracy* evaluation suite with automatic intermediate metrics, error-type tagging, and education-value scoring; we release scoring scripts and formats. (3) A large and image-dense expert-annotated corpus (8,481 cases; 6.51 images per case) that, to our knowledge, is the only one that checks all columns in Table 1. (4) Evidence that cross-image evidence extraction and integration is the current medical VLMs bottleneck.

[2]Eurorad website: `https://www.eurorad.org/`. Note that the original cases are licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license; with Eurorad's permission,

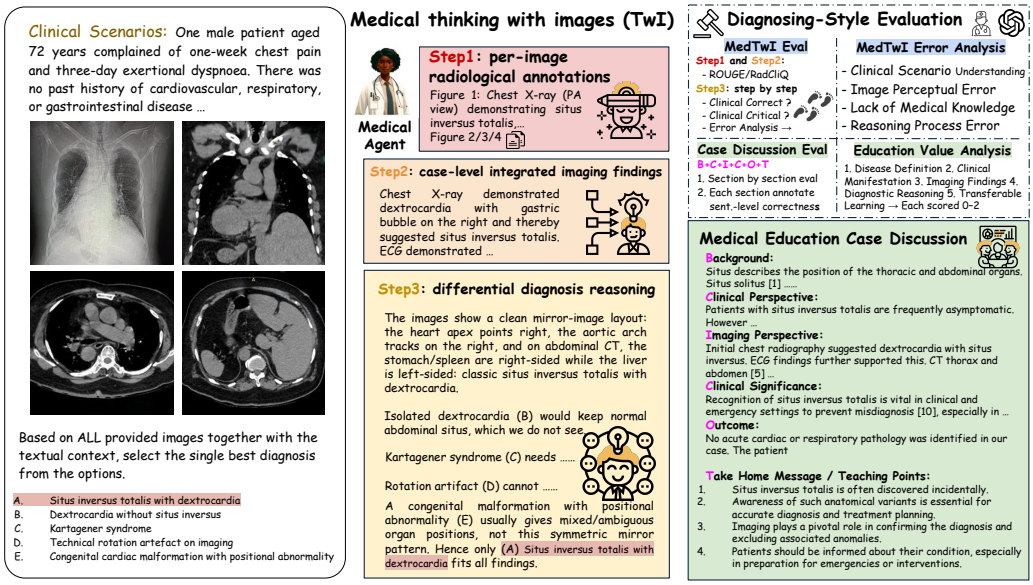

Figure 1: Medical Thinking with Images (TwI): task and diagnosing-style evaluation. Left: a sample case with a clinical scenario, multi-view images (e.g., radiograph + CT), and a five-option single-best-answer diagnosis. Middle: TwI's three supervised steps: (1) *Per-Image Findings* (detect and name key radiological signs for each image, expert-annotated, brief statements); (2) *Case-level Integrated Imaging Summary* (synthesize cross-view evidence into a single case summary); (3) *Differential-Diagnosis (DDx) reasoning* (align the summary with candidate diagnoses, rule out distractors with image-grounded arguments, and pick the most consistent answer). Right: Beyond-accuracy evaluation. Steps 1–2 use automatic metrics (ROUGE / RadCliQ). Step 3 applies step-by-step checks of *clinical correctness* and *clinical criticality* with error-type tags (scenario misunderstanding, missing image evidence, knowledge gap, reasoning error). Human experts and LLM-judges provide complementary assessments (high agreement). Beyond diagnosis, models generate a *Medical Education Case Discussion*; we verify section-level correctness (Background, Clinical Perspective, Imaging Perspective, Clinical Significance, Outcome, Take-Home Notes) and score education value on five rubrics (Disease Overview, Clinical Presentation, Key Imaging Findings, Diagnostic Reasoning, Transferable Learning), each 0–2 (total 10).

2 Related Work

Early MedVQA corpora set task forms but had small scale or shallow reasoning [Ben Abacha et al., 2019, Lau et al., 2018, Liu et al., 2021, He et al., 2020, Zhang et al., 2023]. Later unified benchmarks grew breadth across modalities and specialties [Hu et al., 2024, Ye et al., 2024]. General expert-level suites also add a Health/Medicine subset and try to reduce shortcuts [Yue et al., 2024a,b]. But most questions are single-image or short-context, and many use automatic labels. Many datasets are built from image captions, so labels do not encode diagnostic reasoning or multi-image context. They also lack the detailed clinical information that real cases need. Coverage of medical image types is still limited compared to practice. Within chest radiography, large-scale expansions such as GEMeX and Medical-Diff-VQA (MIMIC-Diff-VQA) increase coverage and add explanations or difference-style question formats, but they still lack clinical scenarios and expert stepwise supervision and are predominantly single-view or pairwise [Liu et al., 2025a, Hu et al., 2025]. So evaluation stays answer-centric and lacks stepwise diagnostic supervision, as reflected in the upper rows of our comparison.

MedXpertQA raises difficulty and realism and has a multimodal track with images and histories [Zuo et al., 2025]. It also provides scenario labels. But it does not release expert *per-image findings* or a *case-level imaging summary*, and it does not annotate option-wise eliminations. Items also use far

we collect, curate, benchmark, and then add human annotation and analysis. We will release the processed data under the same license. More details can be found in Section 3 and Ethical Statement.

Benchmark	# Case	# Img	Annotation	Clinical Scenarios	Think-with-Images Intermediate Signals	Beyond-ACC Evaluation
VQA-Rad Lau et al.	451	0.45	Automatic	✗	✗	✗
VQA-Med Ben Abacha et al.	500	1.00	Automatic	✗	✗	✗
Path-VQA He et al.	6,719	0.13	Automatic	✗	✗	✗
SLAKE-En Liu et al.	1,061	0.09	Automatic	✗	✗	✗
PMC-VQA Zhang et al.	33,430	0.87	Automatic	✗	✗	✗
OmniMedVQA Hu et al.	127,995	0.92	Automatic	✗	✗	✗
GMAI-MMBench Ye et al.	21,281	1.00	Automatic	✗	✗	✗
GEMeX Liu et al.	1,605,575	1.00	Automatic	✗	✗	✗
Medical-Diff-VQA Hu et al.	700,703	1.23	Automatic	✗	✗	✗
GEMeX-ThinkVG Liu et al.	202,384	1.00	Automatic	✗	✓	✓
MMMU (H & M) Yue et al.	1,752	1.14	Expert	✗	✗	✗
MMMU-Pro (H & M) Yue et al.	346	1.25	Expert	✗	✗	✗
MedXpertQA MM Zuo et al.	2,000	1.43	Expert	✓	✗	✗
MedRAX Fallahpour et al.	2,500	1.00	Expert	✓	✗	✗
S-Chain Le-Duc et al.	788,000	1.00	Expert	✗	✓	✓
MedThinkVQA	**8,481**	**6.51**	Expert	✓	✓	✓

Table 1: Comparisons with multimodal medical QA benchmarks. **Case/#Img/Annotation.** *MedThinkVQA* is expert-annotated, averages **6.51** images per case (prior maxima ≤ 1.43; $\geq 4.5\times$ more), and is the largest corpus at the expert-annotation level. **Clinical Scenarios.** Prior work lacks broad, fine-grained coverage of real diagnostic scenarios; only *MedThinkVQA* and *MedXpertQA-MM* include scenario labels. **Think-with-Images Intermediate Signals.** This merged column indicates whether a benchmark provides intermediate supervision for think-with-images reasoning, including *per-image findings*, a *case-level imaging summary*, and a *case discussion* (teaching note). **Beyond-ACC Evaluation.** Leveraging these signals, only *MedThinkVQA* supports fine-grained, end-to-end assessment of think-with-images reasoning and teaching discussions: stepwise checks, error-type tags, education-value scoring, and automatic intermediate metrics, rather than accuracy alone.

fewer images per case (max ≤ 1.43), so cross-view fusion is not stressed. We fill these gaps with expert step labels (per-image findings and a case summary), with option-wise eliminations, and with a reproducible beyond-accuracy suite (step metrics, error types, and education scoring).

Eurorad-based studies often prompt models with textual descriptions from case reports [Kim et al., 2025]. This probes language use, but it does not test reading raw images. Text-only prompting cannot test multi-image fusion or image dependence. In parallel, agent-style evaluation on chest X-rays (MedRAX) orchestrates segmentation, grounding, and report-generation tools on Eurorad-derived multiple-choice cases without releasing expert step traces, complementing but not replacing multi-image diagnostic supervision [Fallahpour et al., 2025]. So our setting requires direct multi-image reading and option-wise, evidence-grounded elimination.

Work on reasoning supervision trains or audits how models explain answers [Gai et al., 2025, Liu et al., 2024, Wang et al., 2025b, Fan et al., 2025]. Prior efforts include chain-of-thought generation, visually grounded reasoning, and cycle consistency. Recent corpora such as GEMeX-ThinkVG and S-Chain further introduce grounded reasoning paths and structured visual chain-of-thought with localization metrics (e.g., A-score, mIoU), moving beyond answer accuracy while still lacking clinical scenarios and multi-view, case-level synthesis [Liu et al., 2025b, Le-Duc et al., 2025]. These help transparency and stability. But most corpora do not release expert, item-specific *diagnostic* traces tied to options. Without option-aligned traces, contrastive fidelity checks and step-level rubrics are hard to standardize. We release expert per-image findings and a case-level summary, and we pair them with option-wise eliminations. This enables contrastive fidelity checks, step-level scoring, and education-oriented evaluation with human and LLM judges. Teaching discussions are also a standard product of medical education, yet benchmarks rarely evaluate this skill.

3 MedThinkVQA

3.1 Source Corpus

MedThinkVQA is adapted from *Eurorad*, a peer-reviewed online database of radiology teaching cases curated by the European Society of Radiology [eur]. The corpus covers major subspecialties (neuro, musculoskeletal, thoracic, abdominal, pediatric, etc.) and common imaging modalities (X-ray, CT, MRI, ultrasound, etc.). Each case includes: (i) a brief clinical history; (ii) a multi-image set

(average 8.3 images per case); (iii) radiologist-annotated, per-image hints; (iv) a case-level *Integrated Imaging Summary* section; (v) an *Expert Reasoning & Teaching Note* that interprets the findings, highlights key diagnostic reasoning, and links to clinical relevance;(vi) the final diagnosis; and (vii) a differential-diagnosis list. Cases are contributed by radiologists and researchers worldwide, typically based on real clinical examinations. Submissions are reviewed by the Eurorad Editorial Board (radiology experts) before publication to ensure authenticity and educational value [eur]. We collected **8,481** cases and curated them into MedThinkVQA. After post-processing, we formed a held-out test set with **751** cases and a training set with **7,730** cases. Details of the MCQ transformation and option policy are provided in §*MCQ Conversion and Option Policy*.

