RADTHINKING: TEACHING VISION-LANGUAGE MODELS TO THINK THE WAY RADIOLOGISTS ARE TAUGHT

Anonymous authors
Paper under double-blind review

ABSTRACT

Radiology education teaches a way of thinking, not a set of answers. A trainee learns to screen each region of an organ, characterize each finding, compare it with earlier scans, and apply a clinical guideline, a rule over the finding's representation: its size, enhancement, change, and a few more. Vision-language models are trained on images paired with answers and judged by the answer alone; the representation in between goes unchecked. We present **RadThinking**, which turns this way of thinking into supervision: the representation a radiologist is taught to build becomes the representation the model learns to write. Each step of reading a scan becomes a question with a reference answer, grounded in a 3D bounding box; questions link into chains; five clinical guidelines become decision tables, explicit rules from a finding's representation to its category. The representation is grounded, so each answer in it is scored at its box on the scan, and explicit, so a radiologist can check it as they check a trainee. RadThinking holds 122,127 questions over 5,365 abdominal CT scans of 2,679 patients from two countries. Across eight models, the best gets every answer right in only 23% of chains, numerical errors aside, and turning thinking on brings no consistent gain. Looking inside open-weight models shows why. Once its intermediate answers are written, Qwen3.8-27B follows that text, not the images, right or wrong. Before it writes, a linear probe on its hidden state predicts the final answer about 9 points more often than the model does. The model holds more than it writes and follows what it writes. So the representation it writes must be correct, and RadThinking supervises every answer in it. Trained on it, a 4B model outperforms zero-shot GPT-6, Claude Opus 5, and Gemini 3.8 Flash: 64% versus 52% on individual questions and 30% versus 23% on chains. Radiology education already wrote this representation down. A small model that learns it outperforms much larger models that have not.

1 INTRODUCTION

A radiologist reading a CT scan for cancer writes down what the decision depends on. For each finding, that is a few answers, such as size, enhancement, and change, which a clinical guideline turns into a category: the finding's *representation* (Chernyak et al., 2018; Silverman et al., 2019). Vision-language models (VLMs) learn differently. They are trained on images paired with answers and judged by the answer alone; what they write in between is checked against nothing, yet it becomes the input for what they write next (Wei et al., 2022; Turpin et al., 2023; Buckley et al., 2026; Bassi et al., 2024). A right conclusion can therefore depend on a wrong intermediate answer that stays hidden. In this paper we ask: *how should the representation a model writes be structured and supervised so that each answer can be checked against the scan?*

We hypothesize that the representation radiology education teaches is the one to supervise: if a model writes those answers correctly, its conclusion follows. To test this, we build RadThinking. Each step of reading a scan becomes a question with a reference answer, the target for scoring and supervision, grounded in a 3D bounding box and linked to the questions it depends on. An endpoint question with every question it depends on forms a *linked-question chain*. The endpoint asks for the conclusion; the other questions ask for intermediate answers. Five clinical guidelines are implemented as decision tables, whose rows map imaging features to categories. Each row is a *rule*. Existing resources supervise one thing each: question and report datasets score a conclusion (Lau et al., 2018; Bai et al., 2024; Johnson et al., 2019), and chain-of-thought and process supervision leave intermediate answers

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

as unchecked free text (Wei et al., 2022; Lightman et al., 2024). RadThinking gives every answer a reference, so a chain is scored as a whole, and one representation serves evaluation, analysis, and training (Figure 1). Because 1,411 of its 2,679 patients have two or more scans, change over time is a question with a reference answer too.

We evaluate eight vision-language models without training them on RadThinking (zero-shot evaluation), five open-weight and three closed-source, on data that were not public when they were trained. *Individual visual question answering (VQA)* asks one question at a time; *linked-question evaluation* asks a whole chain in one prompt. Three results stand out. **(1)** Performance varies across phases, and no model gets every categorical answer of a chain right in more than 23.5% of chains. This share is *categorical chain accuracy*; it excludes numerical answers. **(2)** Native thinking, the model's built-in reasoning mode, gives no consistent gain; turning it on cuts Qwen3.8-27B (Qwen Team, 2026)'s Phase 4 (decision) accuracy in chains from 47.8% to 7.4%. **(3)** Correct conclusions do not imply correct intermediate answers; 63% of Hulu-Med-32B (Jiang et al., 2025)'s correct conclusions come with at least one wrong intermediate answer. Looking inside the open-weight models shows why. Once its intermediate answers are written, Qwen3.8-27B follows that text rather than the images. Yet a linear probe, a classifier trained on its internal activation vector (hidden state) before it writes, predicts the conclusion about 9 percentage points more often than the model does. The model holds more than it writes and follows what it writes. More importantly, supervising every answer in the representation repairs this: fine-tuned on RadThinking, a model with 4 billion parameters (4B) scores above the best zero-shot model in individual VQA and in chains. The conclusion passes through the representation the model writes, so the representation is what to supervise. Radiology wrote this representation down for its trainees, not yet for its models.

To our knowledge, RadThinking is the first dataset in which the steps of a radiology read are questions with reference answers grounded in 3D bounding boxes. We make four contributions.

- **A dataset that records the steps of a radiology read as questions with reference answers.** RadThinking has 122,127 questions over 5,365 abdominal CT scans of 2,679 patients from two countries. Each scan-level question is grounded in a 3D bounding box, each reference answer records its evidence source, each question lists the questions it depends on, and five clinical guidelines are implemented as decision tables (Section 2).
- **A benchmark that scores intermediate and conclusions together.** Across eight vision-language models evaluated zero-shot, categorical chain accuracy does not exceed 23.5%, and enabling native thinking brings no consistent gain (Section 3).
- **Evidence that the conclusion follows the written intermediate answers rather than the images.** In Qwen3.8-27B, swapping two written sizes with the images unchanged lowers reference-option agreement, the share whose highest-scoring label matches the reference answer, from 48.1% to 11.1%, and removing images while retaining reference sizes leaves it at 44.4%. For Qwen3.8-27B and Hulu-Med-32B, a linear probe on the hidden state before writing predicts the conclusion about 9 percentage points more often than the model does (Section 4.1).
- **Fine-tuning on the reference answers raises intermediate and conclusion accuracy.** Supervised fine-tuning (SFT) with linked questions raises categorical chain accuracy from 0.0% to 28.3% for Primus (NV-Reason-CT) (Myronenko et al., 2026) and from 10.9% to 30.6% for Hulu-Med-32B. After SFT and reinforcement learning (RL), the 4B Primus scores above the best zero-shot model on the Phase 1–4 average of individual VQA (59.6% versus 50.2%) and on categorical chain accuracy (30.4% versus 23.5%) (Section 4.2).

Related work. Public radiology datasets give one form of supervision each. Segmentation datasets locate lesions (Bilic et al., 2019; Heller et al., 2019; Antonelli et al., 2021). Report datasets pair a scan with the radiologist's text but ask no questions (Johnson et al., 2019; Hamamci et al., 2026; Blankemeier et al., 2026). Question datasets ask for one attribute of one image or scan (Lau et al., 2018; Liu et al., 2021; Bai et al., 2024; Zhang et al., 2024; Ji et al., 2024; Chen et al., 2025). Chest ImaGenome and 3D-RAD link the scans of a patient over time but have no reasoning traces (Wu et al., 2021; Gai et al., 2025), and CheXthought has reasoning traces from radiologists but no questions and no 3D scans (Sharma et al., 2026). None lets a model be scored on the steps of a radiology read, so a conclusion can be right for the wrong reason. RadThinking is the only dataset that is longitudinal, grounds every one of its 122,127 questions in a 3D box, stores *reference rationales*,

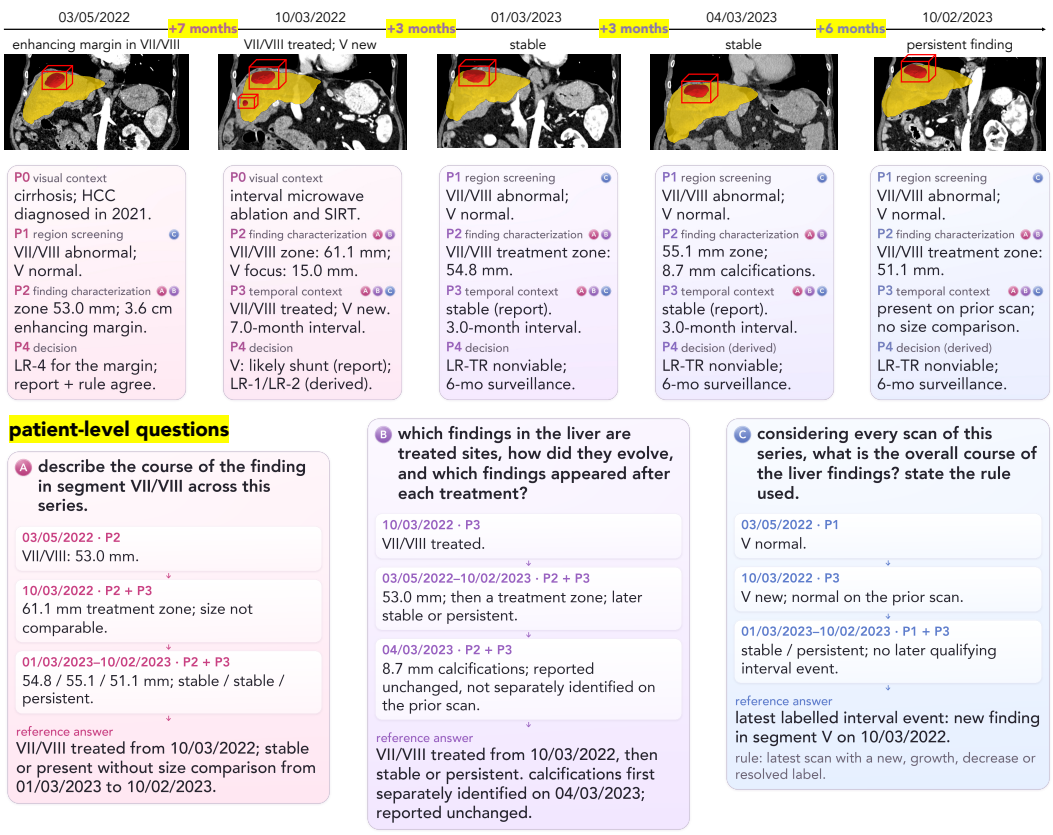

Figure 1: **One patient, five abdominal CT scans, 1.6 years: every answer is grounded in a 3D bounding box.** We show scan-level questions, asked for each scan, and patient-level questions, which span the scans of one organ. Liver in yellow; finding and its box in red. P0 to P4 denote workflow phases from visual context to decision; A to C denote patient-level questions. LR denotes a category in the Liver Imaging Reporting and Data System (LI-RADS) (Chernyak et al., 2018); LR-TR denotes its treatment-response category.

written explanations of organ assessments and diagnoses, and applies five clinical guidelines, and it is the only abdominal CT dataset with patient-level questions across scans. Every step can therefore be scored, probed, and trained on, not only the conclusion (Appendix Table 10 and Appendix C).

2 The RadThinking Dataset

2.1 Source cohort and sources of evidence

RadThinking is built on a longitudinal abdominal CT collection of cancer patients. The CT scans and their radiology reports were collected from institutions in Switzerland and Turkey under institutional review board and ethics approvals. The scans date from 2010 to 2025. RadThinking contains 2,679 patients and 5,365 CT scans. A scan is one CT volume acquired in one contrast phase, with the radiology report of the study it belongs to. A study with several usable volumes contributes each as its own scan. The report describes the whole study, so a reported imaging feature, such as enhancement, may come from a contrast phase that is not in the scan; the reference rationale, the written explanation of an assessment or diagnosis, marks such features as not assessable on the scan's contrast phase (Section 2.3). Of the scans, 241 are control scans with no recorded lesion. In total, 1,411 patients have two or more scans, and for 1,347 of them the scans span more than one day. The median follow-up is 321 days, with an interquartile range (IQR) of 104 to 784 days, and 635 patients are followed for one year or longer. The scans cover four contrast phases: portal venous (2,600), late arterial (1,373), early

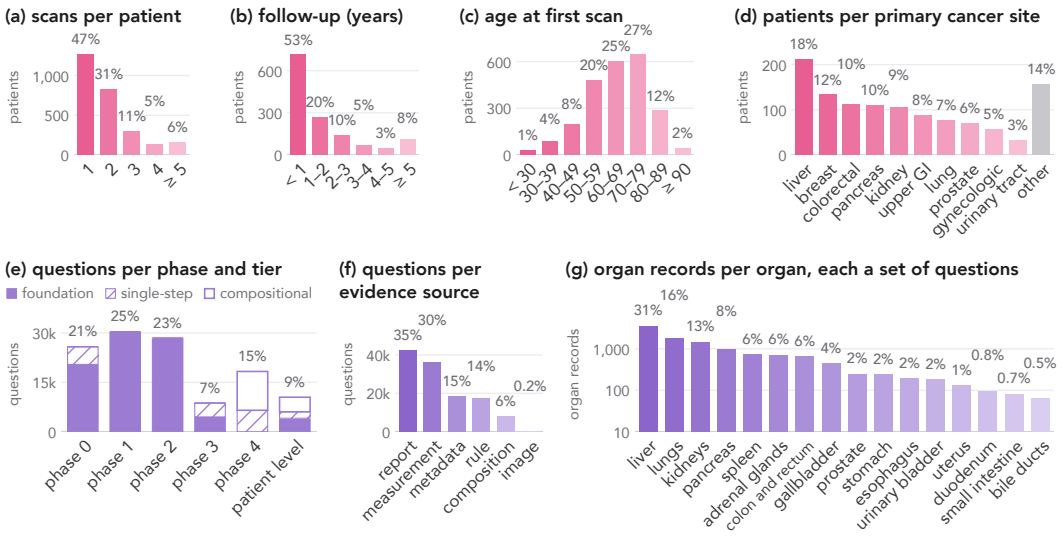

Figure 2: **Most of the 2,679 patients in RadThinking are followed over time, and most questions ask for evidence rather than a conclusion.** Rose bars (a to d) count patients, lavender bars (e to g) count questions or organ records (questions about one organ on one scan); the number above a bar is its share of the panel. (a) Scans per patient: more than half of the 2,679 patients have two or more. (b) Follow-up in whole years for the 1,347 patients whose scans span more than one day (median 321 days). (c) Age at the first scan for the 2,367 patients with a recorded age (median 66 years). (d) Primary cancer site for the 1,153 patients with a recorded diagnosis, one site per patient. (e) The 122,127 questions by workflow phase and tier. Foundation questions read one field; single-step questions make one calculation or decision; compositional questions combine answers. (f) Questions by evidence source: most come from reports and measurements. (g) Organ records per organ, out of 11,741, on a logarithmic axis. Liver stands for hepatocellular carcinoma, kidney for renal cell carcinoma, urinary tract for urothelial carcinoma; GI, gastrointestinal.

arterial (670), and non-contrast (685); 37 scans have other contrast phases. The cohort has 1,360 men and 1,016 women (303 unknown), with a median age of 66 years (IQR 56 to 75) at the first scan. Figure 2 plots the counts in Appendix Table 7.

Each scan comes with three source artifacts: the CT volume, the radiology report, and the scan metadata. The CT carries a 3D bounding box for every organ, organ region, and finding, given as voxel coordinates in the source volume. Every question is grounded by one of these boxes. The de-identified radiology report gives the described findings, the reported categories, and the recommended follow-up. The scan metadata gives the date, the contrast phase, and the clinical history available at the time of the scan. Measurements of a finding (longest diameter, volume, mean attenuation in Hounsfield units (HU), and adjacency to a major vessel) are computed on the CT from lesion segmentations that radiologists reviewed. Every reference answer in RadThinking records its evidence source, which describes how it was obtained from these artifacts. Section 2.5 and Appendix B describe how these artifacts yield questions and reference rationales.

2.2 Dataset composition

RadThinking is organized by patient, scan, and organ. An *organ record* is the set of questions about one organ on one scan. RadThinking contains 11,741 organ records over 16 organs; 6,551 records contain at least one finding and 5,190 are organ-level negatives: the radiologist examined the organ and reported it normal. A *finding* is one lesion or abnormality with its own questions: 10,029 findings in total; 6,234 carry CT-derived measurements (longest diameter, volume, attenuation, and vessel adjacency) and 8,227 carry a description from the report. Every question is grounded by the bounding box of the anatomy that holds its evidence: 18,812 organ boxes, 28,213 region boxes (liver segments, kidney thirds, and pancreatic head, body, and tail), and 8,491 finding boxes. The 122,127 questions span three *tiers*, which describe how much an answer combines: 87,846

foundation questions read one evidence field, 18,049 single-step questions make one calculation or decision, and 16,232 compositional questions combine earlier answers. By *workflow phase*, a step in reading a scan, there are 25,820 Phase 0 visual context questions, 30,369 Phase 1 region screening questions (21,002 normal and 9,367 abnormal regions), 28,415 Phase 2 finding characterization questions, 8,716 Phase 3 temporal context questions, and 18,313 Phase 4 decision questions. These are *scan-level questions*, asked for each scan. Another 10,494 *patient-level questions* span the scans of one organ. Phase 3 includes 4,214 *temporal-change labels*, which describe findings across scans. Phase 4 includes 11,795 reference rationales and 1,283 *guideline questions*, which apply a clinical guideline: LI-RADS (283), Bosniak (141) (Silverman et al., 2019), Fleischner (60) (MacMahon et al., 2017), and the American College of Radiology (ACR) white papers on incidental adrenal masses (503) (Mayo-Smith et al., 2017) and pancreatic cysts (296) (Megibow et al., 2017). Appendix Table 5 gives these counts for each of the 16 organs.

2.3 Questions, tiers, and reference rationales

Every question has a reference answer, a 3D bounding box, an evidence source, a tier, and links to the questions it depends on (Appendix A.4). *Scan-level questions* are asked for each scan in five workflow phases: Phase 0 visual context, Phase 1 region screening, Phase 2 finding characterization, Phase 3 temporal context, and Phase 4 decision. Workflow phases describe steps in reading a scan; contrast phase refers to CT acquisition. Phase 3 asks about time intervals and *temporal-change labels*, which describe findings across scans, such as new, stable, or resolved. *Patient-level questions* span the series of one organ. Figure 1 shows both levels; Appendix A.1 lists every question kind. Three tiers describe how much an answer combines. A *foundation* question reads one evidence field, such as size. A *single-step* question makes one calculation or decision from recorded evidence, such as change from two measurements. A *compositional* question combines earlier answers into a conclusion. These tiers span workflow phases. A *reference rationale* answers a Phase 4 assessment or diagnosis question. Clinical guidelines are written as decision tables that map imaging features to a category or management. A *rule* is the row selected for a finding. The rationale's six sections link findings and temporal context to imaging features, a rule, an assessment, and a plan. Each imaging feature is marked present, absent (the report negates it), not described (the report is silent), or not assessable (the scan's contrast phase cannot show it). The reported category stays the reference answer; the category derived from the rule is stored alongside it. Appendix B.4 defines the sections; Appendix A.3 gives an example. A *linked-question chain* contains an endpoint question and every question it depends on. The endpoint asks for the conclusion; the other questions ask for intermediate answers. Linked-question evaluation asks for all answers in one response. Individual VQA scores each question separately. The reference answers also supervise intermediate answers and conclusions.

