REPORTQA: QA-BASED RADIOLOGY REPORT EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Radiology report evaluation is essential for advancing automated radiology report generation. Common natural language generation metrics have limited clinical relevance. Clinical efficacy (CE) metrics evaluate important medical findings, but focus mainly on presence and cover only a limited set of entities. Due to heavy reliance on manual annotations, it is difficult for CE metrics to extend clinical entities or attributes. In clinical practice, radiology reports serve as a medium for information transfer. Clinicians use them to perform downstream diagnostic tasks without directly inspecting images. Based on this insight, we propose **ReportQA**, a clinical-related and flexible radiology report evaluation framework, supporting **detailed quantitative analysis** of radiology report generation systems. We first collect datasets covering multiple imaging modalities and anatomical regions. We then construct knowledge trees of clinical entities and attributes with radiologist guidance, and use large language models (LLMs) to extract structured information from free-form reports. Next, we generate QA pairs from predefined templates and apply quality control through self-filtering and report-based filtering. During evaluation, the report is treated as context, and an LLM acts as a judge model to answer the QA pairs. Based on the resulting QA accuracy, we introduce **QAScore** metric. Compared with existing metrics, QAScore shows better alignment with radiologist judgments. Experiments on multiple state-of-the-art vision-language models reveal that current report-based inference paradigms struggle to learn fine-grained clinical representations and exhibit strong negative prior biases. In contrast, **question-driven inference** provides a more effective alternative. For reproducibility and extensibility, we release knowledge trees, structured reports, and QA pairs[1], along with the pipeline code[2] for QA construction and evaluation.

1 INTRODUCTION

Medical imaging is indispensable in routine clinical practice (Rao et al., 2025). This has motivated growing interest in automated radiology report generation (RRG), which can improve clinical efficiency and support consistent report quality. However, progress in RRG is constrained by the lack of reliable radiology report evaluation (RRE) frameworks. Existing RRE frameworks fail to provide meaningful feedback on clinical correctness.

Existing RRE frameworks typically combine natural language generation (NLG) metrics with clinical efficacy (CE) metrics. However, they remain limited in clinical relevance and flexibility. NLG metrics, such as BLEU (Papineni et al., 2002), mainly measure lexical overlap between generated and ground-truth reports. This evaluation paradigm is misaligned with clinical requirements. In practice, a report is clinically useful not because it is lexically similar to a reference report, but because it accurately describes key clinical entities and their associated attributes.

CE metrics, such as CheXbert (Smit et al., 2020), partially address this issue by evaluating the presence of selected clinical entities. However, they fail to capture fine-grained attributes such as location, shape and size. Moreover, CE metrics rely on supervised models trained with substantial manual annotations. This makes them difficult to extend to new entities or attributes and thus limiting detailed quantitative analysis of RRG system.

[1]Data: `https://huggingface.co/datasets/anon-iclr-2027/anon-iclr-2027`
[2]Code: `https://anonymous.4open.science/r/anon-iclr-2027-5F83`

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.

In clinical practice, radiology reports serve as a medium for information transfer (Busch et al., 2025; Tanno et al., 2025). Clinicians use report content instead of raw imaging for a wide range of downstream diagnostic tasks. Motivated by this, we propose to evaluate reports by treating them as context for question answering (QA). This design simulates how clinicians interact with reports in real-world settings. The resulting QA-based evaluation framework better aligns RRE with the clinical use of RRG systems and provides a more faithful measure of report utility.

First, we collect diverse radiology report datasets covering multiple imaging modalities and anatomical regions. To better align with real-world clinical scenarios, we transform raw free-form reports into a structured format using dataset-specific knowledge trees. The structured reports are then converted into QA pairs. We use large language models (LLMs) for quality control. During evaluation, each report is treated as contextual input, and an LLM is used as the judge model to answer the corresponding QA pairs. Based on the LLM responses, we introduce QAScore metric.

Compared with existing RRE metrics, QAScore shows stronger alignment with radiologist judgments. Furthermore, through a comprehensive evaluation of state-of-the-art (SOTA) vision-language models (VLMs), we find that current report-based inference paradigms struggle to learn fine-grained clinical representations and exhibit strong negative prior biases. In contrast, question-driven inference provides a more promising alternative. Our main contributions are summarized as follows:

1. We propose **ReportQA**, a QA-based radiology report evaluation framework, supporting detailed quantitative analysis of RRG systems. Additionally, we introduce **QAScore**, which is an aggregated metric based on the judge model responses.
2. ReportQA reveals **hidden failure modes** that are invisible to common report-level metrics. Experiments indicate that existing report-based inference paradigms struggle to learn fine-grained clinical information and exhibit strong negative prior biases. In contrast, **question-driven inference** provides a more effective alternative.
3. We release all knowledge trees, structured reports and QA pairs, along with the pipeline code for QA construction and evaluation. Users can plug in stronger LLMs to improve QA quality and apply the pipeline to more datasets.

2 Related Work

Traditional RRE frameworks can be broadly categorized into NLG metrics and CE metrics. Among NLG metrics, BLEU (Papineni et al., 2002) measures n-gram precision between generated and reference texts, ROUGE (Lin, 2004) focuses on recall-oriented overlap, and METEOR (Banerjee & Lavie, 2005) incorporates synonym matching and alignment-based scoring. Despite their widespread use, NLG metrics primarily rely on surface-level lexical similarity and fail to capture the accuracy of clinically relevant information. CE-based metrics attempt to address this limitation by explicitly evaluating clinical entities. CheXbert (Smit et al., 2020) is a BERT-based classifier trained to predict the presence of predefined clinical findings from reports, while RadGraph (Jain et al., 2021) extracts structured entities and relations to enable more fine-grained evaluation. However, these approaches depend heavily on manually annotated data and predefined label spaces, making them difficult to extend to new clinical entities or attributes.

More recent RRE methods increasingly leverage LLMs or structured clinical representations. GREEN (Ostmeier et al., 2024) uses LLMs to identify and explain clinical errors in generated reports, while RaTEScore (Zhao et al., 2024) combines named entity recognition with embedding-based similarity for entity-level semantic evaluation. ICARE (Dua et al., 2025) further adopts a QA-based formulation by dynamically generating questions for individual report pairs. CLEAR (Jiang et al., 2025b) and LUNGUAGE (Moon et al., 2025) provide structured evaluation for chest X-ray reports, including clinically relevant dimensions such as longitudinal changes and recommendations. Other approaches, such as RadReason (Li et al., 2025) and LLM-RadJudge (Wang et al., 2024), rely on evaluator training or distillation, while RadEval (Xu et al., 2025a) provides a unified toolkit for radiology report evaluation. These methods substantially improve clinical sensitivity compared with lexical metrics, but differ in their evaluation granularity, domain coverage, and dependence on task-specific evaluators or dynamically constructed evaluation instances.

More broadly, evaluating generated text through question answering has been well studied in natural language generation. QAEval (Deutsch et al., 2021) evaluates summary content by generating ques-

tions from reference texts and measuring whether they can be answered from candidate summaries, while QuestEval (Scialom et al., 2021) extends QA-based evaluation toward reference-based and reference-free factual consistency assessment. These works establish QA as a general framework for measuring information preservation in generated text.

Building on these, ReportQA adapts QA-based evaluation to radiology through clinically structured and fine-grained assessment. It emphasizes standardization and extensibility, enabling reproducible comparison across RRG systems while supporting diverse radiology datasets and imaging settings.

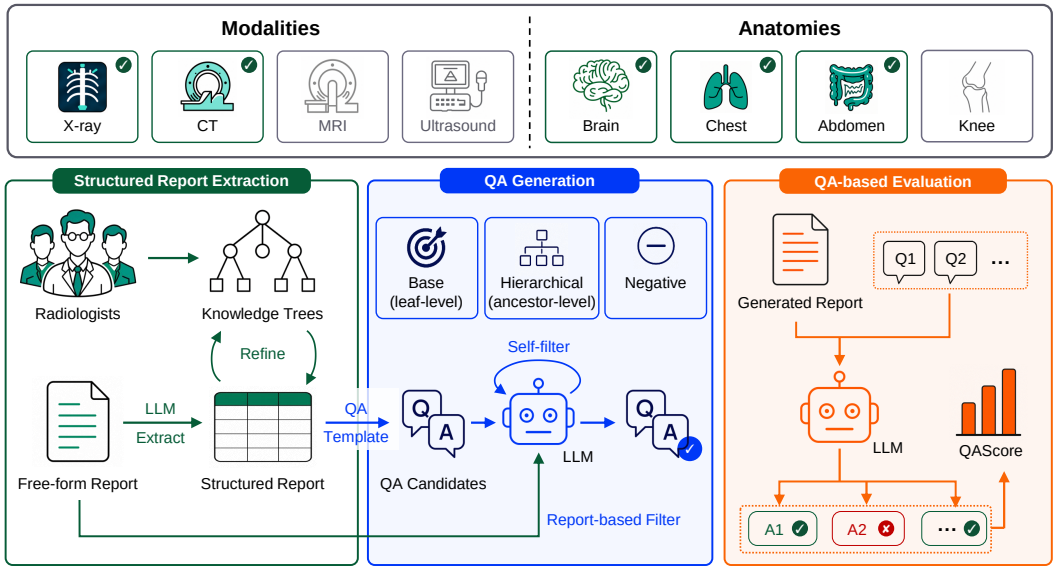


Figure 1: Overview of ReportQA. First, we collect datasets covering multiple imaging modalities and anatomical regions. Second, free-form reports are transformed into a structured format based on radiologist-defined knowledge trees. Third, QA pairs are generated via predefined templates and refined through self-filtering and report-based filtering, resulting in three categories of QA: *base (leaf-level)*, *hierarchical (ancestor-level)*, and *negative*. Finally, the generated report is treated as context, and an LLM answers filtered questions. LLM responses are used to calculate QAScore.

3 METHOD

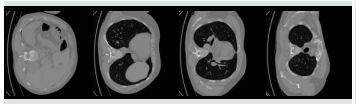
As shown in Figure 1, we construct ReportQA through a four-stage pipeline. First, we collect datasets covering diverse imaging modalities and anatomical regions consisting of free-form radiology reports. Second, we transform these reports into a structured format using dataset-specific knowledge trees, which encode clinically relevant entities and their attributes. Third, we generate QA pairs from the structured reports. Finally, we evaluate reports by answering filtered QA pairs, with comprehensive performance quantified via proposed QAScore.

3.1 DATASET PREPARATION

ReportQA is constructed from diverse radiology report datasets spanning multiple imaging modalities and anatomical regions, including chest X-ray data from MIMIC-CXR (Johnson et al., 2019a), brain CT data from CTRG-Brain (Tang et al., 2024), chest CT data from CT-RATE (Hamamci et al., 2026), and abdominal CT data from AMOS-MM (Ji et al., 2022).

For 2D imaging datasets, images are provided in JPEG format, such as MIMIC-CXR-JPG (Johnson et al., 2019b), while 3D volumes are standardized to the NIfTI format to ensure consistency across datasets. ReportQA is constructed based on the official validation or test splits whenever available. For datasets without predefined splits, such as CTRG-Brain, we adopt a reproducible 4:1 train-validation split.

