

Contents of Technical Appendices and Supplementary Material

A	Limitations and Future Work.	23
B	LLM Scoring Consistency Evaluation Details.	23
B.1	Model Selection	23
B.2	Evaluation Metrics.	23
C	Evaluation Procedure.	24
D	Human Annotation Protocol.	24
D.1	Annotator Background	24
D.2	Annotation Procedure.	24
E	Prompts and Templates.	25
E.1	Question Construction Prompts and Templates.	25
E.2	LLM-Based Scoring Prompt.	25
F	Original Dataset.	25
G	QA Construction Case Visualization.	26
H	Detailed Analysis of Experimental Results.	26
H.1	Finetuned Results of M3D-4B Models.	26
H.2	Finetuned Subtask Performance.	27
H.3	Zero-shot Subtask Performance.	27
H.4	Result Analysis	27
I	Failure Case Analysis	28
J	Word Cloud Visualizations across Tasks.	29
K	Additional Evaluations on General VLMs	29
L	Task-level Ablation Studies.	29

 **Code:** <https://github.com/Tang-xiaoxiao/3D-RAD>

 **Dataset:** <https://huggingface.co/datasets/Tang-xiaoxiao/3D-RAD>

A Limitations and Future Work.

Our Longitudinal Temporal Diagnosis task provides sequence information primarily from a diagnostic label perspective, which captures only one aspect of temporal evolution. However, richer temporal cues—such as spatial and morphological changes observable across full multi-phase 3D scans—remain underutilized. Current model architectures also do not support joint input of multiple 3D volumes across time, limiting comprehensive temporal reasoning. Furthermore, we have not yet introduced open-ended question formats for this task, which could enable deeper and more diverse clinical insights. In future work, we plan to incorporate full-sequence 3D inputs and develop open-ended question generation strategies to better capture longitudinal progression in medical imaging.

B LLM Scoring Consistency Evaluation Details.

We assess the consistency of scoring across four large language models (LLMs). The left panel in [Figure 7](#), and [Figure 8](#) shows the average pairwise agreement scores for each model relative to others, while the right panel presents their rankings based on overall agreement. Notably, GPT-4o-mini achieves the highest consistency, aligning closely with other models, particularly in the high-score range. Based on this result, we adopt GPT-4o as the default evaluator in our scoring pipeline.

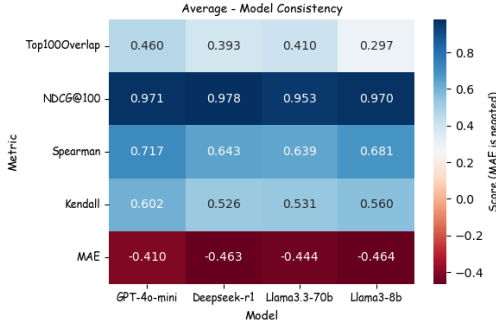


Figure 7: Consistency Heatmap

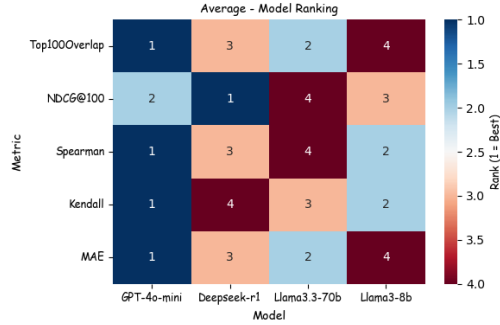


Figure 8: Ranking Heatmap

B.1 Model Selection

To comprehensively evaluate the consistency of scoring across different large language models (LLMs), we select a diverse set of state-of-the-art models varying in architecture, scale, and training methodology:

GPT-4o-mini: A compact version of OpenAI’s GPT-4o, optimized for fast inference with minimal performance trade-off.

DeepSeek-r1: A high-performing open-source LLM tailored for multilingual and multi-domain tasks, emphasizing reliability across medical and scientific content.

LLaMA3-70B: Meta’s latest flagship 70B parameter model trained on a large corpus with improved instruction-following and reasoning capabilities.

LLaMA3-8B: A smaller variant of LLaMA3 designed for cost-efficient deployment while preserving alignment characteristics.

B.2 Evaluation Metrics

To quantify inter-model scoring consistency and ranking agreement, we adopt five metrics widely used in information retrieval and statistical correlation analysis:

Top100Overlap: The number of overlapping QA pairs in the top-100 ranked results between model pairs, measuring agreement on high-quality samples.

NDCG@100 (Normalized Discounted Cumulative Gain): Captures both the ranking position and relevance of QA samples up to the top 100, reflecting graded relevance alignment.

Spearman’s Rank Correlation Coefficient (ρ): Measures the monotonic relationship between two sets of rankings, invariant to scale.

Kendall’s Tau (τ): Evaluates pairwise ranking agreement, more robust to small ranking perturbations than Spearman’s ρ .

MAE (Mean Absolute Error): Computes the average absolute difference in numeric scores between model pairs, capturing score-level discrepancy.

C Evaluation Procedure.

To assess scoring consistency across models, we sample a total of 600 QA pairs from the dataset, proportionally drawn from each of the six tasks. All four models are instructed to rate these samples independently using the same standardized scoring template to ensure fairness. Each model provides five-dimensional scores per QA pair, from which we compute the average score and generate a ranked list of the 600 samples.

We then perform pairwise comparisons across all models using the metrics defined in Section B.2 For each model, we compute its average metric value against the other three models. [Figure 7](#) illustrates the average metric values of each model relative to the others. [Figure 8](#) shows the ranking of each model based on these averaged metrics.

Based on this analysis, we choose the **GPT-4o-mini** model for final QA scoring. It consistently demonstrates the highest agreement with the other large-scale models in terms of ranking and high-score overlap. Since our final QA set is selected based on high-scoring samples, we argue that using a model with high consistency ensures that selected QAs would also be rated highly by other LLMs, thereby preserving the objectivity and robustness of our scoring pipeline.

D Human Annotation Protocol.

D.1 Annotator Background

We recruited eight graduate students with medical domain knowledge to perform manual quality assessments. Each annotator evaluated 75 QA pairs, with an average annotation time of 1.5–2 hours. To ensure consistency with the automated evaluation, all annotations followed the same scoring rubric used by LLMs. In addition, to mitigate subjectivity and provide clearer guidance, two detailed scoring examples were supplied to each annotator.

D.2 Annotation Procedure

Annotators were provided with a comprehensive scoring manual, two reference-scored examples, and a set of sampled QA pairs along with their corresponding clinical reports (see [Figure 10](#)). Each QA pair was scored across five dimensions on a 0–5 scale: **Visual Verifiability**; **Specificity & Clarity**; **Answer Appropriateness**; **Q-A Alignment**; **Linguistic Quality**.

In addition, annotators marked whether the QA pair was consistent with the original clinical report (**binary consistency label: 1 for consistent, 0 for inconsistent**), reflecting the factual accuracy and absence of hallucination.

To ensure reliability, we first identified 53 QA pairs flagged as inconsistent (consistency score = 0). We then computed the average of the five-dimensional scores for each flagged pair and excluded any QA pair that (i) had an average score below 3 or (ii) received a score below 3 in any individual dimension. This process resulted in the retention of only 23 low-quality samples. Consequently, the final dataset achieved a high factual accuracy rate of **96.17%**.

Given the large scale of our dataset (170K QA pairs), we believe our sampling and validation protocol offers a cost-effective yet robust quality assurance mechanism, ensuring the reliability of the benchmark under practical constraints.

Table 5: Results across General VLMs on 3D-RAD Benchmark.