2.4 Clinical coverage and data splits

RadThinking covers 16 organs (Appendix Table 4). The liver (3,617 organ records), lungs (1,834), kidneys (1,485), and pancreas (975) contribute most records; the remaining twelve organs contribute between 63 and 752 records each. Appendix Table 5 lists the records, findings, and questions of every organ. A recorded cancer diagnosis exists for 1,153 patients; hepatocellular, breast, colorectal, pancreatic, and renal cell carcinoma are the most frequent (Figure 2). RadThinking has an official split into training ($N_{\text{patients}} = 1495$), validation ($N_{\text{patients}} = 499$), and test sets. The split is by patient: all scans of a patient are in the same set. Training and validation use Swiss patients only. The test set has two parts: an internal test set of $N_{\text{patients}} = 499$ Swiss patients and an external test set with all 186 Turkish patients. All experiments in Sections 3 to 4.2 use the split.

2.5 How the reference rationales are built and checked

Reference rationales are built from a *fact sheet*: the measurements, report sentences, temporal-change labels, clinical context, and applicable rule for one organ record. Each rationale is checked against these facts and keeps the reported category (Appendices B.2 and B.4). *Eight physicians review random samples to refine report extraction and rationale construction.* The sampled error rate falls below 2% in round five (Appendix B.5). An automated audit checks every file for valid bounding boxes, traceable numbers, and Phase 0 to Phase 3 answers without diagnosis words (Appendix B.6). Zero-shot benchmark inputs exclude reports and reference answers (Appendix A.5).

3 BENCHMARKING EVIDENCE AND CONCLUSIONS

3.1 EVALUATION PROTOCOLS

We evaluate models zero-shot on the official test split. Individual VQA scores questions separately; linked-question evaluation scores intermediate answers and conclusions together.

Individual VQA. Each prompt contains one question and its *evidence crops*, the CT content inside its bounding boxes. The model selects an option or a set of labels, or reports a numerical value. A neutral location cue identifies the organ or finding without revealing whether it is abnormal, malignant, or changing. *Linked-question evaluation.* Each prompt contains a linked-question chain: an endpoint question and every question its answer depends on (Section 2.3). The model gives all intermediate answers and the conclusion in one response, without reference answers between questions. Scans appear in chronological order. An *organ view* contains the organ and surrounding CT context. Organ views use the same physical spacing ($1.5 \times 1.5 \times 3$ mm) and a common canvas. *Answer extraction.* A shared extractor maps responses to each question's answer format. It accepts option text, common synonyms, and equivalent units without using reference answers (Appendix E.3). Missing or unresolved answers receive no credit.

3.2 MODELS

We evaluate five open-weight vision-language models: Hulu-Med-32B (Jiang et al., 2025), Lingshu-I-8B (Xu et al., 2025), Qwen3.8-27B (Qwen Team, 2026), NVIDIA's NV-Reason-CT (Myronenko et al., 2026) (labeled Primus in this paper), and ClinFusion-32B (Yuan et al., 2026). We also evaluate three closed-source models through their APIs: GPT-6 (OpenAI, 2026), Claude Opus 5 (Anthropic, 2026), and Gemini 3.8 Flash (Google DeepMind, 2026). For models with a thinking switch, we compare *w/ thinking* and *w/o thinking* using the same prompts and inputs. For closed-source models, *min. effort* and *max. effort* denote the lowest and highest settings the provider accepts. Appendix E.3 describes model interfaces and generation settings.

3.3 METRICS

We score generated responses against reference answers. Single-choice answers use accuracy; label sets use exact match. Structured answers require every field with a known reference answer to be correct. Numerical answers use field-specific tolerances; *measurements in range* reports the share within tolerance. For diameter, the tolerance is the larger of 2 mm or 50% of the reference value. **Scan-level questions** follow five workflow phases: Phase 0: visual context; Phase 1: region screening; Phase 2: finding characterization; Phase 3: temporal context; and Phase 4: decision. We average task accuracies within each phase. The overall average weights the five phases equally and requires all five. Each scan-level answer in a linked-question chain is scored under the phase of its question. **Patient-level questions** span the scan series of one organ and are scored separately. Categorical chain accuracy is the share of eligible linked-question chains with every categorical intermediate answer and conclusion correct. Numerical errors do not affect this score. We give 95% confidence intervals from 2,000 bootstrap resamples of patients. Appendix E.3 gives scoring details.

3.4 RESULTS

(1) Performance varies across phases. Closed-source models have the highest phase averages: 51.8% for individual VQA and 59.3% for linked-question evaluation (Table 1). The best Phase 3 accuracy is 37.0% and 48.6%, respectively. *Measurements in range* is at most 35.7%. **(2) Thinking gives no consistent gain.** Thinking raises some individual VQA scores and lowers others. In linked-question evaluation, Qwen3.8-27B's categorical chain accuracy falls from 15.9% w/o thinking to 6.6% w/ thinking. For closed-source models, higher effort does not consistently improve the phase average or categorical chain accuracy. **(3) Correct conclusions do not imply correct intermediate answers.** The best Phase 4 accuracy rises from 49.4% for individual VQA to 56.8% for linked-question evaluation. Yet no model exceeds 23.5% categorical chain accuracy. Scoring conclusions alone therefore misses errors in intermediate answers. Section 4.1 tests how written intermediate answers affect the conclusion.

Table 1: **No model exceeds 23.5% categorical chain accuracy.** We compare eight vision–language models on individual VQA and linked-question evaluation. The best Phase 4 decision accuracy is 49.4% for individual VQA and 56.8% for linked-question evaluation. Evaluating intermediate answers exposes errors that Phase 4 decision accuracy alone misses. Scan-level questions follow five workflow phases: Phase 0: visual context; Phase 1: region screening; Phase 2: finding characterization; Phase 3: temporal context; and Phase 4: decision. Each phase reports task-averaged accuracy. The average weights Phase 0 to Phase 4 equally and requires all five phases. Patient-level questions span the scan series of one organ and are scored separately. Categorical chain accuracy requires every categorical intermediate answer and conclusion in an eligible linked-question chain to be correct. Numerical errors are excluded. Measurements in range: the share of numerical answers of Phase 1 to Phase 4 (sizes, volumes, time intervals) that fall within a field-specific tolerance; they are scored separately from the phase columns. All scores are percentages. Table 11 separates intermediate answers and conclusions; Appendix E.3 gives scoring details. Bold marks the column best; shading marks the best phase average in each block. Results include 95% confidence intervals from 2,000 bootstrap resamples of patients, keeping each patient's questions together.

	scan-level accuracy								
model	Phase 0	Phase 1	Phase 2	Phase 3	Phase 4	average	patient-level accuracy	measurements in range	categorical chain accuracy
individual VQA									
open-weight models									
Hulu-Med-32B	33.7 (30.8, 37.0)	**67.7** (64.1, 71.5)	38.7 (32.9, 54.0)	31.5 (14.4, 51.9)	39.4 (33.8, 44.9)	42.2 (38.7, 47.7)	**51.3** (43.2, 71.7)	2.1 (0.8, 4.0)	n/a
Lingshu-I-8B	37.5 (34.0, 41.4)	64.6 (61.4, 67.8)	34.8 (30.4, 48.0)	13.4 (3.6, 26.4)	28.3 (23.1, 33.6)	35.7 (33.3, 40.0)	1.5 (0.0, 4.2)	1.5 (0.6, 2.8)	n/a
Qwen3.8-27B, w/o thinking	38.1 (33.7, 42.4)	63.7 (60.4, 67.0)	30.1 (25.7, 42.0)	26.0 (14.9, 41.1)	41.5 (34.2, 46.6)	39.9 (36.9, 44.1)	72.7 (55.6, 90.4)	11.4 (8.8, 14.7)	n/a
Qwen3.8-27B, w/ thinking‡	41.8 (38.3, 45.1)	33.3 (0.0, 75.0)	30.0 (14.0, 53.0)	33.3 (0.0, 66.7)	38.2 (29.3, 43.2)	35.3 (24.2, 45.7)	51.5 (28.1, 70.0)	0.0 (0.0, 0.0)	n/a
Primus, w/o thinking	16.5 (13.4, 19.9)	10.8 (8.8, 12.9)	4.2 (0.3, 10.9)	**37.0** (21.1, 56.1)	26.3 (19.5, 31.0)	18.9 (15.3, 23.2)	0.0 (0.0, 0.0)	0.4 (0.0, 0.9)	n/a
Primus, w/ thinking	14.5 (11.9, 17.5)	50.3 (47.0, 53.4)	39.9 (17.1, 48.2)	16.4 (8.9, 26.6)	27.5 (23.9, 34.5)	29.7 (25.0, 32.8)	0.0 (0.0, 0.0)	0.1 (0.0, 0.3)	n/a
ClinFusion-32B	38.6 (34.9, 42.3)	37.2 (34.0, 40.5)	31.5 (27.8, 43.2)	28.8 (13.1, 47.1)	43.3 (36.4, 48.1)	35.9 (32.6, 40.4)	65.3 (54.0, 95.0)	2.9 (1.7, 4.4)	n/a
closed-source models									
GPT-6, min. effort†	57.3 (53.6, 60.8)	36.1 (33.0, 39.3)	50.8 (31.4, 58.0)	33.8 (20.8, 50.0)	38.9 (32.5, 45.4)	43.4 (38.4, 46.8)	45.6 (39.6, 67.7)	31.2 (29.8, 32.5)	n/a
GPT-6, max. effort	55.0 (51.7, 58.3)	65.4 (61.9, 69.3)	53.6 (34.2, 59.9)	30.7 (15.2, 49.2)	40.0 (34.9, 46.5)	48.9 (43.5, 52.9)	46.7 (40.2, 65.6)	**32.5** (31.4, 33.3)	n/a
Claude Opus 5, min. effort†	53.7 (50.2, 57.4)	52.8 (49.6, 56.1)	55.6 (38.7, 61.4)	25.0 (13.1, 39.7)	43.6 (37.5, 52.1)	46.1 (41.4, 49.6)	25.7 (12.2, 39.1)	18.1 (15.1, 22.0)	n/a
Claude Opus 5, max. effort	51.5 (47.9, 55.4)	53.5 (50.1, 57.2)	55.0 (38.7, 60.7)	20.6 (9.7, 34.8)	41.2 (35.2, 48.6)	44.4 (40.1, 47.5)	25.2 (11.8, 37.4)	16.8 (14.1, 20.1)	n/a
Gemini 3.8 Flash, min. effort†	56.6 (52.7, 60.1)	63.5 (60.0, 66.9)	44.0 (37.8, 62.7)	29.2 (15.4, 46.9)	**49.4** (41.6, 54.6)	48.6 (45.3, 53.8)	18.3 (11.6, 34.2)	27.7 (24.5, 31.2)	n/a
Gemini 3.8 Flash, max. effort	**58.3** (54.1, 62.0)	65.0 (61.0, 68.8)	**63.2** (47.1, 72.0)	26.2 (10.6, 46.2)	46.4 (38.9, 51.9)	**51.8** (47.5, 55.2)	10.4 (6.4, 21.2)	17.0 (13.7, 20.8)	n/a
linked-question evaluation									
open-weight models									
Hulu-Med-32B	46.2 (40.8, 52.4)	64.8 (59.7, 69.7)	38.8 (26.1, 61.5)	41.7 (24.5, 57.3)	48.9 (38.3, 59.0)	48.1 (42.9, 54.4)	17.1 (7.4, 27.0)	1.1 (0.0, 3.1)	10.9 (6.3, 16.0)
Lingshu-I-8B	37.2 (31.2, 43.3)	19.8 (9.6, 31.3)	2.2 (0.6, 5.2)	5.4 (0.0, 12.1)	9.2 (3.1, 14.2)	14.8 (12.0, 17.7)	1.9 (0.0, 6.1)	0.0 (0.0, 0.0)	2.3 (0.0, 5.4)
Qwen3.8-27B, w/o thinking	36.8 (26.8, 47.2)	65.9 (61.5, 70.8)	32.5 (27.4, 54.7)	**48.6** (32.7, 68.0)	47.8 (35.8, 57.9)	46.3 (41.6, 53.5)	20.6 (10.8, 31.7)	15.4 (7.2, 25.2)	15.9 (10.6, 22.2)
Qwen3.8-27B, w/ thinking	39.1 (28.1, 50.4)	30.7 (22.8, 39.3)	8.7 (4.5, 18.4)	9.7 (3.3, 16.7)	7.4 (2.3, 12.2)	19.1 (3.2, 28.8)	13.3 (4.7, 22.4)	2.9 (0.0, 7.6)	6.6 (3.0, 10.8)
Primus, w/o thinking	6.1 (3.6, 8.9)	2.5 (0.8, 4.6)	1.0 (0.0, 2.9)	16.7 (6.0, 27.2)	3.4 (1.4, 6.0)	5.9 (3.7, 8.1)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
Primus, w/ thinking	9.1 (6.1, 12.2)	2.8 (1.3, 4.5)	1.0 (0.0, 2.9)	22.2 (9.9, 35.1)	4.2 (1.9, 8.8)	7.9 (5.4, 10.8)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
ClinFusion-32B	54.2 (44.4, 66.7)	39.8 (35.5, 45.0)	32.8 (27.5, 55.7)	16.7 (6.5, 28.9)	46.9 (37.8, 57.4)	38.1 (34.5, 44.5)	52.2 (44.5, 73.0)	4.3 (2.2, 7.1)	15.9 (10.8, 21.7)
closed-source models									
GPT-6, min. effort†	**64.3** (58.4, 69.8)	65.9 (57.7, 73.5)	78.0 (63.0, 84.4)	33.8 (17.9, 52.5)	54.7 (43.6, 62.3)	**59.3** (53.0, 63.6)	**79.5** (65.7, 92.6)	**35.7** (33.1, 38.6)	21.4 (14.8, 28.7)
GPT-6, max. effort	60.0 (54.9, 64.0)	64.3 (55.7, 71.8)	**78.5** (62.7, 84.6)	36.8 (19.7, 57.4)	53.2 (41.9, 62.9)	58.5 (52.3, 63.4)	69.3 (51.9, 80.4)	35.0 (31.4, 38.6)	18.9 (12.3, 26.1)
Claude Opus 5, min. effort†	46.1 (41.0, 51.3)	**70.0** (64.1, 75.0)	33.8 (27.5, 57.4)	42.0 (24.5, 59.2)	50.3 (42.9, 60.5)	48.5 (44.0, 55.5)	31.8 (13.0, 47.2)	29.5 (23.1, 36.0)	20.0 (13.4, 27.2)
Claude Opus 5, max. effort	46.3 (41.1, 51.5)	69.7 (63.2, 75.0)	36.1 (29.2, 61.8)	42.0 (26.2, 57.5)	50.5 (42.6, 61.9)	48.9 (45.1, 55.9)	35.6 (18.6, 51.5)	27.5 (19.9, 35.6)	22.1 (15.5, 28.9)
Gemini 3.8 Flash, min. effort†	47.2 (40.8, 53.8)	68.8 (62.7, 74.2)	69.6 (46.7, 83.7)	43.5 (28.9, 59.4)	**56.8** (47.3, 66.8)	57.2 (50.7, 62.7)	30.4 (13.4, 42.5)	20.0 (13.4, 27.3)	**23.5** (16.2, 30.8)
Gemini 3.8 Flash, max. effort	57.4 (53.2, 62.0)	68.8 (62.5, 74.2)	76.7 (58.5, 83.8)	37.3 (19.3, 58.8)	52.2 (41.2, 62.9)	58.5 (52.0, 64.2)	25.9 (8.2, 39.5)	19.5 (9.9, 28.7)	20.2 (13.2, 27.8)

† Thinking cannot be disabled. Min./max. effort is the smallest/largest setting the provider accepts.
‡ Predeclared stratified random sample with equal allocation per task; excluded from ranking.

4 Written Intermediate Answers: Analysis and Fine-tuning

4.1 How written intermediate answers shape conclusions

We test how written intermediate answers affect conclusions in four open-weight models. Hidden states are the model's internal activation vectors. Three experiments measure their changes after image removal, replace intermediate answers, and train classifiers on hidden states before generation. We use training and validation chains, with native thinking disabled for Qwen3.8-27B and Primus and default settings for Hulu-Med-32B and Lingshu-I-8B (Figure 3; Appendix G). We study linked-question chains that conclude with a Phase 3 temporal-change label. For example, a chain gives Phase 2 diameters of 23 mm and 31 mm for the earlier and current scans, then concludes *growing*. The intermediate answers become part of the input for the conclusion. An *answer prefix* is the answer text already written or inserted into this input. The *answer header* is its opening text before the first intermediate answer. *Image removal.* We keep the question and written text fixed and compare hidden states with and without images at each language layer. We measure cosine distance at three points: before generation, after the answer header, and before the conclusion. *Intermediate-answer replacement.* We keep the question and images fixed and compare no intermediate answers, model-generated answers, reference answers, and reference answers with earlier and current sizes swapped. We also test reference answers without images. We score all eight temporal-change labels. *Reference-option agreement* measures how often the highest-scoring label matches the reference answer. *Reference-option probability* is the probability of the reference answer, normalized over the eight labels. *Supervised probe.* We train a linear classifier to predict the reference temporal-change

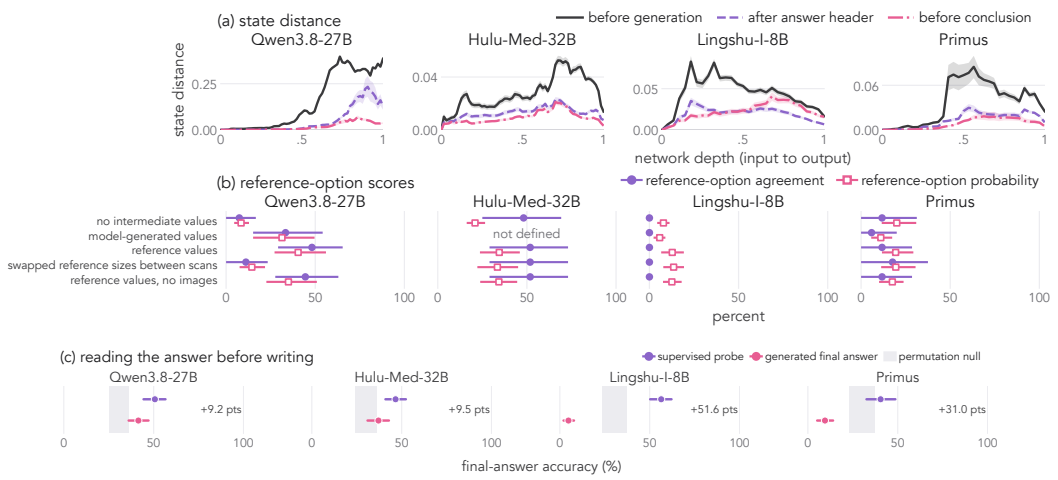

Figure 3: **Swapping written sizes lowers Qwen3.8-27B's reference-option agreement by 37%.** We compare four open-weight models on Phase 3 temporal-change labels. **(a)** Image removal changes hidden states less after intermediate answers for Qwen3.8-27B, Hulu-Med-32B, and Primus. Curves show cosine distance at each language layer before generation, after the answer header, and before the conclusion. **(b)** Swapping earlier and current reference sizes lowers Qwen3.8-27B's agreement from 48.1% to 11.1%. Removing images while retaining reference answers leaves agreement at 44.4%. **(c)** Supervised probes before generation exceed the accuracy of generated conclusions by 9.2 percentage points for Qwen3.8-27B and 9.5 for Hulu-Med-32B on held-out patients. Circles show how often the highest-scoring label matches the reference answer; squares show mean reference-option probability normalized over eight labels. Grey bands mark the permutation null, obtained by shuffling the training labels; other bands show 95% intervals from bootstrap resampling of patients.

label from the last language layer's hidden state before generation. We compare probe accuracy with the accuracy of the generated conclusion on the same chains.