(a) Raw free-form report



Findings: Trachea and both main bronchi are open. No occlusive pathology was detected in the trachea and both main bronchi. **There are linear atelectasis in the right lung middle lobe medial segment and left lung upper lobe lingular segment**. No mass or infiltrative lesion was detected in both lungs. Mediastinal structures cannot be evaluated optimally because contrast material is not given. As far as can be observed: Heart contour and size are normal. There is no pleural or pericardial effusion. The widths of the mediastinal main vascular structures are normal. There are no pathologically enlarged lymph nodes in the mediastinum and hilar regions. No pathological increase in wall thickness was detected in the esophagus within the sections. No upper abdominal free fluid-collection was detected in the sections. Thoracic vertebral corpus heights, alignments and densities are normal. The neural foramina are open. Impression: Linear atelectasis in both lungs.

(b) Structured report (part)


```
{
 "clinical_entity": {
   "raw": "atelectasis",
   "normalized": "Atelectasis/Incomplete Expansion"
 },
 "presence": {
   "raw": "present",
   "normalized": "Present"
 },
 "location": {
   "raw": "lingular segment",
   "normalized": "Inferior lingular segment"
 },
 "distribution": {
   "raw": "left lung",
   "normalized": "Left"
 },
 "shape": {
   "raw": "linear",
   "normalized": "Linear/streaky"
 },
 "evidence": "There are linear atelectasis in the right lung middle lobe medial segment and left lung upper lobe lingular segment"
}
…
```


(c) QA pairs (part)

{"question_type": "base", "attribute": "presence", "question": "Is [Atelectasis/Incomplete Expansion] present?", "options": ["No", "Yes"], "answer": 1}

{"question_type": "base", "attribute": "location", "question": "Is [Atelectasis/Incomplete Expansion] present in the [Right Medial segment]?", "options": ["No", "Yes"], "answer": 1}

{"question_type": "base", "attribute": "shape", "question": "Regarding the [Atelectasis/Incomplete Expansion] in the [Right Medial segment], what is its [Shape]?", "options": ["Mass-like/conglomerate", "Patchy", "Punctate/speckled", "Linear/streaky"], "answer": 3}

{"question_type": "hierarchical", "attribute": "presence", "question": "Is [Abnormal Lung Volume] present?", "options": ["Yes", "No"], "answer": 0}

{"question_type": "negative", "attribute": "presence", "question": "Is [Acute Bronchitis] present?", "options": ["Yes", "No"], "answer": 1}

{"question_type": "negative", "attribute": "presence", "question": "Is [Ankylosing Spondylitis] present?", "options": ["Yes", "No"], "answer": 1}

…

Figure 2: An example from CT-RATE (Hamamci et al., 2026), including raw free-form report, structured report, and QA pairs. The structured report and QA pairs are simplified and shown only partially for visualization.

3.2 Structured report extraction

First, radiologists construct two dataset-specific knowledge trees based on dataset metadata, including imaging modality and anatomical region. One knowledge tree models clinical entities, consisting of findings and diagnoses, where findings are primarily extracted from the Findings section of reports and diagnoses from the Impression section. The other tree defines a comprehensive set of clinical attributes, covering 15 dimensions: *presence*, *location*, *distribution*, *number*, *dimension*, *density*, *shape*, *margin*, *enhancement*, *internal features*, *secondary effects*, *severity*, *chronicity*, *clinical score*, and *certainty*. Detailed examples of knowledge trees are provided in Appendix A.9.

We then employ LLMs to extract structured information from raw free-form reports. To improve consistency and coverage, we adopt a two-stage extraction process. In the first stage, LLMs generate preliminary structured outputs. These outputs are subsequently used to refine the knowledge trees by identifying missing or ambiguous nodes. In the second stage, the refined knowledge trees are used to remap the extracted entities and attributes, ensuring that all structured information is aligned with predefined nodes. Invalid entities and attributes that cannot be mapped to the knowledge trees are filtered out, resulting in a controlled and standardized representation. Figure 2 shows the example of structured reports. The prompts used for structured extraction are provided in Appendix A.9.

3.3 QA generation

Given the structured reports, we construct *base (leaf-level)* multiple-choice QA pairs using predefined templates that systematically cover different attributes of clinical entities. As shown in Figure 2, each QA instance targets a specific (entity, attribute) combination.

To account for *hierarchical (ancestor-level)* clinical entities, we incorporate ontology-aware question generation. In practice, entities may not be described at the most fine-grained level (leaf-level). For example, if the ground-truth report contains diffuse brain atrophy while the generated report predicts brain atrophy, the prediction can be considered partially correct. To reflect this, we additionally generate QA pairs for ancestor nodes of entities based on hierarchy in the clinical entity tree.

We further construct *negative* QA pairs for clinical entities that are absent in the ground-truth report, focusing on the presence attribute. To mitigate overconfident guessing by LLM-based evaluators when evidence is missing, we include an "Insufficient information" option in the answer choices.

Finally, we employ an LLM as the judge model to perform quality control over QA pairs through a two-stage filtering process. First, we directly evaluate QA pairs without providing report context and remove those that can be correctly answered without contextual information. Second, we re-evaluate

Table 1: Statistics of QA construction across datasets. After filtering, a high QA density is maintained, demonstrating that the filtering strategy effectively improves quality without sacrificing coverage.

Dataset	#Reports	#QAs (filtered)	QA density
CTRG-Brain (Tang et al., 2024)	2001	182136	91.0
CT-RATE (Hamamci et al., 2026)	3039	298654	98.3
AMOS-MM (Ji et al., 2022)	400	25433	63.6
MIMIC-CXR (Johnson et al., 2019a)	1417	153821	108.6

QA pairs using the ground-truth report as context and filter out instances with low accuracy. This process ensures that remaining QA pairs are both context-dependent and reliable for evaluation.

After self-filtering and report-based filtering, approximately 660K QAs remain for 6857 radiology reports, averaging nearly 100 high-quality QAs per report. Detailed statistics are provided in Table 1.

3.4 QA-BASED EVALUATION

Following a procedure similar to QA filtering in Section 3.3, an LLM judge answers the filtered questions based on reports generated by RRG systems. QA accuracy supports fine-grained evaluation across question types. We further propose **QAScore**, an entity-level metric that jointly measures the preservation of positive clinical entities and the avoidance of spurious entities.

Let $\mathcal{E}^+$ and $\mathcal{E}^-$ denote the positive and negative entities covered by the filtered QA pairs, respectively. For each positive entity e, we aggregate its *base (leaf-level)* and *hierarchical (ancestor-level)* questions into an entity quality score:

$$q_e = \begin{cases} 0, & \text{if its presence question is answered incorrectly,} \\ \dfrac{1}{|\mathcal{Q}_e^+|} \displaystyle\sum_{j \in \mathcal{Q}_e^+} u_{ej}, & \text{otherwise,} \end{cases} \tag{1}$$

where $\mathcal{Q}_e^+$ contains the questions associated with e, and u_{ej} denotes answer correctness. Repeated evaluations of the same logical question are averaged before aggregation. The presence constraint reflects that entity existence is a prerequisite for attribute reasoning.

For each negative entity a, its false-positive contribution is

$$f_a = \frac{1}{|\mathcal{Q}_a^-|} \sum_{j \in \mathcal{Q}_a^-} v_{aj}, \tag{2}$$

where v_{aj} indicates whether the judge incorrectly answers that the absent entity is present, averaged over repeated evaluations when applicable.

We sum positive-entity errors and false-positive contributions to obtain the report-level error:

$$E_{\text{entity}} = \sum_{e \in \mathcal{E}^+} (1 - q_e) + \sum_{a \in \mathcal{E}^-} f_a, \qquad \text{QAScore} = \frac{1}{1 + E_{\text{entity}}}. \tag{3}$$

Each entity contributes at most one error unit, while correctly rejected negative entities contribute zero. This formulation requires no tunable penalty parameter and prevents entities with more questions from receiving greater weight solely due to question count. Finally, we average QAScore across reports, assigning equal weight to each case.

4 EXPERIMENT

4.1 IMPLEMENTATION DETAILS

For structured report extraction, we employ DeepSeek-V3.2 (Liu et al., 2025) API. For QA filtering and evaluation, we deploy locally a judge model, Qwen3.5-27B, with vLLM (Kwon et al., 2023)

Table 2: Correlation between metrics (Papineni et al., 2002; Zhang et al., 2019; Smit et al., 2020; Jain et al., 2021; Yu et al., 2023; Zhao et al., 2024; Ostmeier et al., 2024; Dua et al., 2025) and radiologist judgments on RadEvalX (Calamida et al., 2024) and CT-RATE (Hamamci et al., 2026). QAScore achieves the best overall alignment with radiologist judgments.

	RadEvalX			CT-RATE		
Metric	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
BLEU-2	0.1851	0.1742	0.1265	0.2128	0.2122	0.1425
BLEU-4	0.1190	0.0964	0.0728	0.2349	0.2028	0.1348
BERTScore	0.2575	0.2525	0.1901	0.4842	0.5114	0.3494
CheXbert	0.3870	0.3779	0.2896	–	–	–
RadGraph F1	0.1743	0.1692	0.1238	–	–	–
RadCliQ	0.2017	0.2153	0.1586	–	–	–
RaTEScore	0.3539	0.3551	0.2651	0.4251	0.4945	0.3424
GREEN	0.4194	0.4022	0.3148	0.5806	0.6013	0.4744
ICARE-GT	0.2073	0.2809	0.2127	0.5944	0.5927	0.4359
ICARE-GEN	0.4770	0.5196	0.4043	0.4753	0.4678	0.3289
ICARE-AVG	0.4657	0.4964	0.3700	**0.6490**	0.6518	0.4728
QAScore	**0.5835**	**0.6153**	**0.5127**	0.5446	**0.8997**	**0.7378**

on a single A800 (80GB) GPU. During inference, we adopt the official inference protocols released with each model if available. Otherwise, we perform inference using SWIFT (Zhao et al., 2025) or TRL (von Werra et al., 2020). All open-source model inference experiments are conducted on a single A800 (80GB) GPU, except for CT-CHAT (70B), which required two A800 GPUs.

For medical image preprocessing, we follow the recommended configurations provided by each model. For API-based VLMs without explicit guidelines, we uniformly sample three slices at equal intervals for inference. For VLMs lacking predefined settings, 2D images are resized to 512×512, and 3D volumes are uniformly resized to 16 slices along the depth dimension.

4.2 QAScore better aligns with radiologist judgments

We evaluate metric-radiologist alignment on RadEvalX (Calamida et al., 2024), which contains 100 chest X-ray report pairs annotated by two radiologists, and a manually annotated CT-RATE subset (Hamamci et al., 2026) containing 144 generated reports. We aggregate the annotated errors as *total_errors* and compare QAScore with common metrics (Papineni et al., 2002; Zhang et al., 2019; Smit et al., 2020; Jain et al., 2021; Yu et al., 2023; Zhao et al., 2024; Ostmeier et al., 2024; Dua et al., 2025). Table 2 shows that QAScore achieves the best overall alignment on both datasets.