Eurorad materials use CC BY-NC-SA 4.0; MedThinkVQA follows the same license and is for research and education only, with attribution and Share-Alike, and no commercial use. We worked with Eurorad and use the materials with permission. Cases are de-identified to the best of our knowledge; we did not collect new personal data; IRB review was not required; we remove items if residual identifiers are suspected. The benchmark is not a clinical device and must not be used for diagnosis, treatment, or triage. To lower leakage risk, we release collection and filtering scripts, run de-duplication, and drop items that text-only models can solve; we also keep a path to refresh held-out items.

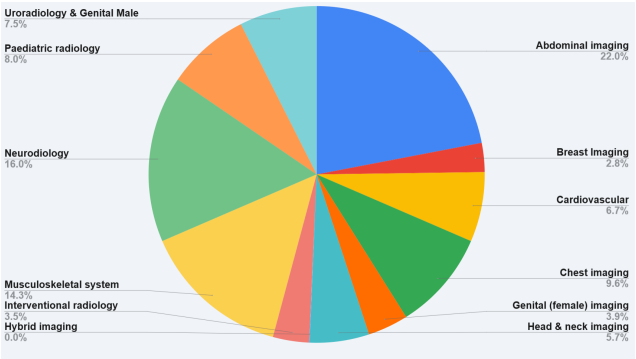

Figure 2: Distribution of radiology imaging main categories

3.2 Dataset Coverage

Task framing. We characterize dataset coverage along two orthogonal axes: (i) a *disease* axis using ICD–10 chapters, and (ii) a *radiology/medical imaging* axis grouped by anatomy and subspecialty. The ICD–10 taxonomy contains 22 chapters. Using GPT-5 to map case labels to ICD–10, our held-out test set covers **20/22** chapters and additionally includes **85** rare-disease cases aligned with *Orphanet*, providing coverage of long-tail conditions.[3]

To assess breadth from an imaging perspective, we aggregated the *full dataset* by radiology subspecialties (*anatomy & subspecialty*). Figure 2 shows the distribution. The cases are not concentrated in a single region but span across all major clinical domains. The largest share comes from *abdominal imaging* (22.0%), followed by *neuroradiology* (16.0%) and *musculoskeletal* (14.3%). Mid-sized categories include chest (9.6%), paediatric (8.0%), and urogenital imaging (7.5%), while cardiovascular (6.7%) and head & neck (5.7%) also make substantive contributions. Smaller but non-negligible proportions are represented in breast and interventional radiology, with hybrid imaging appearing only rarely (<0.1%).

3.3 MCQ Conversion and Option Policy

Each case is presented as a five-choice single-best-answer MCQ: *Given the clinical history and associated radiology images, select the most likely diagnosis from the options.* The ground-truth label is the case's *final diagnosis*. While only the clinical history and images are provided as input context for the QA task, we also retain other curated textual fields (expert caption, Integrated Imaging Summary, and Expert Reasoning & Teaching Note) in the dataset files for potential future use.

If the source differential diagnosis list has ≥ 5 candidates, we *prune* to five using a confusion-aware ranking (keep the correct answer plus four distractors that models most often confuse with the truth). If the list has < 5 candidates, we *augment* with LLM-generated distractors that meet the above rules; duplicates or contradictions are rejected.

[3]The two chapters not present in the test set are *Mental and behavioural disorders* (F01–F99) and *External causes of morbidity and mortality* (V01–Y98), which rarely appear as imaging-target diagnoses. A complete breakdown of ICD–10 chapters and subcategories is reported in the Appendix N.

Detailed statistics of the test subset, including per-sample and option-level metrics, are provided in Appendix C.

Training Set (LLM-augmented options & rationales)

When the differential diagnosis list provides fewer than five plausible options, we expand to five using a GPT-5 prompt adapted from Zuo et al. [2025] (full prompt in Appendix D). GPT-5 receives the case context (clinical history, imaging details, and current options) and proposes additional distractors with short teaching notes that explain: (i) why the distractor might seem reasonable, and (ii) what specific clue rules it out. The resulting **training set** provides five options per case, each with a teaching note.

Test Set (expert-faithful, confusion-preserving, image-dependent)

We design the test split to stay as close as possible to expert reasoning and image-based decision making:

(1) Expert differential diagnosis as starting point. We first use cases where the *expert differential* list has ≥ 5 entries. The final diagnosis serves as the key, and the differential entries form the distractor pool. This ensures all candidate options come directly from experts and filters 2061 data.

(2) Leakage Detection. To ensure the rigor of the dataset, we conducted leakage detection on each clinical history to verify whether it directly revealed the correct diagnosis. Specifically, we examined whether (i) the diagnosis label itself (exact name or ICD-standard term) appeared in the text, (ii) synonyms, abbreviations, or eponyms were explicitly present, or (iii) uncertain mentions of the label or its variants occurred (e.g., "?X," "rule out X," "suspected X," "possible X"). The detailed prompt used for this detection is provided in Appendix F. In total, 35 leaked cases were identified and removed from the dataset.

(3) Confusion-aware pruning. Moreover, if there are more than five distractors, we check which wrong answers preliminary VLM (GPT-4o) models picked mistakenly. We keep these frequently confused distractors when possible, and sample the rest at random. Only deletions are made; the original Expert Reasoning & Teaching Note is lightly edited (via GPT-5 mini) to remove references to deleted options (Appendix E). No new medical content is introduced.

(4) Remove text-solvable cases. To ensure that images are necessary, we test each provisional item with three *text-only* models—Llama-3.3-70B, Qwen-3-32B, and MedGemma-27B-text. Items that all models answer correctly in all 3 runs are removed. This step keeps only problems where imaging is essential or greatly significant. This process removes ~ 611 cases.

(5) Surface Bias Mitigation We observed a surface bias in option length: in 57% of cases the correct answer was the longest choice, far above the uniform expectation of 20%. This likely arises because correct diagnoses are phrased more specifically to a patient, while distractors are shorter and more generic. However, models achieved 5–10 points higher accuracy on such items, suggesting exploitation of this heuristic rather than genuine reasoning. To prevent shortcut learning, we randomly pruned items until the distribution was balanced (≈20%), removing 664 cases.

3.4 Medical Education Case Discussion

We observed that in the original data, some case discussions followed a clear structured format, while others did not. To facilitate evaluation of the long document generation task, we selected only those cases with a well-defined five-section structure: Background, Clinical Perspective, Imaging Perspective, Outcome, and Take-Home Messages. Within the test set, 86 cases strictly conformed to this structure, allowing section-by-section comparison in subsequent evaluation.

4 EXPERIMENTAL SETUP

4.1 Automatic evaluation

Stepwise Reasoning Evaluation We split each model explanation into atomic steps with GPT-5-MINI, then used GPT-5 as an LLM judge to label, per step: factual correctness, whether it is *critical* to the final diagnosis, and an error type when incorrect. Here we only focus on **error types**, where *Image Understanding Err* clearly dominates. Overall, most failures stem from **image misinterpretation /**

information extraction, especially on *critical* steps (69.23%). When answers are wrong, *Reasoning Err* and *Medical Knowledge Err* become more prominent alongside the image errors (details in Appendix H and Appendix L).

Case Discussion Evaluation We implemented a comprehensive automatic evaluation framework to assess the quality of generated case discussions using GPT-5 as evaluators. Each generated discussion contained multiple subsections including background, clinical perspective, imaging perspective, outcome, and take-home messages. Our evaluation employed a two-stage approach: first, we conducted sentence-level factual correctness assessment by splitting each subsection into individual sentences and tasking a prompted LLM (GPT-5) to judge the correctness of each sentence based on the provided case context, imaging findings, differential diagnosis list, image captions, and medical images. The evaluator was instructed to mark sentences as true if explicitly supported or reasonably inferable from the context, and false only if clearly contradictory or incorrect. Second, we performed quality assessment using an expert-curated rubric that scored discussions on five key criteria: disease overview, clinical pathophysiology, imaging analysis, reasoning and differentials, and transferable learning, with each criterion rated on a 0-2 scale. The LLM evaluator provided both numerical scores and brief justifications for each rubric criterion, focusing on medical accuracy, completeness, educational value, and integration of clinical and imaging perspectives. For the automatic evaluation, we randomly sampled 20 case discussions from our dataset for GPT-5 to evaluate using this framework.

4.2 Human evaluation

Stepwise Reasoning Evaluation Two medical experts evaluated 50 cases (202 steps) for step factuality and error types. In total, 44 steps contained errors (21.78%), with **Image Understanding Err** dominant (77.27%), followed by *Reasoning Err*, supporting the automatic evaluation conclusion that image misinterpretation is the primary source of mistakes. Inter-rater agreement was high, confirming the reliability of the LLM judge.

Error type	All error steps (N=1509)	Critical error steps (N=182)
Image Understanding Err	959 (63.55%)	126 (69.23%)
Reasoning Err	583 (38.63%)	71 (39.01%)
Medical Knowledge Err	362 (23.99%)	60 (32.97%)
Clinical Scenario Err	191 (12.66%)	22 (12.09%)

Table 2: LLM-judge error-type coverage. *Note:* categories are multi-label; percentages are step-level coverage over error steps and may sum to >100%. Full per-split (answer-correct vs. wrong) breakdowns are in the Appendix.

Case Discussion Evaluation To validate our automatic evaluation framework, we conducted human evaluation using two medical experts who independently assessed radiology case discussions. Each evaluator was presented with one case discussion randomly selected and generated by three different models, ensuring blinded assessment without knowledge of the generating model. Following the same two-stage methodology as the automatic evaluation, the human evaluators first performed sentence-level factual correctness evaluation and then the evaluators applied the expert-curated rubric to provide quality scores. This human evaluation served as the gold standard for assessing the reliability and validity of our automated evaluation approach.

5 Results and Discussion

5.1 Baseline Results

Table shows representative model accuracy on the held-out test set; detailed experimental settings are omitted by design. We group results into *Inference-Time Scaled Large Multimodal Models* and *Vanilla Large Multimodal Models* (all others). Strong VLMs/VLLMs remain far from expert performance, indicating MedThinkVQA's difficulty. As shown in Fig. 4, when models rely on images alone (Baseline), accuracies are MedGemma-27B: 37.5, GPT-5-nano: 39.5, GPT-5-mini: 49.4, GPT-5: 57.4. Once textual hints are added, accuracy rises sharply, showing that the main bottleneck lies in *image understanding and radiological reasoning*, rather than in language reasoning.

5.2 Image Reasoning Capabilities

Do text hints help? Yes, especially when written by experts. Across all models, expert-derived text extracted from images yields substantial accuracy gains (Fig. 4). Using `Integrated Imaging Summary (expert)` improves performance by +30.8, +30.3, +23.6, and +18.8 points

for MedGemma-27B, GPT-5-nano, GPT-5-mini, and GPT-5, respectively (relative gains of approximately +82%, +77%, +48%, and +33%). Adding `Hint` on top of `Summary` yields only marginal or modest additional gains (+0.8–+3.9 points; the largest on MedGemma-27B), suggesting that once core visual evidence is captured in structured text, language reasoning is largely sufficient; the main bottleneck remains *extracting and structuring radiological evidence from pixels*.