(1) Written measurements change the conclusion. For Qwen3.8-27B, swapping reference sizes lowers reference-option agreement from 48.1% to 11.1%, with the images unchanged (Figure 3b). Removing images while retaining reference answers leaves agreement at 44.4%. The conclusion is more sensitive to the written measurements than to image availability in this experiment. **(2) State differences shrink after intermediate answers.** For Qwen3.8-27B, the last-layer distance before the conclusion is 9% of its value before generation (Figure 3a). Hulu-Med-32B and Primus show the same pattern. Lingshu-I-8B has small last-layer distances at all three points. The decrease also occurs with reference intermediate answers (Appendix Figure 5). **(3) Sensitivity to written measurements differs across models.** With intermediate answers present, Hulu-Med-32B scores *growing* highest and Lingshu-I-8B scores *new* highest in every scored chain, including after sizes are swapped. Primus shows little change in reference-option agreement across reference, swapped, and removed intermediate answers (Appendix G.6). These models show no comparable response to the changed measurements. **(4) Supervised probes predict more accurate conclusions.** On the same chains, probe accuracy exceeds the accuracy of the generated conclusion: 50.7% versus 41.5% for Qwen3.8-27B, and 46.6% versus 37.1% for Hulu-Med-32B (Figure 3c). The paired gains are 9.2 and 9.5 percentage points, respectively; both 95% confidence intervals exclude zero (Appendix G.7). A supervised probe can thus recover label information from the hidden state before generation. These results motivate supervising intermediate answers and conclusions together.

4.2 Learning from linked-question chains

We fine-tune Primus, the smallest evaluated VLM with 4 B parameters, and Hulu-Med-32B, the strongest open-weight model by zero-shot individual VQA phase average. Both use supervised fine-tuning (SFT) on reference intermediate answers and conclusions from the training split. Primus then receives reinforcement learning (SFT + RL). We evaluate on the test split with the protocols and metrics in Section 3. Table 2 reports results; Appendix F.2 has additional Qwen3.8-27B results.

Table 2: **Fine-tuning raises categorical chain accuracy for both models, reaching 30.6%.** We compare zero-shot models with supervised fine-tuning (SFT) on linked questions and SFT followed by reinforcement learning (SFT + RL). In individual VQA, SFT raises the Phase 1–4 average for Primus from 19.6% to 60.8%. In linked-question evaluation, its categorical chain accuracy rises from 0.0% to 28.3%. For Hulu-Med-32B, SFT raises categorical chain accuracy from 10.9% to 30.6%. After SFT + RL, Primus reaches 30.4% categorical chain accuracy and 98.7% patient-level accuracy. All results use the test split and are percentages. Columns and scoring follow Table 1. Bold marks the column best within each model block; shading marks its best Phase 1–4 average. Both require at least two scores. Intervals are 95% confidence intervals from 2,000 bootstrap resamples of patients.

	scan-level accuracy								
model	Phase 0	Phase 1	Phase 2	Phase 3	Phase 4	Phase 1–4 average	patient-level accuracy	measurements in range	categorical chain accuracy
individual VQA									
Primus‡									
zero-shot†	16.5 (13.4, 19.9)	10.8 (8.8, 12.9)	4.2 (0.3, 10.9)	**37.0** (21.1, 56.1)	26.3 (19.5, 31.0)	19.6 (15.1, 24.8)	n/a	0.4 (0.0, 0.9)	n/a
SFT	78.7 (75.8, 81.7)	**72.1** (69.3, 75.2)	**73.7** (62.2, 78.5)	32.9 (19.7, 48.9)	**64.7** (58.4, 68.6)	**60.8** (56.1, 64.2)	n/a	**12.1** (8.6, 15.6)	n/a
SFT + RL	**79.3** (76.5, 82.1)	**72.1** (69.2, 75.4)	72.1 (61.0, 76.5)	30.1 (18.4, 44.1)	64.0 (57.4, 67.7)	59.6 (55.0, 62.6)	n/a	10.8 (7.3, 14.6)	n/a
Hulu-Med-32B§									
zero-shot†	33.7 (30.8, 37.0)	67.7 (64.1, 71.5)	38.7 (32.9, 54.0)	**31.5** (14.4, 51.9)	39.4 (33.8, 44.9)	44.3 (39.9, 51.4)	n/a	2.1 (0.8, 4.0)	n/a
SFT	**79.1** (76.5, 81.6)	**71.0** (68.1, 74.0)	**73.7** (62.3, 78.4)	26.0 (15.7, 39.3)	**62.7** (56.0, 66.8)	**58.3** (54.2, 61.1)	n/a	**22.3** (18.9, 25.6)	n/a
linked-question evaluation									
Primus‡									
zero-shot†	n/a	2.8 (1.3, 4.5)	1.0 (0.0, 2.9)	22.2 (9.9, 35.1)	4.2 (1.9, 8.8)	7.5 (4.4, 11.2)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)	0.0 (0.0, 0.0)
SFT	n/a	**69.7** (65.7, 73.9)	**71.9** (54.5, 76.9)	63.9 (47.7, 76.7)	50.1 (43.2, 61.2)	63.9 (57.7, 68.0)	98.0 (95.1, 99.4)	**10.4** (5.0, 17.3)	28.3 (21.5, 35.2)
SFT + RL	n/a	69.4 (65.2, 73.7)	70.7 (52.4, 75.5)	**65.3** (50.0, 78.1)	**56.8** (48.2, 67.6)	**65.5** (59.5, 69.3)	**98.7** (96.3, 100.0)	5.2 (0.9, 12.4)	**30.4** (23.5, 37.6)
Hulu-Med-32B§									
zero-shot†	n/a	64.8 (59.7, 69.7)	38.8 (26.1, 61.5)	41.7 (24.5, 57.3)	48.9 (38.3, 59.0)	48.5 (41.9, 55.9)	17.1 (7.4, 27.0)	1.1 (0.0, 3.1)	10.9 (6.3, 16.0)
SFT	n/a	**70.6** (66.0, 75.2)	**78.1** (61.8, 83.9)	**63.9** (49.3, 77.9)	**57.6** (49.5, 68.4)	**67.5** (61.3, 72.3)	**91.9** (87.2, 99.6)	**24.4** (16.0, 31.9)	**30.6** (23.2, 38.2)

† Gray zero-shot results repeat Table 1.
‡ Primus: w/o thinking for individual VQA; w/ thinking for linked-question evaluation. SFT uses one epoch; SFT + RL adds reinforcement learning on linked-question chains (Appendix F.1).
§ Hulu-Med-32B: default mode. SFT trains adapters in all language blocks.

(1) SFT improves categorical chain accuracy. SFT raises categorical chain accuracy from 0.0% to 28.3% for Primus and from 10.9% to 30.6% for Hulu-Med-32B (Table 2). In linked-question evaluation, both models reach 63.9% for Phase 3 temporal context. **(2) A 4B model exceeds the best zero-shot scores on three metrics.** After SFT, Primus reaches a Phase 1–4 average of 60.8% in individual VQA, compared with the best zero-shot score of 50.2%. In linked-question evaluation, it exceeds the best zero-shot categorical chain accuracy (28.3% versus 23.5%) and patient-level accuracy (98.0% versus 79.5%; Tables 1 and 2). **(3) RL raises categorical chain accuracy but lowers other scores.** RL raises Primus's categorical chain accuracy to 30.4%. Its individual VQA Phase 1–4 average falls from 60.8% to 59.6%. In linked-question evaluation, measurements in range fall from 10.4% to 5.2%.

5 Discussion and Conclusion

RadThinking turns a radiology read into linked questions, from visual context to the clinical decision, each answered from evidence that the radiologist reported or that was measured on the CT. Scoring every step shows what a single conclusion score hides: no model is strong in every phase, comparing a finding with earlier scans is the hardest step for every model, and few linked-question chains are correct from the first answer to the last (Section 3).

Looking inside the models explains why. Once a model has written its intermediate answers, the images barely shape its conclusion, which follows what was written, including wrong values; the correct answer is often already present in the model's hidden state before it writes, but the written chain does not deliver it (Section 4.1). Supervising the intermediate answers with the reference answers of RadThinking addresses exactly this: it raises categorical chain accuracy for both fine-tuned models, and the 4B Primus exceeds every zero-shot model on individual VQA phase average, categorical chain accuracy, and patient-level accuracy (Section 4.2).

The same data structure enables uses beyond this benchmark. Source-linked findings can retrieve patients with similar imaging trajectories, complementing retrieval of clinical literature, and the phase structure provides a natural curriculum and reward on intermediate answers for training (Appendix D). Extending RadThinking to further modalities and to outcome records is the next step toward models whose conclusions are based on evidence a radiologist can check.

AI USE STATEMENT

Large language models were used in three ways. First, as components of the dataset construction pipeline: GPT-6 and Claude Opus 5 contribute two of the four votes used to extract findings from reports. GPT-6 rewrites template reference rationales, and Claude Opus 5 grades the prose. Automated checks compare the outputs with recorded facts, and physicians review random samples in five rounds (Appendix B). Second, as evaluated models in the benchmark (Section 3). Third, as assistants for writing and editing parts of this paper. The authors reviewed all content and take responsibility for it.

ETHICS STATEMENT

This study analyzes retrospective, de-identified Data. Each contributing site approved the study under its own institutional review board, with a waiver of informed consent for the secondary use of de-identified clinical data. All scans and reports were de-identified at the source before any analysis, and no protected health information entered our pipeline or is reproduced in this paper.

REPRODUCIBILITY STATEMENT

Data, and models will be publicly available. Code is already available in the supplementary.

REFERENCES

Anthropic. Introducing Claude Opus 5, July 2026. URL https://www.anthropic.com/news/claude-opus-5.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, Bram van Ginneken, et al. The medical segmentation decathlon. *arXiv preprint arXiv:2106.05735*, 2021.

Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

Pedro RAS Bassi, Sergio SJ Dertkigil, and Andrea Cavalli. Improving deep neural network generalization and robustness to background bias via layer-wise relevance propagation optimization. *Nature Communications*, 15(1):291, 2024.

Pedro RAS Bassi, Mehmet Can Yavuz, Ibrahim Ethem Hamamci, Sezgin Er, Xiaoxi Chen, Wenxuan Li, Bjoern Menze, Sergio Decherchi, Andrea Cavalli, Kang Wang, Yang Yang, Alan Yuille, and Zongwei Zhou. Radgpt: Constructing 3d image-text tumor datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23720–23730, 2025. URL https://github.com/MrGiovanni/RadGPT.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference on Machine Learning (ICML)*, 2009.

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, Fabian Lohöfer, Julian Walter Holch, Wieland Sommer, Felix Hofmann, Alexandre Hostettler, Naama Lev-Cohain, Michal Drozdzal, Michal Marianne Amitai, Refael Vivanti, Jacob Sosna, Ivan Ezhov, Anjany Sekuboyina, Fernando Navarro, Florian Kofler, Johannes C. Paetzold, Suprosanna Shit, Xiaobin Hu, Jana Lipková, Markus Rempfler, Marie Piraud, Jan Kirschke, Benedikt Wiestler, Zhiheng Zhang, Christian Hülsemeyer, Marcel Beetz, Florian Ettlinger, Michela Antonelli, Woong Bae, Míriam Bellver, Lei Bi, Hao Chen, Grzegorz Chlebus, Erik B. Dam, Qi Dou, Chi-Wing Fu, Bogdan Georgescu, Xavier Giró-i Nieto, Felix Gruen, Xu Han, Pheng-Ann Heng, Jürgen Hesser, Jan Hendrik Moltz, Christian Igel, Fabian Isensee, Paul Jäger, Fucang Jia, Krishna Chaitanya Kaluva, Mahendra Khened, Ildoo Kim, Jae-Hun Kim, Sungwoong Kim, Simon Kohl, Tomasz Konopczynski, Avinash Kori, Ganapathy Krishnamurthi, Fan Li, Hongchao Li, Junbo Li, Xiaomeng Li, John Lowengrub, Jun Ma, Klaus Maier-Hein, Kevis-Kokitsi Maninis, Hans Meine, Dorit Merhof, Akshay Pai, Mathias Perslev, Jens Petersen, Jordi Pont-Tuset, Jin Qi, Xiaojuan Qi, Oliver Rippel, Karsten Roth, Ignacio Sarasua, Andrea Schenk, Zengming Shen, Jordi Torres, Christian Wachinger, Chunliang Wang, Leon Weninger, Jianrong Wu, Daguang Xu, Xiaoping Yang, Simon Chun-Ho Yu, Yading Yuan, Miao Yue, Liping Zhang, Jorge Cardoso, Spyridon Bakas, Rickmer Braren, Volker Heinemann, Christopher Pal, An Tang, Samuel Kadoury, Luc Soler, Bram van Ginneken, Hayit Greenspan, Leo Joskowicz, and Bjoern Menze. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, 2019.

Louis Blankemeier, Ashwin Kumar, Joseph Paul Cohen, Jiaming Liu, Longchao Liu, Dave Van Veen, Syed Jamal Safdar Gardezi, Hongkun Yu, Magdalini Paschali, Zhihong Chen, Jean-Benoit Delbrouck, Eduardo Reis, Robbie Holland, Cesar Truyts, Christian Bluethgen, Yufu Wu, Long Lian, Malte Engmann Kjeldskov Jensen, Sophie Ostmeier, Maya Varma, Jeya Maria Jose Valanarasu, Zhongnan Fang, Zepeng Huo, Zaid Nabulsi, Diego Ardila, Wei-Hung Weng, Edson Amaro Junior, Neera Ahuja, Jason Fries, Nigam H. Shah, Greg Zaharchuk, Marc Willis, Adam Yala, Andrew Johnston, Robert D. Boutin, Andrew Wentland, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, and Akshay S. Chaudhari. Merlin: a computed tomography vision–language foundation model and dataset. *Nature*, 652(8112):1318–1328, March 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10181-8. URL http://dx.doi.org/10.1038/s41586-026-10181-8.

Thomas A Buckley, James A Diao, Cam N Srivastava, Peter G Brodeur, Pranav Rajpurkar, Adam Rodman, and Arjun K Manrai. Multimodal foundation models exploit text to make medical image predictions. *Nature Communications*, 2026.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024.

Yixiong Chen, Wenjie Xiao, Pedro RAS Bassi, Xinze Zhou, Sezgin Er, Ibrahim Ethem Hamamci, Zongwei Zhou, and Alan Yuille. Are vision language models ready for clinical diagnosis? a 3d medical benchmark for tumor-centric visual question answering. *arXiv preprint arXiv:2505.18915*, 2025. URL https://github.com/Schuture/DeepTumorVQA.

Victoria Chernyak, Kathryn J Fowler, Aya Kamaya, Ania Z Kielar, Khaled M Elsayes, Mustafa R Bashir, Yuko Kono, Richard K Do, Donald G Mitchell, Amit G Singal, et al. Liver imaging reporting and data system (li-rads) version 2018: imaging of hepatocellular carcinoma in at-risk patients. *Radiology*, 289(3):816–830, 2018.

Jared Christensen, Ashley Elizabeth Prosper, Carol C Wu, Jonathan Chung, Elizabeth Lee, Brett Elicker, Andetta R Hunsaker, Milena Petranovic, Kim L Sandler, Brendon Stiles, et al. Acr lung-rads v2022: assessment categories and management recommendations. *Journal of the American College of Radiology*, 21 (3):473–488, 2024.

MJJ de Grauw, E Th Scholten, Ewoud J Smit, Matthieu JCM Rutten, Mathias Prokop, Bram van Ginneken, and Alessa Hering. The uls23 challenge: A baseline model and benchmark dataset for 3d universal lesion segmentation in computed tomography. *Medical image analysis*, 102:103525, 2025.

Xiaotang Gai, Jiaxiang Liu, Yichen Li, Zijie Meng, Jian Wu, and Zuozhu Liu. 3d-rad: A comprehensive 3d radiology med-vqa dataset with multi-temporal analysis and diverse diagnostic tasks. *arXiv preprint arXiv:2506.11147*, 2025.

Sushant Gautam, Michael A. Riegler, and Pål Halvorsen. Kvasir-VQA-x1: A multimodal dataset for medical reasoning and robust MedVQA in gastrointestinal endoscopy. *arXiv preprint arXiv:2506.09958*, 2025.

Google DeepMind. Gemini 3.8 Flash model card, September 2026. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering*, pp. 1–19, 2026.

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.

Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Yao Hu, and Shaohui Lin. Vision-R1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Yuanfeng Ji, Chongjian Ge, Ruijiang Li, and Ping Luo. AMOS-MM: Abdominal multimodal analysis challenge: Structured description of the challenge design. Zenodo, 2024.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, Jin Hao, Zijian Chen, Ruijia Wu, Tao Tang, Junhui Lv, Hongxia Xu, Hongwei Wang, Jun Xiao, Bin Feng, Fudong Zhu, Kenli Li, Weidi Xie, Jimeng Sun, Jian Wu, and Zuozhu Liu. Hulu-Med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025. URL https://arxiv.org/abs/2510.08668.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. In *International Conference on Learning Representations (ICLR)*, 2023.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. Med-R1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*, 2025.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

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

Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro RAS Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue, Yixiong Chen, Xiaorui Lin, Yutong Tang, Yining Cao, Haoqi Han, Zheyuan Zhang, Jiawei Liu, Tiezheng Zhang, Yujiu Ma, Jincheng Wang, Guang Zhang, Alan Yuille, and Zongwei Zhou. Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. *Medical Image Analysis*, pp. 103285, 2024. URL https://github.com/MrGiovanni/AbdomenAtlas.

Wenxuan Li, Xinze Zhou, Qi Chen, Tianyu Lin, Pedro RAS Bassi, Xiaoxi Chen, Chen Ye, Zheren Zhu, Kai Ding, Heng Li, Kang Wang, Yang Yang, Yucheng Tang, Daguang Xu, Alan L Yuille, and Zongwei Zhou. Pants: The pancreatic tumor segmentation dataset. In *Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2025. URL https://github.com/MrGiovanni/PanTS.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024.