Notably, QAScore is not the best Pearson metric on CT-RATE because Pearson measures linear agreement in score magnitudes, whereas Spearman and Kendall measure ranking consistency. QAScore first aggregates soft errors at the entity level and maps the total error mass E_{entity} to $1/(1 + E_{\text{entity}})$. This monotonic mapping preserves report rankings but compresses differences among high-error reports. In addition, QAScore uses soft entity-level errors, while radiologists count discrete errors across six categories. Thus, the two quantities need not be linearly proportional, despite their strong agreement in report-quality ranking.

4.3 Current VLMs remain limited in radiology report generation

As shown in Table 3, proprietary VLMs do not consistently outperform open-source VLMs. GPT-5.6 Sol achieves the highest score on CTRG-Brain (Tang et al., 2024), whereas open-source VLMs lead on the other three datasets. InternVL3.5-38B (Wang et al., 2025) outperforms the best proprietary VLMs, Gemini 3.5 Flash, on CT-RATE (Hamamci et al., 2026) (0.2171 vs. 0.1969), as does Hulu-Med-32B (Jiang et al., 2025a) on MIMIC-CXR (Johnson et al., 2019a) (0.2313 vs. 0.1972). On AMOS-MM (Ji et al., 2022), RadFM Wu et al. (2025) and GPT-5.6 Sol obtain nearly identical scores (0.0653 vs. 0.0651). These results highlight the competitiveness of open-source VLMs.

Table 3: ReportQA results of proprietary VLMs, open-source general VLMs (Qwen Team, 2026; Wang et al., 2025), and open-source medical VLMs (Jiang et al., 2025a; Xu et al., 2025b; Sellergren et al., 2026; Wu et al., 2025; Hamamci et al., 2026) across different datasets (Tang et al., 2024; Hamamci et al., 2026; Ji et al., 2022; Johnson et al., 2019a). **Bold** indicates the best performance among all models, while underlined values denote the best performance within each model category.

Model	CTRG-Brain	CT-RATE	AMOS-MM	MIMIC-CXR
Proprietary VLMs				
GPT-5.6 Sol	**0.1471**	0.1592	0.0651	0.1633
Claude Opus 4.7	0.0845	0.1797	0.0602	0.1465
Gemini 3.5 Flash	0.1057	0.1969	0.0569	0.1972
Open-source VLMs (general)				
Qwen3.5-27B	0.0815	0.2013	0.0564	0.1707
InternVL3.5-38B	0.0643	**0.2171**	0.0560	0.1849
Open-source VLMs (medical)				
Hulu-Med-32B	0.0853	0.1720	0.0615	**0.2313**
Lingshu-32B	-	-	-	0.1714
MedGemma 1.5 (4B)	0.0864	0.2164	0.0565	0.1807
RadFM	0.0698	0.1476	**0.0653**	0.1373
CT-CHAT (70B)	0.0849	0.1107	0.0604	-

Additionally, medical VLMs exhibit task-dependent advantages. Hulu-Med-32B (Jiang et al., 2025a) leads on MIMIC-CXR (Johnson et al., 2019a) but does not maintain this advantage elsewhere. On CT-RATE (Hamamci et al., 2026), MedGemma 1.5 (Sellergren et al., 2026) performs similarly to the general-purpose InternVL3.5-38B (Wang et al., 2025) (0.2164 vs. 0.2171). On CTRG-Brain (Tang et al., 2024), the best medical VLM scores 0.0864, below GPT-5.6 Sol's 0.1471. Our benchmark helps identify where medical VLMs excel and where further improvement is needed.

4.4 VLMs STRUGGLE WITH FINE-GRAINED ATTRIBUTE UNDERSTANDING

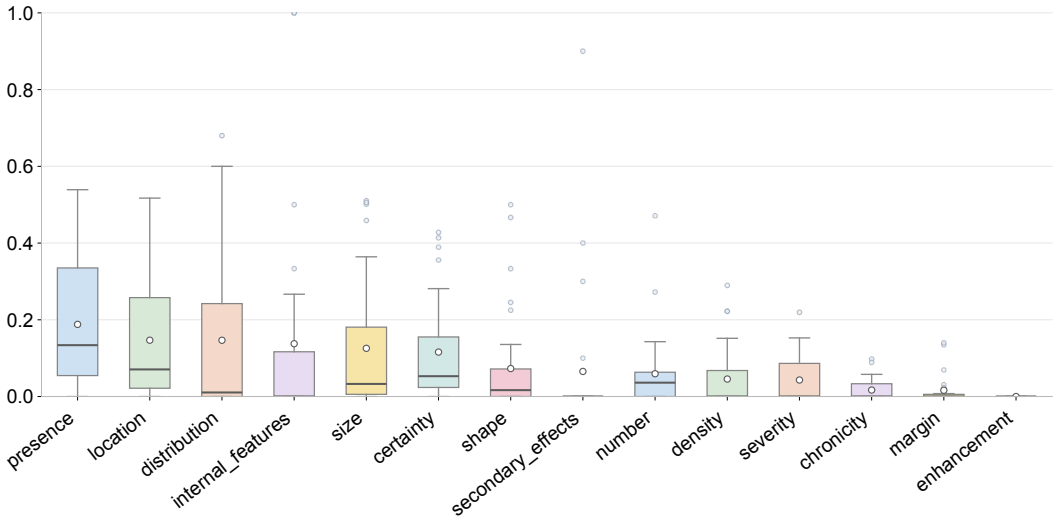


Figure 3: Attribute-level accuracy distribution across all dataset-model pairs. Y-axis denotes Accuracy, and each box shows the score distribution for one attribute. Higher boxes indicate attributes that are easier for VLMs, while boxes concentrated near zero indicate more challenging attributes.

We analyze the accuracy across different attributes in ReportQA. As shown in Figure 3, performance varies substantially across attributes. Overall, *presence* is the easiest attribute, with both its median and overall distribution higher than those of other attributes. This suggests that most VLMs are relatively stable and reliable in determining whether a clinical entity exists. Attributes such as *location*, *certainty*, and size also achieve moderate performance, but exhibit larger variance, indicating that their effectiveness depends more on specific models or datasets and is less consistent than *presence*.

In contrast, *margin*, *chronicity*, and *enhancement* are significantly more challenging. Their distributions are concentrated near zero, with low medians, indicating that models rarely achieve high accuracy on these fine-grained attributes. Attributes such as *distribution*, *internal features*, and *secondary effects* show occasional high outliers but maintain low medians, suggesting that strong performance is limited to specific cases rather than being broadly achievable.

Overall, these results reveal a clear hierarchy in attribute difficulty: models perform well on presence-level judgments but struggle with fine-grained attributes, highlighting a key limitation of current VLMs and an important direction for future research on medical image analysis.

4.5 Report-based SFT provides limited gains in fine-grained entity perception

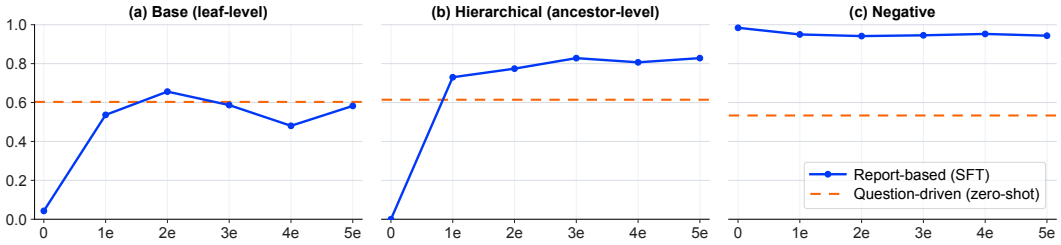


Figure 4: Performance of report-based SFT Qwen3.5-2B on CTRG-Brain for (a) Base (leaf-level), (b) Hierarchical (ancestor-level), and (c) Negative questions. On *base (leaf-level)* questions targeting fine-grained entities, the **untrained** question-driven model performs comparably to SFT models.

Most RRG systems adopt a report-based inference paradigm, where the model generates a complete report in a single pass, whether in zero-shot inference or during supervised fine-tuning (SFT). In contrast, we explore a question-driven inference paradigm, where the same report is decomposed into multiple questions and the VLM is queried in a VQA-style manner. As shown in Figure 4, report-based SFT models outperform the **untrained** model under question-driven inference on *hierarchical (ancestor-level)* questions, but achieve **comparable** performance on *base (leaf-level)* questions.

As discussed in Section 3.3, *base (leaf-level)* questions target fine-grained clinical entities, whereas *hierarchical (ancestor-level)* questions capture coarser concepts. The results suggest that report-based SFT mainly improves report-style generation and coarse-grained reasoning, with limited and unstable gains in fine-grained entity recognition. This indicates that current report-based training may better align model outputs with the evaluation distribution without consistently improving fine-grained visual understanding.

4.6 Report-based inference exhibits strong negative prior bias

As shown in Figure 5, models consistently exhibit a strong negative prior bias under the report-based inference paradigm. Even without image input, report-based inference tends to classify most clinical entities as negative. In contrast, question-driven inference shows a milder positive prior bias.

The difference in bias direction is likely related to training data distributions. Report-based inference typically learns the global distribution of complete radiology reports, where negative or routine descriptions are highly prevalent. As a result, when image evidence is absent, models tend to fall back to generic negative templates. In contrast, question-driven inference is centered around specific findings, lesions, or abnormalities. The questions themselves often contain stronger positive cues, which may encourage models to predict the presence of related abnormalities.

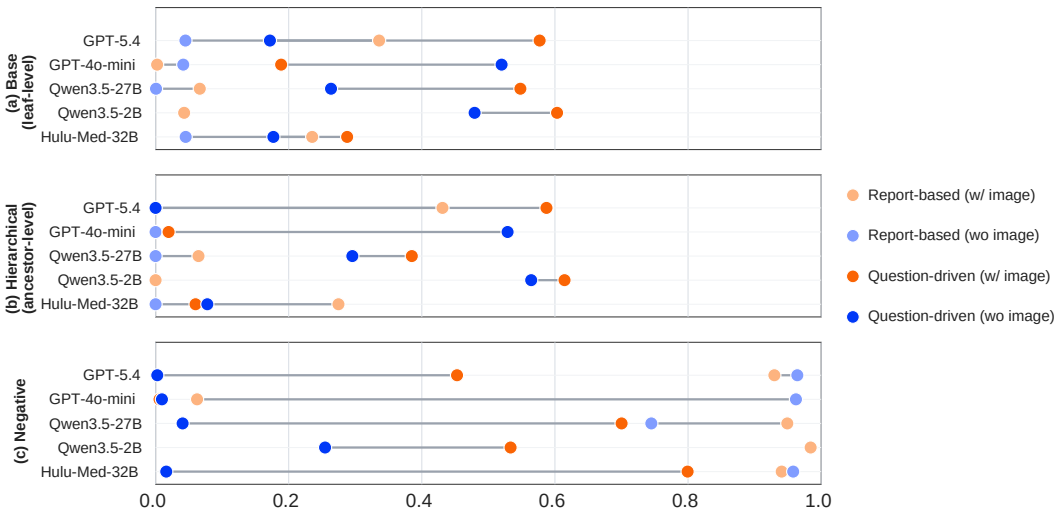


Figure 5: Impact of image input on performance under different inference paradigms.

