MEDICAL THINKING WITH MULTIPLE IMAGES

Anonymous authors
Paper under double-blind review

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 $\leq$ 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 `https://anonymous.4open.science/r/ICLR_DEMO-D35E/`.

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 *think-with-images* (TwI) flow (Fig. 1, middle): models first read individual images (*per-image findings*), then fuse evidence across views into a *case-level imaging summary*, and finally perform *differential-diagnosis reasoning* to select a diagnosis. Beyond this diagnostic core, MedThinkVQA also includes a *Medical Education Case Discussion* task, where models produce teaching-style explanations that mirror how clinicians communicate and share knowledge in practice. Answer accuracy alone cannot reveal where this process breaks down, so MedThinkVQA supports *beyond-accuracy* evaluation that considers both the intermediate imaging and reasoning stages and the educational value of case discussions, enabling a diagnosing-style view that localizes failures in image reading, cross-view fusion, and clinical reasoning (see Fig. 1 and Sections 3, 4).

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

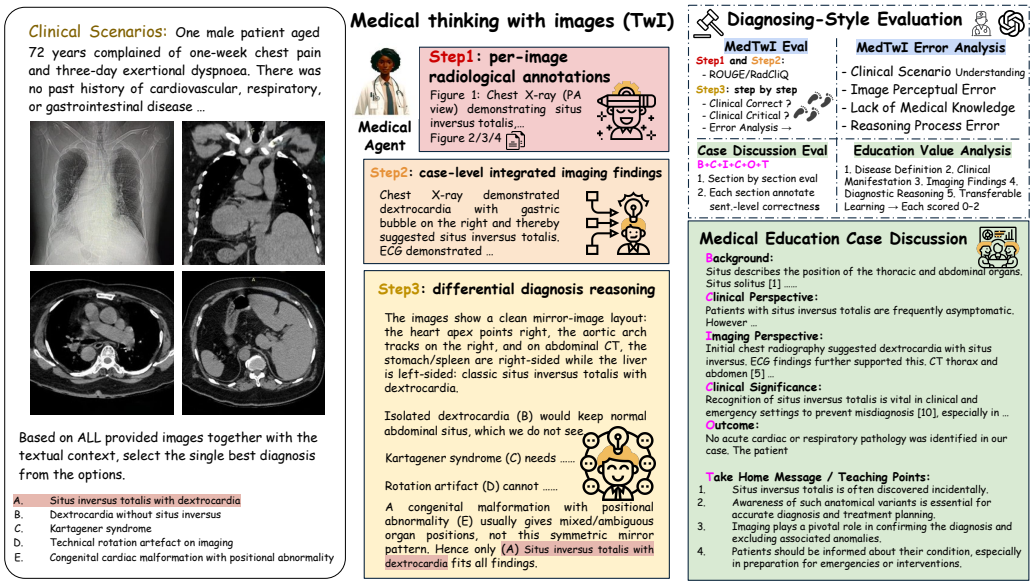


Figure 1: Medical Thinking with Images (TwI): task and diagnosing-style evaluation. Left: a sample case with a clinical scenario, multi-view images (e.g., radiograph + CT), and a five-option single-best-answer diagnosis. Middle: TwI's three supervised steps: (1) *Per-Image Findings* (detect and name 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), while Step 3 and the Medical Education Case Discussion are assessed with structured human- and LLM-judge rubrics that check clinical correctness and educational value, localizing failures in image reading, cross-view fusion, and teaching quality (see Section 4).

recognition. We further design MedThinkVQA so that questions depend on imaging evidence rather than textual shortcuts; Section 3 details the ICD–10 coverage, rare-disease cases, and the filtering and option policies used to control distractors, leakage, and surface biases.

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 ~ 15 percentage points below the strongest result on the hardest prior benchmark of a similar kind and ~ 20 points below clinicians on the same 96-case subset (77.10%), highlighting substantial headroom on MedThinkVQA. Giving *expert* findings and summaries raises accuracy, whereas replacing them with models' *self-generated* ones gives only small gains or even hurts performance. The expert audit further shows that 88.05% of images are judged supportive for the final diagnosis, that test cases typically involve around two distinct imaging modalities per case (mean modalities_count ≈ 2.13), and that 28.2% of test cases are longitudinal follow-up studies, so models must aggregate many informative views across both modality and time rather than relying on a single key image. 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, 44 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, 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.

Benchmark	# Case	Expert Annotation	Clinical Scenarios	# Img per Cas	Multi-Mod Imaging	Longitud Studies	Think-with-Images Intermediate Signals	Beyond-ACC Evaluation
VQA-Rad Lau et al.	451	✗	✗	0.45	✗	✗	✗	✗
VQA-Med Ben Abacha et al.	500	✗	✗	1.00	✗	✗	✗	✗
Path-VQA He et al.	6,719	✗	✗	0.13	✗	✗	✗	✗
SLAKE-En Liu et al.	1,061	✗	✗	0.09	✗	✗	✗	✗
PMC-VQA Zhang et al.	33,430	✗	✗	0.87	✗	✗	✗	✗
OmniMedVQA Hu et al.	127,995	✗	✗	0.92	✗	✗	✗	✗
GMAI-MMBench Ye et al.	21,281	✗	✗	1.00	✗	✗	✗	✗
GEMeX Liu et al.	1,605,575	✗	✗	1.00	✗	✗	✓	✓
Medical-Diff-VQA Hu et al.	700,703	✗	✗	1.23	✗	✓	✗	✗
MedFrameQA Yu et al.	2,851	✗	✗	3.24	✗	✗	✓	✗
ICG-CXR Ma et al.	11,439	✗	✗	2.00	✗	✓	✗	✗
MedRAX[1] Fallahpour et al.	2,500	✗	✗	1.85	✗	✗	✗	✗
GEMeX-ThinkVG Liu et al.	206,071	✗	✗	1.00	✗	✗	✓	✓
MMMU (H & M) Yue et al.	1,752	✓	✗	1.14	✗	✗	✗	✗
MMMU-Pro (H & M) Yue et al.	346	✓	✗	1.25	✗	✗	✗	✗
S-Chain Le-Duc et al.	12,000	✓	✗	1.00	✗	✗	✓	✓
MedXpertQA MM Zuo et al.	2,000	✓	✓	1.43	✗	✗	✗	✗
MedThinkVQA	**8,481**	✓	✓	**6.51**	✓	✓	✓	✓

Table 1: Comparisons with multimodal medical QA benchmarks. **#Case/#Img/Expert 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; a checkmark in the *Expert Annotation* column indicates that items are curated and labeled by clinicians rather than automatically generated. **Clinical Scenarios.** Prior work lacks broad, fine-grained coverage of real diagnostic scenarios; only *MedThinkVQA* and *MedXpertQA-MM* include scenario labels. **Multi-Modal Imaging / Longitudinal Studies.** We mark *Multi-Modal Imaging* when at least some questions require joint interpretation of images from multiple distinct medical imaging modalities for the same case (e.g., radiograph+CT), and *Longitudinal Studies* when questions are built from follow-up imaging of the same patient at different time points (e.g., baseline vs follow-up studies). **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.

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 and longitudinal imaging, large-scale corpora such as GEMeX (Liu et al., 2025a), Medical-Diff-VQA (MIMIC-Diff-VQA) (Hu et al., 2023), ICG-CXR (Ma et al., 2025), and the multi-image MedFrameQA benchmark (Yu et al., 2025) expand data scale and introduce explainable, difference-aware, counterfactual, or explicitly multi-image settings. However, their QAs and rationales are largely produced by rule-based or GPT-style pipelines rather than per-item expert traces, most items remain single-view or image-pair based, and they do not provide the per-image findings, case-level imaging summaries, or teaching notes needed for stepwise diagnostic supervision. 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 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/ChestAgentBench) orchestrates segmentation, grounding, report-generation, and classification tools on Eurorad-derived multiple-choice cases, but the released benchmark exposes only questions and images without the underlying expert step traces, complementing rather than 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., 2024a; 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 (Liu et al., 2025b) and S-Chain (Le-Duc et al., 2025) further introduce step-by-step rationales with explicit visual grounding and localization metrics (e.g., answer–reason scores, A-score, mIoU), moving beyond answer-only evaluation while still focusing mainly on single-image cases without clinical scenarios or multi-view, case-level synthesis. 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 6.51 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.These cases span 13 aggregated imaging modalities in the test split and typically involve around two distinct modalities per case; detailed modality statistics are provided in Appendix J. For concreteness, detailed field-to-annotation examples and six representative Eurorad case studies are provided in Appendix C. Details of the MCQ transformation and option policy are provided in Section 3.3. .

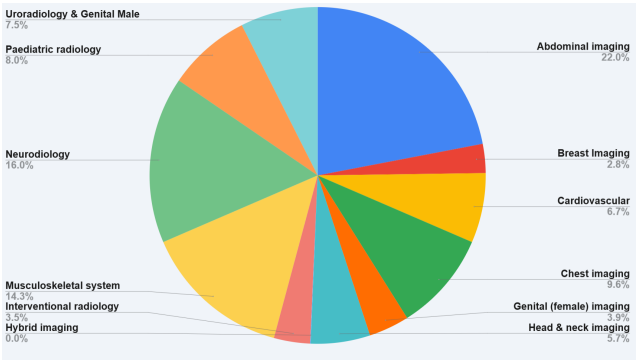


Figure 2: Distribution of radiology imaging main categories

Eurorad materials use CC BY-NC-SA 4.0; MedThinkVQA follows the same license and is for research and education only, with attribution and ShareAlike, 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.[1]

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.[2]

To assess breadth from an imaging perspective, we aggregated the *full dataset* by radiology subspecialties (*anatomy & subspecialty*). Figure 3 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%). From a temporal-structure perspective, roughly one quarter of MedThinkVQA cases are longitudinal follow-up studies (multiple time points for the same patient), so temporal disease evolution is explicitly represented; detailed longitudinal statistics are summarized in Appendix K.

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	2.13
All mod. types	13
Longit. share (%)	28.2

Table 2: Test stats (*Cap/Find/Disc.* = caption, findings, discussion; *Pos. correct* = avg. position of correct option; *Mean mod. cnt* = mean # of imaging modalities; see Appendix J for modality statistics and K for longitudinal-study statistics.

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.

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] For full details on licensing, permissions, privacy, safety, and leakage mitigation, see **Ethical Statement**.

[2] 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 O.

(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 ($\approx$20%), removing 664 cases.

3.4 Medical Education Case Discussion

In clinical practice, difficult or representative cases are often written up as teaching notes and shared with trainees and colleagues, and Eurorad "Discussion" sections already play this role. The human-expert study in Section 4.4 and Tables 3, 4 further shows that even experienced clinicians find a subset of MedThinkVQA cases very difficult, motivating our Medical Education Case Discussion task, where models are asked to generate structured teaching content rather than only predict a single diagnosis. To make this evaluation well-defined, we focus on cases whose discussions follow a clear five-section template—*Background*, *Clinical Perspective*, *Imaging Perspective*, *Outcome*, and *Take-Home Messages*—yielding a subset of test cases that strictly conforms to this structure and supports section-by-section comparison.

4 Experimental Setup

4.1 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).

4.2 Automatic Evaluation

Intermediate imaging metrics For the per-image findings and the case-level integrated imaging summary (Steps 1–2 in Fig. 1), we follow recent radiology–report evaluation work (Yu et al., 2023; Ostmeier et al., 2024) and compute ROUGE as a lexical-overlap baseline together with RadCliQ, which correlates more strongly with radiologist preferences. We apply these metrics to compare model outputs against the expert-written findings and summaries, providing automatic, fine-grained signals for how well models capture clinically salient details.