Katherine E Link, Zane Schnurman, Chris Liu, Young Joon Kwon, Lavender Yao Jiang, Mustafa Nasir-Moin, Sean Neifert, Juan Diego Alzate, Kenneth Bernstein, Tanxia Qu, et al. Longitudinal deep neural networks for assessing metastatic brain cancer on a large open benchmark. *Nature Communications*, 15(1):8170, 2024.

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 *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2021.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-RFT: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.

Heber MacMahon, David P Naidich, Jin Mo Goo, Kyung Soo Lee, Ann N C Leung, John R Mayo, Atul C Mehta, Yoshiharu Ohno, Charles A Powell, Mathias Prokop, Geoffrey D Rubin, Cornelia M Schaefer-Prokop, William D Travis, Paul E Van Schil, and Alexander A Bankier. Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society 2017. *Radiology*, 284(1):228–243, 2017. doi: 10.1148/radiol.2017161659.

William W Mayo-Smith, Julie H Song, Giles L Boland, Isaac R Francis, Gary M Israel, Peter J Mazzaglia, Lincoln L Berland, and Pari V Pandharipande. Management of incidental adrenal masses: a white paper of the acr incidental findings committee. *Journal of the American College of Radiology*, 14(8):1038–1044, 2017.

Alec J Megibow, Mark E Baker, Desiree E Morgan, Ihab R Kamel, Dushyant V Sahani, Elliot Newman, William R Brugge, Lincoln L Berland, and Pari V Pandharipande. Management of incidental pancreatic cysts: a white paper of the acr incidental findings committee. *Journal of the American College of Radiology*, 14(7): 911–923, 2017.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, et al. MM-Eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

Andriy Myronenko, Dong Yang, Yucheng Tang, Baris Turkbey, Benjamin Simon, Stephanie Harmon, Rikhil Makwana, Mariam Aboian, Sena Azamat, Ibrahim Ethem Hamamci, Sezgin Er, Bjoern Menze, Marc Edgar, Yufan He, Pengfei Guo, and Daguang Xu. NV-Reason-CT: 3D visual language model for CT analysis. *arXiv preprint arXiv:2609.27511*, 2026. URL https://arxiv.org/abs/2609.27511.

OpenAI. Model guidance: Using GPT-6. OpenAI API documentation, 2026. URL https://developers.openai.com/api/docs/guides/latest-model.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. MedVLM-R1: Incentivizing medical reasoning capability of vision-language models via reinforcement learning. *arXiv preprint arXiv:2502.19634*, 2025.

Valeria Panebianco, Yoshifumi Narumi, Ersan Altun, Bernard H Bochner, Jason A Efstathiou, Shaista Hafeez, Robert Huddart, Steve Kennish, Seth Lerner, Rodolfo Montironi, et al. Multiparametric magnetic resonance imaging for bladder cancer: development of vi-rads (vesical imaging-reporting and data system). *European urology*, 74(3):294–306, 2018.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*, 2022.

Qwen Team. Qwen3.8-27B. Model card, 2026. URL https://huggingface.co/Qwen/Qwen3.8-27B.

Maximilian R Rokuss, Yannick Kirchhoff, Saikat Roy, Balint Kovacs, Constantin Ulrich, Tassilo Wald, Maximilian Zenk, Stefan Denner, Fabian Isensee, Philipp Vollmuth, et al. Longitudinal segmentation of ms lesions via temporal difference weighting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 64–74. Springer, 2024.

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

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 2024a.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024b. URL https://arxiv.org/abs/2402.03300.

Sonali Sharma, Jin Long, George Shih, Sarah Eid, Christian Bluethgen, Francine L. Jacobson, Emily B. Tsai, Global Radiology Consortium, Ahmed M. Alaa, and Curtis P. Langlotz. CheXthought: A global multimodal dataset of clinical chain-of-thought reasoning and visual attention for chest X-ray interpretation. *arXiv preprint arXiv:2604.26288*, 2026.

Stuart G Silverman, Ivan Pedrosa, James H Ellis, Nicole M Hindman, Nicola Schieda, Andrew D Smith, Erick M Remer, Atul B Shinagare, Nicole E Curci, Steven S Raman, et al. Bosniak classification of cystic renal masses, version 2019: an update proposal and needs assessment. *Radiology*, 292(2):475–488, 2019.

Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual inference via python execution for reasoning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Baris Turkbey, Andrew B Rosenkrantz, Masoom A Haider, Anwar R Padhani, Geert Villeirs, Katarzyna J Macura, Clare M Tempany, Peter L Choyke, Francois Cornud, Daniel J Margolis, et al. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. *European urology*, 76(3):340–351, 2019.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. *Nature Communications*, 2025a.

Joy T Wu, Nkechinyere N Agu, Ismini Lourentzou, Arjun Sharma, Joseph A Paguio, Jasper S Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo A Celi, and Mehdi Moradi. Chest ImaGenome dataset for clinical reasoning. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2021.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, et al. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*, 2025b.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. LLaVA-CoT: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.

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

Hangjie Yuan, Yichen Qian, Zhiwei Tang, Xianzhe Xu, Lirong Wu, Sicheng Yang, Jinwang Wang, Pengju Wang, Zhitao Zeng, Yizeng Han, Yan Xing, Shengxuan Luo, Tao Feng, Qing Xie, Weigen Yao, Yi Yang, Zuozhu Liu, Jiasheng Tang, Shaocheng Wang, Jitao Wang, Jiahong Dong, Weihua Chen, Feng Xu, and Fan Wang. ClinFusion: A vision-centric multimodal LLM system for holistic medical understanding. *arXiv preprint arXiv:2607.24743*, 2026. URL https://arxiv.org/abs/2607.24743.

Jaehoon Yun, Jiwoong Sohn, Jungwoo Park, Hyunjae Kim, Xiangru Tang, Daniel Shao, Yong Hoe Koo, Ko Minhyeok, Qingyu Chen, Mark Gerstein, et al. Med-prm: Medical reasoning models with stepwise, guideline-verified process rewards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 16565–16582, 2025.

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.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Ya Zhang, Yanfeng Wang, and Weidi Xie. RadGenome-Chest CT: A grounded vision-language dataset for chest CT analysis. *arXiv preprint arXiv:2404.16754*, 2024.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations (ICLR)*, 2023.

A Dataset Details

This appendix lists the question kinds of every workflow phase (Section A.1), the clinical guidelines by organ (Section A.2), a complete worked example (Section A.3), the data format (Section A.4), the preparation of benchmark inputs (Section A.5), and the cohort distributions (Section A.6). Appendix B describes how RadThinking was built and checked.

A.1 Question kinds by workflow phase, tier, and evidence source

Scan-level questions are asked for each scan. They follow the order of a radiology read, from Phase 0 visual context to Phase 4 decision. A workflow phase is a step in reading a scan; contrast phase describes CT acquisition. Patient-level questions span the series of one organ. Algorithm 1 lists both levels in order; Table 3 lists every question kind with its count, tier, and evidence source. Phase 3 gives the interval since the previous scan and per-finding temporal-change labels: new, stable, growing, shrinking, resolved, treated, persistent, or unmatched. Persistent means present without a size comparison.

The tier describes how much an answer combines. A *foundation* question reads one evidence field, such as size, contrast phase, or whether a region is abnormal. A *single-step* question makes one calculation or decision from recorded evidence, such as a temporal-change label from two measurements or a category from imaging features. A *compositional* question combines earlier answers into a conclusion, such as a reference rationale or the course of a finding. A tier does not fix the workflow phase: Phase 0 contains foundation and single-step questions, and Phase 4 contains all three tiers.

The evidence source records how the reference answer is obtained from the CT, report, or metadata. *Measurement* is a CT-derived value: size, volume, attenuation, vessel adjacency, or organ volume. *Report* is a statement from the radiology report. *Metadata* is a stored value, such as the scan date, contrast phase, or body region. The stored value `rule` marks an answer computed from recorded values: a temporal-change label from two finding-volume measurements, a category from imaging features, or a series summary from temporal-change labels. *Composition* is a reference rationale written from earlier answers. *Image* marks the 247 contrast-phase answers read from vessel attenuation. An organ without measured findings has a *report* reference rationale.

Answers of Phase 0 to Phase 3 contain no diagnosis words. They describe findings by location, size, attenuation, and morphology, and the bounding-box files carry neutral names. No question uses information from a later scan of the patient.

Algorithm 1 The questions of one patient in RadThinking, in the order of a radiology read. Every question stores its identifier, question text, reference answer, bounding box, evidence source, tier, clinical context, and the questions it depends on (Table 6).

for each scan s of the patient, in time order **do**
    Phase 0: modality and contrast phase; body region; organs and guidelines in focus for the clinical indication; volume and mean attenuation of the large solid organs
    for each organ o in s **do**
        Phase 1: for each region of o: normal or abnormal (regions listed in Table 4)
        Phase 2: for each finding: longest diameter and volume; attenuation relative to the parenchyma; adjacency to a major vessel; report description; enhancement, arterial-phase hyperenhancement, and washout where the report states them
        Phase 3: interval since the previous scan; for each finding matched to the previous scan: temporal-change label
        Phase 4: assessment of o with its reference rationale; category and management under the guideline of o; next step; diagnosis where the report states one
    end for
end for
for each organ with a series **do**
    patient-level questions: scans covered; scans on which the organ was abnormal; course of each finding; first high-suspicion category; treated sites; overall course with the criterion that decides it
end for

Table 3: **Most of the 122,127 questions ask for one piece of evidence: 87,846 are foundation questions, 18,049 single-step, and 16,232 compositional.** We list every question kind by workflow phase with its count, tier, and evidence source. Phase 1 alone screens 30,369 regions. Phase 2 describes 8,227 findings and contains 7,311 size questions; the enhancement questions (409) exist only where the report describes enhancement. Phase 3 gives 4,502 intervals and 4,214 temporal-change labels (new, stable, growing, shrinking, resolved, treated, persistent, or unmatched). Phase 4 has 11,741 assessments with a reference rationale and 1,283 questions that apply a clinical guideline (859 management, 424 category). Patient-level questions summarize 2,694 organ series. The organ measurement of Phase 0 covers the liver, kidneys, pancreas, and spleen.

workflow phase	question kind	count	tier	evidence source
0 visual context	modality and contrast phase	5,365	foundation	metadata or image
	body region	4,536	foundation	metadata
	organ volume and mean attenuation	10,554	foundation	measurement
	organs and guidelines in focus for the indication	5,365	single-step	rule
1 region screening	region normal or abnormal	30,369	foundation	report or measurement
2 finding characterization	description of the finding	8,227	foundation	report
	longest diameter and volume	7,311	foundation	measurement or report
	attenuation relative to the parenchyma	6,234	foundation	measurement
	adjacency to a major vessel	6,234	foundation	measurement
	enhancement, arterial hyperenhancement, washout	409	foundation	report
3 temporal context	interval since the previous scan	4,502	foundation	metadata
	change of a matched finding (temporal-change label)	4,214	single-step	rule or report
4 decision	assessment of the organ with reference rationale	11,741	compositional	composition or report
	recommended next step	5,235	single-step	report or rule
	management derived from the clinical guideline	859	single-step	rule
	category assigned by the clinical guideline	424	single-step	report, rule
	most likely diagnosis with reference rationale	54	compositional	composition
patient-level	organs evaluated across the series	1,411	foundation	metadata
	scans of the organ and period covered	2,694	foundation	metadata
	scans on which the organ was abnormal	1,952	single-step	rule
	course of one finding across the series	1,590	compositional	rule
	treated sites and findings after treatment	101	compositional	rule
	first high-suspicion category and its basis	42	compositional	rule
	from detection to category	10	compositional	rule
	overall course of the organ with the criterion used	2,694	compositional	rule
all		122,127		

A.2 CLINICAL GUIDELINES BY ORGAN

Table 4 lists the 16 organs of RadThinking, the regions into which each organ is screened in Phase 1, and the clinical guideline, if any, that the reference rationales apply. Five guidelines are implemented as decision tables (Appendix B.4). A rationale applies a guideline only when the finding type, the patient context, and the acquired contrast phase make it applicable. Otherwise the rationale states why the guideline does not apply and follows the radiologist's characterization. Guidelines that need another modality are not applied: the Prostate Imaging Reporting and Data System (PI-RADS) (Turkbey et al., 2019) and Vesical Imaging Reporting and Data System (VI-RADS) (Panebianco et al., 2018) apply to magnetic resonance imaging (MRI), and Lung-RADS (Christensen et al., 2024) applies to screening CT rather than to the diagnostic scans of this cohort.

A.3 A WORKED EXAMPLE

Figure 1 summarizes the liver organ record of patient `RTHINK_P002399`. A hypervascular finding at the margin of an earlier ablation defect is categorized LR-4. The finding is ablated. Three follow-up scans show a nonviable treatment zone and no new suspicious lesion. Box 1 lists the organ record of the first scan as stored in RadThinking, grouped by tier, with the Phase 3 questions of the next scan that depend on it. *On* names the questions an answer depends on (the `supported_by` field). Box 2 prints its reference rationale. The answers are reference answers. APHE means arterial-phase hyperenhancement; HU, Hounsfield units; HCC, hepatocellular carcinoma; LR, a LI-RADS category. The stored question and answer text is reproduced verbatim; its use of "reasoning chain" refers to the reference rationale. One convention is visible: the Phase 2 description repeats the report sentence without its category words ("A 3.6 x 1.9 cm arterially hypervascular area . . . "); the report sentence quoted in the reference rationale carries the category. Here the measurement questions describe the segmented ablation zone (53.0 mm) and the description question the LR-4 finding at its margin.

Table 4: **Five of the 16 organs have a clinical guideline implemented as a decision table; the other eleven are assessed from the report.** We list each organ, the regions that Phase 1 screens, the guideline the reference rationales apply, and the imaging features the decision table reads from the recorded evidence. The liver is screened by its eight Couinaud segments, the kidneys and lungs by thirds on each side, the pancreas by head, body, and tail, and the other organs as a whole. The Liver Imaging Reporting and Data System (LI-RADS) v2018 (Chernyak et al., 2018), Bosniak v2019 (Silverman et al., 2019), the Fleischner Society 2017 guideline (MacMahon et al., 2017), and the American College of Radiology (ACR) white papers on incidental adrenal masses (Mayo-Smith et al., 2017) and pancreatic cysts (Megibow et al., 2017) are applied; each row also names when the guideline is not applied. PI-RADS (Turkbey et al., 2019), VI-RADS (Panebianco et al., 2018), and Lung-RADS (Christensen et al., 2024) need another modality or a screening setting and are not applied; the reference rationales of these organs follow the radiologist's characterization.

organ	screening regions	clinical guideline	imaging features read by the decision table
liver	eight Couinaud segments	LI-RADS v2018 (Chernyak et al., 2018)	at-risk context; arterial-phase hyperenhancement; size bracket; nonperipheral washout; enhancing capsule; threshold growth; contrast-phase availability
kidney	left and right, upper, middle, and lower third	Bosniak v2019 (Silverman et al., 2019)	enhancing nodule or solid component; thick or irregular enhancing wall or septa; number and thickness of septa; calcification; attenuation and size; contrast-phase availability
lung	left and right, upper, middle, and lower third	Fleischner Society 2017 (MacMahon et al., 2017)	nodule size; calcified, solid, part-solid, or ground-glass; single or multiple; not applied with a known malignancy or to a lesion above 30 mm
pancreas	head, body, tail	ACR incidental pancreatic cyst (Megibow et al., 2017)	cystic lesion required; main-duct dilatation; mural nodule or solid component; thick or enhancing wall; duct communication; cyst size
adrenal gland	left and right	ACR incidental adrenal mass (Mayo-Smith et al., 2017)	macroscopic fat; unenhanced attenuation (needs a non-contrast scan); absolute washout; stability; size; known malignancy; not applied to hematoma, calcification, abscess, or hyperplasia
spleen	whole organ	none	report-based assessment
gallbladder	whole organ	none	report-based assessment
bile ducts	whole organ	none	report-based assessment
esophagus	whole organ	none	report-based assessment
stomach	whole organ	none	report-based assessment
duodenum	whole organ	none	report-based assessment
small intestine	whole organ	none	report-based assessment
colon and rectum	whole organ	none	report-based assessment
urinary bladder	whole organ	none (VI-RADS is MRI-based)	report-based assessment
prostate	whole organ	none (PI-RADS is MRI-based)	report-based assessment
uterus	whole organ	none	report-based assessment

Table 5: **The liver, lungs, kidneys, and pancreas hold 7,911 of the 11,741 organ records, and the five organs with a clinical guideline hold all 1,283 guideline questions.** We list, per organ, the patients and organ records, the records with at least one finding, the findings, the temporal-change labels, the guideline questions, and the guideline. The liver alone has 3,424 findings and 1,936 temporal-change labels; the adrenal glands have the most guideline questions (503) because the ACR white paper applies to every incidental adrenal mass.

organ	patients	records	with findings	findings	temporal-change labels	guideline questions	clinical guideline
liver	1,929	3,617	1,925	3,424	1,936	283	LI-RADS v2018
lung	1,046	1,834	827	1,080	407	60	Fleischner 2017
kidney	1,033	1,485	1,077	1,542	508	141	Bosniak v2019
pancreas	523	975	499	686	394	296	ACR pancreatic cyst (2017)
spleen	500	752	276	376	85	0	
adrenal gland	430	715	449	520	187	503	ACR adrenal (2017)
colon and rectum	520	673	359	879	397	0	
gallbladder	346	450	324	372	35	0	
prostate	187	250	117	125	40	0	
stomach	152	242	186	353	96	0	
esophagus	114	192	109	172	39	0	
urinary bladder	144	187	149	208	44	0	
uterus	105	131	89	99	17	0	
duodenum	73	93	72	84	15	0	
small intestine	64	82	56	70	12	0	
bile ducts	51	63	37	39	2	0	
all (16 organs)	2,679	11,741	6,551	10,029	4,214	1,283	

Box 1: The liver organ record of the scan of 2022-03-05 (late arterial contrast phase): 16 questions as stored, 13 foundation, 2 single-step, and 1 compositional, plus the two Phase 3 questions of the next scan that depend on them. Clinical context: known liver cirrhosis; hepatocellular carcinoma diagnosed on imaging in 2021; no earlier scan in this series. Every answer is a reference answer; *on* lists the questions it depends on. Question identifiers are in typewriter font.

Foundation questions (one evidence field each).

`seg1` to `seg8`: *Examine liver segment k. Is it normal or does it contain any findings?* Segments 1 to 6: *Normal. No suspicious findings in this segment.* Segments 7 and 8: *Abnormal. Focal hypodense lesion identified.*

`f1_size`: *What is the size of the finding in segment VII/VIII?* 53.0 mm (longest diameter); volume 36.05 mL.

`f1_hu`: *What is the attenuation of the finding in segment VII/VIII?* Mean 46.5 HU versus 71.4 HU parenchyma; hypodense (ratio 0.65).

`f1_vessel`: *Is the finding in segment VII/VIII adjacent to a major vessel?* No.

`f1_desc`: *Describe the finding in segment VII/VIII.* A 3.6 x 1.9 cm arterially hypervascular area at the left margin of the aforementioned defect.