Model	Task1 (ROUGE1 / BERTScore)	Task2 (ROUGE1 / BERTScore)	Task3 (ROUGE1 / BERTScore)	Task4 (ACC)	Task5 (ACC)	Task6 (ACC)
llava-onevision(7b)	22.34 / 86.24	26.30 / 86.79	5.93 / 91.81	29.67	6.86	6.86
qwen2.5-vl(3B)	24.40 / 84.96	22.37 / 83.77	18.21 / 92.31	22.75	24.09	24.09
qwen2.5-vl(32B)	21.77 / 87.66	20.30 / 87.64	11.46 / 90.73	58.60	28.62	28.36
gemma-3(4b)	22.91 / 87.44	22.50 / 87.41	8.56 / 90.50	21.47	10.75	10.74
gemma-3(27b)	26.21 / 88.15	30.43 / 88.89	10.55 / 90.54	28.57	24.09	24.09
gemini-2.0-flash	24.70 / 88.93	27.93 / 89.29	0.53 / 84.36	40.42	23.08	21.33
intern-vl-3(8B)	28.44 / 88.83	34.31 / 89.63	10.65 / 91.64	55.30	26.63	26.61
gpt-4.1-nano	10.99 / 85.00	11.41 / 85.32	4.31 / 83.63	37.95	19.95	25.11
gpt-4.1	15.08 / 86.03	16.55 / 86.50	2.76 / 82.82	61.98	34.08	68.05
Ours-Finetuned (Llama2-7B)	33.76 / 89.16	39.12 / 90.00	36.06 / 94.65	81.09	51.20	74.78
Ours-Finetuned (Phi3-4B)	42.45 / 90.72	50.52 / 92.19	36.46 / 94.86	82.43	49.30	74.77

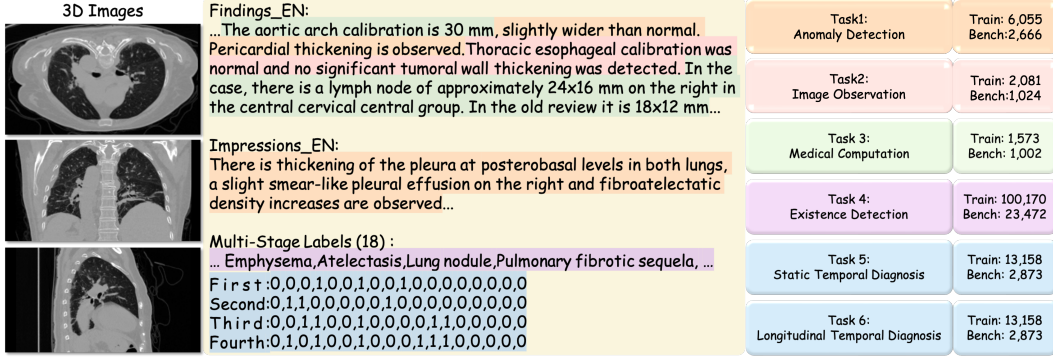


Figure 9: An example of QA Construction

E Prompts and Templates.

E.1 Question Construction Prompts and Templates.

Figure 11, Figure 12, Figure 13, Figure 14, Figure 15, and Figure 16 illustrate the complete set of carefully designed templates used for question construction across different tasks. Each template is tailored to the unique characteristics and objectives of the corresponding task, enabling accurate and diverse generation of clinically meaningful question-answer pairs.

E.2 LLM-Based Scoring Prompt.

Figure 17 illustrates the full prompt template used to evaluate each QA pair across five key dimensions. This prompt guides large language models (LLMs) to assign quality scores, ensuring consistency and comprehensiveness in the evaluation process.

F Original Dataset.

Medical datasets are inherently limited and highly valuable. As illustrated in Figure 9, we present the core components of the original dataset and how we systematically leveraged this information to construct our multi-task, information-rich, high-quality VQA benchmark.

On the left, we show a representative 3D CT volume in .nii.gz format. The central panel contains two key types of information: **clinical text** and structured **labels**. We primarily utilized the *Findings* and *Impression* sections of the clinical text for constructing open-ended VQA pairs (Tasks 1–3). The label annotations include binary indicators for the presence of 18 diagnostic categories per scan; in cases with multiple follow-up scans for the same patient, longitudinal label comparisons are available (shown in four-scan progression examples).

Based on these sources:

- **Tasks 1–3** use clinical text and image data to generate open-ended QA pairs focusing on anomalies, anatomical structures, and measurements.
- **Task 4** uses the image-label mapping to create closed QA pairs (Yes/No) for 18 binary categories.

- **Tasks 5–6** introduce temporal reasoning. After excluding the “medical material” label (e.g., stents, which are not typically lesions), we perform longitudinal comparisons across scans on the remaining 17 labels. Each label is categorized based on its temporal evolution, with corresponding clinical implications, as follows:
 - **Refractory Lesion:** Previously 1 or fluctuating, now 1. Indicates a persistent or recurrent abnormality requiring close monitoring or intensified treatment.
 - **Resolved Lesion:** Previously 1 or fluctuating, now 0. Suggests effective resolution of the lesion, though continued surveillance may be needed to prevent recurrence.
 - **New Lesion:** Previously 0, now 1. Reflects newly emerged pathology, often signaling potential disease progression or relapse, thus necessitating timely clinical attention.
 - **No Abnormality:** Consistently 0. Denotes stable absence of pathology, typically indicating a favorable prognosis with low clinical concern.

To ensure data consistency, we exclude image variants produced by post-processing techniques (e.g., denoising) and only retain raw, unprocessed scans.

G QA Construction Case Visualization.

Figure 9 illustrates an example of how various task-specific information is extracted from a single radiology report. This case highlights the process of transforming complex clinical descriptions into structured QA pairs across different tasks, demonstrating the richness and multi-faceted nature of the original medical content.

In **Figure 3**, we demonstrate representative QA construction processes across the six tasks. This illustration highlights the key distinctions among the six diagnostic tasks in terms of input modality (e.g., text, image, label sequences), question format (open vs. closed), and reasoning requirements (e.g., spatial, numerical, temporal). By contrasting these examples side-by-side, we emphasize the diverse cognitive challenges posed by each task, which collectively test different dimensions of medical VQA capabilities.

- **Task 1:** Focused on identifying and localizing abnormalities in the image.
- **Task 2** extends beyond abnormality detection by including anatomical and structural queries that may involve normal findings, such as “Where is the cardiac pacemaker catheter terminating?”, distinguishing it from Task 1 which focuses exclusively on abnormal observations.
- **Task 3:** Targets clinically relevant numerical values (e.g., diameters).
- **Task 4:** Binary classification of label existence (<Choices_list>: Yes/No) per diagnostic category.
- **Task 5:** Requires inference of current lesion status based solely on the current image, with a four-class <Choices_list>, without longitudinal context.
- **Task 6:** Involves multi-phase diagnosis using both image features and longitudinal label progression, with a four-class <Choices_list>: Refractory Lesion (persistent or recurrent, now present); Resolved Lesion (previously present or recurrent, now absent); New Lesion (absent previously, now present); No Abnormality (always absent).

This detailed visualization supports our data construction strategy and highlights the complexity and diversity of clinical VQA scenarios tackled in our benchmark.

H Detailed Analysis of Experimental Results.

In the main text, we present a summary of the overall experimental results (**Table 3** and **Table 4**). Here, we provide a more detailed analysis of these results.

H.1 Finetuned Results of M3D-4B Models.

Table 6 presents a comprehensive comparison of the finetuned results for the 4B model series across Tasks 1 to 6. The results demonstrate consistent performance improvements after domain-specific finetuning, particularly on temporally-aware tasks (Tasks 5 and 6), highlighting the critical role of tailored 3D medical data in enhancing multi-task and multi-temporal reasoning capabilities of vision-language models.

Table 6: Finetuned Results of 4B Series on Tasks 1–6

Task	Metric	Zero-shot	1%	10%	100%
Task1: Anomaly Detection	BLEU	15.06	21.10	30.54	33.28
	Rouge	23.19	26.03	37.53	42.45
	BERTScore	87.11	88.13	89.78	90.72
Task2: Image Observation	BLEU	16.31	20.54	29.35	39.66
	Rouge	23.19	24.63	38.25	50.52
	BERTScore	86.92	88.23	89.81	92.19
Task3: Medical Computation	BLEU	2.55	7.01	25.47	33.52
	Rouge	5.63	9.95	31.84	36.46
	BERTScore	85.74	85.97	93.06	94.86
Task4: Existence Detection	Accuracy	40.25	80.85	80.93	82.43
Task5: Static Temporal Diagnosis	Accuracy	25.40	41.17	48.11	49.30
Task6: Longitudinal Temporal Diagnosis	Accuracy	24.31	61.01	74.19	74.77

H.2 Finetuned Subtask Performance.

Figure 18, Figure 19, Figure 20, Figure 21, and Figure 22 illustrate the performance comparison across subtasks after fine-tuning. The results highlight consistent improvements in most subtasks, demonstrating the effectiveness of our domain-specific training data in enhancing the model’s task-specific capabilities. Notably, the most significant gains are observed in subtasks requiring temporal reasoning, suggesting that our dataset effectively supports learning nuanced clinical progressions.

H.3 Zero-shot Subtask Performance.

Figure 23, Figure 24, Figure 25, Figure 26, and Figure 27 illustrates the performance of models under the zero-shot setting across different subtasks. The results reveal substantial variation in accuracy, highlighting that some subtasks are more tractable in the absence of fine-tuning, while others remain highly challenging. This emphasizes the varying complexity and reasoning demands posed by each subtask within our benchmark.