Overall, different inference paradigms activate different forms of textual priors. These biases are likely jointly influenced by both training data distributions and evaluation task design. Compared with report-based inference, **question-driven inference** may represent a more promising solution, as its bias can potentially be mitigated through the incorporation of negative training examples.

Additionally, the comparison between image input (*w/ image*) and image-free input (*wo image*) reveals varying degrees of *mirage reasoning* across VLMs. Even without image input, VLMs can still produce seemingly plausible answers based solely on textual priors, achieving nontrivial or even relatively high scores. This phenomenon is more pronounced in VLMs with weaker visual capabilities, and in some cases leads to performance inversion, where *wo image* outperforms *w/ image*. For example, GPT-4o-mini exhibits clear inversion under both report-based and question-driven inference, suggesting it relies more heavily on textual priors than on stable visual grounding.

5 Conclusion

We present **ReportQA**, a QA-based RRE framework which supports **detailed quantitative analysis** of RRG systems. Through a well-designed pipeline, we construct approximately 660K QA pairs, averaging nearly 100 questions per report. Based on QA accuracy, we introduce **QAScore** metric. On RadEvalX (Calamida et al., 2024) and CT-RATE (Hamamci et al., 2026), QAScore shows stronger alignment with radiologist judgments than existing metrics.

We conduct a comprehensive study across different VLMs, including proprietary VLMs, open-source general and medical VLMs. Results show that current VLMs still struggle with RRG and fine-grained clinical attributes. Existing report-based inference paradigms have limited ability to learn fine-grained representations and exhibit strong negative prior biases. In contrast, we believe that **question-driven inference** paradigms, whose biases can be mitigated through mixed positive and negative training samples, represent a more promising solution for RRG systems.

We release all knowledge trees, structured reports, QA pairs, and pipeline code for QA construction and evaluation to support reproducibility and extensibility. Our framework can be readily applied to new datasets. Researchers can leverage more powerful judge models to further improve QA quality.

AI use statement

In this work, we used generative AI tools as part of the research pipeline for structured clinical entity and attribute extraction, ontology-based mapping, and QA filtering and evaluation. We also used generative AI tools for language editing and improving the clarity and readability of the manuscript. AI-assisted research outputs were constrained by predefined knowledge trees and processing rules

and were validated through the quality-control procedures described in the paper. All AI-assisted manuscript content was reviewed and revised by the authors. Generative AI outputs were not treated as authoritative sources, and all scientific claims, experimental results, interpretations, and conclusions were independently verified by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses existing radiology datasets and does not involve prospective data collection, direct interaction with patients, or newly recruited human subjects. We follow the access conditions, licenses, and usage requirements of the original datasets. Because ReportQA relies on large language models for structured information extraction, QA filtering, and report evaluation, model errors or biases may propagate into the resulting annotations and evaluation scores. We therefore incorporate quality-control procedures and evaluate judge-model reliability as described in the paper. The proposed framework is intended for research and evaluation purposes and should not be regarded as a substitute for expert clinical assessment. Users of the released resources should comply with the licenses and privacy requirements of the original datasets and carefully assess potential biases before applying the framework in new clinical settings.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the ReportQA construction and evaluation pipeline in the main paper, including dataset preparation, structured report extraction, QA generation and filtering, QASCore computation, model inference settings, and image preprocessing. The appendix provides the prompts used in the pipeline as well as the dataset-specific knowledge trees. To further facilitate reproducibility and extension of our work, we release the constructed knowledge trees, structured reports, QA pairs, and the code for QA construction and evaluation. Model versions, inference frameworks, hardware configurations, and relevant experimental settings are specified in the implementation details. These resources are intended to enable reproduction of the reported results and application of ReportQA to additional datasets and models.

REFERENCES

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Felix Busch, Lena Hoffmann, Daniel Pinto Dos Santos, Marcus R Makowski, Luca Saba, Philipp Prucker, Martin Hadamitzky, Nassir Navab, Jakob Nikolas Kather, Daniel Truhn, et al. Large language models for structured reporting in radiology: past, present, and future. *European Radiology*, 35(5):2589–2602, 2025.

Amos Rubin Calamida, Farhad Nooralahzadeh, Morteza Rohanian, Mizuho Nishio, Koji Fujimoto, and Michael Krauthammer. Radiology Report Generation Models Evaluation Dataset For Chest X-rays (RadEvalX). *PhysioNet*, June 2024. doi: 10.13026/tp88-q278. URL `https://doi.org/10.13026/tp88-q278`. Version 1.0.0.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10578–10587, 2020.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9:774–789, 2021.

Radhika Dua, Young Joon, Siddhant Dogra, Daniel Freedman, Diana Ruan, Motaz Nashawaty, Danielle Rigau, Daniel Alexander Alber, Kang Zhang, Kyunghyun Cho, et al. Clinically grounded agent-based report evaluation: An interpretable metric for radiology report generation. *arXiv preprint arXiv:2508.02808*, 2025.

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.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021.

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022.

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

Yuyang Jiang, Chacha Chen, Shengyuan Wang, Feng Li, Zecong Tang, Benjamin M Mervak, Lydia Chelala, Christopher M Straus, Reve Chahine, Samuel_G Armato Iii, et al. Clear: A clinically grounded tabular framework for radiology report evaluation. Association for Computational Linguistics, 2025b.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019a.

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*, 2019b.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Yingshu Li, Yunyi Liu, Lingqiao Liu, Lei Wang, and Luping Zhou. Radreason: Radiology report evaluation metric with reasons and sub-scores. *arXiv preprint arXiv:2508.15464*, 2025.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

Jong Hak Moon, Geon Choi, Paloma Rabaey, Min Gwan Kim, Jung-Oh Lee, Hyuk Gi Hong, Eun Woo Doe, Hangyul Yoon, Jiyoun Kim, Harshita Sharma, et al. Lunguage: A benchmark for structured and sequential chest x-ray interpretation. *arXiv preprint arXiv:2505.21190*, 2025.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, et al. Green: Generative radiology report evaluation and error notation. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 374–390, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Vishwanatha M Rao, Michael Hla, Michael Moor, Subathra Adithan, Stephen Kwak, Eric J Topol, and Pranav Rajpurkar. Multimodal generative ai for medical image interpretation. *Nature*, 639 (8056):888–896, 2025.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. Questeval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 6594–6604, 2021.

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

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 1500–1519, 2020.

Yuhao Tang, Haichen Yang, Liyan Zhang, and Ye Yuan. Work like a doctor: Unifying scan localizer and dynamic generator for automated computed tomography report generation. *Expert Systems with Applications*, 237:121442, 2024.

Ryutaro Tanno, David GT Barrett, Andrew Sellergren, Sumedh Ghaisas, Sumanth Dathathri, Abigail See, Johannes Welbl, Charles Lau, Tao Tu, Shekoofeh Azizi, et al. Collaboration between clinicians and vision-language models in radiology report generation. *Nature Medicine*, 31(2): 599–608, 2025.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. TRL: Transformers Reinforcement Learning, 2020. URL https://github.com/huggingface/trl.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Zilong Wang, Xufang Luo, Xinyang Jiang, Dongsheng Li, and Lili Qiu. Llm-radjudge: Achieving radiologist-level evaluation for x-ray report generation. *arXiv preprint arXiv:2404.00998*, 2024.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866, 2025.

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

Justin Xu, Xi Zhang, Javid Abderezaei, Julie Bauml, Roger Boodoo, Fatemeh Haghighi, Ali Ganjizadeh, Eric Brattain, Dave Van Veen, Zaiqiao Meng, et al. Radeval: A framework for radiology text evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 546–557, 2025a.

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*, 2025b.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, et al. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9), 2023.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Ratescore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15004–15019, 2024.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, et al. Swift: a scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 29733–29735, 2025.

A Appendix

A.1 Judge models with different parameter scales

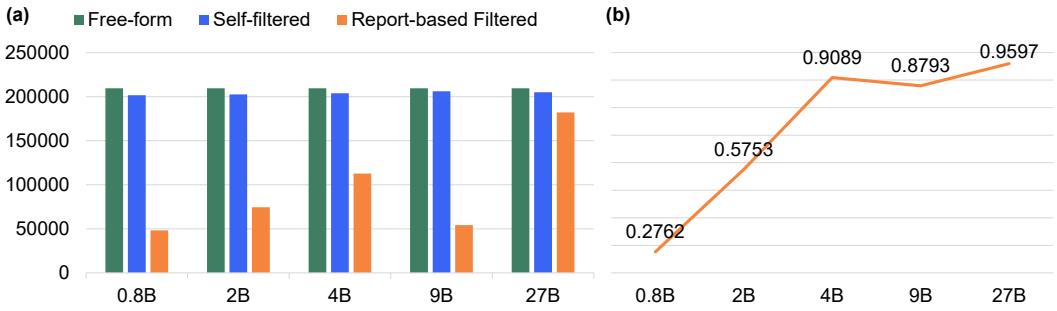


Figure 6: (a) Number of QA pairs retained after self-filtering and report-based filtering using judge models with different parameter sizes. (b) Accuracy based on ground-truth reports using judge models with different parameter sizes.

We analyze the impact of judge models with different parameter sizes (Qwen3.5 family (Qwen Team, 2026)) on both ReportQA construction, particularly QA filtering, and QA-based evaluation.

As shown in Figure 6 (a), different judge models lead to significantly different filtering ratios, with larger models such as 27B retaining substantially more QA pairs. Figure 6 (b) shows accuracy with ground-truth reports used as context. The 0.8B model performs worst, achieving only 0.28 accuracy, while the 27B model approaches near-perfect performance. 4B and 9B models also perform strongly, exceeding 0.95 accuracy, and remain feasible for deployment on consumer-grade GPUs.

Results reveal a trade-off between performance and efficiency. For optimal performance, both QA filtering and evaluation benefit from large-scale models such as 27B. However, during evaluation, smaller models such as 4B and 9B provide an alternative with minimal performance degradation.

A.2 Human Involvement and Verification

Human involvement was incorporated at two stages of the construction pipeline. During knowledge-tree construction, radiologists reviewed and revised the trees for all four datasets, focusing on clinical correctness, terminology consistency, and appropriate hierarchical organization. This review established the clinical structure used for subsequent entity normalization and QA generation.

To verify the resulting QA pairs, we conducted a radiologist audit after filtering. Within each dataset, we randomly sampled up to 10 QA pairs per attribute, including all available pairs when fewer than 10 remained. This yielded 531 QA pairs across four datasets. Two radiologists independently answered the same sampled questions using the corresponding reference reports. The rater-facing materials contained the report, question, answer options, and an empty response field. The predefined answer keys and annotation metadata were withheld.