Findings > captions. For every model, `Integrated Imaging Summary (expert)` outperforms `Image Hint (expert)` by +7.3, +5.4, +7.6, and +2.4 points (for MedGemma-27B and GPT-5 series, respectively). Diagnosis-oriented summaries encode discriminative cues (laterality, location, pattern, extent), offering stronger signals for QA than caption-like descriptions.

Who benefits most? The weaker baselines. Relative improvements from expert text remain inversely correlated with baseline performance: the `Summary (expert)` / baseline ratios are 1.82×, 1.77×, 1.48×, and 1.33× for MedGemma, GPT-5-nano, GPT-5-mini, and GPT-5, respectively. This indicates that once visual evidence is verbalized, language inference is no longer the limiting factor—visual understanding is.

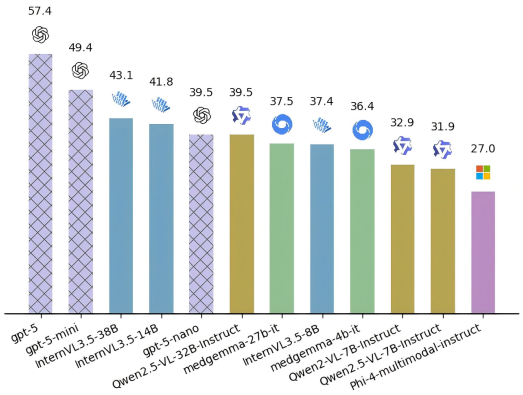

Figure 3: Baseline model accuracy Google DeepMind and Google Health AI [2025], Sellergren et al. [2025], Wang et al. [2024], Bai et al. [2025], Abouelenin et al. [2025], OpenAI [2025a,b], Wang et al. [2025a]

Self-generated text: modest or negative impact. When models first generate their own `Hint`/`Summary` and then reuse it for QA, effects are small and often negative: MedGemma-27B decreases slightly by −1.6, −0.9, −1.0; GPT-5-nano exhibits small gains of +3.6, +1.6, +3.5; GPT-5-mini drops by −4.8, −2.1, and is near baseline with `Both` (−0.1); GPT-5 shows mixed outcomes: −3.9, −1.6, and a slight +0.5 with `Both`. Tab. 3 sheds light: ROUGE-L and RadCliQ-v1 scores for Image→Caption/Findings remain low (≈ 0.13–0.16), meaning self-generated texts often omit decisive clinical details and can introduce noise that distracts the QA stage.

Model	ROUGE-L (↑)		RadCliQ-v1 (↑)	
	Caption	Findings	Caption	Findings
gpt-5-nano	0.1435	0.1585	0.8080	0.6781
gpt-5-mini	0.1510	0.1636	0.8317	0.6931
GPT-5	0.1534	0.16272	0.8341	0.6818
medgemma-27b-it	0.1336	0.1621	0.7810	0.7192

Table 3: Scores of VLMs for Image→Caption and Image→Findings across two metrics (ROUGE-L and RadCliQ).

Why can self-text underperform? We highlight three factors. **(i) Content quality.** Low ROUGE-L (≈ 0.13–0.16) and RadCliQ-v1 shows that self-generated texts often miss laterality, precise location (e.g., "left lung" vs. "left *upper lobe*"), or key patterns. Even small inaccuracies can mislead the QA stage, explaining MedGemma's consistent drops and the mixed results on larger models. **(ii) Token budget and attention dilution.** Multi-view images combined with noisy text increase sequence length; extra tokens draw attention away from informative visual features. Some smaller models (e.g., **GPT-5-nano**) still benefit from a text scaffold, whereas others (e.g., **GPT-5-mini**) can be distracted and lose accuracy. **(iii) Weak image–text grounding.** Current VLMs may over-trust provided text, even when it conflicts with images. Without strong grounding, the model can follow noisy hints rather than pixel evidence, harming accuracy.

Implications. (1) *MedThinkVQA* mainly tests *image reasoning*, with expert summaries yielding large gains.
(2) Further progress should focus on stronger visual encoders, better image–text grounding, and concise, structured hints.

SFT results. The supervised fine-tuning results demonstrate substantial performance improvements for our fine-tuned models when compared to baseline. As shown in Tab. 8, while the GPT-5 series achieved strong baseline performance with GPT-5 at 57.39%, our fine-tuned models showed

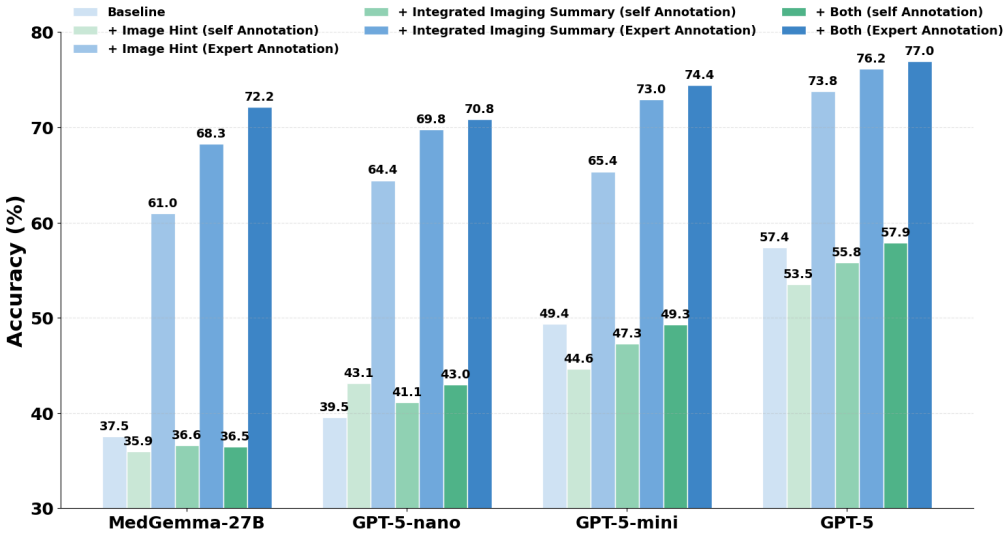

Figure 4: Accuracy on *MedThinkVQA* when augmenting images with text. We compare `Image Hint` (caption-like) and `Integrated Imaging Summary` (diagnosis-oriented findings), each provided either by an *expert* or generated by the *model itself* (self). `Both` combines the two.

competitive or superior results. Notably, our Qwen2.5-VL-7B-Instruct improved dramatically from 31.95% baseline to 61.89%, surpassing even GPT-5's performance. Similarly, InternVL3.5-4B (60.96%) and Qwen2.5-VL-3B-Instruct (60.03%) achieved accuracies comparable to GPT-5, while MedGemma-4B-IT improved from 36.35% to 56.57%. These results indicate that our curated dataset contains high-quality, well-structured training examples that effectively enhance the models' medical reasoning capabilities, enabling smaller fine-tuned models to achieve performance competitive with larger inference-time scaled models like GPT-5. A detailed summary of hyperparameters can be found in Appendix B.

5.3 Medical Education Case Discussion

The generated case discussions demonstrated high factual accuracy across all tested models, with overall correctness rates ranging from 92.81% to 99.22% shown in Tab. 9. The GPT-5 series consistently achieved the highest factual correctness, while the Clinical Perspective subsection scored highest across all models (97.89-100%). The Outcome subsection showed some performance differences, with MedGemma-27B achieving 85.71% compared to other models' which scored above 95%. The rubric-based evaluation revealed GPT-5 achieving the highest overall score of 9.9/10. MedGemma-27B scored 7.05/10, showing particular weakness in clinical pathophysiology (1.15/2) and reasoning differentials (1.1/2), while all models demonstrated consistent strength in disease overview and imaging findings (Tab. 10).

5.4 Data Contamination Analysis

We assess potential test leakage with a strict, sliding-window variant of MELD (Memorization Effects Levenshtein Detector), which measures the character-level overlap between each model's generated answer and its input question on the MEDTHINKVQA test set. Across seven representative LLM/VLMs (Qwen3-32B, Med-Gemma-27B-*it*, Med-Gemma-27B-*text-it*, GPT-4.1-nano, GPT-4.1-mini, Qwen2.5-VL-72B-Instruct, Llama-3.3-70B-Instruct), MELD similarities cluster around $\sim$20–24% with narrow IQRs, and *no* item reaches the commonly used high-risk threshold of $\geq 50\%$. Distributions are similar for text-only and vision-language models, indicating no family-specific effect. Taken together, we find no evidence of severe contamination; details and boxplots appear in Appendix M.

6 Conclusion

MedThinkVQA establishes the first large-scale benchmark for multimodal diagnostic reasoning in radiology, combining authentic multi-image cases with expert-authored reasoning traces. We hope it will serve as a rigorous testbed to advance models that can not only answer correctly but also reason like radiologists, ultimately driving progress toward trustworthy clinical AI.

References

Eurorad – radiology teaching cases. https://www.eurorad.org/. European Society of Radiology, accessed 2025-08-28.

Abdelrahman Abouelenin et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025. doi: 10.48550/arXiv.2503.01743. URL https://arxiv.org/abs/2503.01743.

Shuai Bai et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. doi: 10.48550/arXiv.2502.13923. URL https://arxiv.org/abs/2502.13923.

Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019, 2019.

Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. Medrax: Medical reasoning agent for chest x-ray. *arXiv preprint arXiv:2502.02673*, 2025. URL https://arxiv.org/abs/2502.02673.

Lin Fan, Xun Gong, Cenyang Zheng, Xuli Tan, Jiao Li, and Yafei Ou. Cycle-vqa: A cycle-consistent framework for robust medical visual question answering. *Pattern Recognition*, 165:111609, 2025.

Xiaotang Gai, Chenyi Zhou, Jiaxiang Liu, Yang Feng, Jian Wu, and Zuozhu Liu. Medthink: A rationale-guided framework for explaining medical visual question answering. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7438–7450, 2025.

Google DeepMind and Google Health AI. Medgemma model card, 2025. URL https://developers.google.com/health-ai-developer-foundations/medgemma/model-card.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada, Ronald M. Summers, and Yingying Zhu. Medical-diff-vqa: A large-scale medical dataset for difference visual question answering on chest x-ray images. PhysioNet v1.0.1, 2025. URL https://physionet.org/content/medical-diff-vqa/1.0.1/.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183, 2024.

Di Jin, Edward Pan, et al. What disease does this patient have? a large-scale open-domain medical qa dataset. In *EMNLP*, 2021. MedQA (USMLE).

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.

Su Hwan Kim, Severin Schramm, Lisa C Adams, Rickmer Braren, Keno K Bressem, Matthias Keicher, Paul-Sören Platzek, Karolin Johanna Paprottka, Claus Zimmer, Dennis M Hedderich, et al. Benchmarking the diagnostic performance of open source llms in 1933 eurorad case reports. *npj Digital Medicine*, 8(1):97, 2025.

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):1–10, 2018.