`f1_aphe`: *Does the finding in segment VII/VIII show arterial-phase hyperenhancement?* Yes, hypervascular on arterial phase.

There is no Phase 3 question on this scan because no earlier scan exists in the series.

Single-step questions (one calculation or decision each).

`p4_category_VIIVIII`: *What LI-RADS category applies to the finding in VII/VIII, and why?* LR-4. LI-RADS: at-risk patient, APHE, $\geq$20 mm, 0 additional major features $\rightarrow$ LR-4. Report assigns LR-4, concordant with the feature-derived category (on `f1_size`, `f1_hu`, `f1_aphe`).

`p4_next_step`: *What is the recommended next step?* Post-treatment follow-up imaging in 3 months per LI-RADS treatment-response guidance (on `p4_category_VIIVIII`).

On the next scan (2022-10-03, 7.0 months later), two Phase 3 questions depend on questions in this organ record. `s3_liver_f1_temporal`: *Has the finding in segment VII/VIII changed compared to the prior scan?* Interval treatment: the prior lesion site is now a treatment zone; size not comparable to the untreated lesion. Label TREATED (on `s3_liver_f1_size`, `s2_liver_f1_size`). `s3_liver_f2_temporal`: *Has the finding in segment V changed compared to the prior scan?* New finding: the region was reported normal on the prior scan. Label NEW (on `s2_liver_seg5`, `s3_liver_p3_interval`).

Compositional question.

`p4_assessment`: *What is the assessment of the liver on this scan? Show the reasoning chain.* Answer: the reference rationale in Box 2 (on `f1_size`, `f1_hu`, `f1_vessel`, `f1_desc`, `f1_aphe`, `p4_category_VIIVIII`, `seg7`, `seg8`).

Box 2: The reference rationale of `s2_liver_p4_assessment`, as stored: six sections that restate the evidence of Box 1 and derive LR-4. Every number and report sentence comes from the questions of Box 1; the rule section names the decision-table row that the imaging features select.

[Observations] 53.0 mm liver finding in VII/VIII (mean HU 46.5 vs parenchyma 71.4, volume 36.05 mL). A 3.6 x 1.9 cm arterially hypervascular area at the left margin of the aforementioned defect, LIRADS 4.

[Temporal] No prior scan in this series.

[Features] APHE is present: the report describes the lesion as hypervascular. Nonperipheral washout is not assessable because it requires portal venous or delayed phase; the provided phase is arterial late. Enhancing capsule is not described in the report. Threshold growth is not assessable because no prior measurement is available.

[Rule] LI-RADS v2018 applies given known liver cirrhosis and HCC diagnosed on imaging 2021. The row is at-risk patient, APHE, $\geq$20 mm, 0 additional major features; the derived category is LR-4.

[Assessment] Reported categories are LI-RADS-TR NON-VIABLE and LI-RADS 4. Reported LR-4 is concordant with derived LR-4.

[Plan] No liver-specific follow-up recommendation is stated for this scan.

Patient-level questions of the same organ. The liver series of this patient has five patient-level questions. *How many scans of the liver does this series contain, and over what period?* 5 scans, from 2022-03-05 to 2023-10-02 (1.6 years). *Across the series, on which scans was the liver reported abnormal in at least one screened segment?* Every scan of the series (2-6) reported at least one abnormal segment. *Describe the course of the finding in segment VII/VIII across this series.* First identified on scan 2 (2022-03-05), 53.0 mm; scan 3 (2022-10-03): treated site (treatment zone, size not comparable), 61.1 mm; stable or present without size comparison on scans 4-6 (2023-01-03 to 2023-10-02), 51.1-55.1 mm. *Which findings in the liver are treated sites, how did they evolve, and which findings appeared after each treatment?* Treated site in segment VII/VIII, present from scan 3 (2022-10-03), with the course above; findings first identified after the treated site of scan 3: scan 5 (2023-04-03): finding in segment VII/VIII. *Considering every scan of this series, what is the overall course of the liver findings? State the rule used.* Latest interval event on scan 3 (2022-10-03): new finding in segment V. Interval events: scan 3: new in segment V; scan 3: not separately identified afterwards in segment V; scan 3: not separately identified afterwards in segment VII/VIII; scan 5: not separately identified afterwards in segment VII/VIII. High-suspicion categories assigned: none assigned. Treated sites: segment VII/VIII from scan 3. Rule: the verdict names the latest scan with a non-stable interval label (new, growth, decrease, resolved) across all tracked findings of the organ, listing every such event of that scan; a finding first identified mid-series without a prior match, or not separately identified on later scans, is recorded as an event but only decides the verdict when no labelled event exists; all-stable labels give 'stable'; an organ reported normal on every scan gives 'no finding'. Each of these answers lists the scan-level questions it depends on; the course of the segment VII/VIII finding, for example, depends on the size, description, and Phase 3 questions of that finding on every scan (`s2_liver_f1_size` to `s6_liver_f1_temporal`).

A.4 Data format

Each question stores its reference answer, evidence source, and tier. The records use JavaScript Object Notation (JSON). Its bounding box locates the organ, region, or finding that holds the evidence. When several findings share a box, a neutral location cue names the intended finding without describing it. The question also records the available clinical context and dependency links (Table 6).

The questions and their dependency links form a *question dependency graph*. Each node is a question; each edge says that one question gives evidence for another. A *linked-question chain* contains an endpoint question and every question it depends on, directly or through other questions. Linked-question evaluation asks for all intermediate answers and the conclusion in one response. Each question can also be scored on its own in individual VQA. Intermediate supervision therefore uses the same reference answers as evaluation.

RadThinking is a directory per patient. It holds one `vqa.json` with the patient's questions and one folder per scan with the bounding-box files. Patient, scan, and box identifiers are assigned by a seeded permutation; the mapping to the source is kept private.

```
RTHINK_P002399/                  one folder per patient
  vqa.json                       patient_id, total_scans, date_range,
                                 scans[], longitudinal_vqa
    scans[]                      scan_id, case_id, scan_date,
                                 contrast_phase, clinical_indication, vqa
      vqa.phase0_visual_context  scan-level questions
      vqa.organs.<organ>         phase1_region_screening,
                                 phase2_findings_characterization,
                                 phase3_temporal_context, phase4_decision
    longitudinal_vqa.<organ>     patient-level questions of one organ
  RTHINK_00003210/               one folder per scan (case_id)
    organ_liver_ct.json          organ box: voxel coordinates in the source CT
    liver_segment_7_ct.json      region box
    finding_liver_01_ct.json     finding box (neutral name)
```

A bounding-box file records the voxel bounds of the box in the source CT (`start`, `stop`, `slicing`, `full_ct_shape`), the index convention, the value convention (Hounsfield units as `int16`, no windowing), and the affine of the box. It also stores the SHA-256 digest of the boxed array, so that a crop cut from the source CT can be verified. The question field `volume` gives the Neuroimaging Informatics Technology Initiative (NIfTI) file name of the image crop that holds the question's evidence. We call this an *evidence crop*. The bounding-box file with the same stem describes its voxel bounds, and the accompanying script cuts the crop from the source CT. `prior_volume` names the evidence crop from the earlier scan for Phase 3 questions. Table 6 lists the fields of a question.

A.5 Preparation of benchmark inputs

Reports and reference answers are excluded from zero-shot benchmark inputs. Each request contains only the information allowed for its questions, with reference answers stored separately for scoring.

The individual benchmark uses the questions whose bounding box is prepared and whose answer type has a scorer. Eligibility is decided from the dataset, not from what a model can take as input. For linked-question evaluation, a linked-question chain is eligible only when its endpoint and every question it depends on are prepared and linked to boxes. The eligible linked-question chains cover single-scan and longitudinal reasoning with categorical, multi-label, structured, and numeric answers. A longitudinal linked-question chain keeps the scan history it needs. Evaluation therefore covers every question in the linked-question chain.

A.6 Cohort and coverage

Table 7 lists the numbers behind Figure 2. Switzerland contributes 2,493 patients and Turkey 186. The 241 control scans belong to 241 patients with one scan each and no cancer diagnosis. Age is recorded for 2,367 patients and sex for 2,376.

Table 6: **Every question has the same five fields; the other fields appear where the question kind defines them.** We list each field with the number of questions that carry it and what it holds. The identifier, question, answer, evidence source, and tier are present in all 122,127 questions. Of these, 111,633 name their evidence crop, and 8,716 also name the evidence crop from the earlier scan. The second block holds the measured and reported values: 7,311 sizes, 6,234 volumes and attenuations, 10,554 organ volumes, 4,502 intervals, 4,214 temporal-change labels, 424 reported categories, and 1,283 derived categories or managements. `mask` means CT-derived measurement.

field	questions	content
`id`, `question`	122,127	Identifier and question text. The identifier encodes the scan, organ, finding, and question kind.
`answer` or `answer_chain_of_thought`	122,127	Reference answer; a reference rationale of six sections for assessment and diagnosis questions.
`evidence_basis`	122,127	Evidence source: CT-derived measurement (`mask`), report, metadata, `rule` (computed from recorded values), composition, image, or combinations such as report and `rule`.
`tier`	122,127	foundation, single-step, or compositional.
`volume`, `prior_volume`	111,633 / 8,716	Evidence crop; evidence crop from the earlier scan for Phase 3 questions.
`supported_by`	29,947	Identifiers of the questions the answer depends on.
`region`, `regions`	30,369 / 19,752	Screened region; regions occupied by a finding.
`clinical_context`	11,964	History available at the time of the scan (Phase 4 and patient-level questions).
`size_mm`, `volume_ml`	7,311 / 6,234	Longest diameter; volume of the finding.
`mean_hu`, `parenchyma_hu`, `hu_ratio`	6,234 / 16,780 / 6,233	Attenuation of the finding and of the organ parenchyma, and their ratio.
`organ_volume_ml`, `fov_truncated`	10,554 / 3,243	Segmented organ volume; flag when the organ reaches the edge of the field of view.
`days`, `months`, `prior_scan`	4,502 / 3,728 / 4,502	Interval since the earlier scan and its identifier.
`label`, `prior_finding`, `volume_ratio`	4,214 / 2,367 / 625	Temporal-change label; matched earlier finding; volume ratio when both findings are measured.
`category`, `derived`, `concordant`, `derivation_note`	424 / 1,283	Reported category; rule-derived category or management; whether the two agree, with a note where they do not.
`finding`, `finding_class`	1,283 / 368	Finding a guideline question refers to; `treated_site` marks a finding that is a treatment zone rather than a lesion.
`accepted_answers`, `answer_note`	1,539	Alternative correct answers when the report's plan and the guideline's management differ.
`contrast_phase`, `clinical_indication`	5,365	Contrast phase and clinical indication of the scan (Phase 0).

Table 7: **RadThinking has 2,679 patients, 5,365 CT scans, 11,741 organ records, and 122,127 questions.** We list the cohort on the left and the questions on the right; these are the numbers behind Figure 2. More than half of the patients (1,411) have two or more scans; 1,347 of them are followed for a median of 321 days. Of the organ records, 6,551 contain a finding and 5,190 are organ-level negatives. Of the questions, 87,846 are foundation, 18,049 single-step, and 16,232 compositional; most reference answers come from reports (42,276) and measurements (36,212). An organ record is the set of questions about one organ on one scan; a finding is one lesion or abnormality. Age is known for 2,367 patients and sex for 2,376; IQR, interquartile range; comp., compositional.

cohort		**dataset**	
patients	2,679	organ records (16 organs)	11,741
Switzerland / Turkey	2,493 / 186	with $\geq$1 finding	6,551
CT scans	5,365	organ-level negative	5,190
per patient: 1 / 2 / $\geq$3	1,268 / 827 / 584	findings	10,029
patients with $\geq$2 scans	1,411	with CT-derived measurements	6,234
scans spanning $>$1 day	1,347	with a report description	8,227
follow-up, median (IQR), days	321 (104–784)	bounding boxes: organ / region / finding	18,812 / 28,213 / 8,491
followed for one year or longer	635	questions	122,127
scan years	2010–2025	foundation / single-step / comp.	87,846 / 18,049 / 16,232
contrast phase: portal venous	2,600	Phase 0 / 1 / 2	25,820 / 30,369 / 28,415
late / early arterial	1,373 / 670	Phase 3 / 4 / patient-level	8,716 / 18,313 / 10,494
non-contrast / other	685 / 37	evidence: measurement / report	36,212 / 42,276
sex: men / women / unknown	1,360 / 1,016 / 303	metadata / rule / comp. / image	18,261 / 17,255 / 7,876 / 247
age, median (IQR)	66 (56–75)	temporal-change labels (Phase 3)	4,214
recorded cancer diagnosis	1,153	reference rationales (Phase 4)	11,795
		guideline questions (Phase 4)	1,283
		LI-RADS / Bosniak / Fleischner / ACR	283 / 141 / 60 / 799

B Construction and Quality Control of Reference Rationales

The construction pipeline turns three source artifacts into questions and reference rationales: the CT scan with its organ and lesion segmentations, the de-identified radiology report, and the scan metadata. Every stage is deterministic except two: the extraction of findings from free-text reports and the writing of the reference-rationale prose. Both use large language models, and both are checked by an independent model, by fixed checks, and by physician review of random samples. Table 8 lists the stages in order. The stored source links, not the text below, define the inputs and targets of every experiment.

B.1 Sources and de-identification

A scan enters the pipeline with its CT volume, its organ and lesion segmentations, its de-identified report, and its metadata (date, contrast phase, and the clinical history available at the time). Residual identifying text is removed from every text field: calendar dates of earlier scans, clock times, de-identification placeholders, and sentences that only document how a result was communicated. The scan's own date is kept as a date in YYYY-MM-DD format, because interval questions are computed from it.

B.2 Extracting findings from reports

Free-text reports are the main source of clinical vocabulary, and they are the least structured input. Four extraction sources vote on each finding. A per-lesion table prepared in an earlier extraction pass and a parser with organ-specific vocabularies give two sets of candidates without language models. GPT-6 reads the report together with the table's candidates and confirms, rejects, or adds findings; Claude Opus 5 reads the report alone, so its extraction is independent. Each source gives structured findings: organ, location, finding type, size, change relative to earlier imaging, a reported category where one is stated, and the recommended follow-up. A finding is kept when at least two sources agree. Negations, deferred statements, history, categories, and recommendations are taken verbatim from the report. The same two models handle situations the extraction criteria do not cover, such as two segmentation files that claim the same organ. GPT-6 chooses from a fixed menu of actions, Claude Opus 5 must agree, and the decision is recorded.

Table 8: **Eight stages turn the CT, the report, and the metadata into questions and reference rationales, and every stage records its inputs, so each answer traces to its evidence.** We list the stages in order. Two stages use language models, finding extraction (Section B.2) and rationale prose (Section B.4); the others are deterministic. The physician review (Section B.5) is repeated until the sampled error rate is below 2%; the audit (Section B.6) checks every file.

stage	what it does
ingest and de-identify	Join the scan, its segmentations, its report, and its metadata. Remove residual dates, clock times, placeholders, and communication notes from every text field.
extract findings	Four extraction sources vote on each finding: a per-lesion table, a vocabulary parser, GPT-6 (OpenAI, 2026), and Claude Opus 5 (Anthropic, 2026); a finding needs two votes (Section B.2).
link findings to segmentations	Match each report finding to a segmented component by organ, location, type, and size; refuse a match whose attenuation contradicts the finding type.
link findings over time	Match findings across scans of the same patient and assign a temporal-change label; unmatched findings are labeled as such.
build questions	Write the scan-level questions (Phase 0 to Phase 4) and the patient-level questions with their bounding box, evidence source, tier, and dependency links; remove diagnosis words from Phase 0 to Phase 3 answers.
build reference rationales	Apply the organ's clinical guideline as a decision table to the recorded imaging features; write the rationale from the resulting fact sheet and verify it (Section B.4).
physician review	Sample rationales at random, collect corrections, update extraction criteria and prompts, and rebuild; repeat until the sampled error rate is below 2% (Section B.5).
audit	Check every file automatically (Section B.6) and assign new identifiers.

Each finding is then linked to the segmentations. A report finding is matched to a segmented component by organ, location, type, and size. The match is refused when the component's attenuation contradicts the finding type, for example a stone below 130 HU, a simple cyst above 60 HU on a contrast-enhanced scan, or a fat-containing lesion above 10 HU. The finding then keeps its description but carries no CT-derived measurements. Every finding is backed by a report sentence; a segmented component that no report finding explains is not a finding. Findings are matched across the scans of a patient by organ, location, and type. A matched finding receives a temporal-change label from the report's comparison wording or, where both findings have CT-derived measurements, from the volume ratio. Findings without a match are labeled unmatched, and findings whose earlier size cannot be compared are labeled persistent.

B.3 Building the questions

Templates write the questions from the measurements, the linked findings, and the metadata. Every question records its evidence source (measurement, report, metadata, rule, composition, or image), its tier (foundation, single-step, or compositional), and the questions it depends on. Answers of Phase 0 to Phase 3 questions are written without diagnosis words: a finding is described by its location, size, attenuation, and morphology, and the box files carry neutral names. A diagnosis or category appears only in Phase 4 answers and in the patient-level questions, where it is the conclusion being asked for. When the report's plan and the guideline's management differ, both answers are accepted as correct.

B.4 Reference rationales from clinical guidelines

A *reference rationale* is the reference answer of a Phase 4 assessment or diagnosis question. Its six sections follow the order of a radiology read. The *Observations* section restates the measured and reported findings. The stored *Temporal* section gives temporal context by restating the temporal-change labels. The *Features* section lists the imaging features used by the organ's clinical guideline and whether the scan's contrast phase allows each to be assessed. *Rule* names the clinical guideline and the row of its decision table that the imaging features select. *Assessment* states the reported category and whether the rule-derived category agrees with it. *Plan* states the reported follow-up or the management derived from the clinical guideline. Appendix A.3 gives a complete reference rationale and the questions it depends on.

Five clinical guidelines are implemented as decision tables: LI-RADS v2018 for liver findings in patients at risk of hepatocellular carcinoma (HCC) (Chernyak et al., 2018), Bosniak v2019 for cystic renal masses (Silverman et al., 2019), the Fleischner Society 2017 guideline for incidental pulmonary nodules (MacMahon et al., 2017), and the ACR white papers on incidental adrenal masses (Mayo-Smith et al., 2017) and pancreatic cysts (Megibow et al., 2017). A table applies only when the finding type, patient context, and scan's contrast phase make the clinical guideline applicable. Otherwise the reference rationale states why it does not apply and follows the radiologist's characterization. Table 4 lists the imaging features and applicability conditions for each organ.

Each imaging feature is recorded as present, absent (the report negates it), not described (the report is silent), or not assessable (the scan's contrast phase cannot show it). The decision table counts only imaging features that are present; the reference rationale states the other states explicitly. The reported category stays the reference answer. The rule-derived category and an agreement flag are stored alongside it. When the categories differ, the reference rationale states whether an imaging feature that is not assessable or not described could explain the difference. When the report's plan and the clinical guideline's management differ, both answers are accepted.