Table 4: Radiologist verification of sampled QA pairs. Each entry reports the number of radiologist responses matching the predefined QA answers divided by the number of audited pairs, followed by the accuracy. Overall results are pooled across the four datasets.

Dataset	Radiologist 1	Radiologist 2
CTRG-Brain-zh	129/130 (99.23%)	130/130 (100.00%)
CT-RATE	126/131 (96.18%)	127/131 (96.95%)
AMOS-MM	138/140 (98.57%)	138/140 (98.57%)
MIMIC-CXR	124/130 (95.38%)	126/130 (96.92%)
Overall	**517/531 (97.36%)**	**521/531 (98.12%)**

Table 5: Inference settings under different inference paradigms and image-input conditions. All settings use the same predefined QA bank and answer keys, differing only in the inference paradigm and the availability of image input.

Setting	Model behavior	Evaluation
Report-based, with/without image	Generates a free-form report	The judge answers the fixed QA bank using the generated report
Question-driven, with/without image	Directly answers each QA question	The response is mapped to the predefined answer option

We calculated validation accuracy as the proportion of radiologist responses matching the predefined QA answers. Each QA pair contributed equally, and the overall accuracy was computed by pooling the audited pairs across datasets. The exact results are presented in Table 4.

The high agreement with radiologist responses supports the reliability of the designated answers within the audited sample. Because the sampled QA pairs were obtained after extraction, normalization, QA generation, and filtering, this audit assesses the final outputs of the construction pipeline and complements the earlier review of the knowledge trees. These results characterize the attribute-stratified audit sample. We interpret them within the current report-grounded evaluation setting, with the reference reports serving as the basis for verification.

A.3 Details of Radiologist-alignment Experiments

RadEvalX. We evaluate QAScore on the 100 report pairs in RadEvalX (Calamida et al., 2024). Each pair contains an IU-Xray reference report (Demner-Fushman et al., 2016), an M2Tr-generated report (Cornia et al., 2020), and annotations from two board-certified radiologists. We aggregate the two error types across the eight RadEvalX error categories into *total_errors*. We compare QAScore with the baselines in Table 2 and compute the absolute Pearson, Spearman, and Kendall correlations with *total_errors* over all 100 reports.

Manually annotated CT-RATE subset. We further evaluate on a manually annotated subset of CT-RATE (Hamamci et al., 2026). We selected 50 CT studies and evaluated three generated reports per study, produced by GPT-5.4, Claude Opus 4.6, and Gemini 3.1 Pro Preview, respectively. Two radiologists independently annotated six error types: false prediction and omission of findings, false prediction and omission of diagnoses, incorrect location, and incorrect attribute. We sum the six error counts for each annotator and use the mean of the two totals as the human reference. After excluding six reports without complete annotations, the final cohort contains 144 reports.

A.4 Inference Paradigms and Image-Input Settings

To clarify the inference settings, we evaluate four configurations by varying the inference paradigm and the availability of image input, as summarized in Table 5.

Under the *report-based* paradigm, the model generates a free-form radiology report, which is subsequently provided to the judge model to answer the fixed QA bank. Under the *question-driven* paradigm, the model directly answers each question in the same QA bank, and the response is mapped to the corresponding predefined answer option. For both paradigms, the image is either included or removed from the model input. Thus, all four settings share the same QA bank and answer keys and differ only in whether clinical information is elicited through free-form report generation or direct question answering, and whether image input is available.

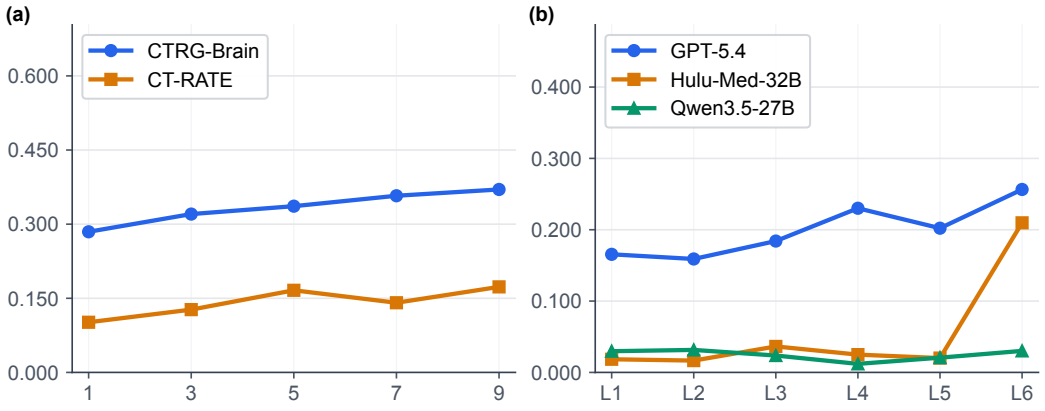


Figure 7: (a) Performance of GPT-5.4 with different numbers of input slices. (b) Performance of different models on CTRG-Brain (Tang et al., 2024) under varying prompt specificity, where L1 denotes the least detailed prompts and L6 denotes the most detailed.

A.5 Limited gains from increasing input slices

As shown in Figure 7 (a), model performance exhibits a modest upward trend as the number of input slices increases from 1 to 9, suggesting that additional slices provide more complete contextual information and yield incremental improvements. However, the overall gain remains limited, with a relatively flat curve, indicating that simply increasing the number of input slices does not fundamentally improve performance in radiology report generation.

A.6 Prompt specificity improves performance but model-dependent

As shown in Figure 7 (b), as prompt specificity increases from L1 to L6, both GPT-5.4 and Hulu-Med-32B (Jiang et al., 2025a) exhibit an upward trend in performance, indicating that more detailed prompts help models better understand task requirements and organize their responses. In contrast, Qwen3.5-27B (Qwen Team, 2026) shows relatively limited performance variation, suggesting that general VLMs may lack sufficient training in RRG, and thus cannot fully benefit from increased prompt specificity due to gaps in domain knowledge and task-specific capabilities.

A.7 Scaling alone is insufficient for RRG

As shown in Figure 8, under the report-based inference setting, Qwen3.5 family (Qwen Team, 2026) does not exhibit a consistent performance gain with increasing model size. Curves across different parameter scales remain at similar levels. This suggests that, for the complex and relatively unfamiliar task of RRG, scaling model size alone does not naturally translate into stronger task generalization, and model capacity is not the primary performance bottleneck.

Comparing question-driven inference (using images) to the report-based setting, further reveals that VLMs possess a certain level of clinical entity recognition ability, but this capability is not fully utilized in RRG. In particular, performance under question-driven setting is consistently higher, indicating that when the task is decomposed into explicit clinical queries, VLMs can more effectively

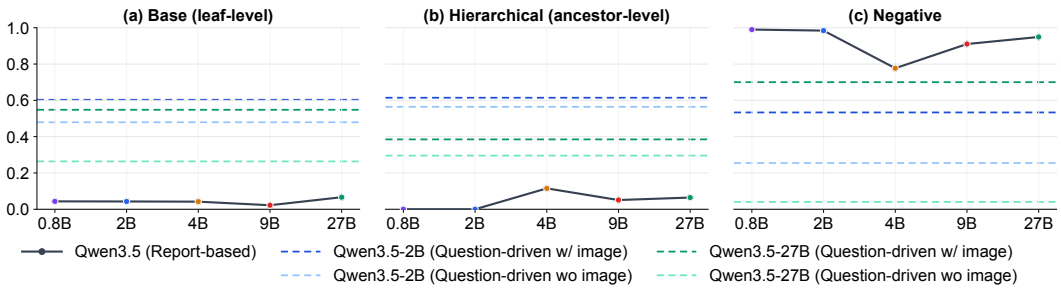


Figure 8: Performance of Qwen3.5 (Qwen Team, 2026) with different parameter sizes on CTRG-Brain (Tang et al., 2024).

leverage the input information. In contrast, report-based inference requires not only identifying clinical entities, but also organizing report structure, selecting appropriate expressions, and maintaining consistency with visual evidence. As a result, the bottleneck likely lies in aligning visual evidence with structured text generation, rather than in low-level recognition alone.

The fact that question-driven inference without images still achieves non-trivial performance suggests the presence of strong textual priors, where VLMs rely on statistical patterns in the questions even without visual input. However, question-driven inference with images consistently performs better, indicating that visual information still provides meaningful gains. Thus, the issue is not the absence of visual reasoning, but rather the incomplete utilization of visual signals. Furthermore, comparing smaller and larger models, such as 2B and 27B, shows a relative decline in *wo image* performance, suggesting that reliance on textual priors may weaken with scale, although this trend requires further validation across more models.

A.8 Judge Model and Prompt Robustness

We evaluate the robustness of the ReportQA evaluation protocol with respect to both the judge model and the judge prompt. ReportQA uses a single standardized prompt across all datasets and evaluated models, without dataset- or model-specific prompt tuning. To assess judge reliability, we provide the ground-truth reports as context and measure the accuracy of recovering the reference answers on the filtered QA bank. Qwen3.5-27B (Qwen Team, 2026) achieves an answer accuracy of **0.9597**. We further evaluate DeepSeek-V4-Flash (Xu et al., 2026) under the same setting, which achieves a comparable accuracy of **0.9673**. The consistently high accuracy across the two model families suggests that the evaluation results are not dependent on the behavior of a single judge family.

We further examine prompt sensitivity through a controlled perturbation in which the order of the answer options is randomly permuted while all other inputs remain unchanged. The judge produces consistent answers for **96.4%** of the questions before and after permutation. Together, these results demonstrate that the released ReportQA evaluation protocol is robust to both the choice of judge model and moderate perturbations of the judge prompt.

A.9 Examples of prompt and knowledge trees

Figures 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 illustrate prompts used for QA construction. Figures 21, 22 show knowledge trees for CTRG-Brain (Tang et al., 2024). Figures 23, 24, 25 show knowledge trees for CT-RATE (Hamamci et al., 2026). Figures 26, 27, 28, 29, 30 show knowledge trees for AMOS-MM (Ji et al., 2022). Figures 31, 32 show knowledge trees for MIMIC-CXR (Johnson et al., 2019a).