Khai Le-Duc, Duy M. H. Nguyen, Phuong T. H. Trinh, Tien-Phat Nguyen, Nghiem T. Diep, An Ngo, Tung Vu, Trinh Vuong, Anh-Tien Nguyen, Mau Nguyen, Van Trung Hoang, Khai-Nguyen Nguyen, Hy Nguyen, Chris Ngo, Anji Liu, Nhat Ho, Anne-Christin Hauschild, Khanh Xuan Nguyen, Thanh Nguyen-Tang, Pengtao Xie, Daniel Sonntag, James Zou, Mathias Niepert, and Anh Totti Nguyen. S-chain: Structured visual chain-of-thought for medicine. *arXiv preprint arXiv:2510.22728*, 2025. URL https://arxiv.org/abs/2510.22728.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE, 2021.

Bo Liu, Ke Zou, Liming Zhan, Zexin Lu, Xiaoyu Dong, Yidi Chen, Chengqiang Xie, Jiannong Cao, Xiao-Ming Wu, and Huazhu Fu. Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis. *arXiv preprint arXiv:2411.16778*, 2025a. URL https://arxiv.org/abs/2411.16778.

Bo Liu et al. Gemex-thinkvg: Towards thinking with visual grounding in medical vqa via reinforcement learning. *arXiv preprint arXiv:2506.17939*, 2025b. URL https://arxiv.org/abs/2506.17939.

Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Tianyi Zhou, and Zuozhu Liu. Medcot: Medical chain of thought via hierarchical expert. *arXiv preprint arXiv:2412.13736*, 2024.

OpenAI. Gpt-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf, 2025a. Version dated 2025-08-13.

OpenAI. gpt-5-mini model card, 2025b. URL https://platform.openai.com/docs/models/gpt-5-mini.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and Jean-Benoit Delbrouck. GREEN: Generative radiology report evaluation and error notation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 374–390, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.21. URL https://aclanthology.org/2024.findings-emnlp.21/.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR, 2022.

Andrew Sellergren et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025. doi: 10.48550/arXiv.2507.05201. URL https://arxiv.org/abs/2507.05201.

Xiangru Tang, Daniel Shao, Jiwoong Sohn, Jiapeng Chen, Jiayi Zhang, Jinyu Xiang, Fang Wu, Yilun Zhao, Chenglin Wu, Wenqi Shi, Arman Cohan, and Mark Gerstein. MedAgentsBench: Benchmarking Thinking Models and Agent Frameworks for Complex Medical Reasoning, 2025. URL https://arxiv.org/abs/2503.07459.

Peng Wang et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. doi: 10.48550/arXiv.2409.12191. URL https://arxiv.org/abs/2409.12191.

Shansong Wang, Mingzhe Hu, Qiang Li, Mojtaba Safari, and Xiaofeng Yang. Capabilities of GPT-5 on Multimodal Medical Reasoning, 2025a. URL https://arxiv.org/abs/2508.08224.

Yuan Wang, Jiaxiang Liu, Shujian Gao, Bin Feng, Zhihang Tang, Xiaotang Gai, Jian Wu, and Zuozhu Liu. V2t-cot: From vision to text chain-of-thought for medical reasoning and diagnosis. *arXiv preprint arXiv:2506.19610*, 2025b.

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

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9):100802, 2023. doi: 10.1016/j.patter.2023.100802. URL https://doi.org/10.1016/j.patter.2023.100802.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024a.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024b.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.

Ethical Statement

Data source, licensing, and legal compliance. All cases are adapted from *Eurorad*, a peer-reviewed educational database maintained by the European Society of Radiology. Eurorad materials are licensed under *Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license*. MedThinkVQA follows the same license. Released data are for research and education only; commercial use is prohibited. Derivative datasets must preserve attribution, non-commercial use, and ShareAlike terms.

Human subjects and privacy. Eurorad cases are intended for education and are de-identified to the best of our knowledge. We did not collect new personal data and did not recruit patients or lay participants; IRB review was not required. We reviewed materials for residual identifiers and removed items when concerns arose.

Evaluation reliability. We combine automatic scripts, expert review, and LLM-judges. On step-level labels, human–human agreement is Cohen's $\kappa = 0.822833$, human1–LLM-judge agreement is $\kappa = 0.838357$, and human2–LLM-judge agreement is $\kappa = 0.701566$. These results support the stability of our automated judging, but LLM-judges do not replace expert oversight.

Bias and fairness. Educational repositories can encode geographic, demographic, and practice-style biases. Rare conditions and certain protocols are unevenly represented. Models trained or tuned on this benchmark may inherit such biases. We encourage stratified analyses and external validation before any deployment.

Safety and misuse. Models evaluated here are research artifacts. They must *not* be used for diagnosis, treatment, triage, or other high-stakes tasks without added clinical validation, regulatory clearance, and domain oversight. Generated discussions may sound authoritative yet still be incomplete or wrong. Any downstream use requires human supervision, documented fail-safes, and monitoring.

Transparency, reproducibility, and environment. We document data construction, metrics, and judging protocols. We release code, scoring scripts, and example data, subject to third-party licenses. No hidden reward models, private test sets, or special samplers were used. We report hardware and runtime where relevant and encourage efficient evaluation to limit environmental impact.

Conflicts of interest and ethics compliance. All authors have read and will adhere to the ICLR Code of Ethics for submission, reviewing, and discussion. Any sponsorships or competing interests will be disclosed in the author checklist.

Data leakage assessment and mitigation. As discussed in Section 5.4, we conducted internal checks for leakage and found no obvious overlap between our test items and publicly released training artifacts that we were aware of. We remove text-only solvable items, strip explicit textual shortcuts, and stress cross-image fusion. Still, the risk of leakage cannot be ruled out. To reduce risk further, we will (i) release the full data collection and processing code for public audit, and (ii) maintain a rolling test set covering the most recent 6–12 months of newly curated cases, with periodic updates and refreshed scores for reported models. We will also publish de-duplication scripts (exact/near-duplicate filters on images and texts) and document all split procedures.

Others. MedThinkVQA is a research benchmark, not a clinical tool. Expert-authored traces are pedagogical; they may overlook interpersonal nuances, local workflows, and institutional contexts. The multiple-choice setting enables standardized scoring; it also simplifies real diagnostic work and stops before treatment planning and longitudinal follow-up. Coverage is broad but not complete across body regions, patient groups, vendors, devices, and acquisition protocols. Although cases span many conditions, some specialties (e.g., pediatrics, psychiatry) and rare diseases remain underrepresented. All cases originate from a single educational repository, so distribution shifts across hospitals, populations, and imaging pipelines are likely. The dataset is currently English-only; multilingual generalization has not been tested. Annotations, while expert-written, can still contain noise or stylistic variation. Our LLM-as-Judge components improve scalability, but they can be prompt-sensitive and may reflect judge-model biases; we therefore report human agreement and keep experts informed. Finally, we evaluate stepwise reasoning for differential diagnosis; reference-free evaluation of clinical reasoning without ground-truth steps is left for future work.

A Model Baseline

We establish baselines using a diverse set of vision–language models (VLMs) to ensure fair and representative evaluation. The selection spans both *Inference-Time Scaled Large Multimodal Models* (e.g., GPT-5 family with nano/mini/full variants) and *Vanilla Large Multimodal Models*, which include open-weight generalist and medical-tuned families such as Qwen2-VL, Qwen2.5-VL, MedGemma, Phi, and InternVL at different parameter scales (4B–38B).

B Supervised Fine-tuning

Supervised Fine-tuning Configuration: We fine-tuned the InternVL3.5-1B, InternVL3.5-2B, InternVL3.5-4B, MedGemma-4B-IT, Qwen2.5-VL-3B-Instruct, and Qwen2.5-VL-7B-Instruct models using QLoRA (Quantized Low-Rank Adaptation). The LoRA configuration employed a rank of 8, alpha value of 16, and dropout rate of 0.05. Training was conducted for 2 epochs with a batch size of 1 per device and gradient accumulation steps of 8, resulting in an effective batch size of 8. We used the AdamW optimizer with a learning rate of 2×10^{-4}, cosine learning rate scheduling, and a 0.03 warmup ratio. The dataset was split 90/10 for training and validation.

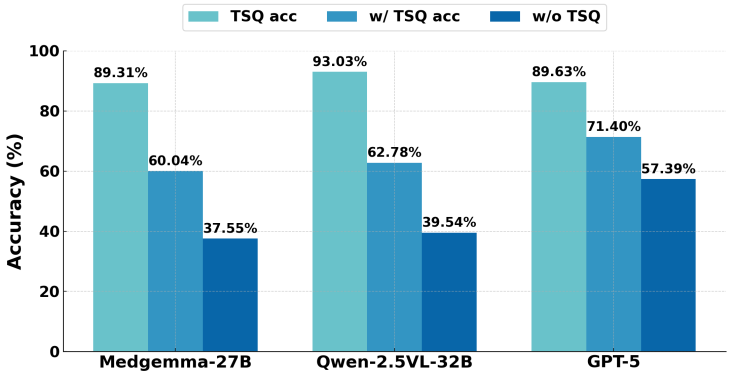

Figure 5: model accuracy across three processed datasets. **TSQ** refers to *Text-Solvable Questions*. The **TSQ acc** corresponds to model performance on the 611 text-solvable cases, where all three models achieved accuracies above 89%. In contrast, the **w/o TSQ** results are computed on the final test set after removing these text-solvable cases, showing a substantial drop in accuracy.

C Dataset Statistics

Overall	
Samples	751
Images	6090
Per-sample	
Imgs/sample	8.11
Cap. length	2444.0
Find. length	857.7
Disc. length	2543.9
Option Length	
Avg.	27.9
Num. Density	
Macro avg.	0.0164
Other	
Pos. correct	2.88
Mean mod. cnt	2.56
All mod. types	12

Table 4: Detailed test statistics (*Cap/Find/Disc.* = caption, findings, discussion; *Pos. correct* = average position of correct option; *Mean mod. cnt* = mean number of imaging modalities).

D Option & Discussion Augmentation Prompt

To ensure reproducibility, we document the exact prompts used for augmenting *Options* and expanding the *Discussion* in the medical multiple-choice QA setting.