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. When a step is incorrect, the judge assigns one of four error types: clinical-scenario misunderstanding, missing or misread image evidence (*Image Understanding Err*), medical knowledge error, or flawed reasoning; this taxonomy is reused for both automatic and human evaluations. 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 M).

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.

4.3 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: Cohen's $\kappa = 0.82$ between the two experts, and human–LLM-judge agreement ranging from $\kappa = 0.70$ to $\kappa = 0.84$, confirming the reliability of the automatic judge.

4.4 EXPERTS PERFORMANCE AND DATA QUALITY ANNOTATION

Annotators and protocol. All annotations were provided by two board-certified clinicians in active clinical practice.[3] We randomly sampled 96 test cases and ran a two-round expert study aligned with our MCQ and stepwise evaluation. In *Round 1*, experts saw only the clinical history and all study images and selected one diagnosis out of five options, matching our VLM setup. In *Round 2*, they additionally received the full reference materials—image captions, per-image findings, the integrated imaging summary, the teaching discussion, and the ground-truth answer—and audited each case for internal consistency, difficulty, and image redundancy (supportive vs. redundant views). The same 96-case subset is used to evaluate VLM baselines for a fair human–model comparison.

Model / Expert	Correct	Incorrect	ACC (%)
Human experts	74	22	77.10
Gemini-2.5-pro	54	42	55.67
GPT-5	53	43	55.21
Claude 4.0	47	49	48.96
Lingshu-32B	42	54	43.75

Table 3: Human expert baseline vs. VLMs on the same 96-case subset.

Round 1: Diagnostic Performance. Experts answered 74/96 cases correctly (**77.10%** accuracy). We also evaluate GPT-5, Claude 4.0, and Lingshu-32B on the same subset with the identical MCQ

[3] One annotator is a diagnostic radiologist at a tertiary academic hospital in Asia with 7 years of post-training experience, and the other is an academic surgeon at a U.S. medical school with 5 years of post-training experience.

Case group	#Cases	#Images	#Supportive imgs	Supportive ratio (%)
All images supportive	65	463	463	100.00
Mixed supportive / redundant	31	315	222	70.48
Overall (96 cases)	**96**	**778**	**685**	**88.05**

Table 4: Round 2 expert audit: image-level redundancy vs. support. Most cases use all images as supportive evidence; even in mixed cases, the majority of views remain supportive rather than redundant.

protocol (Table 3). GPT-5 is the strongest VLM baseline but still trails human experts by about 21.9 percentage points (77.10% vs. 55.21%), and its accuracy on this subset is within 2 percentage points of its full-test performance, suggesting that the expert-study subset is representative of the overall benchmark.

Round 2: Data Quality and Image Redundancy. In the audit phase, experts marked 2/96 cases (2.1%) as *possibly inconsistent* and 18/96 (18.8%) as *very difficult*, indicating that the benchmark largely reflects coherent teaching cases while retaining a non-trivial proportion of challenging items. For image redundancy, experts judged whether each view provided supportive evidence toward the final diagnosis. In 65/96 cases (463 images), all views were deemed supportive (100%). The remaining 31/96 cases contained 315 images, of which 222 (70.48%) were judged supportive. Overall, 685/778 images were rated as supportive (**88.05%**), with the rest considered redundant for determining the final diagnosis.

The expert study shows that experienced clinicians still clearly outperform state-of-the-art VLMs on *MedThinkVQA*, that the curated items are overwhelmingly consistent with only a small fraction flagged as potentially problematic, and that most cases require aggregating evidence from many supportive views. As shown in Fig. 5, when image_ratio $= 0$ the task reduces to choosing one diagnosis out of five options with a random success probability of 20%, and accuracy then rises steadily as a larger proportion of case images is revealed across all models; together with the expert audit in Table 4, where 88% of images are rated supportive, this monotonic gain suggests that additional views are rarely pure redundancy and usually contribute useful diagnostic information, even though overall performance still remains well below human experts. At the same time, the realistic minority of redundant / non-supportive images means models must both integrate multiple supportive views and learn to down-weight redundant ones, mirroring how radiologists select and prioritize key views before forming a diagnosis.

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. 6, 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

Expert imaging summaries sharply boost accuracy. Across all models, providing expert-written text extracted from images leads to large gains on MedThinkVQA (Fig. 6). Feeding the `Integrated Imaging Summary (expert)` raises accuracy by $+30.8$, $+30.3$, $+23.6$, and $+18.8$ points for MedGemma-27B, GPT-5-nano, GPT-5-mini, and GPT-5, respectively, corresponding to $1.82\times$, $1.77\times$, $1.48\times$, and $1.33\times$ their baselines. Once this diagnosis-oriented summary is available, adding an `Image Hint (expert)` provides only modest extra gains ($+0.8$–3.9 points), and the summary consistently outperforms the hint alone by $+2$–8 points. These patterns indicate that *structured findings matter more than caption-like descriptions*: the summary encodes laterality, location, pattern, and extent in a way that gives the language model highly discriminative cues for the final diagnosis. The inverse relation between baseline strength and relative improvement further suggests that, once visual evidence is verbalized, language inference is largely adequate and the main bottleneck is still *extracting and structuring pixel-level radiological evidence*.

Self-generated text is fragile and often hurts. When models first write their own `Hint`/`Summary` and then condition on it, the effects are much smaller and frequently negative. MedGemma-27B and GPT-5-mini generally lose 1–5 points relative to baseline; GPT-5 shows mixed results (around -4 to $+0.5$ points); only GPT-5-nano obtains modest gains of roughly $+1$–4 points. Tab. 6 explains why: Image→Caption/Findings generations achieve low ROUGE-L ($\approx$ 0.13–0.16) and imperfect RadCliQ-v1 scores, meaning that self-produced descriptions often miss laterality, precise locations, or key patterns and may introduce subtle inaccuracies. In addition, noisy text increases sequence length and can dilute attention over multi-view inputs, and current VLMs may overtrust erroneous text when image–text grounding is weak. Together, these factors make self-generated hints a brittle scaffold for reasoning, whereas concise, expert-authored imaging summaries reliably unlock the underlying language capabilities of the models.

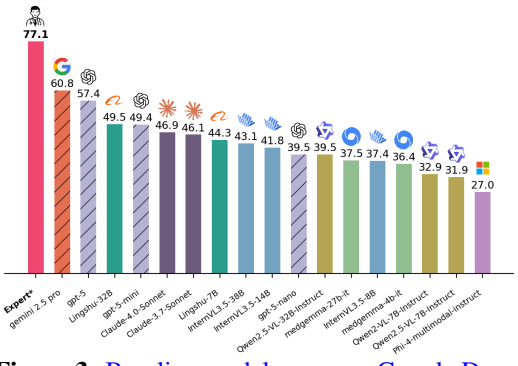


Figure 3: Baseline model accuracy Google DeepMind & Google Health AI (2025); Sellergren et al. (2025); Wang et al. (2024b); Bai et al. (2025); Abouelenin et al. (2025); OpenAI (2025a;b); Wang et al. (2025a); Gemini Team (2025); Anthropic (2024b; 2025); Xu et al. (2025); Chen et al. (2024). Note: Expert scores are computed on a randomly sampled subset of 96 test items; GPT-5 accuracy on this subset differs from its full-test accuracy by less than 2%.

Performance on multimodal and longitudinal subsets. As shown in Fig. 7, model behavior on cases with ≥ 3 imaging modalities and on longitudinal follow-up cases does not uniformly mirror the overall ranking. For highly multimodal cases, performance changes are mixed: some models (e.g., LINGSHU-32B, INTERNVL3_5-38B-HF, HUATUOGPT-VISION-7B) show modest gains over their overall accuracy, while others incur small drops, suggesting that richer modality combinations are helpful only when the model can correctly identify and fuse complementary information across views. By contrast, longitudinal follow-up cases systematically reduce accuracy for most models, often by several percentage points, indicating that current VLMs struggle to reason over temporal trajectories and may implicitly treat serial studies as an unordered set of images, ignoring cues such as interval change, new lesions, and post-treatment evolution. We provide qualitative GPT-5 error analyses for both settings in Appendix C.4.3 (multimodal hydatid disease) and Appendix C.4.2 (longitudinal cystic pulmonary tuberculosis), which illustrate how the model over-focuses on a subset of modalities or a single time point and consequently fails to recover the correct diagnosis.

MedThinkVQA mainly tests *image reasoning*, with expert summaries yielding large gains. Beyond diagnosing an image-fusion bottleneck—models still struggle to read and integrate many views even when language reasoning is strong—We suggest three main directions. First, process-supervised SFT and distillation can use our per-image findings, integrated summaries, and option-wise eliminations as step labels, extending chain-of-thought and SFT ideas from general medical MedVQA rationales to multi-image radiology Zhang et al. (2024); Gai et al. (2024). Second, data-centric and alignment methods can push the model onto the right views: medical-image augmentation plus generative

counterfactual editing and semi-human radiology QA expansion Kebaili et al. (2023); Shoer & Kementchedjhieva (2025); Wang et al. (2024a), combined with preference-based objectives such as DPO and with reinforcement learning that assigns process-level rewards to image-faithful chains and penalizes shortcut solutions Rafailov et al. (2023); Liu et al. (2024b); DeepSeek-AI et al. (2025). Third, test-time "thinking" strategies like Tree-of-Thoughts and frontier multimodal thinking models (Gemini 2.5 Pro, Claude 3.5 Sonnet) motivate architectures that can encode many images while letting the language backbone retrieve only the visual tokens it needs at each step Yao et al. (2023); Kavukcuoglu (2025); Anthropic (2024a), reinforcing the need for stronger visual encoders, better image–text grounding, and concise, structured hints.

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. 14. 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. 15).

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 5: 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.

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 ∼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 N.

	ROUGE-L (↑)		**RadCliQ-v1** (↑)	
Model	**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 6: Scores of VLMs for Image→Caption and Image→Findings across two metrics (ROUGE-L and RadCliQ).

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.

REPRODUCIBILITY STATEMENT

We provide full details to ensure reproducibility. Dataset sources and splits are in Section 3; implementation details and training practices are in Section 3; Hyperparameters for SFT are listed in Appendix B; We attached various prompts for data construction, LLM Judge in Appendix H; We also include an anonymized code repository link in Abstract.

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.

Limitations Our beyond-accuracy evaluation currently relies on a commercial LLM (GPT-5) as the primary judge for stepwise reasoning and case discussions. While we partially mitigate this by reporting human–LLM agreement (Cohen's $\kappa \approx 0.70$–0.84 with two clinicians) and by keeping experts in the loop, the approach still inherits model- and prompt-dependence: neither the GPT-5 API nor any particular snapshot is guaranteed to remain available, and future model updates could change judgments in subtle ways even under identical prompts and data. This limits strict replicability of some scores and means that our automatic annotations should be interpreted as calibrated but not definitive surrogates for expert review. We also experimented with open-weight judges, including

Qwen2.5-VL-72B, but in our setting these models underperformed GPT-5 as evaluators and showed less stable alignment with human experts, even after small-scale distillation on a subset of GPT-5-labeled steps. At present, they do not provide a sufficiently reliable drop-in replacement for the commercial judge. Developing robust, fully open-source evaluation pipelines—e.g., ensembles of open-weight VLM/LLM judges calibrated with human audits and process-level supervision—is an important direction for future work.

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.

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.

Anthropic. Introducing claude 3.5 sonnet—our most intelligent model yet. Anthropic News, 2024a. URL https://www.anthropic.com/news/claude-3-5-sonnet.

Anthropic. Claude 3.5 sonnet model card addendum. Technical report, 2024b. URL https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf. Accessed: 2025-11-24.

Anthropic. Claude 4 system card: Claude opus 4 and claude sonnet 4. Technical report, Anthropic, May 2025. URL https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf. Technical report.

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.