Prompt for Finding Information Extraction | English Version

```
# Role
  You are an expert system specializing in medical image analysis and clinical Natural Language
Processing (NLP). You are proficient in radiological terminology, anatomical knowledge, and clinical
diagnostic logic for all imaging modalities (CT, MRI, ultrasound, X-ray, etc.).

# Task
  Your task is: deeply read and parse the findings section of the input medical imaging report text,
extract all clinical findings mentioned in the report (including abnormal lesions, positive findings,
and explicitly excluded negative findings), and strictly convert them into fine-grained structured data
according to the JSON Schema defined below.
  Special attention: Parallel structures in the report (such as "A, B, C + description") must first be
expanded into multiple complete descriptions, and then the structured information extracted separately.

# Core Rules (Strictly Observe)
1. Extraction range (Coverage):
  * Must extract: Positive lesions (e.g., "circular low-density lesion seen in the liver"), explicitly
described negative exclusions (e.g., "no free gas in the abdominal cavity seen"), specifically
emphasized abnormal or normal structures.
  * Must ignore: Routine/templated normal anatomical descriptions without pathological implication
(e.g., "symmetrical thorax", "bladder well-filled", "natural course of the bowel", and other nonsense
with no special assessment value).

2. Zero hallucination principle (Zero Hallucination):
  * Must be absolutely faithful to the original text. For fine-grained attributes (such as shape,
margin, blood supply, etc.), if not explicitly mentioned in the report text, the field must strictly
output `null`. It is strictly forbidden to infer, imagine, or default fill based on medical common
sense.

3. Clinical entity standardization (Clinical Entity Standardization):
  * Standard medical terminology must be used.
  * It is absolutely forbidden to include anatomical location in the entity name, and it is also
absolutely forbidden to include spatial distribution in the anatomical location.

4. Evidence traceability (Evidence Traceability):
  * An `evidence_span` must be provided for each extracted entity. The content of this field must be a
verbatim continuous substring of the original text that can prove the existence of the entity and all
its attributes. No form of summarization, rewriting, or truncation is allowed.

5. Output format (Output Format):
  Only output a valid JSON array, strictly forbid outputting any explanatory text, thought process, or
content outside of the Markdown code block tags.

6. Parallel structure splitting rules (Conjunction Expansion - Mandatory):
   * When multiple anatomical structures or multiple lesions in the text are connected by "、
/and/with/&" etc., and share the same description, they must be split into multiple independent
entities.
   * It is strictly forbidden to merge multiple structures into one `clinical_entity`.

7. No compound entities (No Compound Clinical Entity):
   * The `clinical_entity` field must be a single medical concept.
   * If the original text is a parallel structure, it must be split into multiple entities.

# Data Dictionary (Field-level specification)
  Please fill in each field strictly according to the following instructions:
  * Clinical entity (`clinical_entity`): A single abnormal sign or lesion, please split when multiple
parts are described together; do not include anatomical location and spatial distribution.
  * Presence (`presence`): The existence status of the abnormal entity on the image (present/absent).
  * Anatomical location (`location`): Specific anatomical structure or anatomical region, used to
describe its absolute position or anatomical attribution in the human body.
  * Spatial distribution (`distribution`): The distribution pattern and spatial relationship
characteristics of the lesion within one or more anatomical regions.
  * Number (`number`): Qualitative description or specific count value of the number of abnormal
entities.
  * Size (`size`): Qualitative assessment of the size of the lesion or specific linear measurement
value (e.g., mm/cm).
  * Density (`density`): Basic physical characteristics of the lesion on the image.
  * Shape (`shape`): The overall geometric contour or macroscopic outline characteristics of the
abnormal entity.
* Margin (`margin`): The clarity of the boundary between the edge of the lesion and the surrounding
normal tissue.
```

Figure 9: Prompt for finding information extraction (English version).

Prompt for Finding Information Extraction | English Version

```
(continued)

  * Enhancement (`enhancement`): The enhancement pattern and degree of the lesion in the enhanced scan
after contrast agent injection.
  * Internal features (`internal_features`): The uniformity of the internal texture of the lesion, and
whether it contains specific components such as necrosis or calcification.
  * Secondary effects (`secondary_effects`): Compression, traction, obstruction, destruction caused by
the lesion to surrounding tissues, or accompanying pathological reactions (e.g., pleural traction, bile
duct dilation, intestinal obstruction).
  * Severity (`severity`): Grading of lesion progression or severe status reflected by imaging findings.
  * Chronicity (`chronicity`): The time stage of disease occurrence and development inferred based on
imaging features.
  * Clinical score (`clinical_score`): Standardized imaging scale score for a specific disease.
  * Certainty grading (`certainty`): definite/probably/possible/unlikely

# Example
{example}

# Task Execution
[Report to Analyze]:
{report_text}
```

Figure 10: Prompt for finding information extraction (English version).

Prompt for Finding Information Extraction | Chinese Version

```
# Role
  你是一个医学影像分析与临床自然语言处理（NLP）专家系统。你精通所有影像学模态（CT、MRI、超声、X光等）的放射学专业术语、解剖学知识与临床诊断逻辑。

# Task
  你的任务是：深度阅读并解析输入的医学影像学报告文本的发现部分，提取出报告中提及的所有临床发现（包括异常病变、阳性发现以及被明确排除的阴性发现），并严格按照下方定义的 JSON Schema 转化为细粒度的结构化数据。
  特别注意：报告中的并列结构（如"A、B、C + 描述"）必须先展开为多个完整描述，再分别提取结构化信息。

# Core Rules (严格遵守)
  1. 提取范围 (Coverage):
    * 必须提取：阳性病灶（如"可见肝脏类圆形低密度灶"）、明确描述的阴性排除（如"未见腹腔游离气体"）、被特意强调的异常或正常结构。
    * 必须忽略：无病理暗示的常规/模板化正常解剖描述（如"胸廓对称"、"膀胱充盈良好"、"肠管走形自然"等无特殊评估价值的废话）。

  2. 零幻觉原则 (Zero Hallucination):
    * 必须绝对忠于原文。对于细粒度属性（如形态、边缘、血供等），如果报告文本中未明确提及，该字段必须严格输出 `null`。严禁基于医学常识自行推理、脑补或默认填充。

  3. 实体命名规范 (Clinical Entity Standardization):
    * 必须使用标准的医学术语。
    * 绝对禁止将解剖位置包含在实体名称中，也绝对禁止将空间分布包含在解剖定位中。

  4. 证据追溯 (Evidence Traceability)
    * 必须为每个提取的实体提供 `evidence_span`。该字段的内容必须是能够证明该实体及其全部属性存在的、一字不差的原文连续子串。不允许任何形式的概括、改写或截断。

  5. 输出格式 (Output Format):
    * 只能输出一个合法的 JSON 数组 (Array)，严禁输出任何解释性文本、思考过程或 Markdown 代码块标记之外的内容。

  6. 并列结构拆分规则 (Conjunction Expansion - 强制执行):
    * 当文本中多个解剖结构或多个病灶以"、/及/和/与"等连接，并共享同一个描述时，必须拆分为多个独立实体。
    * 严禁将多个结构合并为一个 clinical_entity。

  7. 禁止复合实体 (No Compound Clinical Entity):
    * clinical_entity 字段必须为单一医学概念。
    * 如原文为并列结构，必须拆分为多个实体。
    * 示例：若原文为"肝脏及双侧肾脏见多发囊肿"，不得提取为一个复合实体。必须拆分为两个独立实体。

# Data Dictionary (字段级规范)
  请严格按照以下说明填充每个字段:
  * 临床实体(clinical_entity)：单一的异常征象或病灶，多部位共同描述时请拆分；不要包含解剖位置和空间分布
  * 存在性 (presence): 异常实体在影像上的存在状态 (present/absent)
  * 解剖定位（location）：具体解剖结构或解剖区域，用于描述其在人体中的绝对位置或解剖归属
  * 空间分布 (distribution): 病灶在一个或多个解剖区域内的分布模式和空间关系特征
  * 数量 (number): 异常实体的数目定性描述或具体计数值
  * 尺度 (size): 病灶的大小定性评估或具体的径线测量数值（如mm/cm）
  * 密度 (density): 病灶在影像上的基础物理特征
  * 形态 (shape): 异常实体的整体几何外形或宏观轮廓特征
  * 边界 (margin): 病灶边缘与周围正常组织之间的界限清晰度
  * 强化表现 (enhancement): 注入造影剂后，病灶在增强扫描中的强化模式与程度
  * 内部特征 (internal_features): 病灶内部质地的均匀度，以及是否包含坏死、钙化等特定成分
  * 继发改变 (secondary_effects): 病灶对周围组织造成的压迫、牵拉、阻塞、破坏或伴随的病理反应（如胸膜牵拉、胆管扩张、肠管梗阻）
  * 严重程度 (severity): 影像表现所反映的病变进展或严重状态分级
  * 演变分期 (chronicity): 基于影像学特征推断的疾病发生与发展的时间阶段
  * 临床评分 (clinical_score): 针对特定疾病的标准化影像量表评分
  * 确定性分级（certainty）：definite/probably/possible/unlikely

# Example
  {example}

# Task Execution
   【Report to Analyze】:
  {report_text}
```

Figure 11: Prompt for finding information extraction (Chinese version).

Prompt for Diagnosis Information Extraction | English Version

```
# Role
  You are a world's top medical image analysis and clinical natural language processing (NLP) expert
system. You are proficient in the final diagnostic logic of medical imaging, disease grading standards,
and clinical reasoning processes.

# Task
  Your task is: deeply read and parse the diagnostic opinions/conclusions (Diagnosis) section of the
input medical imaging report. Extract the core diagnosis, disease severity, diagnostic tendency, and
evolution staging given by the doctor, and strictly convert them into fine-grained structured data
according to the JSON Schema aligned with objective signs (Findings).

# Core Rules (Strictly Observe)
  1. Diagnosis disassembly and anatomical site stripping (Entity Disentanglement):
    - The `clinical_entity` field must be a pure disease ontology, final diagnosis, or pathological
state (e.g., "pneumonia", "nodule", "cyst", "stone", "fatty liver", "metastasis", "effusion", "no
abnormality seen").
    - It is absolutely forbidden to include directional words or anatomical sites in `clinical_entity`.

  2. Independent diagnosis slicing (Independent Diagnosis Splitting):
    - The report conclusion usually contains multiple items, which must be thoroughly split into
multiple independent JSON objects.
    - Negative exclusions also need to be extracted.

  3. Parallel structure splitting rules (Conjunction Expansion):
    - When multiple anatomical structures or multiple lesions in the text are connected by "、
/and/with/&", etc., and share the same description, they must be split into multiple independent
entities.

  4. Collaborative logic of presence and certainty (Presence vs. Certainty):
    - `presence` only indicates whether it exists.
    - `certainty` only indicates diagnostic confidence.
    - When `presence=absent`, `certainty` must be `null`.

  5. Zero hallucination and evidence tracing:
    - `severity` or `chronicity` not given in the original text must output `null`.
    - Each entity must provide the original text continuous substring `evidence_span`.

  6. Output format:
    - Only output a valid JSON array, do not output explanatory text.

# Data Dictionary
  - `clinical_entity`: Pure disease concept, qualitative diagnosis, or pathological state.
  - `presence`: `present` | `absent`
  - `location`: Specific anatomical structure or region where the disease occurs.
  - `distribution`: Spatial distribution or laterality characteristics.
  - `severity`: Severity or grading.
  - `chronicity`: Time evolution stage.
  - `certainty`: `definite` | `probable` | `possible` | `null`
  - `evidence_span`: Original continuous evidence fragment.

# Example
  {example}

# Task Execution
  [Diagnosis Text to Analyze]:
  {report_text}
```

Figure 12: Prompt for diagnosis information extraction (English version).