D.1 System Prompt

```
You are a careful medical QA assistant.

# Prompt for Option Generation

### Task
Given a medical multiple-choice question of the form "Select the single best diagnosis"
based on CLINICAL_HISTORY, several patient images, the current provided options,
the correct answer, and an existing discussion (including reasoning about the current options), please:

1. Generate additional incorrect options so that the total number of answer choices
   is exactly 5 (no more, no less).
2. Expand and refine the provided discussion, ensuring it thoroughly explains how
   to eliminate all incorrect answers and why the correct answer is most appropriate,
   using reasoning grounded in the CLINICAL_HISTORY and images.

### Suggested Approaches
1. Consider Erroneous Perspectives: Add distractors that misinterpret or
   overemphasize aspects of the CLINICAL_HISTORY or images.
2. Leverage Common Misconceptions: Create distractors based on common diagnostic
   errors or frequently confused conditions.
3. Logical Misdirection: Introduce distractors grounded in logical reasoning
   that appear plausible but are ultimately incorrect.

### General Requirements
1. Maintain Consistency: Ensure new options match the original ones in length,
   structure, and professional wording.
2. Avoid Oversimplified Distractors.
3. Ensure High Plausibility.
4. Expand Discussion:
   - Include reasoning for the newly generated distractors.
   - Strengthen explanations for ruling out incorrect answers.
   - Deepen justification for selecting the correct answer.

5. Final Output Format:
   Return valid JSON with exactly these fields: options (A-E), correct_answer, discussion.
```

```
### Important Output Rules
- Keep all *original* options text unchanged; only add new distractors
  to reach exactly five total options.
- Do NOT reorder existing options; append only the missing letters
  (e.g., add D/E) so that A-E are filled.
- The final correct_answer must correspond to the original correct option's text.
- No extra commentary outside the JSON body.
```

E Discussion Pruning Prompt

This section documents the prompts used to prune *Discussion* paragraphs by removing references to extra differential diagnoses that are not among the allowed answer options.

E.1 System Prompt

```
You are a careful clinical editor. Your job is to MINIMALLY edit a medical DISCUSSION.
Goal: remove references to extra differential diagnoses that appear in
DIF_DIAGNOSIS_LIST but are NOT among the five ALLOWED OPTIONS.
Preserve all content related to ALLOWED OPTIONS.
Keep the original clinical reasoning flow, tone, and meaning. Do not add new facts.

Rules:
1) NEVER delete information that relates to any ALLOWED_OPTIONS
   (even if an EXTRA item partially overlaps).
2) Remove sentences/clauses whose main role is to introduce, justify, or
   list items in EXTRA_TO_REMOVE.
   If a sentence mixes allowed and extra diagnoses, keep the allowed part
   and delete only the extra part, then fix grammar to remain fluent.
3) Keep general disease definitions, imaging/lab reasoning, and conclusions
   that support ALLOWED_OPTIONS.
4) Maintain coherence and clinical correctness; do NOT invent new claims.
5) Output strictly as JSON with one key: discussion_new.
6) If EXTRA_TO_REMOVE is empty, return the original discussion as discussion_new.
```

E.2 User Prompt Template

```
Edit the DISCUSSION by deleting only the parts about the extra differentials.

ALLOWED_OPTIONS (keep anything related to these):
<ALLOWED_OPTIONS_JSON>

DIF_DIAGNOSIS_LIST_CLEAN:
<DIF_DIAGNOSIS_LIST_CLEAN_JSON>

EXTRA_TO_REMOVE (delete content only about these):
<EXTRA_TO_REMOVE_JSON>

DISCUSSION:
‘‘‘text
<DISCUSSION>
Return JSON: {"discussion_new": "..."}
```

F Prompts for Data Leakage Auditing

System Message

```
You are a meticulous clinical QA auditor for multiple-choice diagnosis questions. Your job:
Given ONLY the CLINICAL HISTORY text and the list of candidate diagnosis OPTIONS, decide
whether the history text DIRECTLY REVEALS any option(s).

Definition of DIRECT REVEAL (diagnosis label appears in the text itself, not inferred):
• L3 Explicit label: the exact diagnosis name or ICD/standard label appears, or patterns like
  "Diagnosis: X", "biopsy-proven X".
• L2 Explicit synonym/acronym/eponym/foreign-language variant of the diagnosis label appears
  (e.g., "MI" for myocardial infarction; "Osler-Weber-Rendu" for HHT).
• L1 Explicit but uncertain mention of the diagnosis label (or its synonym/acronym/eponym):
  e.g., "?X", "r/o X", "rule out X", "query X", "suspected X", "possible/probable X",
  "consistent with X", "concern for X", "Hx of/known case of X".

NOT a leak: symptoms, signs, risk factors, imaging descriptors, or lab patterns that merely
SUGGEST a diagnosis. Only mark a leak if the diagnosis LABEL itself (or its standard
synonym/acronym/eponym) occurs in the text.

Use the OPTIONS solely as a dictionary of candidate labels and their widely-used
synonyms/acronyms/eponyms to search for DIRECT textual mentions. Do NOT infer diagnoses
from context. Do NOT mark based on reasoning.

For each leaked option, return:
- option_id, option_text
- overall leak_level (max severity across its evidences; L3>L2>L1)
- evidences: verbatim snippet(s) with [start,end) character indices into the EXACT Clinical
  history string
- a brief justification

If no option is leaked, set has_leak=false and provide non_leak_reason.

Return ONLY valid JSON following the required schema. No extra prose.
```

User Message (Template)

```
CLINICAL HISTORY (use this exact string when computing char spans):
<<<<HISTORY>>>>
{CLINICAL_HISTORY}
<<<<END_HISTORY>>>>

OPTIONS (candidate diagnoses; DO NOT infer-use only as label dictionary):
A) {option_A_text}
B) {option_B_text}
C) {option_C_text}
D) {option_D_text}
E) {option_E_text}
... (continue as needed, preserving order)

Task: Identify ALL options (if any) that are directly revealed by the HISTORY text
under L1/L2/L3 definitions. Extract verbatim evidence snippet(s) and 0-based [start,end)
char spans into the exact HISTORY string above. If none, set has_leak=false.
```

G Prompts for Discussion Generation

System Prompts

```
You are a board-certified radiologist. Given clinical history, imaging
findings, a differential diagnosis list, the final diagnosis, and one or
more images (with captions), write a Discussion with five sections:
Background; Clinical Perspective; Imaging Perspective; Outcome; Take Home
Message. Be accurate, concise, and grounded in the provided info.

Return strict JSON with keys exactly:
{
  "Background": "...",
  "Clinical Perspective": "...",
  "Imaging Perspective": "...",
  "Outcome": "...",
  "Take Home Message": "..."
}

Example of tone/structure (content is just an example; DO NOT copy text):

{
  "Background": "May and Thurner described for the first time in 1956
  a spur-like formation on the left common iliac vein in 22% of autopsies.
  May-Thurner syndrome, also known as Iliac Venous Compression Syndrome
  (IVCS), is a condition of venous compression by the overlying artery,
  usually the left common iliac vein by the right common iliac artery.",

  "Clinical Perspective": "This disease is reported to be more frequent
  in women and the main clinical presentation is deep vein thrombosis.
  The true prevalence of this condition is unknown, but some autopsies
  series reported 22% to 33%. May-Thurner syndrome is a progressive
  vascular disease with long-term disabling complications.",

  "Imaging Perspective": "Iliac vein compression, with or without
  thrombosis, should be treated if symptomatic. The procedure includes
  an ascending venogram through the iliac vein to show the stenotic area.
  A guidewire is advanced through the lesion and a stent is than placed
  over-the-wire.",

  "Outcome": "Since 1995 venous stents have been placed into the narrowed
  vein area. Stents seem to be beneficial, improving the clinical outcome
  and the quality of life of these patients.",

  "Take Home Message": "If a patient has discomfort, swelling or deep
  venous thrombosis (DVT), in the iliofemoral vein territory, especially
  on the left side think about May-Thurner syndrome."
}
```

H LLM Judge Prompt

H.1 System Prompt

```
You are an evaluator for radiology case analyses. Judge the correctness of each step
based on the provided context (Clinical history, Captions, Imaging findings, Discussion)
and relevant teaching value/domain knowledge.
Rules:
1) Evaluate whether each step is correct or reasonably supported; reasonable analysis counts as c
2) Mark True if the step is explicitly supported, correctly implied, or logically reasonable give
```

```
   and your teaching value/domain knowledge.
3) Mark False only if the step is clearly wrong, contradictory, or cannot be reasonably inferred
   the context or standard domain knowledge.
4) Ignore style, redundancy, or reasoning quality-focus only on correctness.
5) Provide exactly one concise 1-2 sentence explanation per step.
6) Return ONLY JSON following the provided schema; one verdict per step, same order.
```

H.2 User Prompt (Template)

```
Task: For each step below, judge if it is supported by the provided context and relevant
teaching value/domain knowledge.

- Title: {{title}}
- Clinical history: {{clinical_history}}
- Imaging findings: {{imaging_findings}}
- Discussion: {{discussion}}
- Captions (all):
{{captions_block}}   # e.g., lines like "- {{caption_i}}"; if none, use "(none)"

Steps to judge (in order):
{{steps_block}}      # e.g., "1. {{step_1}}\n2. {{step_2}}\n..."

Output strictly as JSON; one verdict per step in the same order, using this schema:
{
  "verdicts": [
    {
      "is_factual": true,
      "explanation": "A brief, self-contained justification (1-2 sentences). If true, mention sup
    }
    // ... one object per step, in order
  ]
}
```

H.3 LLM as Judge for Case Discussions

```
You are a board-certified radiologist tasked with evaluating the factual
correctness of radiology case discussions.

Judge the correctness of each sentence from the Discussion section
(Background / Clinical Perspective / Imaging Perspective / Outcome /
Take-Home) based on the provided case context (Clinical history, Imaging
findings, Differential list), the image captions, and the images themselves.

Rules:
1) Mark True if the sentence is explicitly supported, correctly implied,
   or logically reasonable given the context and standard domain knowledge.
2) Mark False only if clearly wrong, contradictory, or not reasonably
   inferable.
3) Ignore style and redundancy-focus only on correctness.
4) Provide exactly one concise 1-2 sentence explanation per sentence.
5) Return ONLY JSON for the schema below.

Return STRICT JSON with this schema:
{
  "sentence_judgments": {
    "<sentence_key>": {
      "text": "<original sentence>",
      "factual": true|false,
      "explanation": "<ONE concise 1-2 sentence explanation>"
```

```
    }
  }
}
```

H.4 Rubric Evaluation Prompt

```
You are a board-certified radiologist tasked with evaluating the quality
of radiology case discussions.

TASK: Evaluate the Discussion section of the provided radiology case
using a standardized rubric.

MATERIALS PROVIDED:
- Clinical history and imaging findings
- Differential diagnosis list
- Medical images with captions
- Discussion section (containing: Background, Clinical perspective,
  Imaging perspective, Outcome, Take-Home messages)

EVALUATION INSTRUCTIONS:
1. Read the entire Discussion section carefully
2. Score each of the 5 rubric criteria on a 0-2 scale.
3. For each rubric score, provide a brief 1-2 sentence justification
4. Calculate total score (sum of all 5 rubrics, range 0-10)

FOCUS ON:
- Medical accuracy and evidence-based content
- Completeness of information
- Educational value for radiology trainees
- Clear communication of key concepts
- Integration of clinical and imaging perspectives

OUTPUT FORMAT:
Return ONLY a valid JSON object following the specified schema.
Do not include any additional text or explanations outside the
JSON structure.

Return STRICT JSON with this schema:
{
  "rubric_scores": {
    "rubric_1_disease_overview": {"score": 0|1|2, "explanation": "<1-2 sentences>"},
    "rubric_2_clinical_pathophysiology": {"score": 0|1|2, "explanation": "<1-2 sentences>"},
    "rubric_3_imaging": {"score": 0|1|2, "explanation": "<1-2 sentences>"},
    "rubric_4_reasoning_differentials": {"score": 0|1|2, "explanation": "<1-2 sentences>"},
    "rubric_5_transferable_learning": {"score": 0|1|2, "explanation": "<1-2 sentences>"},
    "total": 0-10
  }
}
```

I Additional Evaluation Tables for Text-Solvable Cases1

All results below are evaluated on the same **raw test set of 2,159 items** . For each model we perform three independent runs using the same evaluation protocol and report per-run accuracy (*Correct/Total*), along with the *joint-correct* statistic—i.e., the size of the intersection of items answered correctly by *all three runs* of the same model. Small variations across runs are expected due to non-determinism in decoding. Where the third-run line is not available in the input data, we report the provided runs and the reported joint-correct number as-is.