The rationale text is produced in two stages. A template first writes the rationale from the fact sheet of the organ on that scan: the measurements, the verbatim report sentences, the temporal-change labels, the clinical context, and the decision-table trace. GPT-6 then rewrites the rationale as prose under fixed constraints: every number must appear in the fact sheet, every report sentence must be quoted verbatim, every temporal-change label must be restated, and the reported category must be kept. Claude Opus 5 grades the prose against a fixed rubric (facts only, section order, contrast-phase availability respected, rule applied faithfully, assessment kept, no omissions). A deterministic pass removes any sentence about the pipeline instead of the patient and rejects a rationale that quotes a number not in the fact sheet.

B.5 Physician review

The extraction criteria, the prompts of the two models, and the decision tables were developed with physicians in the loop. In each round, the pipeline was run end to end, and 5% of the resulting reference rationales were sampled at random. Eight physicians marked each rationale as correct or recorded the error (a missed or invented finding, a wrong temporal-change label, a misapplied guideline, or a wrong conclusion). Each error was traced to its cause and fixed at the source: a missing pattern in the extraction vocabulary, a prompt instruction for one of the two models, a wrong or missing row in a decision table, or a matching criterion. The pipeline was then rerun on all scans, and a new random sample was reviewed. We stopped when the error rate of the sample fell below 2%, in the fifth round; the rationales of that round are the ones in RadThinking. Samples were drawn independently in every round, so a rationale could be reviewed more than once.

B.6 Automated checks

Every file is checked by an automated audit that repeats, on the JSON files alone, the checks the pipeline made during construction. The audit checks seven things. Every bounding box referenced by a question exists. Every dependency link points to a question of the same or an earlier scan. Every number in a reference rationale traces to a linked question, the report, the clinical context, or a decision table. Every linked measurement, temporal-change label, category, and report description is restated in the rationale. Phase 0 to Phase 3 answers contain no diagnosis words. No free-text field contains a calendar date, clock time, placeholder, or communication note. No rationale describes the pipeline itself. Patient, scan, and bounding-box identifiers are assigned by a seeded random permutation, and the mapping to the source is kept private.

B.7 Clinical history and organ vocabulary

Clinical history. The clinical context of a question is the history available at the time of the scan: known cancer diagnoses, treatments, and resections from the metadata, and the findings of earlier scans of the same patient. Information from later scans never enters an earlier question. A category or diagnosis word in the context is written as history (for example, a history of hepatocellular carcinoma), never as a statement about the current scan.

Organ normalization. Organ names in reports and segmentation files are mapped to the 16 canonical organs of RadThinking (Table 4); the original wording is kept in the description. Exact matches come before broader matches, and a compound label that spans organs is not turned into a single-organ finding. A finding that names a region (a liver segment, a kidney pole) is linked to that region's bounding box; the box, not the wording, decides which evidence crop the model sees.

C Relationship to Existing Data and Reasoning Resources

Evaluating and supervising the steps of a radiology read needs three things from a dataset: the evidence of each step located in the image, the steps linked to one another and across the scans of a patient, and the clinical guideline that maps imaging features to a decision. Table 9 groups the resources that informed the design of RadThinking into six families by what they give. Each family gives one form of supervision; RadThinking connects them. The grouping describes the focus of each resource, and the members of a family do not all have the same annotations.

Datasets. Public radiology datasets give one form of supervision each. Segmentation datasets and lesion benchmarks locate organs and lesions (Heller et al., 2019; Bilic et al., 2019; Antonelli et al., 2021; de Grauw et al., 2025; Li et al., 2024; 2025). Report datasets pair each scan with the radiologist's text (Johnson et al., 2019; Hamamci et al., 2026; Blankemeier et al., 2026; Bassi et al., 2025). Question datasets ask for one attribute of one image and score each question alone (Lau et al., 2018; Liu et al., 2021; He et al., 2020; Zhang et al., 2023; Hu et al., 2024; Bai et al., 2024; Ji et al., 2024; Chen et al., 2025; Gautam et al., 2025); RadGenome-Chest CT grounds its answers in segmentation masks (Zhang et al., 2024). Few link the scans of a patient over time. Chest ImaGenome relates findings across chest radiographs (Wu et al., 2021), 3D-RAD asks temporal questions about CT scans (Gai et al., 2025), and longitudinal lesion benchmarks track lesions across MRI scans without questions (Link et al., 2024; Rokuss et al., 2024). One records reasoning: CheXthought stores the written steps and the gaze of radiologists reading chest radiographs (Sharma et al., 2026). Table 10 compares the closest of these datasets with RadThinking property by property. RadThinking keeps all of them in one data structure: the box of every answer, the evidence source behind it, the link between the scans of a patient, and the reference rationale of each assessment or diagnosis.

Models. Vision-language models for CT answer questions or write reports about a CT scan (Wu et al., 2025a; Bai et al., 2024; Blankemeier et al., 2026; Hamamci et al., 2026; Xu et al., 2025; Sellergren et al., 2025). They are scored on the answer to each question or on the report text. Section 3 scores such models on the intermediate answers and the conclusion in one response, and Section 4.1 measures whether the conclusion follows the images or the written intermediate answers.

Reasoning steps. Methods for reasoning treat the steps as free text. Chain-of-thought prompting lets a model write its steps before the answer (Wei et al., 2022; Xu et al., 2024), and visual chain of thought adds the image region to look at (Shao et al., 2024a). Free text has two limits. No step is checked against the image, and the answer need not follow the written steps: for language models, altering the steps or the prompt shows that the answer often does not depend on them (Turpin et al., 2023; Lanham et al., 2023). Medical reasoning models train on written steps (Chen et al., 2024; Wu et al., 2025b) or reward the conclusion with reinforcement learning, as general vision-language models do (Liu et al., 2025; Meng et al., 2025; Huang et al., 2025; Pan et al., 2025; Lai et al., 2025). Process rewards score each step instead (Lightman et al., 2024); Med-PRM checks each step of a text answer against clinical guidelines (Yun et al., 2025). In all of these, a step has no reference answer in the image. Steps with a reference exist in general vision: CLEVR and GQA attach a program of steps to each question (Johnson et al., 2017; Hudson & Manning, 2019). Decomposition at inference splits a question into sub-questions or primitives, but these have no reference answers (Press et al., 2022; Zhou et al., 2023; Khot et al., 2023; Gupta & Kembhavi, 2023; Surís et al., 2023). RadThinking records the steps of a radiology read in the data: each is a question with a reference answer, a bounding box, and an evidence source, so it is scored on its own, its effect on the conclusion is measured, and it is a training target.

Clinical guidelines. A clinical guideline states which imaging features decide a category and which management follows: LI-RADS v2018 for liver findings in patients at risk of hepatocellular carcinoma (HCC) (Chernyak et al., 2018), Bosniak v2019 for cystic renal masses (Silverman et al.,

Table 9: **Six families of existing resources each give one form of supervision; RadThinking connects all six in one data structure.** We group the resources that informed the dataset design by what they give and state, for each family, what RadThinking adds. Spatial annotation, reports, and visual questions each cover one step of a radiology read; longitudinal resources relate the scans of a patient over time; reasoning resources supervise written steps, mostly in text; clinical guidelines state the rules a radiologist applies. RadThinking keeps the box of every answer, the evidence source behind it, the link between the scans of a patient, the reference rationale of each assessment or diagnosis, and the clinical guideline it applies. The table is about roles; a resource's full collection and its evaluation subset need not share every annotation type.

resource family	representative resources	what it gives and what we add
spatial annotation	KiTS, LiTS, Medical Segmentation Decathlon, ULS23, AbdomenAtlas, and PanTS (Heller et al., 2019; Bilic et al., 2019; Antonelli et al., 2021; de Grauw et al., 2025; Li et al., 2024; 2025)	Masks say where an organ or lesion is. Every question in RadThinking names the box of the organ, region, or finding that holds its evidence.
images and reports	MIMIC-CXR, CT-RATE, Merlin, and AbdomenAtlas 3.0 (Johnson et al., 2019; Hamamci et al., 2026; Blankemeier et al., 2026; Bassi et al., 2025)	Reports give the radiologist's findings and conclusions as free text. RadThinking turns the report of each scan into typed answers that keep their source sentence.
medical visual questions	VQA-RAD, SLAKE, PathVQA, PMC-VQA, OmniMedVQA, M3D-VQA, RadGenome-Chest CT, AMOS-MM, 3D-RAD, DeepTumorVQA, and Kvasir-VQA-x1 (Lau et al., 2018; Liu et al., 2021; He et al., 2020; Zhang et al., 2023; Hu et al., 2024; Bai et al., 2024; Zhang et al., 2024; Ji et al., 2024; Gai et al., 2025; Chen et al., 2025; Gautam et al., 2025)	Each question asks for one attribute of one image and is scored alone. RadThinking links its questions, so a linked-question chain is scored as a whole.
longitudinal imaging	Chest ImaGenome, 3D-RAD, and longitudinal lesion benchmarks (Wu et al., 2021; Gai et al., 2025; Link et al., 2024; Rokuss et al., 2024)	Findings are related across the scans of a patient. RadThinking matches findings across scans and stores a temporal-change label for each match (Phase 3).
reasoning supervision	chain-of-thought prompting, CheXthought, HuatuoGPT-o1, MedReason, MedVLM-R1, Med-R1, and Med-PRM (Wei et al., 2022; Xu et al., 2024; Shao et al., 2024a; Sharma et al., 2026; Chen et al., 2024; Wu et al., 2025b; Pan et al., 2025; Lai et al., 2025; Yun et al., 2025)	Written steps, a reward on the conclusion, or a process reward supervise the steps, mostly in text. The steps of RadThinking are questions with reference answers and boxes, so each is scored, probed, and trained on its own.
clinical guidelines	LI-RADS v2018, Bosniak v2019, the Fleischner Society 2017 guideline, and the ACR white papers on incidental adrenal masses and pancreatic cysts (Chernyak et al., 2018; Silverman et al., 2019; MacMahon et al., 2017; Mayo-Smith et al., 2017; Megibow et al., 2017)	A guideline states which imaging features decide a category and its management. RadThinking implements five as decision tables and stores the rule-derived category next to the reported one.

2019), the Fleischner Society 2017 guideline for pulmonary nodules (MacMahon et al., 2017), and the ACR white papers on incidental adrenal masses and pancreatic cysts (Mayo-Smith et al., 2017; Megibow et al., 2017). RadThinking implements these five guidelines as decision tables, applies them to the recorded imaging features, and stores the rule-derived category next to the reported one (Section 2.3, Table 4).

Table 10: **RadThinking is the only public dataset that is longitudinal, grounds every question in a 3D box, stores a reference rationale with every decision, and applies clinical guidelines.** We compare RadThinking with public chest radiograph and CT datasets that come with reports, questions, or reasoning annotations, grouped by region and modality. Each has only one or two of these. CT-RATE (Hamamci et al., 2026) and Merlin (Blankemeier et al., 2026) have reports but no questions. RadGenome-Chest (Zhang et al., 2024) and DeepTumorVQA (Chen et al., 2025) have 1.3M and 395k questions about single scans and are not longitudinal; AMOS-MM (Ji et al., 2024) has questions but no grounding. 3D-RAD (Gai et al., 2025) is longitudinal but has no grounding and no reasoning traces. Chest ImaGenome (Wu et al., 2021) is longitudinal with 2D boxes, and CheXthought (Sharma et al., 2026) has 104k reasoning traces from radiologists, but both are 2D and have no questions. RadThinking has all of these on abdominal CT: 2,679 patients with 2.0 scans each on average, four contrast phases, 122k questions, 11.8k reference rationales, and five clinical guidelines. Contrast phases: those in the data (NC, non-contrast only); longitudinal: the scans of a patient are linked and every matched finding carries a temporal-change label (Phase 3); grounding: the image region of each answer (masks, 2D or 3D boxes, or radiologist gaze); rationales: written reasoning annotations per answer. All datasets except CheXthought and DeepTumorVQA come with reports; the AbdomenAtlas 3.0 (Bassi et al., 2025) reports are generated from segmentations. ✓ present, parentheses partial (MIMIC-CXR (Johnson et al., 2019) links the studies of a patient in time but has no temporal-change labels), – absent or not stated; k, thousands; M, millions.

dataset	patients	scans	scans per pt.	contrast phases	questions	longitudinal	grounding	rationales	guidelines
chest									
2D radiograph									
MIMIC-CXR	65k	228k	3.5	–	–	(✓)	–	–	–
Chest ImaGenome	–	242k	–	–	–	✓	2D boxes	–	–
CheXthought	–	50k	–	–	–	–	gaze	104k	–
3D CT									
CT-RATE	21k	26k	1.2	NC	–	–	–	–	–
RadGenome-Chest	21k	26k	1.2	NC	1.3M	–	masks	–	–
3D-RAD	11k	16k	1.4	NC	170k	✓	–	–	–
abdomen									
3D CT									
Merlin	18k	25k	1.4	–	–	–	–	–	–
AbdomenAtlas 3.0	–	9.3k	–	–	–	–	masks	–	–
AMOS-MM	–	2.3k	–	–	✓	–	–	–	–
DeepTumorVQA	–	9.3k	–	–	395k	–	–	–	–
RadThinking (ours)	2,679	5,365	**2.0**	**4**	122k	✓	**3D boxes**	**11.8k**	**five**

D Dataset Reuse and Learning Pathways

The question hierarchy enables several training and analysis settings on the same evidence. The fine-tuning and reinforcement-learning results in Section 4.2 are the training experiments reported in this paper. The uses below are possible directions, not completed comparisons.

D.1 A curriculum from findings to composition

Foundation questions can supervise image-conditioned recognition of individual attributes. Single-step questions add one relation, eligibility check, or specified operation. Compositional questions then connect the annotated intermediate answers to an endpoint. This gives a natural curriculum (Bengio et al., 2009): learn the evidence fields first and their dependencies next, or sample the tiers together. Which schedule is better needs a matched experiment; the hierarchy alone does not show it.

The same questions give both (image, question, answer) examples and (image, question, intermediate answers, conclusion) examples. Intermediate targets tied to a source differ from explanations a model distills. Because both use the same fields, recognition, composition, and the effect of supervision can be compared directly. Image–text alignment or broader medical instruction training can precede these stages but does not replace the provenance of each question.

D.2 Possible reward-based extensions

The typed fields offer several rewards. An endpoint is checked against its reference. A category question can be checked against the reported category, with the derived one as a second reference. A management answer can be checked against its accepted answers. Temporal-change labels can

check whether the answer names a new, stable, growing, shrinking, or resolved finding. Rewards must account for unavailable, uncertain, or inapplicable evidence. Missing values are not negative answers.

These rewards could be combined with intermediate supervision in further reward-based experiments. Format validity should be measured apart from agreement with the clinical evidence, and agreement with a reference answer is not verification against pathology. Penalties and reward weights per category need to be specified and validated; the fields alone do not imply them.

D.3 Independent evaluation axes

Tier and evidence source describe different things. The tier says what operation the question asks for; the evidence source says where the reference answer comes from (measurement, report, metadata, rule, composition, or image). Crossing the two separates the difficulty of reading a piece of evidence from the difficulty of combining evidence that is already available. Questions across scans versus questions about one scan is a third axis. These breakdowns complement the workflow-phase averages and linked-question metrics of Section 3; the benchmark table is computed from the workflow-phase fields, not the tiers.

D.4 Retrieval and longitudinal evidence reuse

An index of typed findings could retrieve similar patient histories by anatomy, appearance, change over time, and context. Such a system would return patient evidence, not a diagnosis based on surface similarity. Literature retrieval would add general clinical knowledge. A retrieval experiment would need patient-disjoint evaluation and controls for overlapping documentation, missing follow-up, and outcome leakage. This paper provides the basis for that direction, not a retrieval system.

E Dataset and Evaluation Details

E.1 Question records and linked-question chains

Each evaluated question retains a source-question identifier, patient data split, scan and evidence-crop association, answer type, options when applicable, and evidence source. Each linked-question chain additionally records its endpoint question and prerequisite questions. Zero-shot benchmark inputs and reference answers are stored separately. Identifiers used for joining records do not provide a clinical answer and are not original accession numbers.

E.2 Interpretation of spatial and temporal evidence

Organ and finding bounding boxes locate the evidence without prescribing the answer. Temporal view labels and scan dates establish order; an elapsed-time question can therefore test use of the given metadata rather than discovery of an imaging feature. Image-derived and provided-context targets are distinguished in representation analyses. Temporal-change labels are scored as given in the dataset; we do not reinterpret a generic size or volume change as a complete clinical response criterion.

E.3 Evaluation details

Inputs. Each individual question keeps the dataset's question–evidence-crop association and uses closed categorical options, explicit label sets, or typed numerical fields, as appropriate to the task. In linked-question chains each field keeps its own answer type and original options; the same source question can appear in more than one chain, and patient and source-question identity are retained. Longitudinal organ views are resampled at $1.5 \times 1.5 \times 3$ mm and padded to a common physical canvas; the crop retains the organ and surrounding context, earlier and current organ views are not independently stretched, and every planned axial plane is kept. Neutral target coordinates are mapped through each model's actual preprocessing. Matched scale does not imply anatomical registration. Questions that need broader anatomical coverage keep their source associations rather than being answered from an unrelated organ view. Phase 0 visual-context questions enter the phase average

over Phase 0 to Phase 4 when all five workflow phases are scored. Patient-level questions are reported separately and do not enter that average.

Answer extraction. Shared, model-independent extractors extract explicit answers to individual and linked questions for a common task scorer. Individual extraction accepts complete answer wrappers, case-normalized public options, whole option descriptions, singleton label sets, and explicit typed fields; an unmapped letter alone is not assigned an option by position. Linked-question extraction additionally accepts direct step maps, nested answer wrappers, option descriptions that differ only in case, dash style, or one terminal punctuation mark, an option code paired with its own description, explicit step assignments inside an answer string, single-key value wrappers, and one allowed option marked true; a bare value is never bound to a step. Both extractors then map a remaining whole answer to the question's published options with a fixed, reference-blind synonym lexicon (for example, yes or no for present/absent options, "no appreciable change" for STABLE, LR5 for LR-5, an option letter followed by that option's own text, and a length with units to the matching size category); hedged, negated, multi-option and misspelled answers stay unresolved. Dimension-compatible, field-bound units, such as centimetres, are converted using public field names and options, never reference values. JSON syntax alone is not penalized. Unlabelled narrative, uncertain measurements, missing values and conflicting answers remain unresolved, and answers are neither inferred from reasoning text nor copied between questions. An unresolved answer is therefore not evidence that every clinical statement in the response is incorrect.