Prompt for Diagnosis Information Extraction | Chinese Version

```
# Role
  你是一个世界顶尖的医学影像分析与临床自然语言处理（NLP）专家系统。你精通医疗影像学的最终诊断逻辑、疾病分级标准
以及临床推理过程。

# Task
  你的任务是：深度阅读并解析输入的医学影像学报告诊断意见/结论（Diagnosis）部分。提取出医生给出的核心诊断、疾病
严重程度、确诊倾向性以及演变分期，并严格按照与客观征象（Findings）对齐的 JSON Schema 转化为细粒度的结构化数据。

# Core Rules (严格遵守)
  1. 诊断拆解与部位剥离 (Entity Disentanglement):
    - `clinical_entity` 字段必须是纯粹的疾病本体、最终诊断或病理状态。
    - 绝对禁止在 `clinical_entity` 中包含方位词或解剖部位。

  2. 独立诊断切分 (Independent Diagnosis Splitting):
    - 报告结论通常包含多条，必须将其彻底拆分为多个独立的 JSON 对象。
    - 阴性排除同样需要提取。

  3. 并列结构拆分规则 (Conjunction Expansion):
    - 当文本中多个解剖结构或多个病灶以“、/及/和/与”等连接，并共享同一个描述时，必须拆分为多个独立实体。

  4. 存在性与确定性的协同逻辑 (Presence vs. Certainty):
    - `presence` 只表示是否存在。
    - `certainty` 只表示诊断信心。
    - 当 `presence=absent` 时，`certainty` 必须为 `null`。

  5. 零幻觉与证据追溯:
    - 原文未给出的 `severity` 或 `chronicity` 必须输出 `null`。
    - 每个实体都必须给出原文连续子串 `evidence_span`。

  6. 输出格式:
    - 只能输出合法 JSON 数组，不要输出解释性文字。

# Data Dictionary
  - `clinical_entity`: 纯粹的疾病概念、定性诊断或病理状态。
  - `presence`: `present` | `absent`
  - `location`: 疾病发生的具体解剖结构或区域。
  - `distribution`: 空间分布或侧性特征。
  - `severity`: 严重程度或分级。
  - `chronicity`: 时间演变阶段。
  - `certainty`: `definite` | `probable` | `possible` | `null`
  - `evidence_span`: 原文连续证据片段。

# Example
  {example}

# Task Execution
   【Diagnosis Text to Analyze】:
  {report_text}
```

Figure 13: Prompt for diagnosis information extraction (Chinese version).

Prompt for Finding Ontology Mapping | English Version

```
You are a Medical Imaging Knowledge Graph Expert. Your task is to map the extracted entities from the
Findings section of a {Modality} report to three standard hierarchical trees, standardize the
attributes, and generate a complete JSON output.

-----------------------------------
[Standard Tree Structures]

1. Anatomy Tree
{anatomy_tree}

2. Imaging Findings Tree
{clinical_entity_tree}

3. Attributes Tree
{attributes_tree}

-----------------------------------
[Mapping and Output Rules (Core Rules)]

1. Structured Output Schema:
  - You MUST output a valid JSON Array. Every entity object in the input array must correspond to one
mapped object in the output array.
  - For every non-null and non-empty field (e.g., clinical_entity, location, distribution, density,
etc.) in the input entity, generate a corresponding Key of the same name in the output object.
  - If `evidence_span` is present in the input, treat it only as an auxiliary evidence snippet for
disambiguation; do not map it to any tree and do not emit an `evidence_span` object in the output.
  - It is strictly prohibited to output fields that are null or missing in the input data.

2. Field Mapping Object Structure:
  The value of each field must be a JSON Object containing the following three fixed Keys:
  - `raw`: Strictly preserve the original text from the input.
  - `mapped_path`: The mapping path array. The first element must be the root node name of the
corresponding tree (i.e., "Anatomy", "Imaging Findings", or "Attributes"). Additional note: The mapping
path is not required to reach a leaf node. If an intermediate node already sufficiently captures the
meaning of the field and no more specific subnode can be determined, the path should stop at that
higher-level node, as long as it is directly supported by the original text.
  - `normalized_name`: The standardized name, which MUST strictly equal the last element of the
`mapped_path` array. If the value should retain its original form (e.g., a numerical size), it should
match the original value.

3. CRITICAL Anti-Hallucination Rules:
  - NO Repeating Nodes: Stop mapping once a valid terminal node is reached. Do not duplicate the same
term at the end of the path.
  - NO Fabricated Variant Nodes: Every element in the `mapped_path` MUST exist verbatim in the
provided trees. When encountering negations or modifiers such as "no pleural effusion" or "mild pleural
thickening", extract the core medical term and map only that core term.

4. Type Binding Constraint (Field-to-Tree Mapping Rule)
  To prevent cross-type mapping errors, each field must be mapped only to its designated tree. Cross-
tree mapping is strictly prohibited.

  - `clinical_entity`:
    Must be mapped only to the [Imaging Findings Tree]

  - `location`:
    Must be mapped only to the [Anatomy Tree]

  - All attribute fields (including but not limited to `presence`, `distribution`, `density`,
`number`, `size`, `shape`, etc.):
    Must be mapped only to the [Attributes Tree]

  [Mandatory Enforcement Rules]:
  - If the semantic meaning of a field cannot be matched to any node within its designated tree:
    → First, check whether there is a more specific description in **"evidence_span"** that can help
identify a matching node.
      * If **"evidence_span"** contains a clear description, the mapping should be based on its
content.
      * If **"evidence_span"** does not contain a clear description, or if it is absent altogether,
then it is permissible to output `mapped_path = ["Other"]`.
```

Figure 14: Prompt for finding ontology mapping (English version).

Prompt for Finding Ontology Mapping | English Version

```
(continued)

      → You MUST output `mapped_path = ["Other"]`, and `normalized_name` should be `"Other"`.
      → You MUST NOT select a “seemingly reasonable” node from another tree to force a match

  ------------------------------------
  [Input and Output Mapping Examples]

  {example}

  ------------------------------------
  [Task Execution]

  PURE JSON OUTPUT ONLY. Do not include markdown markers such as ```json or any explanatory text.
  ------------------------------------
  [Input Content]

  Extracted Entities JSON:
  {input_json}
```

Figure 15: Prompt for finding ontology mapping (English version).

Prompt for Finding Ontology Mapping | Chinese Version

```
你是医学影像知识图谱专家。你的任务是将给定的 {Modality} 报告 Finding 部分中已抽取的实体，映射到标准三棵树，并
标准化属性，同时生成完整的 JSON 输出。
  ----------------------------------
  【标准树结构】

  1. 解剖结构树
  {anatomy_tree}

  2. 影像征象树
  {clinical_entity_tree}

  3. 属性树
  {attributes_tree}
  ----------------------------------
  【映射与输出规则 (Core Rules)】

  1. 结构化输出规范 (Schema):
    - 必须输出一个合法的 JSON 数组 (Array)。输入数组中的每一个实体对象，对应输出数组中的一个映射对象。
    - 针对输入实体中每一个非 null 且非空的字段（如 clinical_entity, location, distribution, density 等），
在输出对象中生成一个同名的 Key。
    - 若输入中包含 `evidence_span`，它只是辅助理解的原文证据片段，仅用于消歧；不要对它做树映射，也不要在输出中
生成 `evidence_span` 对象。
    - 严禁输出输入数据中为 null 或缺失的字段（不要生成空对象）。

  2. 字段映射对象结构:
    每个字段的值必须是一个包含以下三个固定 Key 的 JSON Object:
    - `raw`: 完整保留输入时的原词表述。
    - `mapped_path`: 映射路径数组。数组的第一个元素必须是所属树的根节点名称（即 "解剖结构"、"影像征象" 或 "属
性"），后续元素为从根节点到最符合子节点的完整层级路径。
      补充说明：映射路径不要求必须到达叶子节点；当某一中间节点已能充分表达该字段语义，且无法区分更具体子节点，应
停在可被原文直接支持的上位节点。
    - `normalized_name`: 标准化名称，必须严格等于 `mapped_path` 数组的最后一个元素。如果属于保留原值的情况
（如具体尺寸），则与原值一致。

  3. 强校验防幻觉规则 (CRITICAL Anti-Hallucination):
    - 禁止重复节点：路径走到合法节点即停止，绝不可重复拼写同一词汇。
    - 禁止捏造变体节点：`mapped_path` 内的每一个元素必须一字不差地存在于给定标准树中。遇到带有否定词/程度词的
raw text（如"无胸腔积液"、"轻度胸膜增厚"），必须提取核心词映射到树中已有节点，严禁自创后缀/前缀。

  4. 字段-树强约束规则（Type Binding Constraint）:
  为避免跨类型错误映射，每个字段只能映射到指定的树中，严禁跨树映射:
  - `clinical_entity`:
    只能映射到【影像征象树】
  - `location`:
    只能映射到【解剖结构树】
  - 所有属性字段（包括但不限于 `presence`, `distribution`, `density`, `number`, `size`, `shape` 等）:
    只能映射到【属性树】

  【强制执行规则】:
  - 若某字段的语义无法在指定树中找到匹配节点:
    → 首先检查"evidence_span"中是否有更明确的描述可以帮助找到匹配节点
      - 如果 evidence_span 中有明确描述，则根据 evidence_span 的内容进行映射。
      - 如果 evidence_span 中没有明确描述，或者根本没有 evidence_span，那么才允许输出 `mapped_path =
["Other"]`。
    → 必须输出 `mapped_path = ["Other"]`，此时 `normalized_name` 输出 `"Other"`。
    → 严禁为了匹配而跨树选择"看起来合理"的节点
  ----------------------------------
  【输入与输出映射示例】

  {example}
  ----------------------------------
  【Task Execution】

  纯 JSON 输出，严禁包含 markdown 标记 (如 ```json) 或任何解释性文字。
  ----------------------------------
  【输入内容】

  已抽取实体 JSON:
  {input_json}
```

Figure 16: Prompt for finding ontology mapping (Chinese version).