Table 5: Llama-3.3-70B-Instruct: per-run and joint-correct results on the 2,159-item raw test set.

Run	Total	Correct	Accuracy
Run 1	2,159	1,199	0.555 (55.53%)
Run 2	2,159	1,207	0.559 (55.91%)
Run 3	2,159	1,197	0.554 (55.44%)
Joint-correct	2,159	1,172	0.543 (54.28%)

Mean across 3 runs: 55.63% $\pm$ 0.25 (std. dev., in percentage points).

Table 6: medgemma-27b-text-it: per-run and joint-correct results on the 2,159-item raw test set.

Run	Total	Correct	Accuracy
Run 1	2,159	1,236	0.572 (57.25%)
Run 2	2,159	1,212	0.561 (56.14%)
Run 3	2,159	1,213	0.562 (56.18%)
Joint-correct	2,159	975	0.452 (45.16%)

Mean across 3 runs: 56.52% $\pm$ 0.63 (std. dev., in percentage points).

Table 7: Qwen3-32B: per-run and joint-correct results on the 2,159-item raw test set.

Run	Total	Correct	Accuracy
Run 1	2,159	1,193	0.553 (55.26%)
Run 2	2,159	1,184	0.548 (54.84%)
Run 3	2,159	1,183	0.548 (54.79%)
Joint-correct	2,159	1,118	0.518 (51.78%)

Mean across 3 runs: 54.96% $\pm$ 0.26 (std. dev., in percentage points).

J Test Data Modalities

The test set encompasses a broad spectrum of imaging modalities commonly used in clinical radiology and medical practice. Specifically, it includes: X-ray, fluoroscopy, computed tomography (CT), magnetic resonance imaging (MRI), ultrasound, positron emission tomography (PET), single-photon emission computed tomography (SPECT), nuclear medicine studies, mammography, angiography, endoscopy, and echocardiography. This diversity ensures that the evaluation captures performance across both routine and specialized imaging techniques.

K Prompts for Stepwise Explanation Extraction

K.1 System Prompt

```
You are a meticulous clinical reasoning editor. Convert a given explanation paragraph
into an ordered list of numbered steps that preserves the original meaning and evidence.
Rules:
1) Preserve content: do NOT introduce facts not present in the explanation.
2) Decompose into atomic inferences or observations -- each step one concise sentence
   (<= ~30 words).
3) Order steps to reflect the reasoning flow (e.g., findings -> interpretation -> decision).
4) Rewrite references like 'option A/B/C' into plain statements; avoid option letters.
5) If the explanation contrasts entities (e.g., 'X not Y'), separate them into distinct steps.
6) Use the same language as the explanation text (typically English).
7) If the explanation is very short, return a single clear step.
Return ONLY the JSON that matches the provided schema.
```

K.2 User Prompt (Template)

```
Task: Convert the following explanation into an ordered list of steps.

Context (for referent clarity only - do NOT add facts not present in the explanation):
- Title: {title}
- Clinical history: {clinical_history}
- Imaging findings: {imaging_findings}

Explanation to convert (source of truth):
<<<
{explanation}
>>>

Output strictly as JSON following the schema (no extra text).
```

Model	Accuracy (%)
InternVL3.5-1B	43.96
InternVL3.5-2B	58.96
InternVL3.5-4B	60.96
MedGemma-4B-it	56.57
Qwen2.5-VL-3B-Instruct	60.03
Qwen2.5-VL-7B-Instruct	61.89

Table 8: Supervised fine-tuning results on the 751-item test set.

Model	Background (%)	Clinical (%)	Imaging (%)	Outcome (%)	Take-Home (%)	Overall (%)
gpt-5	100.0	100.0	97.81	98.70	100.0	99.08
gpt-5-mini	98.59	98.65	99.10	100.0	100.0	99.22
gpt-5-nano	97.87	98.99	97.39	95.89	98.46	97.76
medgemma-27b-it	89.0	97.89	94.93	85.71	93.65	92.81

Table 9: Sentence-level factual correctness evaluation across discussion subsections

Model	Total	Disease Overview	Clinical Pathophys.	Imaging	Reasoning Different.	Transfer Learning
gpt-5	9.9	2.0	1.9	2.0	2.0	2.0
gpt-5-mini	9.4	1.95	1.6	2.0	1.85	2.0
gpt-5-nano	8.4	1.7	1.25	2.0	1.45	2.0
medgemma-27b-it	7.05	1.4	1.15	1.85	1.1	1.55

Table 10: Rubric evaluation scores across different models

Pairwise comparison	Cohen's κ
Expert 1 vs. Expert 2	0.822833
Expert 1 vs. LLM judge	0.838357
Expert 2 vs. LLM judge	0.701566

Table 11: Inter-rater reliability on step factuality (Cohen's κ). High agreement with Expert 1 and substantial agreement with Expert 2 support the reliability of the LLM judge.

L LLM Judge Stats

GPT-5 was evaluated on the **entire** test set, whereas the other three models were evaluated on a **random sample of 100** test cases due to cost and time constraints. *Error-type coverage is computed over erroneous steps; since a step may bear multiple error labels, the percentages can exceed* 100%.

L.1 GPT-5 (Full Test Set with 6,425 steps)

Correctly answered (*is_correct=True*).

- Steps (with valid `is_factual`): **3,903**
- Step factual accuracy: **3311/3903** (**84.83**%)
- Critical steps: **1,264**
- Critical-step factual accuracy: **1212/1264** (**95.89**%)
- Erroneous steps (all): **592**
- Error-type coverage (among erroneous steps):
  - Reasoning Err: **167/592** (28.21%)
  - Image Understanding Err: **374/592** (63.18%)
  - Clinical Scenario Err: **53/592** (8.95%)
  - Medical Knowledge Err: **91/592** (15.37%)
  - Other/Unspecified: **60/592** (10.14%)
- Erroneous *critical* steps only: **52**
- Error-type coverage (among erroneous critical steps):
  - Reasoning Err: **14/52** (26.92%)
  - Image Understanding Err: **37/52** (71.15%)
  - Clinical Scenario Err: **6/52** (11.54%)
  - Medical Knowledge Err: **11/52** (21.15%)

Incorrectly answered (*is_correct=False*).

- Steps (with valid `is_factual`): **2,522**
- Step factual accuracy: **1605/2522** (**63.64**%)
- Critical steps: **520**
- Critical-step factual accuracy: **390/520** (**75.00**%)
- Erroneous steps (all): **917**
- Error-type coverage (among erroneous steps):
  - Reasoning Err: **416/917** (45.37%)
  - Image Understanding Err: **585/917** (63.79%)
  - Clinical Scenario Err: **138/917** (15.05%)
  - Medical Knowledge Err: **271/917** (29.55%)
  - Other/Unspecified: **9/917** (0.98%)
- Erroneous *critical* steps only: **130**

- Error-type coverage (among erroneous critical steps):
  - Reasoning Err: **57/130** (43.85%)
  - Image Understanding Err: **89/130** (68.46%)
  - Clinical Scenario Err: **16/130** (12.31%)
  - Medical Knowledge Err: **49/130** (37.69%)

L.2 `InternVL3_5-14B_100_sample` (100-Sample Subset)

Overall. Total number of steps (all samples): **607**.

Correctly answered (*is_correct=True*).

- Steps (with valid `is_factual`): **247**
- Step factual accuracy: **189/247** (**76.52**%)
- Critical steps: **91**
- Critical-step factual accuracy: **88/91** (**96.70**%)
- Erroneous steps (all): **58**
- Error-type coverage (among erroneous steps):
  - Reasoning Err: **23/58** (39.66%)
  - Image Understanding Err: **29/58** (50.00%)
  - Clinical Scenario Err: **9/58** (15.52%)
  - Medical Knowledge Err: **24/58** (41.38%)
- Erroneous *critical* steps only: **3**
- Error-type coverage (among erroneous critical steps):
  - Reasoning Err: **0/3** (0.00%)
  - Image Understanding Err: **3/3** (100.00%)
  - Clinical Scenario Err: **0/3** (0.00%)
  - Medical Knowledge Err: **0/3** (0.00%)

Incorrectly answered (*is_correct=False*).

- Steps (with valid `is_factual`): **360**
- Step factual accuracy: **195/360** (**54.17**%)
- Critical steps: **61**
- Critical-step factual accuracy: **52/61** (**85.25**%)
- Erroneous steps (all): **165**
- Error-type coverage (among erroneous steps):
  - Reasoning Err: **104/165** (63.03%)
  - Image Understanding Err: **84/165** (50.91%)
  - Clinical Scenario Err: **33/165** (20.00%)
  - Medical Knowledge Err: **81/165** (49.09%)
- Erroneous *critical* steps only: **9**
- Error-type coverage (among erroneous critical steps):
  - Reasoning Err: **4/9** (44.44%)
  - Image Understanding Err: **6/9** (66.67%)
  - Clinical Scenario Err: **2/9** (22.22%)
  - Medical Knowledge Err: **2/9** (22.22%)

L.3 `medgemma27b_100_sample` (100-Sample Subset)

Overall. Total number of steps (all samples): **1,074**.

Correctly answered (*is_correct=True*).

- Steps (with valid `is_factual`): **376**
- Step factual accuracy: **285/376** (**75.80**%)
- Critical steps: **102**
- Critical-step factual accuracy: **97/102** (**95.10**%)
- Erroneous steps (all): **91**
- Error-type coverage (among erroneous steps):
  - Reasoning Err: **22/91** (24.18%)
  - Image Understanding Err: **50/91** (54.95%)
  - Clinical Scenario Err: **15/91** (16.48%)
  - Medical Knowledge Err: **36/91** (39.56%)
- Erroneous *critical* steps only: **5**
- Error-type coverage (among erroneous critical steps):
  - Reasoning Err: **3/5** (60.00%)
  - Image Understanding Err: **4/5** (80.00%)
  - Clinical Scenario Err: **0/5** (0.00%)
  - Medical Knowledge Err: **1/5** (20.00%)

Incorrectly answered (*is_correct=False*).