Execution. Each model uses its checkpoint-specific input interface: physical coverage is shared, while the final tensor resolution and native context capacity differ. Table 1 combines original and recovery outputs by unique source identity, keeping accepted responses and adding only disjoint recovery outputs; recovery runs use extended-context or full fixed-spacing tiled inputs where needed, so the comparison is not restricted to native interfaces. ClinFusion-32B (Yuan et al., 2026) uses memory-bounded execution with all visual branches retained. API models receive the same views rendered as image panels; chains and questions beyond the provider image limit are not generated, and API calls that ended without an answer are excluded and never re-sent. With thinking on, individual questions receive a larger generation allowance than with thinking off so that the reasoning can finish; linked-question chains use the same allowance in both modes. Hulu-Med-32B (Jiang et al., 2025) and Lingshu-I-8B (Xu et al., 2025) use default generation, because a native thinking switch has not been verified; this does not prevent them from writing intermediate answers. Greedy decoding of Qwen3.8-27B (Qwen Team, 2026) on our inference stack is not bitwise reproducible across runs, as near-tied tokens can flip, although the parsed answers were unchanged in repeated-run checks. Generated answers and any exposed reasoning are stored separately.

Numerical tolerances. A numerical answer is correct only when every known scalar lies within its field-specific tolerance, the same for every model: the larger of 2 mm or 50% for diameter, 25% for finding volume, 10% for organ volume, 10 HU for attenuation, and 0.10 absolute for ratios; elapsed days must match exactly and months allow 0.05 month for rounding. These are exploratory benchmark tolerances, not clinically validated equivalence bounds, and the diameter tolerance was revised after inspecting initial results. Unknown reference scalars are omitted, answers with a wholly unknown reference are excluded, and missing values are never replaced by zero.

Categorical chain accuracy: eligibility. Conclusion accuracy does not require correct intermediate answers. A chain enters the denominator of categorical chain accuracy when it has a categorical conclusion, at least one categorical intermediate answer, and known references for all required categorical answers; chains with a numerical conclusion or only numerical intermediate answers are not scored for it. These three categorical metrics exclude numerical errors.

Linked-question Phase 0. Phase 0 of a linked-question chain asks about the whole scan, so the inputs of each chain include a whole-CT view for its Phase 0 answers. The API models see every plane of the whole CT; 124 of the 198 chains runnable for them fit within the provider limit of 50 images with this view; showing the whole CT as evenly spaced planes instead, to fit the other 74 chains into 50 images, leaves body-region coverage nearly never correct (0 of 60 for GPT-6, min. effort), so the table uses the full-view chains. Open-weight models whose context cannot hold every plane see a compact whole-CT view instead: the coronal and sagittal midline reformats and 8 axial

Table 11: **Categorical chain accuracy remains below 23.5% when intermediate answers and conclusions are scored together.** We compare categorical intermediate-answer accuracy, conclusion accuracy, and categorical chain accuracy, with the rows and scoring of Table 1. All values are percentages. Interm. and concl. are task-averaged accuracies of categorical intermediate answers and conclusions; categorical chain accuracy and measurements in range are defined as in Table 1. Min./max. effort denote the lowest/highest provider settings; ‡, thinking cannot be disabled; §, predeclared stratified sample excluded from ranking.

model	interm.	concl.	categorical chain accuracy	measure-ments in range
Hulu-Med-32B	34.09	37.91	10.93	1.09
Lingshu-I-8B	2.58	8.74	2.33	0.00
Qwen3.8-27B, w/o thinking	51.69	36.89	15.93	15.36
Qwen3.8-27B, w/ thinking§	6.64	9.60	6.59	2.88
Primus, w/o thinking	4.77	2.64	0.00	0.00
Primus, w/ thinking	4.80	3.65	0.00	0.00
ClinFusion-32B	28.81	48.37	15.85	4.34
GPT-6, min. effort‡	66.21	60.70	21.38	35.69
GPT-6, max. effort	66.78	56.95	18.88	34.98
Claude Opus 5, min. effort‡	39.84	44.88	20.00	29.54
Claude Opus 5, max. effort	44.32	45.92	22.07	27.53
Gemini 3.8 Flash, min. effort‡	64.46	50.02	23.45	20.00
Gemini 3.8 Flash, max. effort	59.10	44.47	20.15	19.46

planes (Hulu-Med-32B, Lingshu-I-8B). Primus sees the whole CT volume in 3D, tiled into native 192-voxel cubes. Completed of runnable chains: Hulu-Med-32B 232 of 239, excluding 1 chain over its context and 6 in which the compact view would change how its organ views are processed; Lingshu-I-8B 167 of 239, all chains that fit its context; Primus, w/o thinking 228 of 239; Primus, w/ thinking 228 of 239; GPT-6, min. effort 122 of 124; GPT-6, max. effort 116 of 124; Claude Opus 5, min. effort 122 of 124; Claude Opus 5, max. effort 122 of 124; Gemini 3.8 Flash (Google DeepMind, 2026), min. effort 121 of 124; Gemini 3.8 Flash, max. effort 113 of 124. API calls that ended without an answer are excluded, and an unresolved Phase 0 answer receives no credit. Phase 1 to Phase 4 are scored on the chain answers with the organ views. Rows without a completed run show n/a.

F Reproducibility of the Matched Training Comparison

The matched supervision comparison has two conditions: supervision of conclusions alone, and supervision of intermediate answers and conclusions. Both use the same prepared images and reference answers for conclusions. Intermediate targets are reference answers to intermediate questions, not model-generated explanations. Complete tile coverage and mapped target locations are checked after physical preprocessing. Pooling combines groups of image-token vectors and changes the number of tokens seen by the language model but not the input voxel spacing. Comparisons to native unpooled model results are therefore labeled by input condition.

F.1 Reinforcement learning

Starting from the supervised fine-tuning (SFT) adapter, reinforcement learning uses group relative policy optimization (GRPO) (Shao et al., 2024b) on linked-question chains of the training set. Two rewards are defined: an answer-only reward, which scores the conclusion with a small format term, and a per-step reward, which also scores the intermediate answers along the chain against the reference answers. The SFT + RL row of Primus (NV-Reason-CT) (Myronenko et al., 2026) in Table 2 uses the answer-only reward, selected on validation linked-question chains by a criterion fixed before training.

F.2 Fine-tuning Qwen3.8-27B

For Qwen3.8-27B, SFT trains a low-rank adaptation (LoRA) adapter (Hu et al., 2022) (rank 8, $\alpha = 16$) on the attention projections of the last 4 of its 64 language blocks, a much smaller update than the adapters of Primus and Hulu-Med-32B, which cover every language block. Table 12 reports it with the scoring of Table 2.

Table 12: Qwen3.8-27B fine-tuned with linked questions (%), scored as in Table 2.

model	scan-level accuracy						patient-level accuracy	measurements in range	categorical chain accuracy
	Phase 0	Phase 1	Phase 2	Phase 3	Phase 4	average			
individual VQA									
Qwen3.8-27B‡									
zero-shot†	38.1 (33.7, 42.4)	63.7 (60.4, 67.0)	**30.1** (25.7, 42.0)	26.0 (14.9, 41.1)	41.5 (34.2, 46.6)	39.9 (36.9, 44.1)	n/a	11.4 (8.8, 14.7)	n/a
SFT	**59.1** (54.6, 63.5)	**64.3** (61.1, 67.5)	21.4 (19.7, 28.8)	**27.4** (16.0, 43.4)	**49.0** (41.6, 55.4)	**44.3** (41.0, 48.7)	n/a	**17.0** (14.4, 19.9)	n/a
SFT + RL	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
linked-question evaluation									
Qwen3.8-27B‡									
zero-shot†	n/a	30.7 (22.8, 39.3)	8.7 (4.5, 18.4)	9.7 (3.3, 16.7)	7.4 (2.3, 12.2)	n/a	13.3 (4.7, 22.4)	2.9 (0.0, 7.6)	6.6 (3.0, 10.8)
SFT	n/a	**66.8** (62.8, 71.3)	**28.5** (23.3, 48.4)	**34.7** (18.9, 54.4)	**52.2** (42.4, 62.2)	n/a	**48.3** (35.5, 65.6)	**25.4** (17.6, 33.9)	**15.4** (10.2, 21.8)
SFT + RL	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

† Gray zero-shot results repeat Table 1.
‡ Qwen3.8-27B: w/o thinking for individual VQA; w/ thinking for linked-question evaluation.

G Hidden-State Analysis Details

G.1 Frozen inputs and supervised probes

The hidden-state analysis uses a development panel of 34 validation chains ordered by original plane count, not a representative test sample. Training and validation patients are disjoint. Qwen3.8-27B (Qwen Team, 2026), complete image planes, fixed $1.5 \times 1.5 \times 3$ mm spacing, neutral location cues, chronological view binding, and thinking-off setting remain unchanged. Image-conditioned inputs include images; image-omitted inputs remove them. Teacher forcing inserts fixed answer text instead of generating it. Before-generation extraction contains no reference answer or teacher-forced intermediate response. Image-omitted inputs retain the original question, geometry, and metadata context; they are not context-free controls.

We extract states from the visual encoder, visual-to-language projector, language input, every language block, and final normalization. For numerical probes, we average the hidden states at the corresponding organ view's image-token positions. View binding uses each field's view identifier and the physical chronological scan, rather than interpreting the field-local role `current` as global chronology. Categorical curves reported here use the final prompt-token state. The extractor also saves a prompt-tail mean that includes template/suffix tokens; it is not a pure question-span mean. Because image tokens precede the question in the causal decoder, their states are not question-conditioned reasoning features.

A supervised probe is a classifier or regressor trained on frozen model activations. Categorical probes use logistic regression with fixed $C = 1$, and numerical probes use ridge regression with fixed $\alpha = 10$. Normalization is fitted exclusively on training features. The equal-dimensional analysis uses train-standardized, train-fitted principal component analysis (PCA) with 16 components, full singular value decomposition (SVD), and no whitening. Constant layer-input vectors are baseline-only rather than assigned artificial PCA components. Hyperparameters are not selected on validation performance. Native-dimensional probes use all components of the extracted vectors; PCA-16 probes use 16 principal components. Common validation chains and source-deduplicated numerical targets are retained across stages, and no best layer is selected for a held-out performance claim. A fixed training-label permutation contextualizes probe accuracy; a single permutation is a diagnostic, not a calibrated significance test.

G.2 Probe training set and generated-answer comparison

Expanding only the probe training set changes the last-block PCA image-conditioned/image-omitted accuracies from 23.53%/41.18% to 58.82%/38.24% on identical validation features. The expanded native-dimensional comparison is 47.06%/44.12%. This reversal demonstrates sensitivity to the probe training set, including label coverage, rather than a change to the frozen model. Against image-conditioned generation, the native-dimensional probe gap is 11.76 points (95% interval: -15.63 to 36.37) and the PCA-16 gap is 23.53 points (-6.07 to 51.35); neither is resolved. Paired outcomes for the PCA-16 probe are 14 probe-correct/generation-incorrect, 6 generation-correct/probe-incorrect, 6 both correct, and 8 both incorrect. The generated answers, by contrast, are scored without a trained probe using the unchanged typed parser and saved conclusion field. No content is recovered by searching the reasoning text for a more favorable answer.

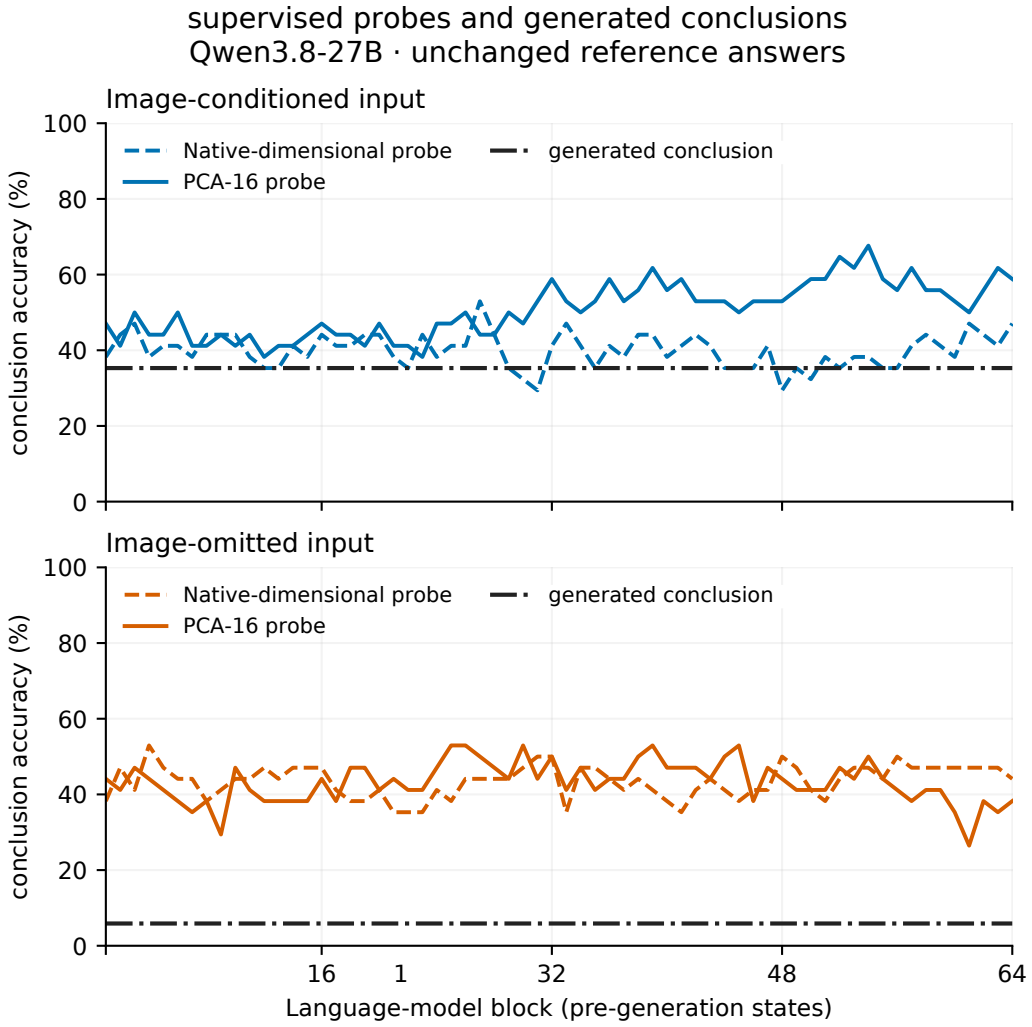

Figure 4: **Supervised probe accuracy and generated-conclusion accuracy are different measurements.** Curves show conclusion accuracy of fixed native-dimensional and PCA-16 probes at each pre-generation language stage, against the dataset's reference answers. Horizontal lines show the actual generated-conclusion accuracy for each input condition; they are not layer-specific generation results. All language blocks are displayed. The probe–generation contrast differs in supervision and available answer labels and does not identify information loss.

Image-conditioned generation favors `STABLE` (64.71% of conclusions), whereas image-omitted generation favors `UNMATCHED` (82.35%). The latter is a valid option indicating that no prior match is established and newness remains uncertain, not a parser failure or generic refusal. The generator retains the eight original allowed options, including options absent from probe-training labels. We do not remove these outputs from scoring. All conclusion labels are valid and all sequences reach the end-of-sequence (EOS) token without hitting the generation limit.

Numerical comparisons require matching both source identity and availability. On common sources with valid generated measurements, image-conditioned diameter mean absolute error (MAE) is 19.27 mm for generation and 19.00 mm for the PCA probe; the probe-minus-generation difference is -0.28 mm (95% interval: -4.27 to 3.37). Volume MAE is 20.18 mL and 20.74 mL, respectively, a difference of 0.56 mL (-3.68 to 4.70). Neither comparison resolves a probe advantage. Full-source probe errors are not substituted into these availability-matched comparisons. Image-omitted

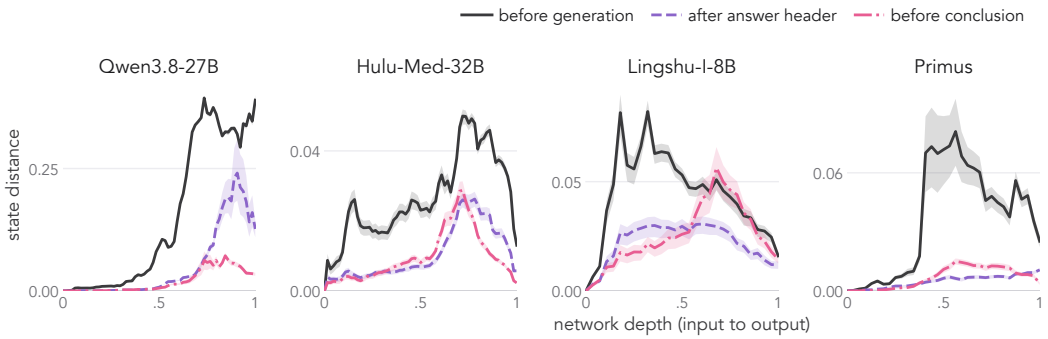

Figure 5: **State distance also drops with inserted reference intermediate answers.** Same measurement as Figure 3a, but reference intermediate answers are inserted for every model. At the last layer, the value right before the conclusion is 9% of the value before writing for Qwen3.8-27B, 17% for Hulu-Med-32B, 14% for Primus and 91% for Lingshu-I-8B. Bands give 95% intervals from bootstrap resampling of patients.

generation gives no diameter or volume values, so its corresponding MAE is undefined, not zero. Whole-view pooling may obscure localized evidence, but no cue-local recovery result is claimed.

G.3 Prefix conditioning and uncertainty

Each replay appends an exact saved token prefix generated with image-conditioned input to the original prompt with or without images. The answer generated separately with image-omitted input is never substituted. Boundaries before the first intermediate answer contain no emitted field values; boundaries before the conclusion contain the preceding intermediate fields but not the conclusion value. Boundary selection follows the structured response and does not consult reference targets. Fresh cache-free forwards measure the last input-token state before any candidate continuation.

For hidden states h_ℓ^I with images and h_ℓ^T without images at layer ℓ, we report cosine distance $d_{\cos}$ and relative Euclidean distance d_{rel}:

$$d_{\cos} = 1 - \frac{\langle h_\ell^I, h_\ell^T \rangle}{\|h_\ell^I\| \, \|h_\ell^T\|}, \qquad d_{\text{rel}} = \frac{2\|h_\ell^I - h_\ell^T\|}{\|h_\ell^I\| + \|h_\ell^T\|}. \tag{1}$$

All plotted distances are finite and use common chains across the three answer positions. Means are chain-weighted. Pointwise percentile intervals and paired endpoint changes use 2,000 patient-cluster bootstrap resamples with seed 20260923. These intervals hold extracted features and fitted predictions fixed; they do not include probe-refitting uncertainty or simultaneous coverage across layers. Image omission changes token positions, and advancing an answer boundary changes both the last token and its context. The fixed-layout interventions address different aspects of those confounds without converting the prefix curves into a causal information-loss measure.

G.4 Intervention construction and option scoring

The visual intervention selects the first equal-depth pair in the frozen order and exchanges complete processed organ views, preserving all original planes and plane order within each organ view. Original location cues, dates, prompt tokens, model-generated prefix tokens, and structural layout remain fixed. This tests view assignment with fixed context, without anatomical registration or clinical reversal.