H.4 Result Analysis

We evaluate several vision-language models (VLMs) on the M3D-RAD benchmark, a comprehensive suite of six carefully designed medical visual question answering (VQA) tasks. All tasks follow the same image-question-answering paradigm but differ in form and difficulty: Tasks 1–3 are open-ended generation tasks, Tasks 4–6 are closed-form classification tasks, and among them, Tasks 5 and 6 specifically evaluate temporal reasoning—Task 5 based on single-phase (static) inputs and Task 6 involving longitudinal multi-phase understanding.

Fine-tuned Performance. Supervision Boosts Temporal Understanding Fine-tuning significantly improves performance across all tasks, with Phi3-4B consistently outperforming LLaMA2-7B. In open-ended tasks like Anomaly Detection and Image Observation, BLEU and Rouge scores increase by 20–30 points, indicating improved alignment with radiology-style clinical descriptions.

The most dramatic gains occur in Tasks 5 and 6, which test the model’s ability to infer temporal lesion status. Phi3-4B achieves 49.30% accuracy on Static Temporal Diagnosis (Task 5) and 74.77% on Longitudinal Temporal Diagnosis (Task 6), outperforming LLaMA2-7B by more than 25%–50%. These results clearly show that temporal reasoning in 3D medical data benefits greatly from task-

specific supervision, yet also reveal that existing models still struggle with this type of complex inference.

Notably, Medical Computation (Task 3) remains challenging across the board. Even with fine-tuning, generative scores (BLEU/Rouge) remain low despite high BERTScore, pointing to persistent limitations in handling structured quantitative reasoning.

Zero-shot Performance. Descriptive VQA Transfers Well, Temporal Reasoning Does Not In the zero-shot setting, OmniV (Qwen2.5–1.5B) achieves the strongest performance on open-ended descriptive tasks (Tasks 1–2), demonstrating strong cross-modal generalization. In contrast, RadFM outperforms all others on temporal classification tasks, with 44.11% and 42.99% on Tasks 5 and 6, respectively. This suggests that domain-specific inductive bias remains crucial for temporally grounded reasoning tasks in medical imaging.

However, even the best-performing zero-shot models exhibit significant drops in Temporal Diagnosis tasks compared to their fine-tuned counterparts—indicating that temporal lesion understanding is not emergent in current VLMs, and must be explicitly taught.

Key Insights and Benchmark Value. Our benchmark reveals several critical insights into current model capabilities:

- All tasks benefit from fine-tuning, but the largest improvements occur in temporally grounded VQA (Tasks 5 and 6), where supervised adaptation yields +50% accuracy gains. This indicates that temporal lesion reasoning is learnable but not captured in pretraining.
- Descriptive tasks (Tasks 1–2) transfer better to zero-shot settings, while temporal and binary classification tasks (Tasks 4–6) require explicit supervision or domain priors to perform reliably.
- Medical Computation (Task 3) consistently exposes model limitations in numerical reasoning, motivating future work on inference-aware VQA models.
- Most importantly, our benchmark is the first to systematically expose these weaknesses across diverse VQA tasks, especially in multi-phase lesion diagnosis. By explicitly separating static and longitudinal reasoning, and grounding all tasks in real 3D CT scans with clinical language, M3D-RAD offers a fine-grained diagnostic lens into the limitations of current multimodal models—and a clear path forward for future method development.

I Failure Case Analysis

To better understand model limitations in temporal reasoning, we provide a detailed analysis of representative failure cases in [Figure 28](#). We focus on Tasks 5 and 6—*Static Temporal Diagnosis* and *Longitudinal Temporal Diagnosis*—and compare the performance of zero-shot and fine-tuned models.

Case 1: Both Task 5 and Task 6 failed in zero-shot, and fine-tuning only improved Task 6. In this case, the lesion had resolved, but both tasks were misclassified as *Refractory Lesion* by the zero-shot model. Fine-tuning improved longitudinal reasoning (Task 6), correctly identifying the lesion as *Resolved*, while Task 5 remained incorrect. This highlights the challenge of inferring temporal status from a single image without explicit sequential cues.

Case 2: Zero-shot failed on both tasks; fine-tuning successfully corrected both. This case demonstrates that fine-tuning effectively leverages subtle spatial features in Task 5 and temporal patterns in Task 6. The consistent improvement suggests the model benefits from domain-adapted learning to distinguish patterns like disappearance of subtle abnormalities.

Case 3: Task 5 failed in both conditions; Task 6 succeeded without fine-tuning. Here, the model incorrectly inferred a *No Abnormality* outcome in Task 5. Despite this, the zero-shot model correctly answered Task 6, likely due to explicit access to the label sequence. This underscores that Task 6 provides more external structure, making it more robust even without fine-tuning.

Case 4: Task 5 failed in zero-shot but succeeded after fine-tuning; Task 6 succeeded in both. This case illustrates the potential of fine-tuning to teach the model to recognize spatial indicators of stable lesions in a single image. The strong performance on Task 6 across both settings affirms that longitudinal cues (label sequences) serve as a stabilizing prior.

Table 7: Evaluation accuracy on Task4–Task6 under different finetuning settings.

Finetuned Task(s)	Evaluate Task4 (ACC)	Evaluate Task5 (ACC)	Evaluate Task6 (ACC)
Only Task1	82.03	38.78	72.61
Only Task2	81.88	48.73	73.55
Only Task3	81.94	48.30	72.78
Only Task4	81.96	38.97	73.52
Zero-Shot M3D (Llama2-7B)	18.00	25.47	24.17
Zero-Shot M3D (Phi3-4B)	40.25	25.40	24.31
All Tasks (Ours)	82.43	49.30	74.77

Table 8: Evaluation metrics on Task1 (BLEU / ROUGE1 / F1) under different finetuning settings.

Finetuned Task(s)	Evaluate Task1 (BLEU / ROUGE1 / F1)
Only Task5	31.60 / 42.01 / 90.48
Only Task6	32.45 / 42.87 / 90.73
Zero-Shot M3D (Llama2-7B)	9.10 / 18.64 / 86.07
Zero-Shot M3D (Phi3-4B)	15.06 / 23.19 / 87.11
All Tasks (Ours)	33.28 / 42.45 / 90.72

Summary. Our analysis reveals distinct challenges for Tasks 5 and 6: Task 5 requires spatial-temporal inference from a single snapshot, which proves difficult without targeted training; in contrast, Task 6 benefits from the presence of explicit sequential labels, making it more amenable to reasoning even in zero-shot. Fine-tuning significantly improves performance, particularly for Task 5, by embedding temporal priors into image understanding.

J Word Cloud Visualizations across Tasks.

To illustrate the diversity of our benchmark, we present word cloud visualizations for each individual task. These visualizations highlight the varying distributions of clinical concepts across tasks, reflecting the distinct focuses and linguistic patterns inherent to each QA setting. [Figure 29](#) shows the task-wise word clouds.

K Additional Evaluations on General VLMs

We also evaluate our benchmark on mainstream general-purpose vision–language models. As illustrated in [Table 5](#), though some models perform better—especially on tasks 4–6 with more reliable metrics, none achieve strong results across all tasks. Models fine-tuned on 3D-RAD consistently outperform general models, highlighting the dataset’s value in promoting domain-specific learning.

L Task-level Ablation Studies

We added task-level ablation studies to examine cross-task effectiveness in [Table 7](#) and [Table 8](#). The table below presents results where each task is fine-tuned individually and then evaluated on all other tasks. As shown, fine-tuning on a single task can still improve performance on other tasks. However, fine-tuning on the full dataset consistently yields the best results. This demonstrates that the effectiveness of our dataset stems not only from its size but also from the inclusion of high-quality domain knowledge.

Guideline

Please score the following radiology visual question and answer pair, using the 5 dimensions below. Each should be scored from 1 (very poor) to 5 (excellent). Be strict, specific, and consistent in your evaluation.