- Steps (with valid `is_factual`): **698**
- Step factual accuracy: **383/698** (**54.87**%)
- Critical steps: **114**
- Critical-step factual accuracy: **78/114** (**68.42**%)
- Erroneous steps (all): **315**
- Error-type coverage (among erroneous steps):
  - Reasoning Err: **156/315** (49.52%)
  - Image Understanding Err: **221/315** (70.16%)
  - Clinical Scenario Err: **72/315** (22.86%)
  - Medical Knowledge Err: **119/315** (37.78%)
- Erroneous *critical* steps only: **36**
- Error-type coverage (among erroneous critical steps):
  - Reasoning Err: **16/36** (44.44%)
  - Image Understanding Err: **22/36** (61.11%)
  - Clinical Scenario Err: **9/36** (25.00%)
  - Medical Knowledge Err: **14/36** (38.89%)

L.4 `qwen2.5vl-32b_100` (100-Sample Subset)

Overall. Total number of steps (all samples): **781**.

Correctly answered (*is_correct=True*).

- Steps (with valid `is_factual`): **337**
- Step factual accuracy: **274/337** (**81.31**%)
- Critical steps: **103**
- Critical-step factual accuracy: **100/103** (**97.09**%)
- Erroneous steps (all): **63**
- Error-type coverage (among erroneous steps):
  - Reasoning Err: **22/63** (34.92%)
  - Image Understanding Err: **36/63** (57.14%)
  - Clinical Scenario Err: **6/63** (9.52%)
  - Medical Knowledge Err: **31/63** (49.21%)
- Erroneous *critical* steps only: **3**
- Error-type coverage (among erroneous critical steps):
  - Reasoning Err: **0/3** (0.00%)
  - Image Understanding Err: **3/3** (100.00%)
  - Clinical Scenario Err: **0/3** (0.00%)
  - Medical Knowledge Err: **0/3** (0.00%)

Incorrectly answered (*is_correct=False*).

- Steps (with valid `is_factual`): **444**
- Step factual accuracy: **236/444** (**53.15**%)
- Critical steps: **67**
- Critical-step factual accuracy: **35/67** (**52.24**%)
- Erroneous steps (all): **208**
- Error-type coverage (among erroneous steps):
  - Reasoning Err: **130/208** (62.50%)
  - Image Understanding Err: **113/208** (54.33%)
  - Clinical Scenario Err: **52/208** (25.00%)
  - Medical Knowledge Err: **109/208** (52.40%)
- Erroneous *critical* steps only: **32**
- Error-type coverage (among erroneous critical steps):
  - Reasoning Err: **21/32** (65.62%)
  - Image Understanding Err: **26/32** (81.25%)
  - Clinical Scenario Err: **4/32** (12.50%)
  - Medical Knowledge Err: **11/32** (34.38%)

M MELD-Based Data Contamination Analysis (Full Details)

Detector. We use MELD (Memorization Effects Levenshtein Detector) in a stricter, sliding-window form. For a model output y and its corresponding question x, we compute normalized Levenshtein similarity over fixed-width windows on the longer string and take the maximum across windows. Scores are reported as percentages; higher values indicate longer, more verbatim copying. Following prior medical-QA practice Tang et al. [2025], samples with similarity $\geq 50\%$ are flagged as high-risk for contamination.

Protocol. We run the exact inference setup used in our main experiments on the MEDTHINKVQA test set and apply MELD between each generated answer and its input question. We evaluate seven models spanning both LLMs and VLMs: Qwen3-32B, Med-Gemma-27B-*it*, Med-Gemma-27B-*text-it*, GPT-4.1-nano, GPT-4.1-mini, Qwen2.5-VL-72B-Instruct, and Llama-3.3-70B-Instruct.

Results. Appendix Figure 6 plots the full distributions. Across all models, medians lie near $\sim$20–24% with tight interquartile ranges, and the upper tails are short. Importantly, we do not observe any case with MELD similarity $\geq 50\%$; the largest outliers remain below that threshold. Text-only LLMs and VLMs exhibit highly similar distributions, suggesting that the presence of images does not drive overlap behavior.

Context vs. prior benchmarks. MedAgentsBench Tang et al. [2025] reports broader spreads and heavier right tails (with many outliers above 50%) on several widely used QA datasets (e.g., MMLU, MedQA, MedMCQA). In contrast, MEDTHINKVQA shows uniformly low overlap and no high-similarity spikes, indicating a substantially lower contamination risk.

Limitations. MELD is a surface-form detector; heavy paraphrasing or template-level memorization may evade detection. Our analysis should therefore be viewed as strong negative evidence for verbatim leakage rather than a proof of absence of all forms of contamination.

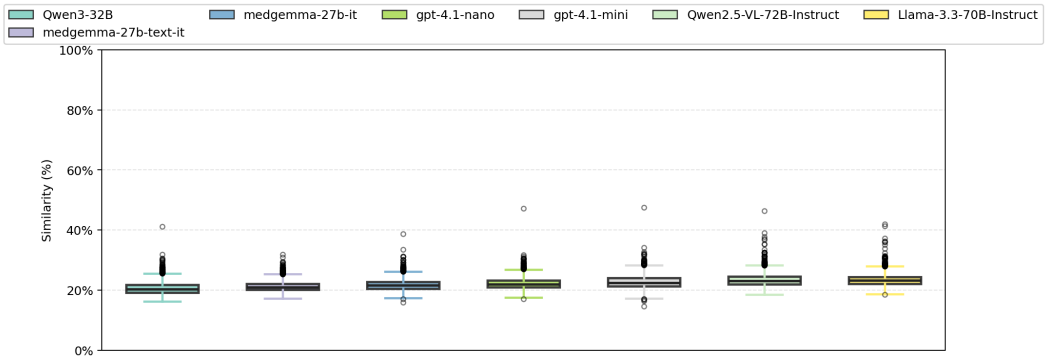

Figure 6: MELD data leakage test results on LLMs and VLMs for EuroRadQA. Boxplots show the distribution of similarity (%) between generated text and question text.

M.1 MELD and Our Windowed Variant

We first restate the original MELD procedure (Algorithm 1), and then present our implementation (Algorithm 2), which adds (i) a fixed-denominator Levenshtein ratio with respect to $|q_2|$, (ii) a length-$|q_2|$ sliding window over the model's continuation restricted to its early prefix, and (iii) length-aware bucketing and generation caps for efficient parallel decoding.

Algorithm 1: MELD (original reproduction)

Data: Generative model g; dataset D of question–answer pairs; tokenizer T; threshold $Y \in [0, 1]$.
Result: Z: percentage (or average strength) of completions with overlap above Y.

Initialize an empty list L
foreach $(q, a) \in D$ **do**
  Split q into two halves: q_1 and q_2
  Tokenize: $t_1 \leftarrow T(q_1)$ and $t_2 \leftarrow T(q_2)$
  Set sampling temperature to 0 and pass q_1 as context to g
  Let $k \leftarrow |t_2|$ and generate a continuation x consisting of k tokens from g
  Compute the (paper-style) Levenshtein-based overlap ratio

$$\ell \;=\; \frac{\mathrm{int}\big(\mathrm{round}\big(\frac{2.0\times M}{|q|}\times 100\big)\big)}{100},$$

  where $|q|$ is the total number of characters in both strings and M is the number of matches.
  if $\ell > Y$ **then**
    append ℓ to L
$Z \leftarrow \mathrm{mean}(L)$
return Z

Algorithm 2: MELD (ours, concise): windowed Levenshtein with length-aware batching

Data: Model g; dataset D; tokenizer T; threshold Y; cap multiplier $c \geq 1$; min gen tokens m; batch size B.
Result: Z (near-exact rate), $\bar{\ell}$ (mean similarity).

Build items. For each $r \in D$: form text $q \leftarrow \mathsf{build}(r)$; if empty, continue. Tokenize ids $\leftarrow T(q)$; split at $h = \max(1, \lfloor |\text{ids}|/2 \rfloor)$; set $q_1 = T^{-1}(\text{ids}[:h])$, $q_2 = T^{-1}(\text{ids}[h:])$, $k = |\text{ids}| - h$. Collect tuples $(q_1, q_2, k, |q_2|)$.
Bucket. Group tuples into batches of size $\leq B$ with similar k (length-aware).
foreach *batch* b **do**
  $G \leftarrow \max\big(m,\ c \cdot \max_{i \in b} k_i\big)$; set decoding $(\mathrm{temp} = 0,\ \text{top-}p = 1,\ \text{max tokens} = G)$
  Generate in parallel $x_i \leftarrow g(q_{1,i})$ for all $i \in b$
  foreach *item* i *in* b **do**
    $L \leftarrow |q_{2,i}|$,
    region $\leftarrow$ first cL characters of x_i
    $\rho_i \leftarrow \max\limits_{0 \leq j \leq |\textit{region}| - L} \left(1 - \dfrac{\mathrm{Lev}(\textit{region}[j{:}j{+}L],\ q_{2,i})}{L}\right)$;
    $s_i \leftarrow \mathbf{1}[\rho_i \geq Y]$
$Z \leftarrow \frac{1}{n}\sum_i s_i$; $\quad \bar{\ell} \leftarrow \frac{1}{n}\sum_i \rho_i$;
return $Z, \bar{\ell}$

N Disease Category Break Down

Test set size: $n = 751$ ($n = 680$ *common* + $n = 71$ *rare*).
Rare cases: $n = 71$ (~9.5% of total). Rare cases are a cross-category tag and are *not* double-counted in the chapter breakdown below.