The value intervention selects the first feasible chain in that order using only the saved prefix and tokenization. It exchanges earlier/current diameter and volume together, using model-generated positive finite scalar values. Corresponding scalar spans must have equal token lengths and exact boundaries. Original donor token IDs are copied without retokenization, padding, truncation, or reference substitution. All non-value tokens, the final prefix token, actual position IDs, images, cues, dates, and elapsed-time fields are unchanged. Applying the exchange twice restores the original prefix exactly. The intervention changes temporal attribution of written values, not syntax alone.

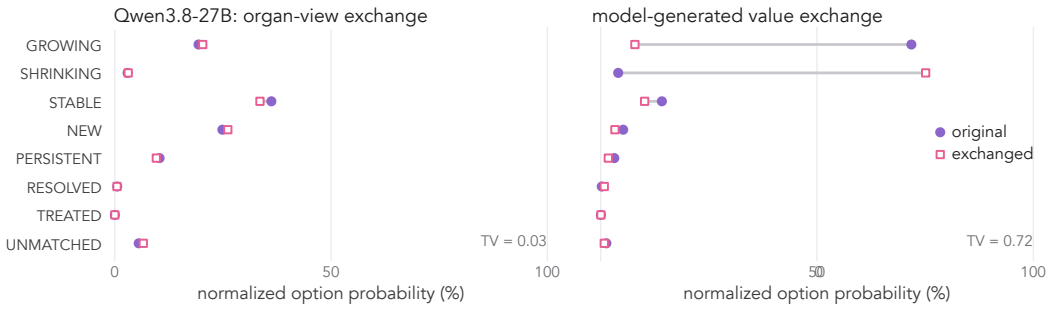

Figure 6: **The conclusion of Qwen3.8-27B follows its written measurements, not the order of the images.** Normalized probability of each of the eight conclusion options in one chain each, before (filled circles) and after (open squares) an exchange with all token positions fixed. Left: exchanging the earlier and current organ views barely moves the probabilities (total variation 0.03). Right: exchanging the model-generated earlier and current measurements reverses its preferred conclusion from GROWING to SHRINKING (total variation 0.72). TV denotes total variation between the option distributions.

Let x denote the fixed input, $a_{<\text{final}}$ the saved answer prefix, and $\mathcal{Y}$ the original allowed options. For each quoted JSON option y, its score is

$$s(y) = \sum_{t=1}^{|y|} \log p(y_t \mid x, a_{<\text{final}}, y_{<t}), \qquad q(y) = \frac{\exp s(y)}{\sum_{y' \in \mathcal{Y}} \exp s(y')}. \tag{2}$$

There is no length normalization. Total variation (TV) is $\frac{1}{2}\sum_{y\in\mathcal{Y}} |q_{\text{original}}(y) - q_{\text{exchange}}(y)|$. Complete-option probability mass is 99.512%/99.426% for original/exchanged organ views and 81.457%/92.002% for original/exchanged written values. Reporting this mass distinguishes restricted-option redistribution from calibrated confidence over all possible continuations. Next-token probabilities at a prefix boundary may mainly reflect syntax; we therefore score complete options.

Original-condition states and logits reproduce the saved prefix runs exactly. Actual hooked/unhooked logits and candidate first-position logits agree exactly in both conditions. Saved visual stages reproduce the expected view reversal under visual exchange and remain identical under value exchange; all stage features and candidate logits are finite. Independent likelihood recomputation agrees to floating-point precision. These execution controls establish the measured perturbations, not correctness of a counterfactual clinical answer. The two interventions use different chains and admit no population confidence interval or factorial effect-size comparison.

G.5 Teacher-forced reference and swapped intermediate values

This experiment is counterfactual conditioning. Reference values are inserted into the prefix on purpose, so its outcomes are never generated-answer or benchmark accuracy. The design, exclusion criteria, perturbation and analysis were frozen before any forward pass. It uses the publicly available Qwen3.8-27B with the inputs of the development panel: complete planes, $1.5 \times 1.5 \times 3$ mm spacing, chronological view binding, neutral location cues, and thinking off. Every prefix ends at the model's own boundary just before the conclusion value.

Conditions. C0 removes every intermediate member from the model-generated answer prefix, cutting at the model's own whitespace boundaries. C1 is the saved model-generated answer prefix. C2 replaces each intermediate scalar with its reference value. C3 applies the pre-registered perturbation to C2: for finding size, the earlier and current reference diameter and volume are swapped together, with chronology taken from the physical view binding; elapsed-time values are left unchanged because they imply no direction. C4 is the C2 prefix without images. Substitution touches only exact scalar value-token spans, and every other token of the model-generated answer is copied and checked token by token. Reference values are rendered as shortest round-trip JSON numbers without rounding.

Table 13: **Swapping reference sizes lowers reference-option agreement by 37.0 percentage points.** We compare five conditions that insert or remove intermediate answers; these are counterfactual option scores, not benchmark accuracy. Agreement abbreviates reference-option agreement: the share of chains whose highest-scoring conclusion option equals the reference answer. Reference-option probability: mean probability assigned to the reference option after normalization over the eight allowed options. Values are percentages or percentage points; parentheses give 95% bootstrap intervals over patients.

condition or contrast	reference-option agreement	reference-option probability
C0 no intermediates	7.4 (0.0, 16.7)	8.4 (4.5, 12.8)
C1 model-generated values	33.3 (15.2, 54.2)	31.3 (15.1, 49.6)
C2 reference values	48.1 (29.2, 65.4)	40.4 (27.2, 56.0)
C3 swapped reference values	11.1 (0.0, 23.3)	14.4 (7.6, 21.8)
C4 reference values, no images	44.4 (27.6, 63.0)	35.0 (22.6, 50.9)
C2 − C1	14.8 (−15.4, 40.6)	9.1 (−12.3, 28.1)
C1 − C3	22.2 (4.0, 43.5)	16.9 (1.0, 35.8)
C2 − C3	37.0 (17.4, 54.5)	26.0 (13.8, 40.8)
C2 − C0	40.7 (21.4, 59.1)	32.0 (18.4, 47.5)
C2 − C4	3.7 (−9.5, 15.0)	5.5 (−0.8, 12.1)
Exploratory: growing or shrinking reference answer		
C1 model-generated values	21.1 (4.5, 46.7)	19.6 (5.9, 39.6)
C2 reference values	57.9 (33.3, 80.0)	48.1 (32.0, 66.5)
C2 − C1	36.8 (0.0, 66.7)	28.5 (5.3, 48.5)
C2 − C3	52.6 (27.8, 73.7)	37.4 (21.5, 57.6)

Chains. Of the 34 chains in the frozen development panel, 7 are excluded because their intermediates are region-screening or description attributes, for which no perturbation was pre-registered. In 6 of the remaining 27 chains the model wrote `null` for both measurements, and the tokenizer merges the preceding space into that token. A declared extension treats the whole token as the value span and substitutes the space plus the value. Without it, 5 of 14 growing chains would be dropped, and exactly those whose model-generated intermediate answers are missing. All 27 chains from 18 patients are scored under all five conditions, with no failures and no further exclusions. Reference answers are 14 growing, 5 stable, 5 shrinking, 2 treated and 1 resolved.

Scoring and controls. The eight original conclusion options are scored by summed complete-option log-likelihood and normalized over the options, exactly as in the value intervention. Mean complete-option probability mass is 74.8%, 92.9%, 92.4%, 92.9% and 93.8% for C0–C4. All 135 forwards passed these checks:

- hooked and unhooked logits agree exactly, and candidate first-position parity holds with total variation 0;
- all stage features, logits and option scores are finite;
- visual-stage features are bitwise identical across the four image conditions;
- the C1 prefix-only logits reproduce the earlier replay of the model-generated answer prefix bitwise in every chain;
- on the value-intervention chain, the C1 option probabilities reproduce that run with a maximum difference of 0.

Intervals are 95% percentile intervals from 10,000 patient-cluster bootstrap resamples (seed 20260924).

Model-generated intermediate-answer correctness. Model-generated values are scored with the unchanged numerical scorer and the current tolerances: diameter within $\max(2\,\mathrm{mm}, 50\%)$ of the reference, volume within 25%, days exact, and months within 0.05. A null model-generated value counts as incorrect. Only 1 of the 27 chains has all model-generated intermediate answers,

and all model-generated measurements, within tolerance. The planned stratification of C1 by model-generated intermediate-answer correctness is therefore uninformative. In the other 26 chains (18 patients), reference-option agreement is 30.8% with model-generated values (12.1–52.4), 46.2% with reference values (28.0–63.2), and 7.7% with swapped values (0.0–19.0). The paired differences are $+15.4$ points for C2 $-$ C1 (-15.8 to 42.9) and $+23.1$ for C1 $-$ C3 (4.2 to 45.5).

Limits. The reference answers for conclusions and intermediate questions derive from the same source reports. Agreement therefore partly measures whether the conclusion is consistent with the written values, not clinical correctness. The swap is one error type, and it reverses the implied direction of change only for size measurements. C0 leaves an unusual prefix, and its complete-option probability mass is lower. The small, single-model panel gives a wide C2 $-$ C1 interval.

G.6 Replication on other open-weight models

Models and inputs. Hulu-Med-32B (Jiang et al., 2025), Lingshu-I-8B (Xu et al., 2025), and Primus (Myronenko et al., 2026) are run on the same frozen panel of 34 validation chains, in the same order, with each model's publicly available weights and the input construction of its own benchmark runner: complete organ views in 3D through the model's three-dimensional processor for Hulu-Med-32B, all axial planes rendered as 448-pixel images for Lingshu-I-8B, and organ views on the native 192^3 grid with air padding for Primus. For Hulu-Med-32B, the two views of a chain are air-padded to a common in-plane size so that both receive the same processed millimetres per patch, as its benchmark check requires; no plane is cropped, resized or dropped. The benchmark answer reserve and context limits are kept (Hulu-Med-32B: 32,768 positions and a 4,096-token reserve; Lingshu-I-8B: 65,536 positions with its default dynamic rotary scaling; Primus: 8,192-token reserve, thinking off), as are the neutral location cues rendered for each model's presentation. Hulu-Med-32B ran on one H200 GPU; the other two ran on L40S GPUs.

Model-generated answers. Each model first generated its own image-conditioned answer with greedy decoding. Primus uses w/o thinking; Hulu-Med-32B and Lingshu-I-8B use default settings. None of these answers yields the exact model-generated answer boundaries used for Qwen3.8-27B. Hulu-Med-32B wrote a valid, compact answer object in all 34 chains, but its tokenizer merges the delimiters with the neighbouring keys, so every boundary falls inside a token (all 34 chains, all boundaries). The other two did not follow the requested answer object: Lingshu-I-8B repeated the given geometry context in 32 of 34 answers, and Primus returned text that is not a JSON object in all 21 chains it could process. Primus cannot process 13 chains whose organ views exceed its native 192^3 grid at the fixed spacing; tiling was not used here. The analysis of model-generated answer prefixes and the "model-generated values" condition are therefore not defined for these three models.

Teacher-forced reference prefixes. For the 27 chains of the reference-prefix experiment, we instead use the frozen prefix texts of that experiment, which were built before any of these runs: the answer up to the conclusion field with no intermediates, with reference values, and with swapped reference values. Each text is re-tokenized with the model's own tokenizer and must decode exactly, and the eight conclusion options must tokenize identically after every prefix. The same texts give the three answer positions of the image-conditioned/image-omitted comparison: no answer prefix, the answer header before the first intermediate key, and the reference-filled answer prefix up to the conclusion field. The hidden state is the output of the last language block at the last input position, as for Qwen3.8-27B. For Qwen3.8-27B, the matching values on these 27 chains are computed from saved states, without new forwards: 0.390, 0.127 and 0.034, against 0.039 with its model-generated values before the conclusion field.

Controls and exclusions. Every forward is fresh and cache-free, and all logits and states are finite. Before any option is scored, the first-position distribution of each option forward must reproduce the prefix forward (same top token, total variation at most 10^{-3}), as for Qwen3.8-27B; a chain that fails is excluded from option scoring without a relaxed retry, but its prefix states are kept. For Lingshu-I-8B, 21 of the 27 chains fail this check. Its default dynamic rotary scaling recomputes the position frequencies when a sequence exceeds 32,768 tokens, so the longer option forward changes earlier positions; 6 chains from 6 patients are scored. The same scaling applies to image-conditioned inputs above that length but not to their shorter image-omitted counterparts, which adds to the layout

Table 14: **Three other open-weight models show little sensitivity to swapped written sizes.** We compare four models with inserted reference answers. Distances: mean cosine distance between the last-block states with and without images at the three prefixes. Agreement abbreviates reference-option agreement: the share of chains whose highest-scoring conclusion option equals the reference answer (counterfactual, not benchmark accuracy). Complete-option probability mass: mean complete-option probability of the eight options, in percent. Values in parentheses are 95% patient-cluster bootstrap intervals. Qwen3.8-27B distances are computed from saved states on the same chains.

	Qwen3.8-27B	Hulu-Med-32B	Lingshu-I-8B	Primus
chains (patients), distances	27 (18)	27 (18)	27 (18)	18 (15)
before generation	0.390 (0.377, 0.400)	0.0130 (0.0126, 0.0134)	0.016 (0.015, 0.017)	0.025 (0.022, 0.028)
after the answer header	0.127 (0.109, 0.145)	0.0057 (0.0053, 0.0061)	0.012 (0.010, 0.013)	0.011 (0.010, 0.012)
before the conclusion, reference values	0.034 (0.031, 0.037)	0.0023 (0.0020, 0.0025)	0.015 (0.013, 0.016)	0.004 (0.003, 0.004)
before the conclusion − before generation	−0.356 (−0.366, −0.345)	−0.0108 (−0.0113, −0.0103)	−0.001 (−0.003, 0.000)	−0.022 (−0.025, −0.019)
chains (patients), options	27 (18)	27 (18)	6 (6)	17 (14)
agreement, no intermediates	7.4 (0.0, 16.7)	48.1 (25.0, 69.2)	0.0 (0.0, 0.0)	11.8 (0.0, 31.2)
agreement, reference values	48.1 (29.2, 65.4)	51.9 (29.0, 73.1)	0.0 (0.0, 0.0)	11.8 (0.0, 28.6)
agreement, swapped values	11.1 (0.0, 23.3)	51.9 (29.0, 73.1)	0.0 (0.0, 0.0)	17.6 (0.0, 37.5)
agreement, reference, no images	44.4 (27.6, 63.0)	51.9 (29.0, 73.1)	0.0 (0.0, 0.0)	11.8 (0.0, 28.6)
option mass, reference values	92.4	0.01	0.04	0.6

difference between the two conditions. For Primus, 9 of the 27 chains exceed its grid and 1 fails the check, leaving 18 chains from 15 patients for the distances and 17 chains from 14 patients for option scoring. Hulu-Med-32B and Lingshu-I-8B contribute all 27 chains (18 patients) to the distances; for Hulu-Med-32B all 27 chains also pass the check and are scored (maximum first-position total variation 0). Intervals use the same patient-cluster bootstrap as for Qwen3.8-27B (2,000 resamples for distances, 10,000 for options).

Depth profiles. For Figure 3a, the image-conditioned and image-omitted forwards of the three answer positions are repeated with the same chains, prefixes and inputs, saving the last-position state of every language block and of the input embedding; intervals use the same patient-cluster bootstrap. The last-block values reproduce the single-readout runs exactly (maximum absolute difference 0, same chain counts). For Qwen3.8-27B with the reference-filled answer prefix, the per-layer states were already stored by the earlier runs and are read from them, again reproducing the published last-layer values exactly. For the model-generated answers of Hulu-Med-32B, whose tokenizer merges each field delimiter with the neighbouring key, every cut point is the longest token prefix that ends at or before the exact character boundary. In all 34 chains this is exactly one token short: the omitted text is only the closing quotation mark and colon of the preceding key (plus an opening brace for the opening text), never a written value. The states before generation reproduce the saved run with model-generated answers exactly.

Results. Table 14 lists the state distances with teacher-forced reference prefixes (Appendix Figure 5) and the corresponding option scores in Figure 3b. Hulu-Med-32B and Primus show the decrease seen in Qwen3.8-27B at a smaller scale; Lingshu-I-8B shows small state distances at the last layer for any prefix. No other model's conclusion follows the written values: Hulu-Med-32B scores `GROWING` highest in every chain once any intermediate is present (reference, swapped, or without images; 25 of 27 chains without intermediates), Lingshu-I-8B scores `NEW` highest in every scored chain and condition, and Primus agrees with the reference answer about equally often under reference, swapped and removed values. For all three, the eight quoted options receive about 1% or less of the complete continuation probability (Hulu-Med-32B at most 0.03%, Lingshu-I-8B at most 0.05%, Primus 0.6–1.4%, against 74.8–93.8% for Qwen3.8-27B), so their option scores describe a format they rarely produce.

G.7 Supervised probes before generation

Setup. We ask whether a supervised probe can recover the reference answer from the model's hidden state before it writes any answer. The probe set contains every training and validation linked-question chain whose endpoint is the Phase 3 temporal-change label: 294 chains from 176 patients (247 training, 47 validation); the test set is not used. Reference answers are GROWING (111), STABLE (92), SHRINKING (53), NEW (29), TREATED (7) and RESOLVED (2). For each chain

we run the model once with the images and the question and take the output of every language block at the last input position, before any answer token. A multinomial logistic regression (fixed $C = 1$, standardization fitted on training folds only) is trained with 5-fold patient-grouped cross-validation, and every chain is scored by the fold that did not see its patient. The last language block gives the probe input and is fixed before any result is seen; no layer or hyperparameter is selected on held-out chains. The generated conclusion is obtained by greedy decoding on the same linked-question chain, scored with the same content-based scorer as the benchmark; an unresolved generated conclusion counts as incorrect. Chance is a 200-permutation null with labels shuffled and the probe refitted. Intervals and the paired probe−generation difference use 2,000 patient-cluster bootstrap resamples (seed 20260923). Lingshu-I-8B excludes 44 chains whose input exceeds its 65,536-token context, and Primus excludes 126 chains whose organ views exceed its 192^3 grid.

Results. For Qwen3.8-27B and Hulu-Med-32B the supervised probe predicts the reference answer more often than the model generates it: 50.7% against 41.5% for Qwen3.8-27B (difference $+9.2$ percentage points, 95% interval $+1.5$ to $+16.8$) and 46.6% against 37.1% for Hulu-Med-32B ($+9.5$, $+1.9$ to $+17.6$), both above the permutation null (97.5th percentile 36.1%, $p = 0.005$). Both models generated a resolved conclusion in all 294 chains, and their generated conclusions concentrate on one label (Qwen3.8-27B STABLE in 194 chains, Hulu-Med-32B GROWING in 220), while the probe separates the labels. The probe is correct where the generated conclusion is wrong in 81 chains for Qwen3.8-27B and 77 for Hulu-Med-32B, and the reverse holds in 54 and 49 chains. For Lingshu-I-8B (56.4% probe, 4.8% generated; 250 chains) and Primus (40.5% and 9.5%; 168 chains) the gap mainly reflects generated conclusions that cannot be resolved (229 of 250 and 104 of 168); among resolved generated conclusions, 57.1% of 21 and 25.0% of 64 are correct. The hidden state contains information about the conclusion before writing, but the written chain does not preserve it.