Question: {question}
Answer: {answer}

Scoring Dimensions:

1. *Visual Verifiability: Can this question be answered just by looking at the image, without requiring medical knowledge, inference, or external context? Does the answer also rely on the image for validation?*
 - 5: Can be answered purely from the image (e.g., "Is there a fracture?") and the answer is clearly image-based (e.g., "Yes, there is a fracture in the left femur.")
 - 1: Requires background knowledge, inference, or external context (e.g., "What type of tumor?") and the answer similarly relies on external context rather than the image.
 - If the answer requires knowledge of clinical context or interpretation beyond the image itself, it should be scored lower.
2. *Specificity & Clarity: Is the question precise and unambiguous? Is the answer also specific and clear, without ambiguity?*
 - 5: The question and answer are both specific and clear with one unambiguous interpretation (e.g., "Which lobe is involved in the lesion?" and the answer directly states the affected lobe).
 - 1: The question or the answer is vague, ambiguous, or open to multiple interpretations (e.g., "What is abnormal?" and the answer gives an unclear response like "There is something unusual.")
 - Avoid vague terms like "a few," "several," "some," "few millimeters," or similar imprecise quantities. If these terms appear in the answer, it should be scored lower.
 - Both the question and the answer must be precise. If either is unclear, reduce the score.
3. *Answer Appropriateness: Is the answer correct, medically appropriate, specific, and directly relevant to the question? Does it match the expected format/type?*
 - 5: Accurate, specific, and directly answers the question with no irrelevant details (e.g., "There is a mass in the right upper lobe.>").
 - 1: Inaccurate, vague, or only loosely related to the question (e.g., "There might be something abnormal.")
 - If the answer contains errors, vague descriptions, or misinterprets the question, it should be scored lower. The answer should directly match the content of the question.
4. *Q-A Alignment: Does the answer format/type match the question format/type? Is the answer logically aligned with the type of information requested in the question?*
 - 5: Perfect match (e.g., question asks "Where", answer gives a location; question asks "What size", answer gives a size).
 - 1: Mismatch in type or category (e.g., question asks "What", but the answer gives "When").
 - Ensure the answer matches the expected type (e.g., answering a "Where" question with a location, answering a "What" question with a description of the object or condition).
5. *Linguistic Quality: Are both the question and the answer grammatically correct, fluent, and easy to understand? Are there any issues with language clarity in either the question or the answer?*
 - 5: Both question and answer are fluent, clear, and easy to understand without grammatical issues or awkward phrasing.
 - 1: Question or answer is awkward, grammatically incorrect, or confusing.
 - If the answer contains significant linguistic errors that cause confusion or misunderstandings, it should be scored lower.

Also, check for spelling and grammar.

Example 1:

VolumeName, Visual Verifiability, Specificity & Clarity, Answer Appropriateness, Q-A Alignment, Linguistic Quality
test_806_a_2.nii.gz,4,4,5,5,5

Question: Which structures are midline?

Answer: Trachea, bronchi

Clinic Text: Trachea and both main bronchi were in the midline and no obstructive pathology was observed in the lumen...

Reasons:

Visual Verifiability (4): "Trachea" is clearly midline and visually verifiable; "bronchi" is ambiguous without specifying main bronchi.

Specificity & Clarity (4): The term "bronchi" is too general and could refer to various branches, reducing clarity.

Answer Appropriateness (5): The answer includes anatomically relevant and reasonable structures.

Q-A Alignment (5): The answer directly responds to the question about midline structures.

Linguistic Quality (5): Fluent and grammatically correct.

Example 2:

VolumeName, Visual Verifiability, Specificity & Clarity, Answer Appropriateness, Q-A Alignment, Linguistic Quality
test_63_b_2.nii.gz,5,5,5,5,5

Question: What is the diameter of the main pulmonary artery?

Answer: 36mm

Clinic Text: ...The diameter of the main pulmonary artery was 36mm, the diameter of the right pulmonary artery was 28mm, and the diameter of the left pulmonary artery was 25mm, showing dilatation...

Reasons:

Visual Verifiability (5): Main pulmonary artery diameter is a directly measurable anatomical feature.

Specificity & Clarity (5): The answer is precise ("36mm") and unambiguous.

Answer Appropriateness (5): Value is correctly extracted and clinically appropriate.

Q-A Alignment (5): The answer directly matches the question.

Linguistic Quality (5): Answer is concise and grammatically correct.

Figure 10: Guideline

Task 1-2 Prompts

system_text = ""

You are a medical AI visual assistant that can analyze a single CT image. You receive the medical diagnosis report. The report describes multiple abnormal lesions in the image.

The task is to use the report information to create 6 questions and answers about the image.

The first five questions and answers must be based strictly on the Findings text, and the sixth question and answer must be based strictly on the Impressions text.

These questions come from the following 6 aspects:

- 1). Anatomical observation (based on Findings)
- 2). Pathological observation (based on Findings)
- 3). Abnormality type (based on Findings)
- 4). Abnormality feature (based on Findings)
- 5). Abnormality position (based on Findings)
- 6). Abnormality or normality diagnosis (based on Impressions)

""

PROMPT = ""

Clinical text: {

Findings: <Findings>

Impressions: <Impressions>

}

Please generate a set of exactly 6 clinical image-based question-answer pairs, strictly following these constraints:

- Focus the questions on image features related to 6 aspects, with the first five questions generated strictly based on "Findings", and the sixth question based strictly on "Impressions".
- Do not reference, mention, or imply the words "findings" or "impressions" in any part of the question. Questions must not use phrases like "assessment", "based on findings", or "what is noted" that imply summary or interpretation.
- Treat the content as if it comes from direct image observation, not from a text report.
- Avoid overly broad or vague questions. Ensure each question is specific, objective, visually verifiable, and based solely on image evidence.
- Questions must only be answerable by directly observing the image — not from general knowledge, not through inference, and not through assumptions.
- Do not generate a question that can be answered without the image.
- Avoid generating questions that require medical calculations such as number, size, volume, or specific coordinate locations.
- The answer must directly correspond to the question asked, based on the content of the clinical text.
- The answers and the questions must conform to correct medical knowledge. Ensure that the answer is both clinically relevant to the image and accurate.
- The answers should describe observable visual aspects of the image in concise phrases, using no more than 3 words.
- Please do not ask directly what organs or abnormalities are visible in the image, as the answers are not unique.
- Ensure that the question only has one clear, unique answer that would be consistently given by different people analyzing the same image.
- Avoid overly broad or vague questions. Ensure each question is specific, objective, and visually verifiable.
- The questions and answers should assume that the task is to be performed based on the image alone. Do not mention the report in the questions and answers.
- The generated questions should begin with <Starter> in a way that is natural and coherent, avoiding awkward or forced phrasing.
- The questions should not be overly complicated and should be easy for both AI models and doctors to answer accurately.
- The questions and answers must be strictly aligned; the answer must match the question type.
- For "Where" questions, the answer must be a specific anatomical location, not an appearance or visibility description.
- The answers should be directly accurate.

Desired format:

1). Anatomical observation (based on Findings)

Question-1: <Starter1> ...? Answer: ...

2). Pathological observation (based on Findings)

Question-2: <Starter2> ...? Answer: ...

3). Abnormality type (based on Findings)

Question-3: <Starter3> ...? Answer: ...

4). Abnormality feature (based on Findings)

Question-4: <Starter4> ...? Answer: ...

5). Abnormality position (based on Findings)

Question-5: <Starter5> ...? Answer: ...

6). Abnormality or normality diagnosis (based on Impressions)

Question-6: <Starter6> ...? Answer: ...

""

Figure 11: Prompts of Task 1-2

Task 3 Prompts (First Stage)

```

system_text = """
    You are a medical AI visual assistant that can analyze a single CT image. You receive the medical diagnosis report. The
    report describes multiple abnormal lesions in the image.
    The task is to extract all complete sentences from the text that contain numerical values.
    """

PROMPT = """
    Please extract all complete sentences that contain specific numerical values from the following radiology finding report.
    - Keep the original expression of the sentences without making any modifications.
    - The extracted sentences should be applicable for creating a medical quantitative question and answer high-quality dataset
    from these three perspectives: 1). Size 2). Diameter 3). Thickness.
    - Quantities should be specific, estimates such as "a few" should not be included.
    - If no suitable sentences are found, no output is necessary.

    Radiology finding report:
    <Findings>

    Desired format:
    1). Sentence-1: ...\n
    2). Sentence-2: ...\n
    ...
    """

```

Figure 12: Prompts of Task 3 (First Stage)

Task 3 Prompts (Second Stage)

system_text = ""

You are a medical AI visual assistant that can analyze a single CT image. You receive the medical diagnosis report. The report describes multiple abnormal lesions in the image.

The task is to use the report information to choose the most related aspect and create one question and answer about the image.

Each question must include a specific descriptor or location to uniquely identify the lesion or nodule being referred to. Answers must be exact and not estimations or imaginary numbers.

The question and answer come from one of the following 3 aspects:

- 1). *Size*
- 2). *Diameter*
- 3). *Thickness*

""

PROMPT = ""

Based on the following radiology description sentence, choose the most appropriate aspect from the four options and generate one quantitative medical question and its corresponding answer.