Junying Chen, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale, 2024. URL https://arxiv.org/abs/2406.19280.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

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: Explaining medical visual question answering via multimodal decision-making rationale. *arXiv preprint arXiv:2404.12372*, 2024. URL https://arxiv.org/abs/2404.12372.

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*, pp. 7438–7450, 2025.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. URL https://arxiv.org/abs/2507.06261.

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. Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23)*, pp. 4156–4165, New York, NY, USA, 2023. Association for Computing Machinery. doi: 10.1145/3580305.3599819.

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*, pp. 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.

Koray Kavukcuoglu. Gemini 2.5: Our most intelligent AI model. Google DeepMind Blog, 2025. URL https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/.

Aghiles Kebaili, Jérôme Lapuyade-Lahorgue, and Su Ruan. Deep learning approaches for data augmentation in medical imaging: A review. *Journal of Imaging*, 9(4):81, 2023. doi: 10.3390/jimaging9040081. URL https://www.mdpi.com/2313-433X/9/4/81.

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)*, pp. 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*, 2024a.

Ziyu Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Conghui He, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. MIA-DPO: Multi-image augmented direct preference optimization for large vision-language models. *arXiv preprint arXiv:2410.17637*, 2024b. URL https://arxiv.org/abs/2410.17637.

Chenglong Ma, Yuanfeng Ji, Jin Ye, Lu Zhang, Ying Chen, Tianbin Li, Mingjie Li, Junjun He, and Hongming Shan. Towards interpretable counterfactual generation via multimodal autoregression. In *Proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15961 of *Lecture Notes in Computer Science*, pp. 611–620. Springer Nature Switzerland, September 2025. doi: 10.1007/978-3-032-04937-7_58.

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 (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 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*, pp. 248–260. PMLR, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.18290.

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.

Belal Shoer and Yova Kementchedjhieva. A simple data augmentation strategy for text-in-image scientific vqa. In *Proceedings of the 9th Widening NLP Workshop*, pp. 100–105, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.winlp-main.17. URL https://aclanthology.org/2025.winlp-main.17.

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.

Junda Wang, Yujan Ting, Eric Z. Chen, Hieu Tran, Hong Yu, Weijing Huang, and Terrence Chen. Semihvision: Enhancing medical multimodal models with a semi-human annotated dataset and fine-tuned instruction generation. *arXiv preprint arXiv:2410.14948*, 2024a. URL https://arxiv.org/abs/2410.14948.

Peng Wang et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b. 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.

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. arXiv preprint arXiv:2305.10601, 2023. URL https://arxiv.org/abs/2305.10601.

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.

Suhao Yu, Haojin Wang, Juncheng Wu, Cihang Xie, and Yuyin Zhou. Medframeqa: A multi-image medical vqa benchmark for clinical reasoning. *arXiv preprint arXiv:2505.16964*, 2025. URL https://arxiv.org/abs/2505.16964.

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*, pp. 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.

Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. Chain of preference optimization: Improving chain-of-thought reasoning in llms. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.09136.

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.

A LLM USAGE

In accordance with the ICLR 2026 policies on LLM usage, we disclose how LLMs were used in this work. LLMs were employed to assist with grammar polishing, wording improvements, and drafting text during paper preparation. All technical content, proofs, experiments, and analyses were conceived, implemented, and validated by the authors. Authors remain fully responsible for the correctness of the claims and results.

No LLMs were used to generate research ideas, write code for experiments, or produce results. No confidential information was shared with LLMs, and no prompt injections or other inappropriate uses were involved.

This disclosure aligns with the ICLR Code of Ethics: contributions of tools are acknowledged, while accountability and verification rest entirely with the human authors.

B SUPERVISED FINE-TUNING

Training setup. We fine-tuned `InternVL3.5-1B`, `InternVL3.5-2B`, `InternVL3.5-4B`, `MedGemma-4B-IT`, `Qwen2.5-VL-3B-Instruct`, and `Qwen2.5-VL-7B-Instruct` on MedThinkVQA using QLoRA (Quantized Low-Rank Adaptation). All models adopted a LoRA rank of 8, $\alpha = 16$, and a dropout rate of 0.05. We trained for 2 epochs with a per-device batch size of 1 and 8 gradient accumulation steps (effective batch size 8), using AdamW with a learning rate of 2×10^{-4}, cosine learning-rate scheduling, and a warmup ratio of 0.03. The dataset was split $90/10$ into training and validation sets. These choices were kept fixed across models to enable a controlled comparison of fine-tuning gains.

SFT results. Supervised fine-tuning yields substantial accuracy improvements over the zero-shot or instruction-tuned baselines. As summarized in Tab. 13, the GPT-5 series provides a strong reference point with GPT-5 achieving 57.39% accuracy. After SFT, several smaller open-source models become competitive with or even surpass this level: `Qwen2.5-VL-7B-Instruct` improves from 31.95% to 61.89%, outperforming GPT-5; `InternVL3.5-4B` (60.96%) and `Qwen2.5-VL-3B-Instruct` (60.03%) reach accuracies comparable to GPT-5; and `MedGemma-4B-IT` improves from 36.35% to 56.57%. Taken together, these gains indicate that the MedThinkVQA training split provides high-quality, well-structured supervision that substantially enhances medical reasoning, enabling compact vision–language models to close most of the gap to much larger inference-time–scaled systems.

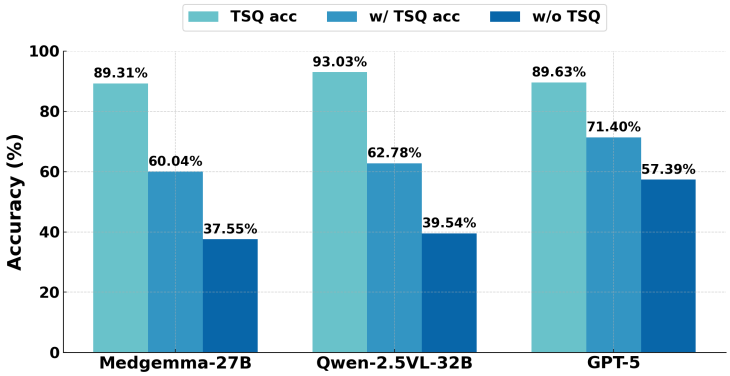


Figure 4: model accuracy across three processed datasets. **TSQ** refers to *Text-Solvable Questions*. The **TSQ acc** corresponds to model performance on the 611 text-solvable cases, where all three models achieved accuracies above 89%. In contrast, the **w/o TSQ** results are computed on the final test set after removing these text-solvable cases, showing a substantial drop in accuracy.

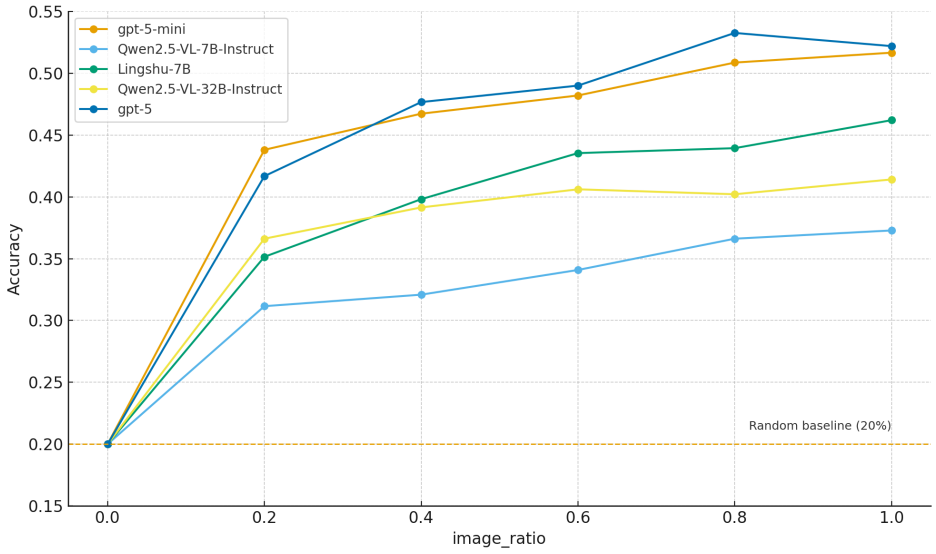


Figure 5: Accuracy on the test set as a function of the proportion of case images visible to the model (*image_ratio*). When no images are provided (image_ratio = 0), performance is close to the random baseline of 20% (dashed line); as the visible image proportion increases, accuracy consistently improves across models.

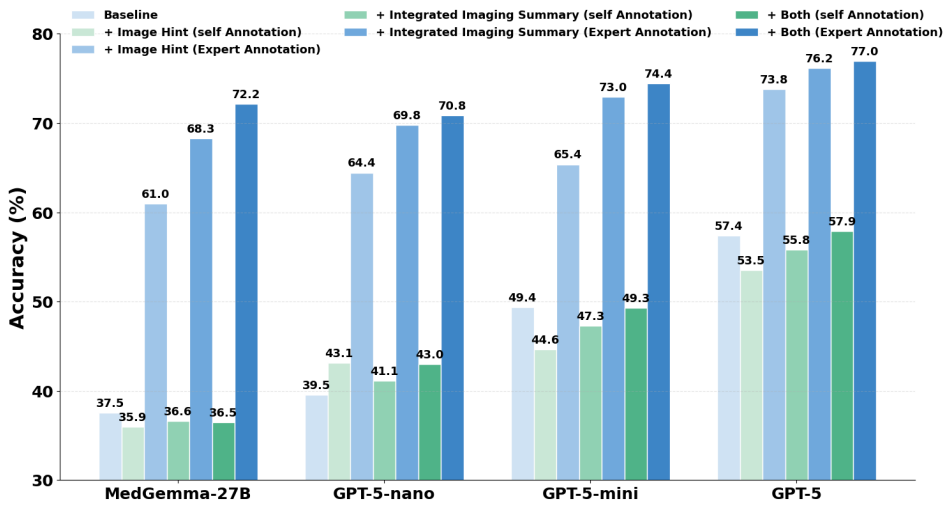


Figure 6: Accuracy on *MedThinkVQA* when augmenting images with text. We compare `Image Hint` (caption-like) and `Integrated Imaging Summary` (diagnosis-oriented findings), each provided either by an *expert* or generated by the *model itself* (self). `Both` combines the two.

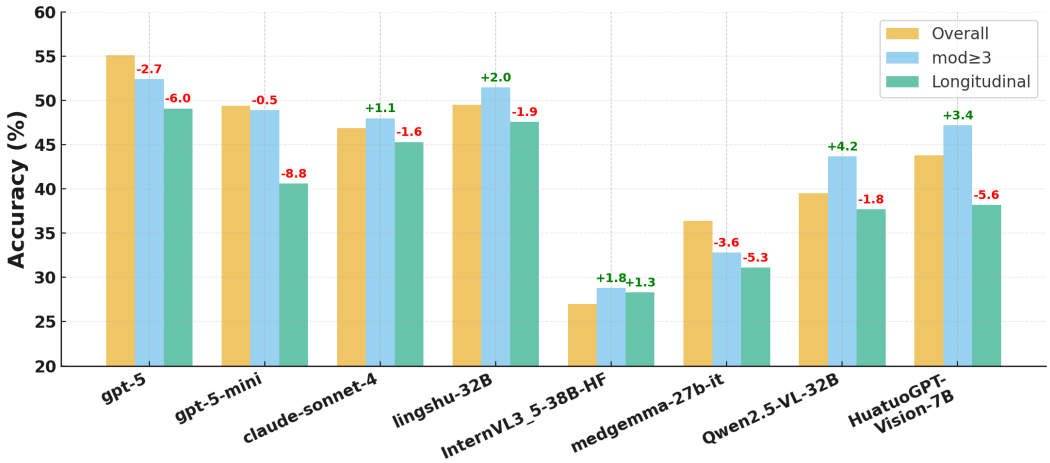


Figure 7: Comparison of accuracy for each model on the full test set (Overall), on cases with ≥ 3 imaging modalities, and on longitudinal follow-up cases. Accuracy on highly multimodal cases fluctuates around the overall level across models, whereas accuracy on longitudinal cases decreases for most models.