Subcategory Detail (within each ICD-10 Chapter)

1. Certain infectious and parasitic diseases ($n = 35$; 4.7% of total)

- **1.1** A00–A09 Intestinal infectious diseases — 4
- **1.2** A15–A19 Tuberculosis — 13
- **1.3** A20–A28 Certain zoonotic bacterial diseases — 2
- **1.4** A50–A64 Infections with a predominantly sexual mode of transmission — 2
- **1.5** B15–B19 Viral hepatitis — 1
- **1.6** B20 Human immunodeficiency virus [HIV] disease — 2
- **1.7** B65–B83 Helminthiases — 11

2. Neoplasms ($n = 241$; 32.1% of total)

- **2.1** C00–C14 Malignant neoplasms of lip, oral cavity and pharynx — 2
- **2.2** C15–C26 Malignant neoplasms of digestive organs — 11
- **2.3** C30–C39 Malignant neoplasms of respiratory and intrathoracic organs — 10
- **2.4** C40–C41 Malignant neoplasms of bone and articular cartilage — 6
- **2.5** C45–C49 Malignant neoplasms of mesothelial and soft tissue — 13
- **2.6** C50 Malignant neoplasms of breast — 1
- **2.7** C51–C58 Malignant neoplasms of female genital organs — 9
- **2.8** C60–C63 Malignant neoplasms of male genital organs — 3
- **2.9** C64–C68 Malignant neoplasms of urinary tract — 4
- **2.10** C69–C72 Malignant neoplasms of eye, brain and other parts of CNS — 10
- **2.11** C73–C75 Malignant neoplasms of thyroid and other endocrine glands — 3
- **2.12** C76–C80 Malignant neoplasms of ill-defined, other secondary and unspecified sites — 17
- **2.13** C7A Malignant neuroendocrine tumors — 5
- **2.14** C81–C96 Malignant neoplasms of lymphoid, hematopoietic and related tissue — 20
- **2.15** D00–D09 In situ neoplasms — 1
- **2.16** D10–D36 Benign neoplasms (except benign neuroendocrine tumors) — 98
- **2.17** D37–D48 Neoplasms of uncertain behavior, polycythemia vera and MDS — 22
- **2.18** D49 Neoplasms of unspecified behavior — 6

3. Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism ($n = 14$; 1.9% of total)

- **3.1** D55–D59 Hemolytic anemias — 1
- **3.2** D70–D77 Other disorders of blood and blood-forming organs — 6
- **3.3** D80–D89 Certain disorders involving the immune mechanism — 7

4. Endocrine, nutritional and metabolic diseases ($n = 12$; 1.6% of total)

- **4.1** E00–E07 Disorders of thyroid gland — 1
- **4.2** E20–E35 Disorders of other endocrine glands — 4
- **4.3** E70–E88 Metabolic disorders — 7

5. Diseases of the nervous system ($n = 16$; 2.1% of total)

- **5.1** G00–G09 Inflammatory diseases of the central nervous system — 3
- **5.2** G20–G26 Extrapyramidal and movement disorders — 1
- **5.3** G30–G32 Other degenerative diseases of the nervous system — 2
- **5.4** G35–G37 Demyelinating diseases of the CNS — 2
- **5.5** G50–G59 Nerve, nerve root and plexus disorders — 3
- **5.6** G70–G73 Diseases of myoneural junction and muscle — 2
- **5.7** G89–G99 Other disorders of the nervous system — 3

6. Diseases of the eye and adnexa ($n = 2$; 0.3% of total)

- **6.1** H00–H05 Disorders of eyelid, lacrimal system and orbit — 1
- **6.2** H25–H28 Disorders of lens — 1

7. Diseases of the circulatory system ($n = 32$; 4.3% of total)

- **7.1** I20–I25 Ischemic heart diseases — 2
- **7.2** I26–I28 Pulmonary heart disease and diseases of pulmonary circulation — 1
- **7.3** I30–I5A Other forms of heart disease — 3
- **7.4** I60–I69 Cerebrovascular diseases — 5
- **7.5** I70–I79 Diseases of arteries, arterioles and capillaries — 12
- **7.6** I80–I89 Diseases of veins, lymphatic vessels and lymph nodes, NEC — 9

8. Diseases of the respiratory system ($n = 27$; 3.6% of total)

- **8.1** J00–J06 Acute upper respiratory infections — 1
- **8.2** J09–J18 Influenza and pneumonia — 5
- **8.3** J30–J39 Other diseases of upper respiratory tract — 4
- **8.4** J40–J47 Chronic lower respiratory diseases — 3
- **8.5** J60–J70 Lung diseases due to external agents — 1
- **8.6** J80–J84 Other respiratory diseases principally affecting the interstitium — 6
- **8.7** J90–J94 Other diseases of the pleura — 3
- **8.8** J96–J99 Other diseases of the respiratory system — 4

9. Diseases of the digestive system ($n = 81$; 10.8% of total)

- **9.1** K00–K14 Diseases of oral cavity and salivary glands — 4
- **9.2** K20–K31 Diseases of esophagus, stomach and duodenum — 10
- **9.3** K35–K38 Diseases of appendix — 4
- **9.4** K40–K46 Hernia — 5
- **9.5** K50–K52 Noninfective enteritis and colitis — 2
- **9.6** K55–K64 Other diseases of intestines — 20

- **9.7** K65–K68 Diseases of peritoneum and retroperitoneum — 8
- **9.8** K70–K77 Diseases of liver (note: viral hepatitis → Chapter 1, B15–B19) — 8
- **9.9** K80–K87 Disorders of gallbladder, biliary tract and pancreas — 20

10. Diseases of the skin and subcutaneous tissue ($n = 2$; 0.3% of total)

- **10.1** L60–L75 Disorders of skin appendages — 2

11. Diseases of the musculoskeletal system and connective tissue ($n = 43$; 5.7% of total)

- **11.1** M05–M14 Inflammatory polyarthropathies — 7
- **11.2** M20–M25 Other joint disorders — 6
- **11.3** M30–M36 Systemic connective tissue disorders — 3
- **11.4** M45–M49 Spondylopathies — 1
- **11.5** M50–M54 Other dorsopathies — 2
- **11.6** M60–M63 Disorders of muscles — 1
- **11.7** M65–M67 Disorders of synovium and tendon — 5
- **11.8** M70–M79 Other soft tissue disorders — 5
- **11.9** M80–M85 Disorders of bone density and structure — 3
- **11.10** M86–M90 Other osteopathies — 9
- **11.11** M91–M94 Chondropathies — 1

12. Diseases of the genitourinary system ($n = 40$; 5.3% of total)

- **12.1** N10–N16 Renal tubulo-interstitial diseases — 6
- **12.2** N25–N29 Other disorders of kidney and ureter — 6
- **12.3** N30–N39 Other diseases of the urinary system — 4
- **12.4** N40–N53 Diseases of male genital organs — 6
- **12.5** N60–N65 Disorders of breast — 2
- **12.6** N70–N77 Inflammatory diseases of female pelvic organs — 4
- **12.7** N80–N98 Noninflammatory disorders of female genital tract — 11
- **12.8** N99 Intraoperative and postprocedural complications and disorders of genitourinary system, NEC — 1

13. Pregnancy, childbirth and the puerperium ($n = 5$; 0.7% of total)

- **13.1** O00–O08 Pregnancy with abortive outcome — 3
- **13.2** O30–O48 Maternal care related to the fetus and amniotic cavity and possible delivery problems — 1
- **13.3** O94–O9A Other obstetric conditions, NEC — 1

14. Congenital malformations, deformations and chromosomal abnormalities ($n = 82$; 10.9% of total)

- **14.1** Q00–Q07 Congenital malformations of the nervous system — 7
- **14.2** Q10–Q18 Congenital malformations of eye, ear, face and neck — 1
- **14.3** Q20–Q28 Congenital malformations of the circulatory system — 20

- **14.4** Q30–Q34 Congenital malformations of the respiratory system — 10
- **14.5** Q38–Q45 Other congenital malformations of the digestive system — 13
- **14.6** Q50–Q56 Congenital malformations of genital organs — 4
- **14.7** Q60–Q64 Congenital malformations of the urinary system — 10
- **14.8** Q65–Q79 Congenital malformations and deformations of the musculoskeletal system — 11
- **14.9** Q80–Q89 Other congenital malformations — 6

15. Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified ($n = 5$; 0.7% of total)

- **15.1** R40–R46 Symptoms and signs involving cognition, perception, emotional state and behavior — 1
- **15.2** R50–R69 General symptoms and signs — 1
- **15.3** R90–R94 Abnormal findings on diagnostic imaging and in function studies, without diagnosis — 3

16. Injury, poisoning and certain other consequences of external causes ($n = 37$; 4.9% of total)

- **16.1** S00–S09 Injuries to the head — 2
- **16.2** S20–S29 Injuries to the thorax — 3
- **16.3** S30–S39 Injuries to the abdomen, lower back, lumbar spine, pelvis and external genitals — 7
- **16.4** S40–S49 Injuries to the shoulder and upper arm — 2
- **16.5** S80–S89 Injuries to the knee and lower leg — 1
- **16.6** T15–T19 Effects of foreign body entering through natural orifice — 3
- **16.7** T51–T65 Toxic effects of substances chiefly nonmedicinal as to source — 1
- **16.8** T80–T88 Complications of surgical and medical care, NEC — 18

17. Factors influencing health status and contact with health services ($n = 4$; 0.5% of total)

- **17.1** Z00–Z13 Persons encountering health services for examinations — 2
- **17.2** Z77–Z99 Family/personal history and certain other factors influencing health status — 2

18. Codes for special purposes ($n = 2$; 0.3% of total)

- **18.1** U00–U49 Provisional assignment of new diseases of uncertain etiology or emergency use (incl. U07.x) — 2

Note: Subcategory counts within each chapter sum to the chapter total for the *common* set ($n = 680$). Rare-tagged cases ($n = 71$) are reported separately and are not included in the subcategory lines.
Abbreviations: NEC = not elsewhere classified.

O Rubric for Discussion Evaluation

O.1 Rubric 1: Disease Overview & Core Definition (0–2 points)

Focus: Understanding of the disease's fundamental attributes, including: nomenclature, classification, and etiology.

- **0 points:** Unable to identify or define the disease.
- **1 point:** States the disease name, but classification or core etiology is vague or inaccurate.
- **2 points:** Accurately states the standard medical name, clearly defines its essential nature, and identifies principal etiologies or key risk factors.

O.2 Rubric 2: Clinical Presentation & Pathophysiology (0–2 points)

Focus: How the disease manifests and its underlying mechanisms.

- **0 points:** Unable to describe any clinical features.
- **1 point:** Describes some common symptoms/signs but cannot explain the underlying pathophysiology, or omits critical features.
- **2 points:** Systematically outlines the typical clinical presentation and clearly explains the core pathophysiologic mechanisms.

O.3 Rubric 3: Key Imaging Findings & Interpretation (0–2 points)

Focus: Recognition, description, and interpretation of disease-specific imaging features across modalities.

- **0 points:** Unable to describe any imaging characteristics.
- **1 point:** Provides only generic descriptors (e.g., "mass," "opacity") without modality-specific features (CT, MRI, radiography, ultrasound), or fails to distinguish key benign versus malignant signs.
- **2 points:** Clearly and accurately describes characteristic findings on one or more relevant modalities (e.g., morphology, attenuation/signal characteristics, margins, enhancement pattern, diffusion restriction), and interprets their clinical significance (e.g., stage, aggressiveness, complication risk).

O.4 Rubric 4: Diagnostic Reasoning & Differential Diagnosis (0–2 points)

Focus: Integrating clinical and imaging data to reach a diagnosis and distinguish differential considerations.

- **0 points:** Unable to articulate a diagnostic approach.
- **1 point:** Arrives at the correct diagnosis but does not present a coherent, integrated reasoning process, or does not propose appropriate differential considerations.
- **2 points:** Clearly demonstrates how clinical information and imaging findings are synthesized to close the diagnostic loop, and lists at least two high-priority differential considerations with brief imaging discriminators (key features that separate each mimic from the index diagnosis).

O.5 Rubric 5: Transferable Learning & Generalization (0–2 points)

Focus: Lessons that extend beyond a single case.

- **0 points:** Teaching points are confined to this case.
- **1 point:** Some generalizability is suggested but remains vague and lacks actionable takeaways.
- **2 points:** Clearly summarizes transferable learning points and explains how to avoid misinterpretation or improve diagnostic accuracy in similar future scenarios.