- Include a specific description or location in the question to uniquely identify the nodule or lesion being referred to.
- Do not reference, mention, or imply the words “findings” or “impressions” in any part of the question. Questions must not use phrases like “assessment”, “based on findings”, or “what is noted” that imply summary or interpretation.
- Treat the content as if it comes from direct image observation, not from a text report.
- Avoid overly broad or vague questions. Ensure each question is specific, objective, visually verifiable, and based solely on image evidence.
- Questions must only be answerable by directly observing the image — not from general knowledge, not through inference, and not through assumptions.
- Do not generate a question that can be answered without the image.
- Avoid generating questions that require medical calculations such as number, size, volume, or specific coordinate locations.
- The answer must directly correspond to the question asked, based on the content of the clinical text.
- The answers and the questions must conform to correct medical knowledge. Ensure that the answer is both clinically relevant to the image and accurate.
- The answers should describe observable visual aspects of the image in concise phrases, using no more than 3 words.
- Please do not ask directly what organs or abnormalities are visible in the image, as the answers are not unique.
- Ensure that the question only has one clear, unique answer that would be consistently given by different people analyzing the same image.
- Avoid overly broad or vague questions. Ensure each question is specific, objective, and visually verifiable.
- The questions and answers should assume that the task is to be performed based on the image alone. Do not mention the report in the questions and answers.
- The questions should not be overly complicated and should be easy for both AI models and doctors to answer accurately.
- The questions and answers must be strictly aligned; the answer must match the question type.
- For "Where" questions, the answer must be a specific anatomical location, not an appearance or visibility description.
- The answers should be directly accurate.

The question comes from the following 3 aspects:

- 1). *Size*
- 2). *Diameter*
- 3). *Thickness*

Radiology description sentence:

<Sentence>

Desired format:

Aspect(Size, Diameter or Thickness)

Question: ...? Answer: ...

""

Figure 13: Prompts of Task3 (Second Stage)

Task 4 Prompts	
<pre> question_aspects = ["Medical material", "Arterial wall calcification", "Cardiomegaly", "Pericardial effusion", "Coronary artery wall calcification", "Hiatal hernia", "Lymphadenopathy", "Emphysema", "Atelectasis", "Lung nodule", "Lung opacity", "Pulmonary fibrotic sequela", "Pleural effusion", "Mosaic attenuation pattern", "Peribronchial thickening", "Consolidation", "Bronchiectasis", "Interlobular septal thickening"] </pre>	<pre> question_patterns = ["Is there any presence of {aspect} on the CT scan?", "Can {aspect} be detected on the CT scan?", "Is {aspect} visible on the CT images?", "Does the CT scan suggest {aspect}?", "Can {aspect} be observed on the CT scan?", "Is there evidence of {aspect} on the CT images?", "Is {aspect} identifiable on the CT scan?", "Does this CT image show {aspect}?", "Is {aspect} present on the CT scan?", "Does the CT scan demonstrate {aspect}?", "Can {aspect} be clearly seen on the CT scan?", "Is {aspect} apparent on the CT images?", "Does the CT scan reveal {aspect}?", "Is {aspect} confirmed by the CT findings?", "Can signs of {aspect} be recognized on the CT scan?", "Does this CT image present {aspect}?", "Is {aspect} noted on the CT images?", "Does the CT scan exhibit {aspect}?"] </pre>

53

Figure 14: Prompts of Task4

Task 5 Prompts	
<pre> question_aspects = ["Arterial wall calcification", "Cardiomegaly", "Pericardial effusion", "Coronary artery wall calcification", "Hiatal hernia", "Lymphadenopathy", "Emphysema", "Atelectasis", "Lung nodule", "Lung opacity", "Pulmonary fibrotic sequela", "Pleural effusion", "Mosaic attenuation pattern", "Peribronchial thickening", "Consolidation", "Bronchiectasis", "Interlobular septal thickening"] </pre>	<pre> question_patterns = ["Based on the current CT scan, how would you classify the {aspect}?", "How would you classify the {aspect} observed in the current CT scan?", "Given the present CT findings, what is the nature of the {aspect}?", "According to the CT scan, how would you categorize the {aspect}?", "Based on the CT scan, what is the current status of the {aspect}?", "After reviewing the CT scan, how would you interpret the presence of {aspect}?", "In light of the current CT imaging, how should we classify the {aspect}?", "With reference to the CT scan, what is the classification for {aspect}?", "How should we interpret the {aspect} from the current CT findings?", "Based on the CT scan, what is the condition of the {aspect}?", "How would you categorize the {aspect} based on the CT scan?", "Considering the CT scan, what is your assessment of the {aspect}?", "How would you classify the {aspect} seen in the CT imaging?", "According to the CT scan, what is the current status of {aspect}?", "How would you classify the {aspect} based on the current CT scan?", "In light of the CT findings, how should we interpret the {aspect}?", "Based on the CT scan, what is your evaluation of the {aspect}?"] </pre>
<pre> label_map = { "A": "Refractory Lesion (Persistent or recurrent, now present)", "B": "Resolved Lesion (Previously present or recurrent, now absent)", "C": "New Lesion (Absent previously, now present)", "D": "No Abnormality (Always absent)" } </pre>	

Figure 15: Prompts of Task5

Task 6 Prompts

```

question_aspects = [
  "Arterial wall calcification",
  "Cardiomegaly",
  "Pericardial effusion",
  "Coronary artery wall calcification",
  "Hiatal hernia",
  "Lymphadenopathy",
  "Emphysema",
  "Atelectasis",
  "Lung nodule",
  "Lung opacity",
  "Pulmonary fibrotic sequela",
  "Pleural effusion",
  "Mosaic attenuation pattern",
  "Peribronchial thickening",
  "Consolidation",
  "Bronchiectasis",
  "Interlobular septal thickening"
]

fluctuation_prefixes = [
  "Based on sequence history and current CT.",
  "From multi-stage sequences and present CT.",
  "Sequence trend plus current CT defines status.",
  "History and CT decide lesion category.",
  "Stage-wise sequences + current CT = status.",
  "Past sequences and CT determine lesion type.",
  "Multi-phase sequences guide CT-based judgment.",
  "Classification uses sequences and current image.",
  "Lesion state from timeline and CT.",
  "CT reflects pattern of prior sequences.",
  "Lesion judged by sequences and scan.",
  "Sequence evolution and CT define outcome.",
  "Diagnosis combines sequence history and CT.",
  "CT confirms what sequences suggest.",
  "Lesion behavior from past to CT.",
  "CT status follows sequence progression.",
  "Final label from history and CT."
]

sequence_prefixes = [
  "{aspect} sequence was: {seq}.",
  "Past sequences of {aspect}: {seq}.",
  "Previous {aspect} states: {seq}.",
  "Earlier {aspect} sequences: {seq}.",
  "Historical {aspect} status: {seq}.",
  "Sequence history for {aspect}: {seq}.",
  "Prior {aspect} timeline: {seq}.",
  "Scans showed {aspect} as: {seq}.",
  "Old sequences for {aspect}: {seq}.",
  "Earlier CTs showed {aspect} as: {seq}.",
  "Before now, {aspect} showed {seq}.",
  "{aspect} had: {seq} in earlier scans.",
  "{aspect} progression: {seq}.",
  "In past, {aspect} was: {seq}.",
  "{aspect} over time: {seq}.",
  "{aspect} condition history: {seq}.",
  "Recorded sequences for {aspect}: {seq}."
]

question_patterns = [
  "What is the current status of {aspect} based on previous sequences?", "Based on past sequences, what is the current condition of {aspect}?",
  "How does {aspect} appear now compared to its sequence history?", "What type of lesion is {aspect} now, given its temporal sequence?",
  "From sequence history to now, what best describes {aspect}?", "How has {aspect} changed from past sequences to the current CT?",
  "What does the current CT show about {aspect} based on earlier sequences?", "Given its sequence history, what is {aspect} in the current CT?",
  "How has {aspect} evolved from earlier scan sequences to now?", "What lesion category applies to {aspect} in the current CT?",
  "How is {aspect} classified now using past sequence data?", "What does the CT show for {aspect} considering its sequence history?",
  "What is the current CT assessment of {aspect} based on prior sequences?", "How has {aspect} progressed according to its temporal sequence?",
  "What does {aspect} currently represent, based on past sequences?", "Based on the temporal sequence, what is the present status of {aspect}?",
  "From prior sequences to now, what is the CT-based status of {aspect}?"
]

label_map = {
  "A": "Refractory Lesion (Persistent or recurrent, now present)", "B": "Resolved Lesion (Previously present or recurrent, now absent)",
  "C": "New Lesion (Absent previously, now present)", "D": "No Abnormality (Always absent)"
}

```

Figure 16: Prompts of Task6

Check

system_text = ""

You are a medical expert specializing in radiology, particularly in chest CT imaging. Your task is to score the following radiology question and answer pairs based on specific criteria. You will evaluate each pair using the five dimensions listed below. Be strict, specific, and consistent in your evaluation, and ensure the scores reflect the quality of the content in the context of chest CT findings.

PROMPT = ""

Please score the following radiology visual question and answer pair, using the 5 dimensions below. Each should be scored from 1 (very poor) to 5 (excellent). Be strict, specific, and consistent in your evaluation.

Question: {question}

Answer: {answer}

Scoring Dimensions:

1. Visual Verifiability: Can this question be answered just by looking at the image, without requiring medical knowledge, inference, or external context? Does the answer also rely on the image for validation?