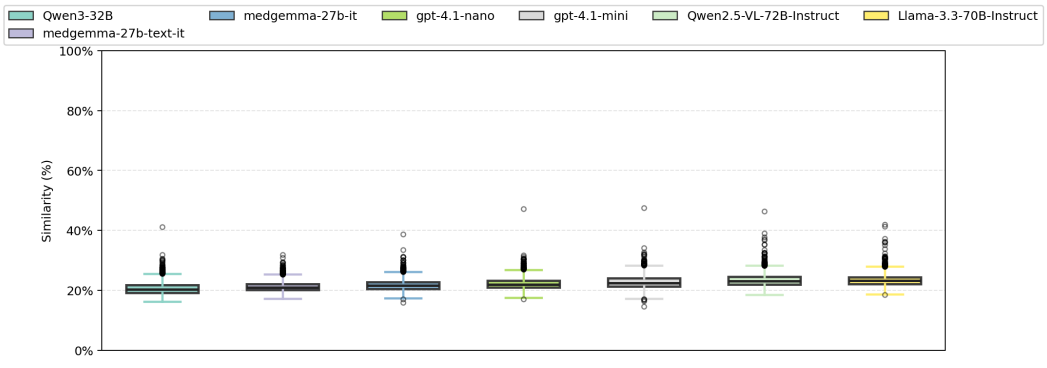


Figure 8: MELD data leakage test results on LLMs and VLMs for EuroRadQA. Boxplots show the distribution of similarity (%) between generated text and question text.

C Example Mapping from Eurorad Fields to MedThinkVQA Annotations

To make the supervision signals in MedThinkVQA concrete, this section uses a single Eurorad case from the test split, *"Ureteropelvic junction laceration following blunt trauma"*, whose processed JSON is stored at `cases/general/case_219/case.json`. We list the main JSON fields and show how they correspond to the supervision concepts used in the main text.

Clinical scenario.

- JSON field: `CLINICAL_HISTORY` (string).
- Content: brief free-text description of the presenting complaint and relevant history (for case 219, an elderly patient with cardiovascular comorbidities presenting with right-sided thoraco-abdominal trauma and microscopic haematuria).
- Main-text concept: this field is used verbatim as the *Clinical Scenario* shown to models before any images or answer options (Fig. 1, left).

Per-image hints (*Image Hint* / per-image findings).

- JSON field: `img` (list). Each element is a dictionary with keys `img_id`, `img_path`, `img_alt` (short legend), and `img_alt2` (full descriptive caption).
- Content structure (case 219):
  - Images 1–3: prior multidetector CT study 8 months earlier, showing bilateral peripelvic renal cysts with otherwise normal renal morphology.
  - Images 4–5: current CT for abdominal trauma, with right perirenal and fascial fluid and dependent hyperattenuation in the renal pelvis compatible with acute blood.
  - Image 6: arterial-phase CT and MIP reconstructions, without contrast extravasation, again emphasising hyperattenuation in the renal pelvis and a peripelvic cyst.
  - Images 7–8: delayed excretory-phase images, showing medial perirenal extraluminal opacified urine and normal parenchymal/collecting-system opacification.
  - Images 9–10: delayed images showing extraluminal opacified urine arising from a focal breach at the ureteropelvic junction and an opacified proximal ureter.
- Main-text concept: `img_alt2` provides the expert *per-image hint* used in Step 1 (*Image Hint* / per-image findings). In the TwI setting, models are asked to produce concise radiological finding sentences that are consistent with these captions.

Case-level Integrated Imaging Summary.

- JSON field: `IMAGING_FINDINGS` (string).

- Content: a case-level narrative integrating all imaging examinations (prior CT, current CT, delayed acquisitions), key abnormalities (peripelvic cysts, perirenal fluid, extraluminal opacified urine from a UPJ breach), and absence of other traumatic lesions, plus immediate management (e.g., ureteral stenting).
- Main-text concept: this field is the expert reference for the *Integrated Imaging Summary* (Step 2 in Fig. 1); models must fuse per-image findings into a single summary that matches this cross-view evidence.

Differential diagnosis and MCQ construction.

- JSON field: `DIF_DIAGNOSIS_LIST` (string with comma-separated diagnoses).
- Content (case 219, simplified): contains the target diagnosis "Ureteropelvic junction laceration following blunt trauma" and related entities such as "Ureteropelvic avulsion", "Renal parenchymal laceration with calyceal disruption", "Urinoma", "Perinephric haematoma", and "Subcapsular haematoma".
- Additional JSON fields used for the MCQ:
  - `options`: dictionary mapping option letters (`"A"`–`"E"`) to diagnosis strings.
  - `correct_answer`: the correct option letter (e.g., `"C"`).
  - `correct_answer_text`: the correct diagnosis string (e.g., "Ureteropelvic junction laceration following blunt trauma.").
- Main-text concept: these fields instantiate the five-option single-best-answer MCQ used in Step 3 (*Differential-Diagnosis Reasoning*); models compare their Integrated Imaging Summary to the `options` and must select `correct_answer_text`.

Medical Education Case Discussion.

- JSON field: `DISCUSSION` (string).
- Content: long-form teaching text covering epidemiology (e.g., rarity of UPJ injuries), mechanisms, imaging pitfalls, management strategies, and prognosis.
- Main-text concept: this field is the expert reference for the *Medical Education Case Discussion* task, where models generate a structured explanation (Background, Clinical Perspective, Imaging Perspective, Clinical Significance, Outcome, Take-Home Notes) that is graded against `DISCUSSION` for clinical correctness and educational value.

Overall, this case illustrates how raw Eurorad sections and figure captions are mapped onto the *Clinical Scenario*, *Image Hint*, *Integrated Imaging Summary*, *Differential Diagnosis*, and *Medical Education Case Discussion* supervision signals defined in the main text and implemented as JSON fields in MedThinkVQA.

C.1 Multimodal Case Study: Primary Carcinoma of the Rectovaginal Septum

This MedThinkVQA case is a 61-year-old woman with pelvic pain and inguinal lymphadenopathy, ultimately diagnosed with *primary carcinoma of the rectovaginal septum*. In the JSON, all four modalities (Endoscopy, CT, MRI, Histology / pathology) share the same `CLINICAL_HISTORY`, `IMAGING_FINDINGS`, `DISCUSSION`, and each image is referenced by its `img_id` and stored under `images/cases/modality/{img_id}.jpg` with `img_alt` and `img_alt2` captions.

Endoscopy. The endoscopic modality contains a single sigmoidoscopy frame (`img_id = l9iMrGt3`) that documents both focal bulging of the sigmoid wall and a 3 cm vegetative rectal lesion; these findings are encoded in the corresponding `img_alt` and `img_alt2` fields.

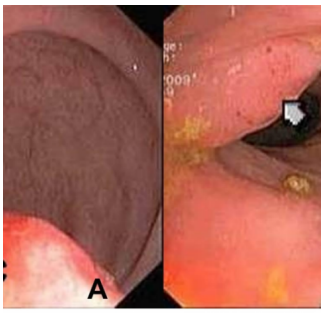
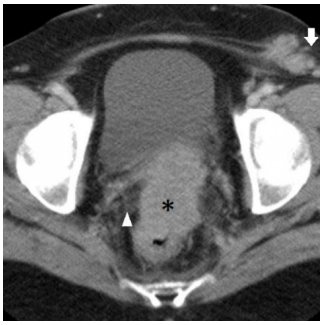
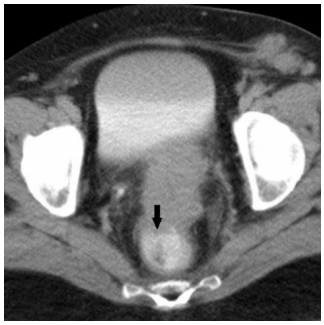


Figure 9: Endoscopy image for this case (`img_id = l9iMrGt3`).

CT. The CT modality consists of two axial contrast-enhanced CT images (`3uIJtKe-`, `pfx97TC8`) that show a heterogeneous mass in the pouch of Douglas, invasion of adjacent structures, and inguinal lymphadenopathy; the excretory-phase scan with rectal contrast further clarifies rectal wall involvement.



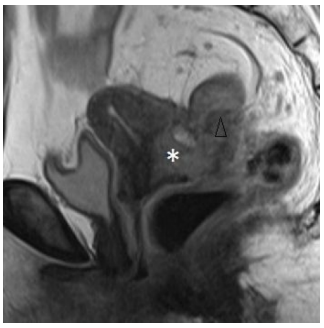
3uIJtKe-. Axial contrast-enhanced CT with pelvic mass and enlarged left inguinal nodes.



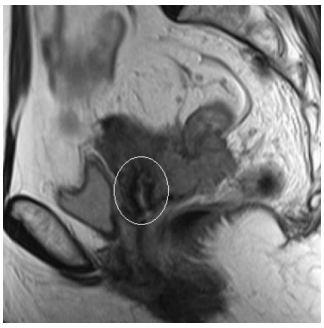
pfx97TC8. Excretory-phase CT with rectal contrast, better defining rectal wall involvement.

Figure 10: CT images (`img_id = 3uIJtKe-, pfx97TC8`) stored under `images/cases/modality/`.

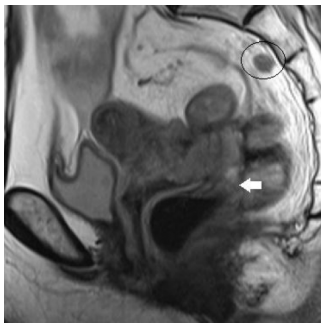
MRI. The MRI modality includes five T2-weighted images (`O5kEGVZq`, `IAV1h4UN`, `FjWYFzXB`, `F0KIjEeq`, `U14cWn_5`) that jointly characterise mass location (rectovaginal septum), extension to cervix and myometrium, intimate contact with the rectal wall, nodal disease, and preservation of the right ovary and inner cervical stromal layer.



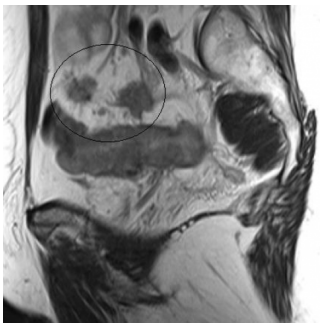
O5kEGVZq. Sagittal T2: mass in pouch of Douglas extending to cervix and myometrium.



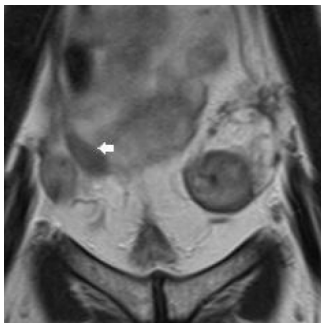
IAV1h4UN. Preserved low-signal inner cervical stroma.



FjWYFzXB. Mass inseparable from the anterior rectal wall.



F0KIjEeq. Multiple enlarged pelvic and abdominal lymph nodes.



U14cWn_5. Coronal T2: normal right ovary separately identified from the mass.

Figure 11: MRI images (`img_id = O5kEGVZq, IAV1h4UN, FjWYFzXB, F0KIjEeq, U14cWn_5`) with uniform image size and aligned top edges.

Histology / pathology. The histology / pathology modality contains a single composite slide (`3xrCMRPY`) showing solid tumour growth with marked atypia and immunostaining for CAM5.2, CK7, and WT1, all encoded in the `img_alt2` description.

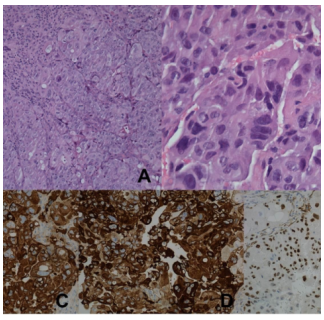


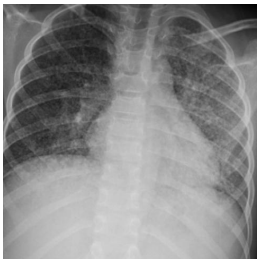
Figure 12: Histology / pathology image for this case (`img_id = 3xrCMRPY`).

Multimodal reasoning signal. In the dataset, this case is formatted as a five-option diagnostic MCQ with ground-truth label *primary carcinoma of the rectovaginal septum*. Endoscopy and CT highlight an extraluminal pelvic mass with rectal involvement; MRI localises the tumour to the rectovaginal septum and shows preserved cervix and ovaries with nodal spread; histology confirms a Müllerian-type carcinoma. A model must integrate all four modalities together with the shared textual fields in the JSON to distinguish this entity from rectal, cervical, and ovarian primaries.

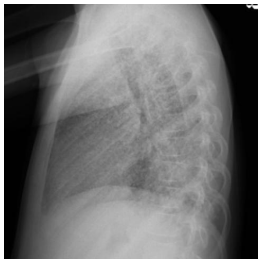
C.2 Longitudinal Case Study: Cystic Pulmonary Tuberculosis

This MedThinkVQA case is a nine-year-old boy with severe cystic pulmonary tuberculosis, followed radiologically over almost a year. In the JSON, the shared `CLINICAL_HISTORY`, `IMAGING_FINDINGS`, and `DISCUSSION` fields are linked to chest radiographs and CT scans at multiple time points. Below, we group images by clinical time point to illustrate longitudinal disease evolution.

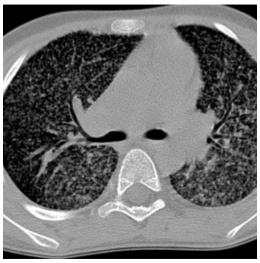
Baseline imaging at admission. Baseline chest radiographs (posteroanterior and lateral views) show a bilateral diffuse micronodular acinar infiltrate. A same-day chest CT (lung and mediastinal windows) reveals a diffuse micronodular pattern with random distribution throughout both lungs, suggestive of an inflammatory or infectious process.



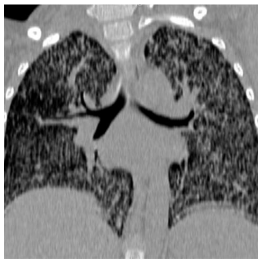
PA view. Baseline chest radiograph with diffuse micronodular infiltrates.



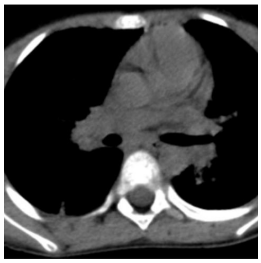
Lateral view. Baseline chest radiograph confirming bilateral involvement.



CT, lung window. Diffuse micronodular pattern in both lungs.



CT, lung window. Randomly distributed nodules throughout the parenchyma.



CT, mediastinal window. No large focal mass; diffuse micronodular disease.

Figure 13: Baseline chest radiographs and CT at admission, all displayed with a uniform relative size.

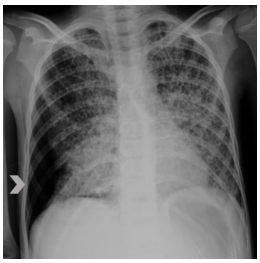
Early course with pneumothoraces. During the ICU stay, the patient develops spontaneous pneumothoraces requiring chest drainage. Serial radiographs show persistent diffuse micronodular infiltrates with evolving unilateral and bilateral pneumothoraces and multiple chest tubes in place.

Development of confluent cystic disease. A subsequent contrast-enhanced CT demonstrates extensive confluent cystic lesions predominantly in the upper lobes and posterior regions, consistent with cystic pulmonary tuberculosis and explaining the recurrent pneumothoraces.

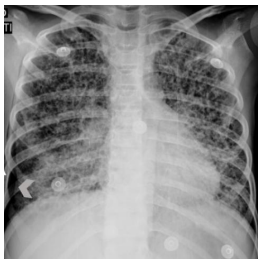
Persistent cysts and larger pneumothoraces. A further CT shows similar cystic disease but larger bilateral pneumothoraces, pneumomediastinum, and multiple chest drains in place, underscoring the mechanical complications of cystic tuberculosis.

Pre-discharge CT. Before discharge, CT still shows cystic lesions and residual pneumothorax, but with improved overall ventilation. The patient tolerates these sequelae after chest tube removal and can leave the hospital.

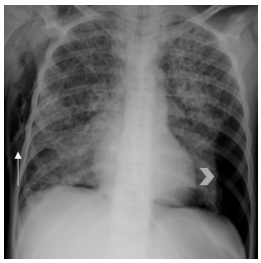
Late follow-up. At eight months after discharge, follow-up CT shows near-complete resolution of the cystic and nodular lesions, with only subtle residual cysts and fibrotic sequelae.



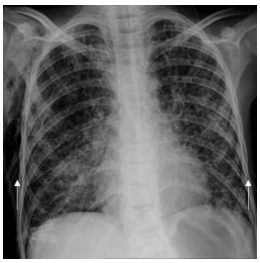
Chest radiograph. Right pneumothorax with chest tube in situ.



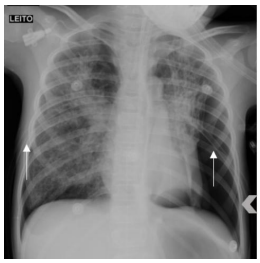
Chest radiograph. Persistent diffuse micronodular opacities.



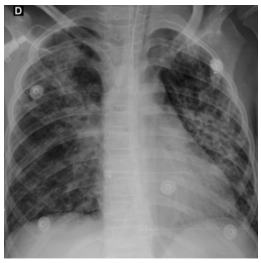
Chest radiograph. Bilateral lung involvement with scattered nodules.



Chest radiograph. Multiple thoracic drains for recurrent pneumothoraces.

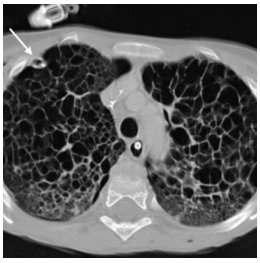


Chest radiograph. Persistent residual pneumothorax.

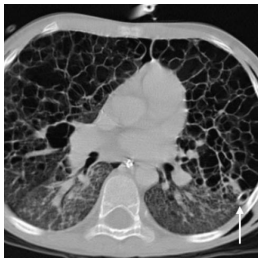


Chest radiograph. Diffuse cystic–nodular changes on a background of severe disease.

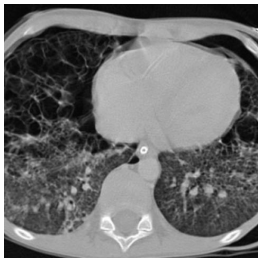
Figure 14: Serial chest radiographs during ICU stay, showing pneumothoraces and multiple chest drains.



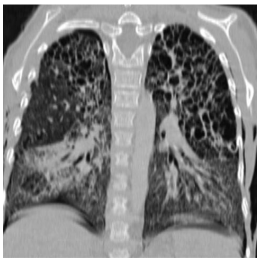
CT, lung window. Multiple cystic lesions in both upper lobes.



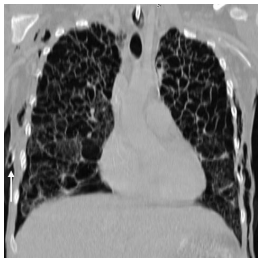
CT, lung window. Coalescent cysts forming large air-filled spaces.



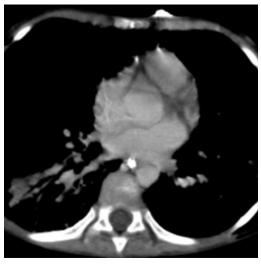
CT, lung window. Cystic lesions with surrounding ground-glass opacities.



CT, coronal view. Upper-lobe predominance of cystic disease.



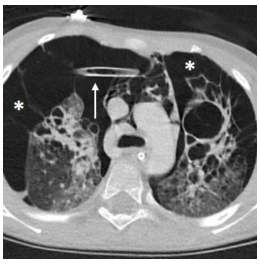
CT, lung window. Posterior lung involvement with confluent cysts.



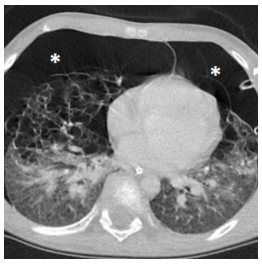
CT, mediastinal window. Multiple large cystic spaces without solid mass.

Figure 15: CT during peak disease severity, with confluent cystic lesions and diffuse parenchymal involvement.

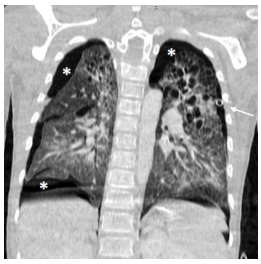
Longitudinal reasoning signal. In MedThinkVQA, this case is encoded as a five-option diagnostic MCQ with the correct answer *cystic pulmonary tuberculosis*. A model must integrate longitudinal information across all time points—progression from micronodular infiltrates to confluent cystic dis-



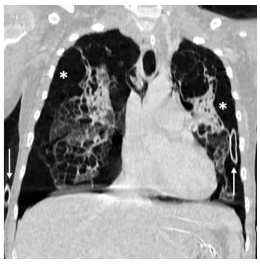
CT, lung window. Large right pneumothorax on a cystic background.



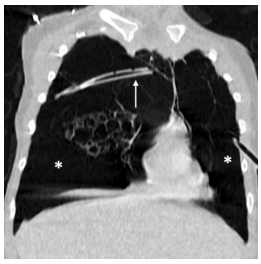
CT, lung window. Extensive bilateral cystic changes.



CT, coronal view. Bilateral pneumothoraces with upper-lobe cysts.

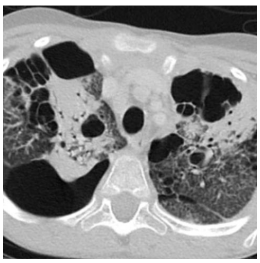


CT, mediastinal window. Pneumomediastinum and thoracic drains.

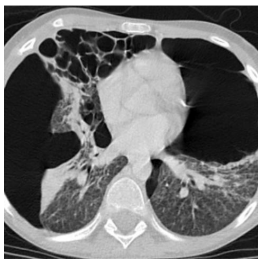


CT, lung window. Persistent cystic lesions despite drainage.

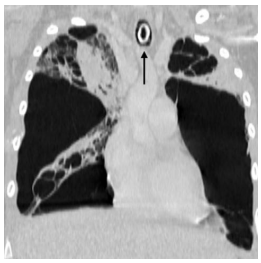
Figure 16: CT with larger pneumothoraces and pneumomediastinum, on a background of cystic pulmonary tuberculosis.



CT, lung window. Residual cystic changes with improved aeration.



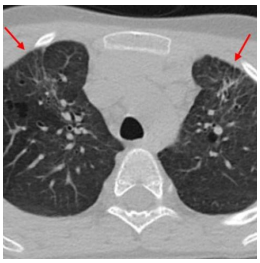
CT, lung window. Decreased extent of parenchymal disease.