- 5: Can be answered purely from the image (e.g., "Is there a fracture?") and the answer is clearly image-based (e.g., "Yes, there is a fracture in the left femur.")

- 1: Requires background knowledge, inference, or external context (e.g., "What type of tumor?") and the answer similarly relies on external context rather than the image.

- If the answer requires knowledge of clinical context or interpretation beyond the image itself, it should be scored lower.

2. Specificity & Clarity: Is the question precise and unambiguous? Is the answer also specific and clear, without ambiguity?

- 5: The question and answer are both specific and clear with one unambiguous interpretation (e.g., "Which lobe is involved in the lesion?" and the answer directly states the affected lobe).

- 1: The question or the answer is vague, ambiguous, or open to multiple interpretations (e.g., "What is abnormal?" and the answer gives an unclear response like "There is something unusual.")

- Avoid vague terms like "a few," "several," "some," "few millimeters," or similar imprecise quantities. If these terms appear in the answer, it should be scored lower.

- Both the question and the answer must be precise. If either is unclear, reduce the score.

3. Answer Appropriateness: Is the answer correct, medically appropriate, specific, and directly relevant to the question? Does it match the expected format/type?

- 5: Accurate, specific, and directly answers the question with no irrelevant details (e.g., "There is a mass in the right upper lobe.").

- 1: Inaccurate, vague, or only loosely related to the question (e.g., "There might be something abnormal.")

- If the answer contains errors, vague descriptions, or misinterprets the question, it should be scored lower. The answer should directly match the content of the question.

4. Q-A Alignment: Does the answer format/type match the question format/type? Is the answer logically aligned with the type of information requested in the question?

- 5: Perfect match (e.g., question asks "Where", answer gives a location; question asks "What size", answer gives a size).

- 1: Mismatch in type or category (e.g., question asks "What", but the answer gives "When").

- Ensure the answer matches the expected type (e.g., answering a "Where" question with a location, answering a "What" question with a description of the object or condition).

5. Linguistic Quality: Are both the question and the answer grammatically correct, fluent, and easy to understand? Are there any issues with language clarity in either the question or the answer?

- 5: Both question and answer are fluent, clear, and easy to understand without grammatical issues or awkward phrasing.

- 1: Question or answer is awkward, grammatically incorrect, or confusing.

- If the answer contains significant linguistic errors that cause confusion or misunderstandings, it should be scored lower.

Also, check for spelling and grammar.

Example response format:

```
{{
  "Visual Verifiability": 3,
  "Specificity & Clarity": 4,
  "Answer Appropriateness": 5,
  "Q-A Alignment": 5,
  "Linguistic Quality": 4
}}
```