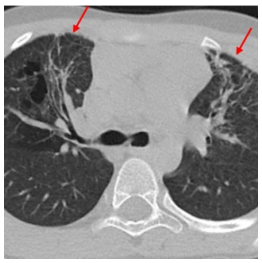
CT, coronal view. Tracheostomy in place with residual cysts.

Figure 17: Pre-discharge CT: persistent cystic lesions but improved clinical tolerance and removal of chest drains.

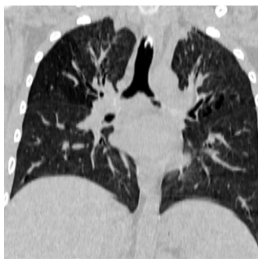
ease, recurrent pneumothoraces requiring multiple drains, and eventual radiologic recovery—together with the clinical text to distinguish this entity from other cystic lung diseases (e.g., *Pneumocystis jirovecii* pneumonia, Langerhans cell histiocytosis, Birt–Hogg–Dubé syndrome). The unified, uniformly sized image panels highlight how temporal evolution in a single patient can be represented as a structured longitudinal multimodal item in our dataset.



CT, lung window. Marked improvement with near-normal parenchyma.



CT, lung window. Few residual discrete cysts.

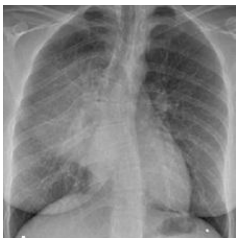


CT, coronal view. Almost complete radiologic recovery.

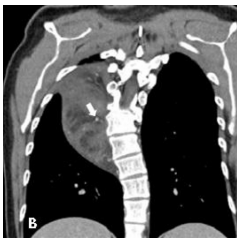
Figure 18: Late follow-up CT eight months after discharge, demonstrating almost complete recovery with limited sequelae.

C.3 GPT-5 Correct Case Study: Hibernoma of the Chest Wall

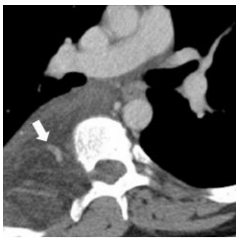
A 28-year-old woman underwent a conventional chest examination for suspected pneumonia, which incidentally revealed a right paraspinal chest-wall mass. Radiography showed a paraspinal opacity with scoliosis and rib deformities. CT demonstrated a solid, non-mineralised paravertebral lesion with fatty components slightly denser than subcutaneous fat and prominent internal serpiginous vessels. MRI confirmed predominantly fatty signal intensity that was slightly less bright than subcutaneous fat, with incomplete fat suppression, streaky soft-tissue components and slow, inhomogeneous enhancement after contrast—features typical of a hypervascular brown-fat tumour (hibernoma).



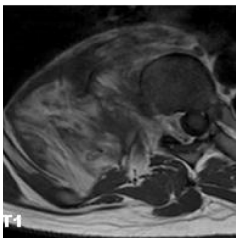
Chest radiograph. Right paraspinal mass with scoliosis and rib deformities.



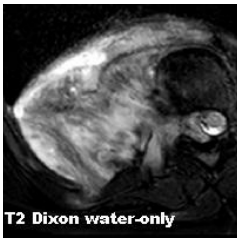
CT, coronal view. Fatty mass slightly denser than subcutaneous fat.



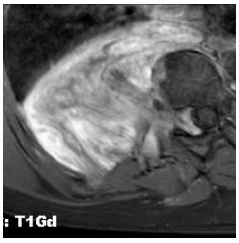
CT, axial view. Solid lesion with internal serpiginous enhancing vessels.



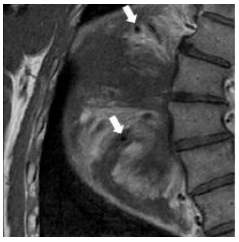
MRI T1-weighted. Predominantly fatty high signal, not as bright as subcutaneous fat.



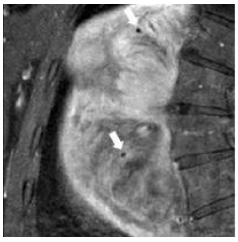
MRI T2 with fat suppression. Incomplete fat suppression with heterogeneous appearance.



MRI T1 post-contrast. Slow, inhomogeneous enhancement of soft-tissue components.



MRI T1-weighted. Fatty tissue with streaky muscle-like strands.



MRI T1 post-contrast. Prominent internal vessels with flow voids and enhancement.

Figure 19: Imaging from a paraspinal chest-wall mass later diagnosed as hibernoma: chest radiograph, contrast-enhanced CT, and multiparametric MRI (T1, T2 with fat suppression, and post-contrast). All images are stored in the dataset as individual files associated with this case.

GPT-5 answer. For the five-option question (Atypical lipomatous tumour; Clear cell sarcoma of the soft tissues; Hibernoma of the chest wall; Liposarcoma; Resolving haematoma), GPT-5 selected *Hibernoma of the chest wall* and produced the following explanation:

> Young woman with a paraspinal chest-wall mass showing predominantly fatty density with prominent internal serpiginous vessels on CT and marked, heterogeneous enhancement on MRI. The lesion is slightly less T1-hyperintense than subcutaneous fat and demonstrates avid enhancement with flow voids—features characteristic of a hypervascular brown-fat tumor (hibernoma). Atypical lipomatous tumor / well-differentiated liposarcoma and other liposarcomas typically occur in older patients, are less vascular and show thick septa or nodular non-fatty components rather than prominent intralesional vessels. Clear cell sarcoma lacks macroscopic fat and classically arises in distal extremities. A resolving hematoma would not show this degree of internal vascularity or persistent avid enhancement. Thus, hibernoma of the chest wall is most likely.

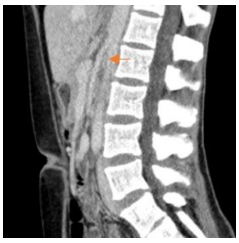
Why this reasoning is correct. GPT-5's explanation closely matches the expert interpretation of this case. It correctly recognises a fatty but *not quite normal* fat signal pattern (slightly denser than

subcutaneous fat on CT, slightly less T1-bright on MRI, incomplete fat suppression) and emphasises the key hallmark of hibernoma: a hypervascular intramuscular lesion with prominent serpiginous internal vessels and slow, inhomogeneous enhancement rather than a homogeneous pure-fat mass. It then uses these imaging cues, plus the patient's young age and paraspinal chest-wall location, to rule out the main differentials: atypical lipomatous tumour and other liposarcomas (typically less vascular, in older patients, with thick septa and nodular non-fatty components), clear cell sarcoma (no macroscopic fat, usually distal extremities), and resolving haematoma (lacking persistent vascular flow voids and avid enhancement). This stepwise, modality-aware reasoning is consistent with the teaching point for hibernoma and leads to the correct diagnosis for this MedThinkVQA item.

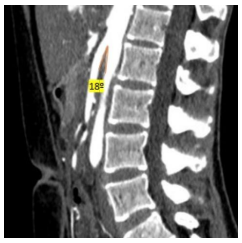
C.4 GPT-5 Error Case Study

C.4.1 GPT-5 Error Case Study: Combined Wilkie, Nutcracker, and May–Thurner Syndromes

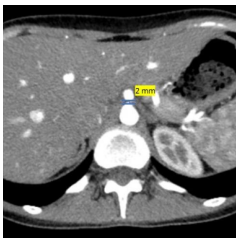
A 26-year-old woman with a three-year history of weight loss and postprandial abdominal discomfort, prior anorexia nervosa, and known pelvic congestion syndrome underwent contrast-enhanced abdominal CT. Imaging demonstrated: (i) compression of the third portion of the duodenum between the superior mesenteric artery (SMA) and aorta with reduced aortomesenteric angle and distance (Wilkie / SMA syndrome), (ii) compression of the left renal vein between the aorta and SMA with the classic "beak sign," proximal left renal vein dilatation and engorged left ovarian and pelvic veins (Nutcracker syndrome), and (iii) compression of the left common iliac vein by the right common iliac artery against the lumbar spine (May–Thurner syndrome), together with gastric and proximal duodenal dilatation.



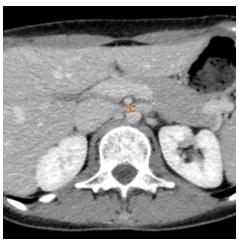
Sagittal CT. Third portion of the duodenum compressed between aorta and SMA.



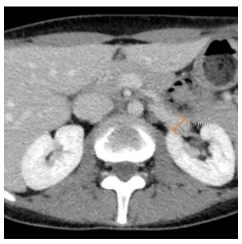
Aortomesenteric angle. Markedly reduced SMA–aorta angle.



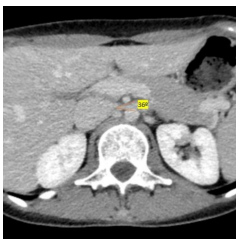
Aortomesenteric distance. Narrowed aortomesenteric distance with duodenal compression.



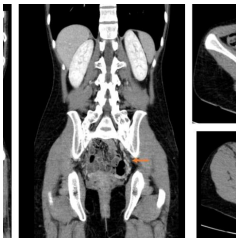
LRV beak sign. Abrupt narrowing of the left renal vein between aorta and SMA.



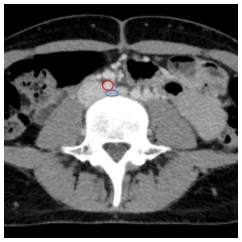
LRV dilatation. Proximal enlargement of the left renal vein.



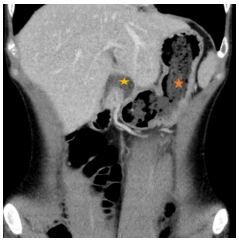
Beak geometry. Angulated, tapered appearance of the compressed left renal vein.



Pelvic veins. Tortuous left ovarian vein and dilated pelvic venous plexus.



May–Thurner. Compression of the left common iliac vein by the right common iliac artery.



Global effect. Dilatation of stomach and proximal duodenum from outflow obstruction.

Figure 20: Nine CT slices from a single case demonstrating three concurrent vascular compression syndromes: Wilkie (SMA) syndrome, Nutcracker syndrome with pelvic venous congestion, and May–Thurner syndrome. All images are stored as separate files but belong to one MedThinkVQA item.

GPT-5 answer (incorrect). For the five-option question (iliac vein compression by an occult pelvic mass; hereditary megaduodenum; scleroderma with duodenal involvement; superior mesenteric artery syndrome; Nutcracker syndrome and May–Thurner syndrome), GPT-5 selected *superior mesenteric artery syndrome* and produced the following explanation:

> CT demonstrates a markedly reduced aortomesenteric angle ($\approx 18^\circ$) and aortomesenteric distance ($\sim 2\,\mathrm{mm}$), which are diagnostic measurements for compression of the third portion of the duodenum by the SMA, that is, superior mesenteric artery (SMA) syndrome. The clinical context of a young woman with significant weight loss and anorexia nervosa, plus postprandial abdominal discomfort, is classic for SMA syndrome due to loss of the mesenteric fat pad. While other vascular compression findings may be present, the imaging and symptoms most directly support SMA syndrome as the primary diagnosis.

Why GPT-5 is wrong: failure of multi-image reasoning. Although this explanation is internally coherent for an *isolated* Wilkie / SMA syndrome, it systematically ignores the additional information carried by the other six images. Three CT slices do quantify the reduced aortomesenteric angle and distance and show duodenal compression, but the remaining views encode two further, independent haemodynamic abnormalities:

- Several slices through the renal hilum show an abrupt, "beak"-shaped narrowing of the left renal vein between the aorta and SMA with proximal dilatation and engorged ovarian and pelvic veins, which is classic for Nutcracker syndrome with pelvic congestion.
- A slice at the iliac bifurcation demonstrates focal compression of the left common iliac vein by the right common iliac artery against the spine, characteristic of May–Thurner syndrome.