Figure 17: Prompts of Check

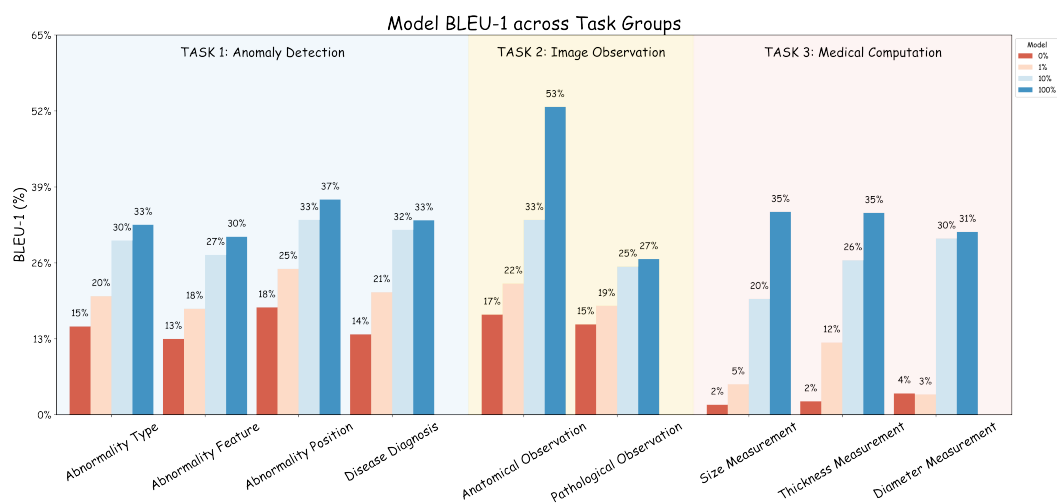


Figure 18: Finetuned BLEU Result of Task 1-3

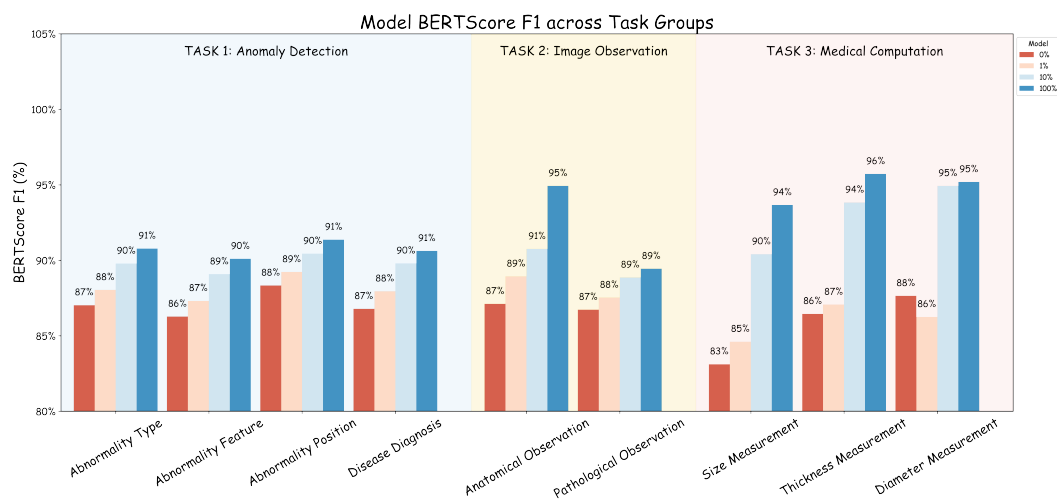


Figure 19: Finetuned F1 Result of Task 1-3

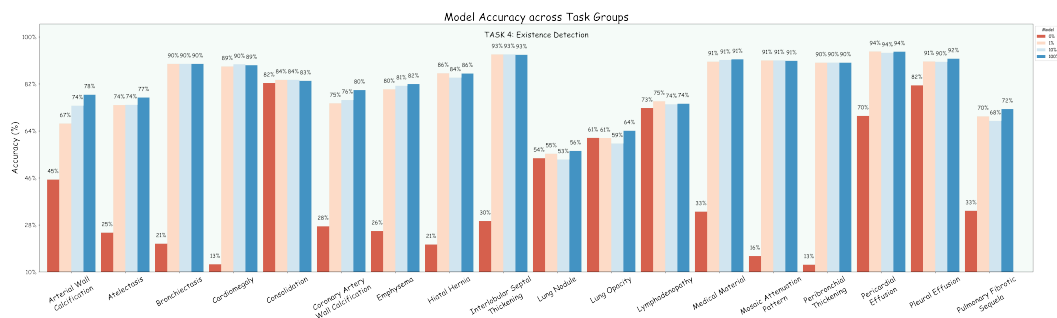


Figure 20: Finetuned results of models on Task 4

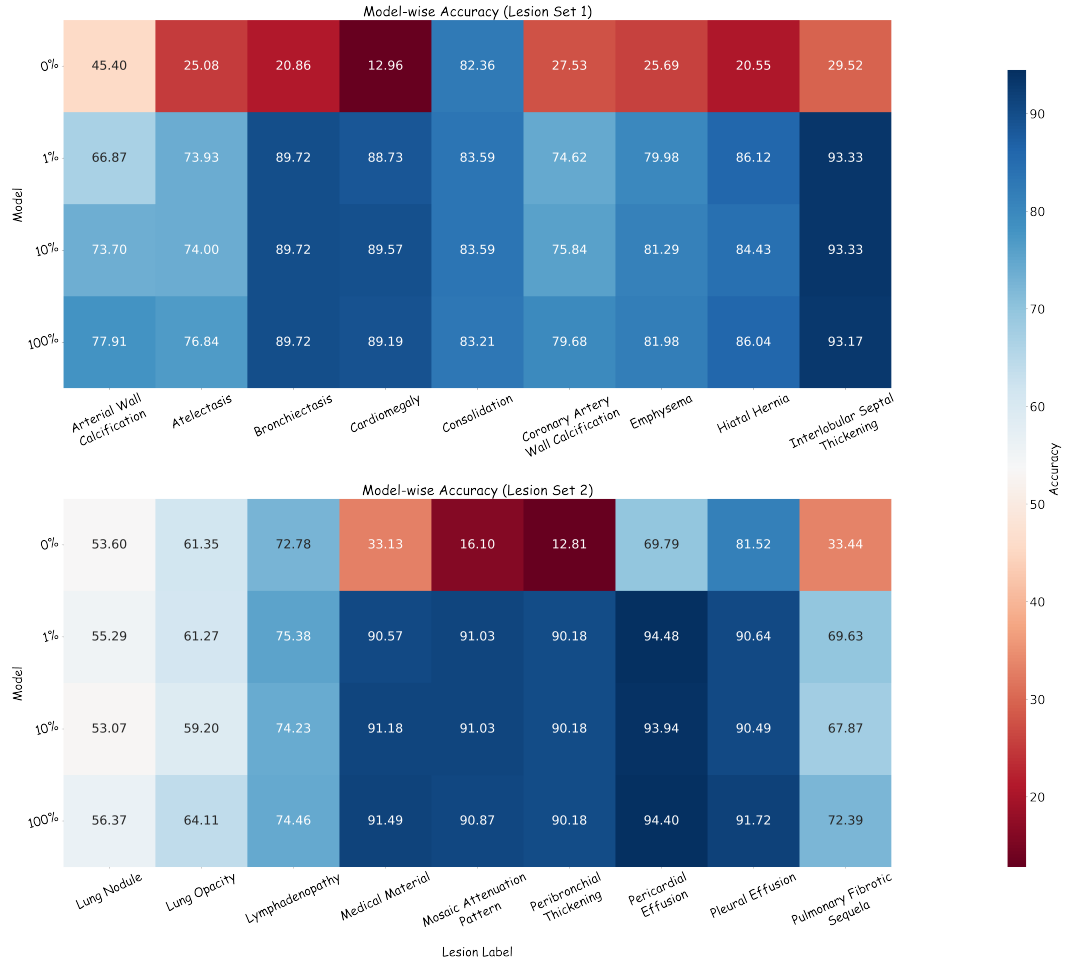


Figure 21: Finetuned Result of Task 4

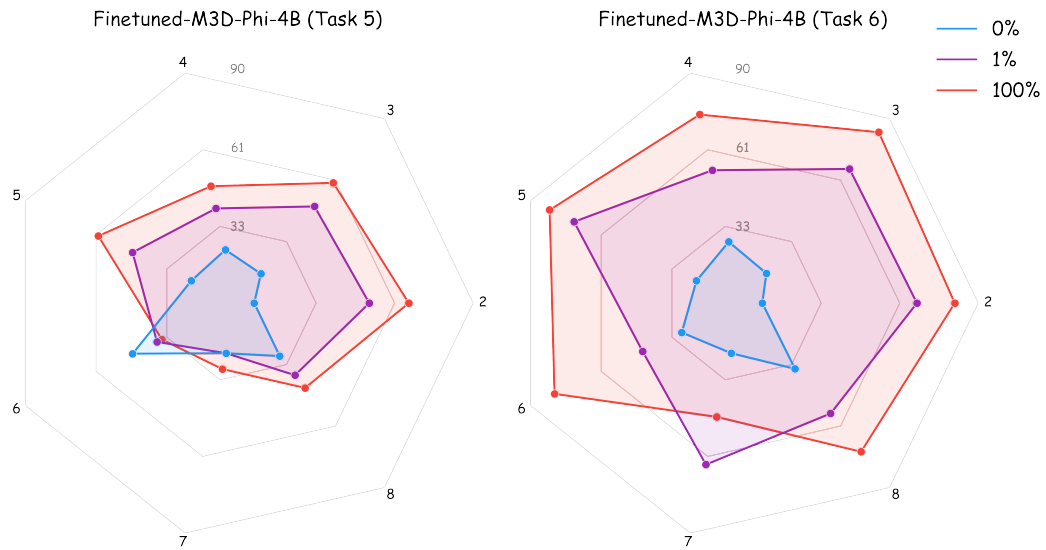


Figure 22: Finetuned Result of Task 5-6

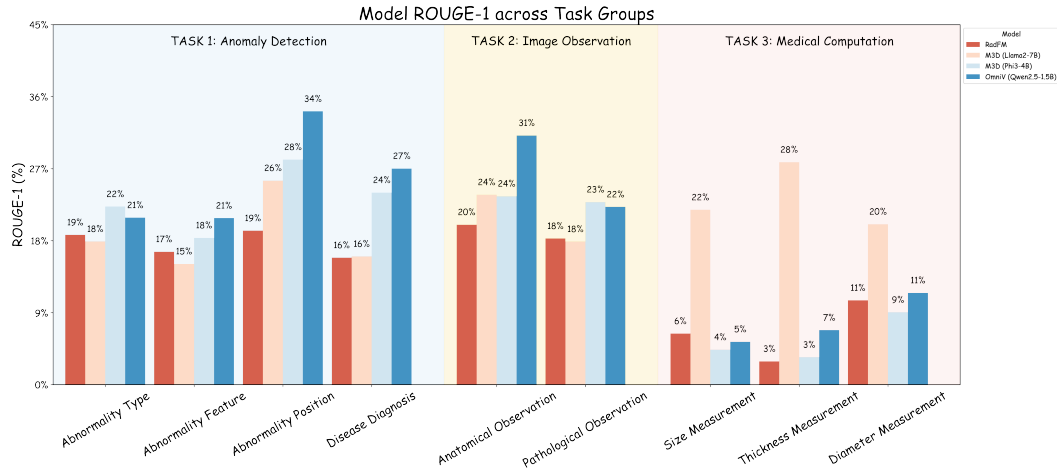


Figure 23: Zero-Shot Rouge Result of Task 1-3

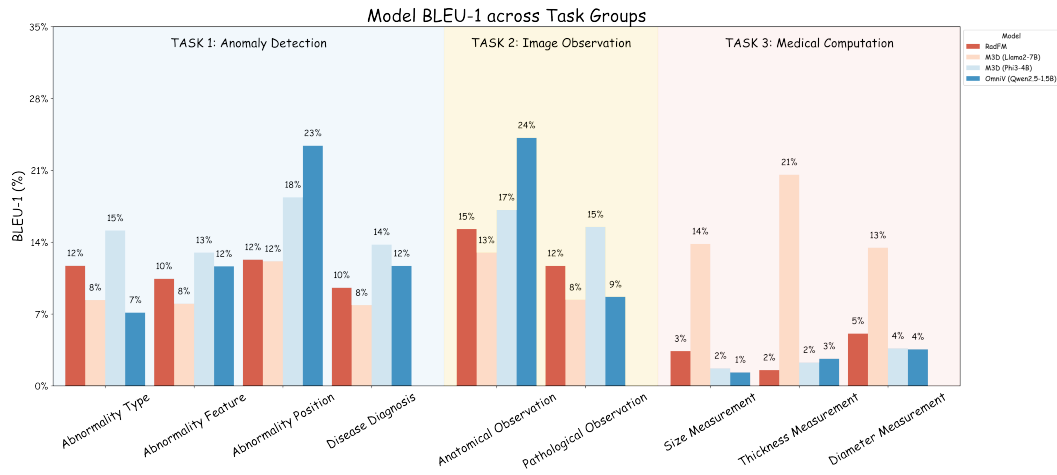