The ground-truth label for this MedThinkVQA item reflects the *combination* of vascular compression syndromes documented in the full nine-image set, not just SMA syndrome. However, GPT-5 effectively behaves as if it were solving a single-image task: it focuses on the most salient slice showing SMA–duodenum compression, maps this to the textbook pattern of superior mesenteric artery syndrome, and then dismisses the rest with the remark that other compression findings "may be present."

This behaviour illustrates a central limitation of current multimodal LLMs on genuinely multi-image cases. Instead of integrating heterogeneous evidence across different anatomical levels and mechanisms (duodenal obstruction, renal venous outflow obstruction, iliac venous compression), the model implicitly treats the images as redundant views of one problem and optimises for a single, locally consistent story. In other words, GPT-5 correctly explains one mechanism, but it fails at joint, cross-region reasoning over a curated set of complementary images, which is precisely what this multi-syndrome case is designed to test.

C.4.2 GPT-5 Error Case Study: Longitudinal Cystic Pulmonary Tuberculosis

This MedThinkVQA item is a longitudinal case of a previously healthy nine-year-old boy with cystic pulmonary tuberculosis. The clinical course spans almost eleven months: initial presentation with fever, cough, dyspnoea and diffuse micronodular infiltrates; multiple bilateral spontaneous pneumothoraces requiring repeated chest drainage and intensive care; evolution on CT from diffuse micronodules and ground-glass opacities to extensive thin-walled cystic change with recurrent pneumothoraces; and, after prolonged anti-tuberculous therapy, near-complete radiologic recovery with only a few residual cysts and fibrotic sequelae. All time points are encoded as separate images but linked to a single question whose ground-truth diagnosis is *cystic pulmonary tuberculosis*.

GPT-5 answer (incorrect). For the five-option question (Pneumocystis *jirovecii* pneumonia; Lymphoid interstitial pneumonia; Langerhans cell histiocytosis; Cystic pulmonary tuberculosis; Birt–Hogg–Dubé syndrome), GPT-5 chose *Pneumocystis* jirovecii *pneumonia* and produced the following explanation:

> The imaging shows an initial diffuse interstitial/ground-glass process (crazy-paving appearance) followed by the development of numerous thin-walled cysts/pneumatoceles throughout both lungs and recurrent bilateral spontaneous pneumothoraces. This evolution – diffuse GGO progressing to widespread pneumatoceles with high risk of pneumothorax – is classic for *Pneumocystis jirovecii* pneumonia, especially in children with underlying or occult immunodeficiency. Langerhans cell histiocytosis typically affects young adult smokers with upper-lobe-predominant, irregular cysts and costophrenic angle sparing; lymphoid interstitial pneumonia is a chronic disease with lower-lobe cysts and not an acute febrile presentation; tuberculosis produces cavities rather than numerous thin-walled cysts; Birt–Hogg–Dubé syndrome presents in adults with basilar subpleural cysts without acute diffuse ground-glass opacities. Thus, *Pneumocystis jirovecii* pneumonia best fits the overall picture.

Error analysis: ignoring the longitudinal time axis. GPT-5's reasoning is internally consistent for an acute, immunodeficiency-related *Pneumocystis* pneumonia, but it fails as a longitudinal interpretation of this case. Several key aspects of the time series are either misread or ignored:

- **Disease duration and follow-up.** The patient is an immunocompetent nine-year-old boy followed over almost eleven months, with documented near-complete radiologic recovery after prolonged anti-tuberculous therapy. This long clinical evolution with structured follow-up CTs is far more typical of tuberculosis than of uncontrolled *Pneumocystis* infection, which in an undiagnosed immunodeficient child would be expected to progress or relapse rather than steadily resolve.
- **Full temporal chain, not a single snapshot.** GPT-5 effectively compresses the sequence "diffuse micronodules/ground-glass → extensive thin-walled cysts → gradual resolution" into a standard short-course template for *Pneumocystis* pneumonia. It focuses on the middle phase (GGO with cysts and pneumothoraces) and treats the early and late time points as redundant, rather than evidence of a slowly evolving, ultimately reversible granulomatous infection under long-term treatment.
- **Misconception about tuberculosis morphology.** The explanation assumes that tuberculosis "produces cavities rather than numerous thin-walled cysts," implicitly excluding cystic or pneumothorax-prone forms of TB. However, the discussion explicitly describes cystic pulmonary tuberculosis as a rare but recognised entity in which diffuse nodules and ground-glass change can evolve into confluent thin-walled cysts with recurrent pneumothoraces, followed by gradual radiologic improvement once therapy is effective.
- **Underuse of sequential improvement as a diagnostic cue.** A core signal in this longitudinal case is the progressive improvement across follow-up CTs: cysts shrink, pneumothoraces resolve, and only sparse cystic or fibrotic sequelae remain. Such a trajectory is strongly supportive of treated infection (here, tuberculosis) and less compatible with chronic interstitial or cystic diseases (such as LCH, LIP, Birt–Hogg–Dubé) that typically show persistent or progressive structural damage.

C.4.3 GPT-5 Error Case Study: Multimodal-Hydatid

This MedThinkVQA item is a multimodal, multi-organ case of disseminated hydatid disease in a 19-year-old man presenting with seizures and low-grade fever. Cranial CT and MRI show a solitary parietal cystic lesion with mild peripheral enhancement and no diffusion restriction; chest X-ray and body CT demonstrate multiple, well-defined peripheral pulmonary nodules with water attenuation; thoracic ultrasound confirms the cystic nature of the lung lesions; and abdominal CT reveals a cystic hepatic lesion with peripheral calcification. Serology for *Echinococcus granulosus* confirms the diagnosis, and the ground-truth label for the item is *disseminated hydatid disease*.

GPT-5 answer (incorrect). For the five-option question (Disseminated hydatid disease; Lung: pulmonary septic emboli; Pulmonary metastases; Brain: cystic brain tumour; Brain abscess), GPT-5 selected *Brain abscess* and produced the following explanation:

> The clinical presentation of seizures and febricula along with laboratory findings of leukocytosis and neutrophilia suggests an infectious or inflammatory process. The absence of respiratory symptoms makes options related to pulmonary pathology less likely as the primary diagnosis. Given the neurological symptoms, the most plausible diagnosis would involve the brain. Between options D and E, brain abscess (E) is more likely to present acutely with seizures and fever, and is often associated with leukocytosis and neutrophilia, whereas a cystic brain tumor (D) might not typically present with such acute infectious symptoms. The imaging findings, although not described, would likely show a localized area of infection in the brain for an abscess, which aligns with the clinical picture provided.

Error analysis: failure to integrate multimodal, multi-organ evidence. GPT-5's reasoning treats this case as if it were a purely neuroradiologic vignette: it focuses on seizures, low-grade fever and inflammatory markers, implicitly assumes the key information is "ring-enhancing brain lesion plus infection," and then chooses brain abscess by comparing only options D and E. This single-modality shortcut neglects almost all of the structured multimodal evidence presented:

- **Pulmonary imaging is downgraded to "background".** Chest X-ray and chest CT clearly show numerous, well-defined, peripheral pulmonary nodules with water attenuation and no features of suppurative consolidation or infarction. Thoracic ultrasound further confirms that these nodules are true cysts (anechoic with posterior acoustic enhancement), a pattern much more typical of hydatid disease than septic emboli or metastases. GPT-5's statement that pulmonary options are "less likely" because of absent respiratory symptoms ignores that hydatid cysts are often asymptomatic in the lungs and that imaging, not symptoms, carries the main diagnostic weight here.
- **Hepatic cyst is completely ignored.** Body CT demonstrates a classic hydatid cyst in the left hepatic lobe with a well-defined cystic lesion and peripheral calcification of the pericyst. This second non-brain, non-lung organ involvement is a strong clue for systemic parasitic disease. GPT-5's explanation does not mention the liver at all, indicating that this modality and organ channel are effectively dropped from its reasoning.
- **Brain MRI is interpreted through a generic "ring-enhancement = abscess" template.** The brain lesion in this case is a solitary, CSF-like cyst with a thin rim, mild peripheral enhancement, no diffusion restriction, and only moderate oedema. These features, particularly the absence of diffusion restriction and the characteristic low-signal rim on T2-weighted images, are more consistent with a hydatid cyst than with a pyogenic abscess. GPT-5 instead imagines a typical abscess pattern and even states that the imaging findings "would likely" show a focal infection, revealing that it is reasoning from a mental template rather than actually integrating the provided MRI sequences.
- **Cross-organ pattern is never assembled.** The correct diagnosis requires noticing a triad: (i) multiple cystic pulmonary lesions, (ii) a calcified hepatic cyst, and (iii) a solitary brain cyst with hydatid-like MRI characteristics. Taken together, these represent a classic multi-organ, haematogenously disseminated parasitic infection. GPT-5 never composes this cross-organ, cross-modality picture; instead, it chooses the most salient single modality (brain MRI/CT) and maps the entire case to a focal intracranial infection.

C.5 Step-level Evaluation Case Study: Bilateral Subareolar Abscesses

C.5.1 Error Taxonomy for Model Responses

- **Reasoning Error (Reasoning Err).** The imaging and clinical facts themselves are correctly stated, but the model misconstructs the causal chain or diagnostic logic, reaches an incorrect conclusion, selects an inappropriate differential diagnosis, or uses correct facts to support an incorrect judgment.
- **Image Understanding Error (Image Understanding Err).** The model misreads or hallucinates objective visual information that is directly apparent on the images (e.g., location, morphology, signal intensity, enhancement pattern, extent), and the error originates from image interpretation rather than downstream reasoning.
- **Clinical Scenario Error (Clinical Scenario Err).** The model misunderstands, misquotes, or inaccurately restates clinical information provided in the stem (e.g., symptoms, age, duration, physical findings, laboratory data), or treats information that is not given as if it were explicitly provided.
- **Medical Knowledge Error (Medical Knowledge Err).** The model violates general medical knowledge or textbook-level facts—such as typical manifestations of a lesion, epidemiology, key features for differential diagnosis, or common complications—in ways that are independent of how the specific images or clinical history are read.

Clinical scenario and imaging summary. A 39-year-old woman presented with unilateral, single-orifice nipple discharge (galactorrhoea) from the right breast and no fever, palpable mass, or skin inflammatory changes. Initial mammography (craniocaudal and oblique views) and ultrasound of both breasts were reported as normal. Galactography of the right breast showed an intraductal filling defect in a canalised nipple duct and raised the suspicion of an intraductal papilloma. Subsequent contrast-enhanced breast MRI revealed bilateral subareolar fluid-filled cavities without internal enhancement, with intensely enhancing peripheral capsules and surrounding inflammatory tissue. A repeat ultrasound, performed after the patient clarified that the discharge was actually bilateral, showed bilateral subareolar hypoechoic nodules with anechoic centres, consistent with abscesses. Cytology confirmed bilateral subareolar abscesses.

GPT-5 diagnosis and explanation. The multiple-choice options included galactocele, fat necrosis, intracystic papillomas, and subareolar abscesses (correct answer: subareolar abscesses). GPT-5 instead chose *intracystic papillomas* and produced the following free-text explanation:

> Unilateral, single-duct nipple discharge suggests a focal intraductal process. Mammography is essentially negative. Ultrasound shows a dilated retroareolar duct containing an intracystic solid nodule. MRI demonstrates a small enhancing subareolar intraductal lesion without surrounding inflammatory change. Findings are classic for an intraductal/intracystic papilloma, not a galactocele (no fat–fluid level), abscess (no pain, erythema or inflammatory changes), or fat necrosis.