Figure 24: Zero-Shot BLEU Result of Task 1-3

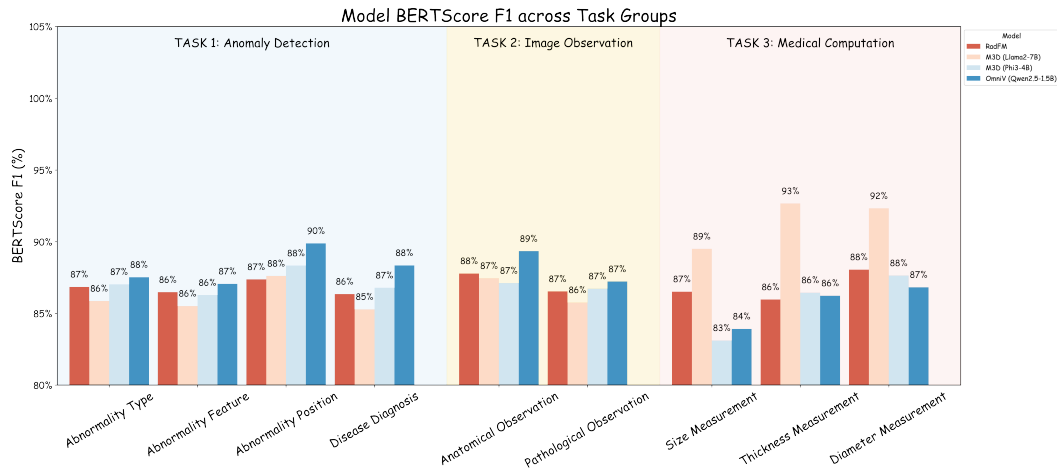


Figure 25: Zero-Shot F1 Result of Task 1-3

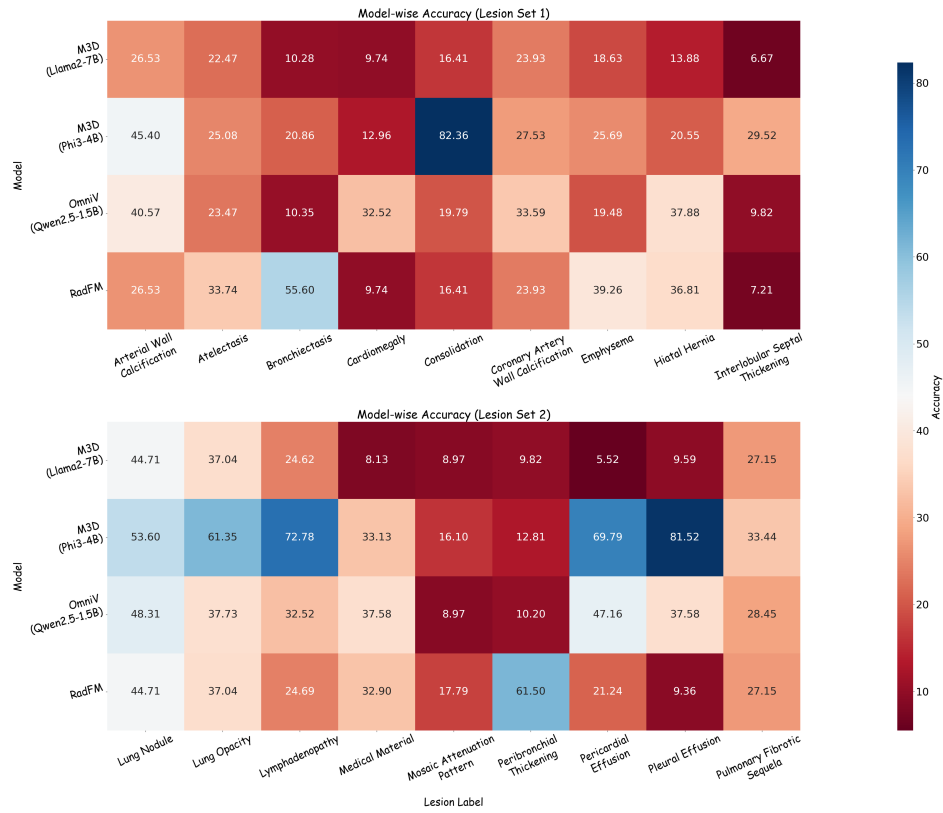


Figure 26: Zero-Shot Result of Task 4

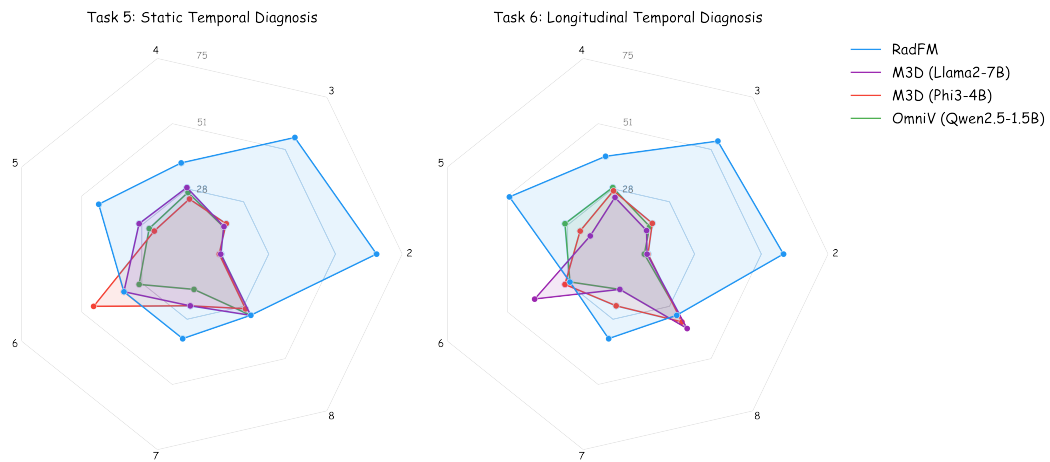


Figure 27: Zero-Shot Result of Task 5-6

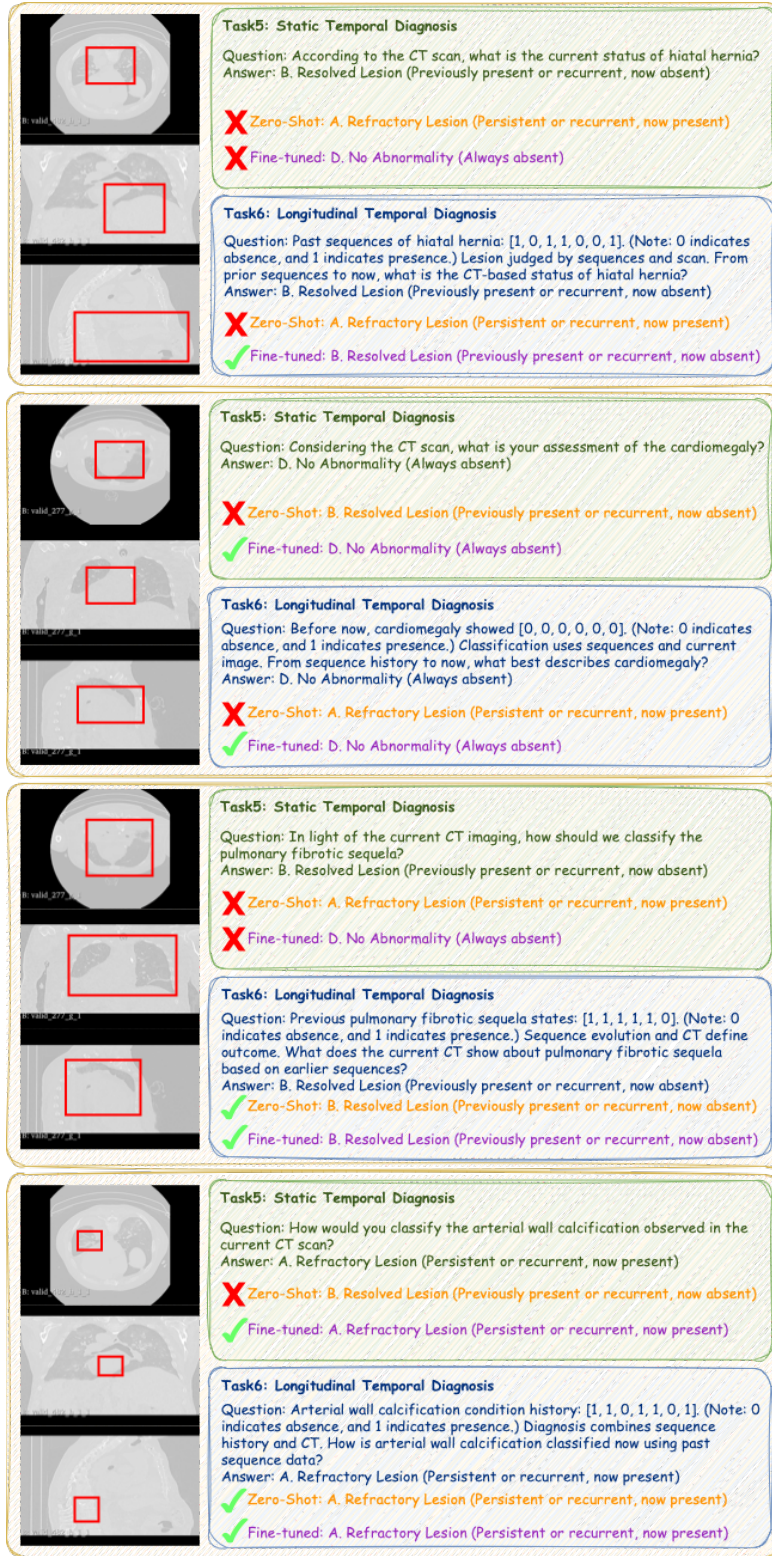


Figure 28: Failure case visualization for Tasks 5 and 6. Green checkmarks indicate correct answers; red crosses indicate failures. Each row pair shows performance of zero-shot (top) and fine-tuned (bottom) models for both task types.