Step-level parsing and expert error annotation. We use GPT-5 to decompose its own explanation into discrete reasoning steps, then ask a breast imaging expert to judge the factual correctness of each step and assign an error type when applicable. Error types follow our taxonomy: *Clinical Scenario Err*, *Image Understanding Err*, *Reasoning Err*, and *Medical Knowledge Err* as defined in Section **??**.

1. **Step 1:** "Unilateral, single-duct nipple discharge suggests a focal intraductal process."
   Expert factual judgment: **Incorrect**. *Error type:* **Clinical Scenario Err**.
   Although the original history was recorded as unilateral, single-orifice discharge, the case narrative later clarifies that the galactorrhoea is actually bilateral. GPT-5 treats the initial, incomplete history as definitive and over-anchors on a classic teaching pattern for intraductal papilloma, misrepresenting the true clinical scenario.
2. **Step 2:** "Mammography is essentially negative."
   Expert factual judgment: **Correct**. *Error type:* None.
   The report states that mammography did not show any pathological findings, so this step accurately reflects the imaging description.

3. **Step 3:** "Ultrasound shows a dilated retroareolar duct containing an intracystic solid nodule."
   Expert factual judgment: **Incorrect**. *Error type:* **Image Understanding Err**.
   In reality, the initial ultrasound was normal, and the follow-up ultrasound showed bilateral subareolar hypoechoic nodules with anechoic centres (abscesses), not a duct with an intracystic solid nodule. GPT-5 hallucinates a very specific retroareolar papilloma pattern directly from the images, so the primary error is misreading the imaging rather than subsequent reasoning.
4. **Step 4:** "MRI demonstrates a small enhancing subareolar intraductal lesion without surrounding inflammatory change."
   Expert factual judgment: **Incorrect**. *Error type:* **Image Understanding Err**.
   The MRI actually showed bilateral fluid-filled cavities without internal enhancement, surrounded by an intensely enhancing capsule and inflammatory tissue with moderate enhancement. GPT-5 inverts this pattern (solid enhancing nodule, no inflammation), again indicating a failure to correctly translate the described MRI findings into its intermediate reasoning.
5. **Step 5:** "These imaging findings are classic for an intraductal/intracystic papilloma."
   Expert factual judgment: **Incorrect**. *Error type:* **Reasoning Err**.
   Given the true imaging pattern (bilateral subareolar fluid collections with enhancing capsule and surrounding inflammation), the correct diagnosis is abscess. Here GPT-5 generalises from its misread ultrasound/MRI steps and commits a diagnostic reasoning error: it asserts a papilloma pattern even though the combination of bilateral disease, fluid-filled cavities, and inflammatory enhancement better matches abscess.
6. **Step 6:** "The findings are not compatible with a galactocele because there is no fat–fluid level."
   Expert factual judgment: **Correct as stated**. *Error type:* None.
   The expert accepts that there is no explicit fat–fluid level described. While the accompanying textual explanation overstates how definitively this excludes a galactocele, the literal step ("no fat–fluid level is seen") is factually consistent with the report, so no image or scenario error is assigned at the step text level.
7. **Step 7:** "The findings are not compatible with an abscess because there is no pain, erythema, or inflammatory changes."
   Expert factual judgment: **Incorrect**. *Error types:* **Reasoning Err**, **Clinical Scenario Err**, **Medical Knowledge Err**.
   This step bundles three distinct problems:
   - ***Clinical Scenario Err***: It equates the absence of external pain, erythema, or skin signs with the absence of inflammation, ignoring the MRI description of surrounding inflammatory tissue and the later clarification of bilateral symptoms.
   - ***Medical Knowledge Err***: It assumes that breast abscesses cannot present without overt systemic or cutaneous signs, which contradicts the broader discussion that non-puerperal central abscesses can be subtle.
   - ***Reasoning Err***: It uses this flawed premise to exclude the true diagnosis (abscess), despite direct imaging and cytology evidence to the contrary.
8. **Step 8:** "The findings are not compatible with fat necrosis."
   Expert factual judgment: **Correct**. *Error type:* None.
   The described lesions are fluid-filled cavities with enhancing capsules and inflammatory tissue, without oil cysts or internal fat signal; this pattern is more characteristic of abscess than fat necrosis, so excluding fat necrosis at this step is reasonable.

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)

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. You
Given ONLY the CLINICAL HISTORY text and the list of candidate diagnosis OPTIONS, dec
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 patte
  "Diagnosis: X", "biopsy-proven X".
• L2 Explicit synonym/acronym/eponym/foreign-language variant of the diagnosis label
  (e.g., "MI" for myocardial infarction; "Osler-Weber-Rendu" for HHT).
• L1 Explicit but uncertain mention of the diagnosis label (or its synonym/acronym/ep
  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
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 diagnos
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 Cl
  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,e
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, Discussi
and relevant teaching value/domain knowledge.
Rules:
1) Evaluate whether each step is correct or reasonably supported; reasonable analysis
2) Mark True if the step is explicitly supported, correctly implied, or logically rea
```

```
   and your teaching value/domain knowledge.
3) Mark False only if the step is clearly wrong, contradictory, or cannot be reasonab
   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 relev
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,
    }
    // ... 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 sente
    "rubric_3_imaging": {"score": 0|1|2, "explanation": "<1-2 sentences>"},
    "rubric_4_reasoning_differentials": {"score": 0|1|2, "explanation": "<1-2 senten
    "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 7: 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 8: 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 9: 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 MODALITY STATS IN DATASET

We analyze imaging modality statistics of our dataset using the GPT-5.1-mini model. Each case consists of one or more images and associated textual metadata. For every image, we infer a fine-grained `imaging_technique` from the alt-text and, when necessary, from the imaging findings. These fine-grained techniques are then mapped into a set of aggregated modality groups stored in `imaging_technique_group`. At the case level, we define `modalities_count` as the number of distinct aggregated modality groups present among all images belonging to a given case.

J.1 AGGREGATED MODALITY CATEGORIES

Across the training split, we obtain 14 aggregated modality groups: *X-ray / plain radiograph*, *CT*, *Ultrasound*, *Other / Unknown*, *Histology / pathology*, *MRI*, *Clinical photo*, *Mammography*, *Fluoroscopy*, *PET-CT*, *Nuclear medicine*, *Angiography*, *Endoscopy*, and *PET*. The test split covers the same set of categories except Mammography, i.e., 13 aggregated modalities in total.

J.2 PER-IMAGE MODALITY DISTRIBUTION

Table 10 reports the frequency of each aggregated modality computed over all images in the training and test splits. The training split contains 49,159 images, while the test split contains 6,090 images.

Table 10: Per-image distribution of aggregated imaging modalities in the training and test splits. Counts are absolute image counts, and percentages are relative to the total number of images in each split.

Modality	Train images	Train (%)	Test images	Test (%)
X-ray / plain radiograph	3,803	7.74	398	6.54
CT	20,766	42.24	2,510	41.22
Ultrasound	4,092	8.32	505	8.29
Other / Unknown	312	0.63	44	0.72
Histology / pathology	929	1.89	176	2.89
MRI	15,060	30.64	2,156	35.40
Clinical photo	407	0.83	79	1.30
Mammography	393	0.80	0	0.00
Fluoroscopy	1,059	2.15	57	0.94
PET-CT	238	0.48	55	0.90
Nuclear medicine	243	0.49	13	0.21
Angiography	1,623	3.30	56	0.92
Endoscopy	180	0.37	32	0.53
PET	54	0.11	9	0.15
Total	49,159	100.00	6,090	100.00

Overall, CT and MRI dominate both splits, together accounting for approximately 73% of training images and 76% of test images, followed by Ultrasound and X-ray / plain radiograph. The remaining modalities (e.g., Angiography, Histology / pathology, Fluoroscopy, PET-CT) appear less frequently but provide additional multimodal diversity.

J.3 PER-CASE MODALITY DIVERSITY

Beyond per-image counts, we characterize the multimodal diversity of each case using `modalities_count`. This quantity measures how many distinct aggregated modality groups are present in a case's image set. Table 11 summarizes the distribution of `modalities_count` for the training and test splits.

The training split contains 7,729 cases with an average of 1.84 modalities per case, while the test split contains 751 cases with an average of 2.13 modalities per case. In both splits, most cases involve one or two modalities, but a non-trivial fraction of cases exhibit higher multimodal diversity.

Table 11: Distribution of the number of distinct aggregated modalities per case (`modalities_count`) in the training and test splits.

# Modalities	Train cases	Train (%)	Test cases	Test (%)
1	3,359	43.46	230	30.63
2	2,775	35.90	291	38.75
3	1,178	15.24	152	20.24
4	345	4.46	61	8.12
5	64	0.83	15	2.00
6	5	0.06	2	0.27
7	2	0.03	0	0.00
8	1	0.01	0	0.00
Total	7,729	100.00	751	100.00

In the training split, 79.4% of cases contain at most two modalities. The test split is slightly more multimodal on average: 69.4% of cases have one or two modalities, and around 30.6% contain three or more modalities.

J.4 Common Modality Combinations at the Case Level

We also examine which combinations of modalities co-occur at the case level. Here, a modality combination is defined as the set of distinct aggregated modality groups present in a given case. We report statistics over these sets without regard to the number of images per modality.

For the test split (751 cases), the five most frequent modality combinations are:

- CT alone (128 cases, 17.0%),
- MRI alone (84 cases, 11.2%),
- CT + MRI (84 cases, 11.2%),
- CT + X-ray / plain radiograph (54 cases, 7.2%),
- CT + Ultrasound (41 cases, 5.5%).

For the training split (7,729 cases), the most frequent combinations are:

- CT alone (1,496 cases, 19.4%),
- MRI alone (1,100 cases, 14.2%),
- CT + X-ray / plain radiograph (679 cases, 8.8%),
- CT + MRI (610 cases, 7.9%),
- CT + Ultrasound (372 cases, 4.8%).

K LONGITUDINAL STUDIES IN MEDTHINKVQA

We additionally track whether each case contains longitudinal follow-up imaging (i.e., multiple time points for the same patient). In the held-out test set, 212 out of 751 cases are longitudinal ($\approx 28.2\%$). In the training set, 1,947 out of 7,729 cases are longitudinal ($\approx 25.2\%$). Aggregating across both splits, MedThinkVQA contains 2,159 longitudinal cases out of 8,480 total cases ($\approx 25.5\%$), indicating that roughly one quarter of the dataset requires reasoning over temporal disease evolution.

Table 12: Prevalence of longitudinal studies in MedThinkVQA.

Split	# Cases	# Longitudinal	Share (%)
Train	7729	1947	25.2
Test	751	212	28.2
Overall	8480	2159	25.5

L PROMPTS FOR STEPWISE EXPLANATION EXTRACTION

L.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 evider
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 -> dec
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 distir
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.
```

L.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 13: 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 14: 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 15: 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 16: 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.

M 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%.

M.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%)

M.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%)

M.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%)

M.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%)

N 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 8 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.

N.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 $h = \max(1, \lfloor |\mathrm{ids}|/2 \rfloor)$; set $q_1 = T^{-1}(\mathrm{ids}[:h])$, $q_2 = T^{-1}(\mathrm{ids}[h:])$, $k = |\mathrm{ids}| - h$. 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
    $\rho_i \leftarrow \max_{0 \leq j \leq |\textit{region}| - L} \left(1 - \frac{\mathrm{Lev}(\textit{region}[j{:}j{+}L],\ q_{2,i})}{L}\right)$;
    $s_i \leftarrow \mathbf{1}[\rho_i \geq Y]$
$Z \leftarrow \frac{1}{n}\sum_i s_i$; $\quad \bar{\ell} \leftarrow \frac{1}{n}\sum_i \rho_i$;
return $Z, \bar{\ell}$

O 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.

P RUBRIC FOR DISCUSSION EVALUATION

P.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.

P.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.

P.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).

P.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).

P.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.