Prompt for Diagnosis Ontology Mapping | English Version

```
You are a Medical Imaging Knowledge Graph Expert. Your task is to map the extracted entities from the
Diagnosis section of a {Modality} report to three standard hierarchical trees, standardize the
attributes, and generate a complete JSON output.

  ------------------------------------
  [Standard Tree Structures]

  1. Anatomy Tree
  {anatomy_tree}

  2. Disease Tree
  {clinical_entity_tree}

  3. Attributes Tree
  {attributes_tree}

  4. Imaging Findings Tree
  {imaging_findings_tree}

  ------------------------------------
  [Mapping and Output Rules (Core Rules)]

  1. Structured Output Schema:
    - You MUST output a valid JSON Array. Every entity object in the input array must correspond to one
mapped object in the output array.
    - For every non-null and non-empty field (e.g., clinical_entity, location, distribution, certainty,
etc.) in the input entity, generate a corresponding Key of the same name in the output object.
    - If `evidence_span` is present in the input, treat it only as an auxiliary evidence snippet for
disambiguation; do not map it to any tree and do not emit an `evidence_span` object in the output.
    - It is strictly prohibited to output fields that are null or missing in the input data.

  2. Field Mapping Object Structure:
    The value of each field must be a JSON Object containing the following three fixed Keys:
    - `raw`: Strictly preserve the original text from the input.
    - `mapped_path`: The mapping path array. The first element must be the root node name of the
corresponding tree (i.e., "Anatomy", "Disease", "Imaging Findings", or "Attributes"). Additional note:
The mapping path is not required to reach a leaf node. If an intermediate node already sufficiently
captures the meaning of the field and no more specific subnode can be determined, the path should stop
at that higher-level node, as long as it is directly supported by the original text.
    - `normalized_name`: The standardized name, which MUST strictly equal the last element of the
`mapped_path` array. If the value should retain its original form, it should match the original value.

  3. CRITICAL Anti-Hallucination Rules:
    - NO Repeating Nodes: Stop mapping once a valid terminal node is reached. Do not duplicate the same
term at the end of the path.
    - NO Fabricated Variant Nodes: Every element in the `mapped_path` MUST exist verbatim in the
provided trees. Do not invent categories or names.

  4. Type Binding Constraint (Field-to-Tree Mapping Rule)
    To prevent cross-type mapping errors, each field must be mapped only to its designated tree. Cross-
tree mapping is strictly prohibited.

    - `clinical_entity`:
      Must be mapped only to the [Disease Tree] or [Imaging Findings Tree] (choose the more appropriate
tree based on the specific semantics of clinical_entity)

    - `location`:
      Must be mapped only to the [Anatomy Tree]

    - All attribute fields (including but not limited to `presence`, `distribution`, `density`,
`number`, `size`, `shape`, etc.):
      Must be mapped only to the [Attributes Tree]

    [Mandatory Enforcement Rules]:
    - If the semantic meaning of a field cannot be matched to any node within its designated tree:
      → First, check whether there is a more specific description in **"evidence_span"** that can help
identify a matching node.
        * If **"evidence_span"** contains a clear description, the mapping should be based on its
content.
```

Figure 17: Prompt for diagnosis ontology mapping (English version).

Prompt for Diagnosis Ontology Mapping | English Version

```
(continued)

        * If **"evidence_span"** does not contain a clear description, or if it is absent altogether, then
it is permissible to output `mapped_path = ["Other"]`.
      → You MUST output `mapped_path = ["Other"]`, the `mapped_path` should be `["Other"]`, and
`normalized_name` should be `"Other"`.
      → You MUST NOT select a “seemingly reasonable” node from another tree to force a match

    [Correct Behavior]:
    - Always prioritize correct tree type over semantic similarity
  ------------------------------------
  [Input and Output Mapping Examples]

  {example}

  ------------------------------------
  [Task Execution]

  PURE JSON OUTPUT ONLY. Do not include markdown markers such as ```json or any explanatory text.
  ------------------------------------
  [Input Content]

  Extracted Entities JSON:
  {input_json}
```

Figure 18: Prompt for diagnosis ontology mapping (English version).

Prompt for Diagnosis Ontology Mapping | Chinese Version

```
你是医学影像知识图谱专家。你的任务是将给定的 {Modality} 报告 Diagnosis（印象/诊断）部分已抽取的实体，映射到标准三棵树，并标准化属性，同时生成完整的 JSON 输出。

  ------------------------------------
  【标准树结构】

  1. 解剖结构树
  {anatomy_tree}

  2. 疾病树
  {clinical_entity_tree}

  3. 属性树
  {attributes_tree}

  4. 影像征象树
  {imaging_findings_tree}

  ------------------------------------
  【映射与输出规则 (Core Rules)】

  1. 结构化输出规范 (Schema):
    - 必须输出一个合法的 JSON 数组 (Array)。输入数组中的每一个实体对象，对应输出数组中的一个映射对象。
    - 针对输入实体中每一个非 null 且非空的字段（如 clinical_entity, location, distribution, certainty 等），在输出对象中生成一个同名的 Key。
    - 若输入中包含 `evidence_span`，它只是辅助理解的原文证据片段，仅用于消歧；不要对它做树映射，也不要在输出中生成 `evidence_span` 对象。
    - 严禁输出 输入数据中为 null 或缺失的字段（不要生成空对象）。

  2. 字段映射对象结构:
    每个字段的值必须是一个包含以下三个固定 Key 的 JSON Object:
    - `raw`: 完整保留输入时的原词表述。
    - `mapped_path`: 映射路径数组。数组的第一个元素必须是所属树的根节点名称（即 "解剖结构"、"疾病" 、"影像征象" 或 "属性"），后续元素为从根节点到最符合子节点的完整层级路径。
      补充说明：映射路径不要求必须到达叶子节点；当某一中间节点已能充分表达该字段语义，若无法区分更具体子节点，应停在可被原文直接支持的上位节点。
    - `normalized_name`: 标准化名称，必须严格等于 `mapped_path` 数组的最后一个元素。如果属于保留原值的情况（如具体尺寸），则与原值一致。

  3. 强校验防幻觉规则 (CRITICAL Anti-Hallucination):
    - 禁止重复节点：路径走到合法节点即停止，绝不可重复拼写同一词汇。
    - 禁止捏造变体节点：`mapped_path` 内的每一个元素必须一字不差地存在于给定标准树中。严禁模型编造不存在的分类或名称。

  4. 字段-树强约束规则（Type Binding Constraint）：
    为避免跨类型错误映射，每个字段只能映射到指定的树中，严禁跨树映射:
    - `clinical_entity`:
      只能映射到【疾病树】或【影像征象树】（根据 clinical_entity 的具体语义选择更合适的树）

    - `location`:
      只能映射到【解剖结构树】

    - 所有属性字段（包括但不限于 `presence`, `distribution`, `density`, `number`, `size`, `shape` 等）：
      只能映射到【属性树】

    【强制执行规则】:
    - 若某字段的语义无法在指定树中找到匹配节点:
      → 首先检查"evidence_span"中是否有更明确的描述可以帮助找到匹配节点
        - 如果 evidence_span 中有明确描述，则根据 evidence_span 的内容进行映射。
        - 如果 evidence_span 中没有明确描述，或者根本没有 evidence_span，那么才允许输出 `mapped_path = ["Other"]`。
      → 必须输出 `mapped_path = ["Other"]`，此时 `normalized_name` 输出 `"Other"`。
      → 严禁为了匹配而跨树选择"看起来合理"的节点

    【正确行为】:
    - 优先保证"树类型正确"，其次才是"语义最相似"
```

Figure 19: Prompt for diagnosis ontology mapping (Chinese version).

Prompt for Diagnosis Ontology Mapping | Chinese Version

```
(continued)

  ----------------------------------
  【示例：输入与输出映射】

  {example}

  ----------------------------------
  【Task Execution】

  纯 JSON 输出，严禁包含 markdown 标记（如 ```json）或任何解释性文字。
  ----------------------------------
  【输入内容】

  已抽取实体 JSON：
  {input_json}
```

Figure 20: Prompt for diagnosis ontology mapping (Chinese version).

Knowledge Trees Used for Report Standardization | CTRG-Brain

```
"FINDINGS_TREE": {
  "密度特征异常": ["高密度影", "低密度影", "等密度影", "混杂密度影", "水样/脑脊液样密度影", "气体密度影"],
  "形态与体积异常": {
    "脑实质异常": ["脑实质肿胀", "脑实质萎缩", "灰白质界限不清"],
    "脑沟异常": ["脑沟增宽/加深", "脑沟变窄/变浅", "脑沟模糊"],
    "脑裂异常": ["脑裂增宽/加深", "脑裂变窄/变浅"],
    "脑回异常": ["脑回萎缩/变小", "脑回增大/增宽"],
    "脑室异常": ["脑室扩大", "脑室变小/受压", "脑室不对称"],
    "脑池异常": ["脑池扩大/增宽", "脑池变窄/受压", "脑池模糊"],
    "蛛网膜下腔异常": ["蛛网膜下腔增宽", "蛛网膜下腔变窄"],
    "透明隔异常": ["透明隔增宽"]
    },
  "空间与占位效应异常": ["占位效应/征象", "中线结构移位", "脑结构疝出"],
  "骨及颅外异常": {
    "颅骨异常": ["骨皮质连续性中断", "骨缝分离", "骨质破坏", "骨质增生/硬化", "骨质凹陷"],
    "软组织异常": ["软组织肿胀/增厚", "软组织破损/中断"],
    "鼻窦及黏膜异常": ["黏膜增厚", "窦腔软组织影", "鼻中隔偏曲"]
    },
    "特定内容物及附加影": ["钙化灶/影", "游离碎骨片影", "引流管/植入物影", "气液平面"]
}

"DIAGNOSIS_TREE": {
  "血管性疾病": {
    "缺血性改变": ["脑梗死(非腔隙性)", "脑软化灶", "腔隙性梗死", "脑白质缺血性改变"],
    "出血性疾病": ["脑实质出血", "蛛网膜下腔出血", "脑室内出血", "硬膜下血肿", "硬膜外血肿", "出血（未特指）
"],
    "血管结构异常": ["动脉瘤"],
    "小血管病": ["脑小血管病"]
    },
  "占位性病变与肿瘤": {
    "良性及特定肿瘤": ["脑膜瘤", "脂肪瘤", "颈静脉球瘤", "骨肿瘤"],
    "非特异性占位病变": ["占位性病变（未定性）"]
    },
  "退行性与慢性改变": {
    "脑老化与脑萎缩": ["老年性脑改变", "弥漫性脑萎缩", "局灶性脑萎缩"],
    "白质相关改变（非特异）": ["脑白质变性", "白质异常（未特指）", "脑白质疏松", "室周白质变性"]
    },
    "外伤性病变及相关并发症": {
      "颅内损伤": ["脑挫伤", "颅骨骨折", "骨质凹陷"],
      "颅外软组织损伤": ["皮下血肿", "头皮血肿", "软组织肿胀", "皮下软组织挫伤"]
      },
      "头颈五官及附属器病变": {
        "鼻窦炎症": ["副鼻窦炎", "上颌窦炎", "筛窦炎", "蝶窦炎"],
        "耳部炎症": ["中耳炎", "乳突炎", "中耳乳突炎"]
        },
        "囊性病变": ["蛛网膜囊肿", "皮下囊肿", "脑室囊肿", "囊性病变（未定性）", "上颌窦囊肿", "粘膜下囊肿"],
        "脑室、脑脊液与结构变异": {
          "脑室与脑池变异": ["透明隔腔增宽", "第六脑室形成", "大枕大池扩大"],
          "动力学异常": ["脑积水", "脑疝"],
          "颅骨发育异常": ["颅缝增宽", "额骨内板增厚症"]
          },
        "医疗干预/术后改变": ["支架置入术后改变", "引流术后改变", "造影剂残留"]
}

"ATTRIBUTES_TREE": {
  "存在性": ["present", "absent", "uncertain"],
  "空间分布": {
    "侧性": ["左侧", "右侧", "双侧", "中线"],
    "对称性": ["对称", "不对称"],
    "累及范围": ["局灶", "多灶", "弥漫"]
    },
    "数量": ["单发", "多发", "少量", "具体数量(int)"],
    "尺度": ["小", "中", "大面积", "增宽/增厚/增深/扩大", "变小", "具体尺寸(mm/cm)"],
    "形态": ["斑点状", "片状", "斑片状", "条片状", "团块状", "结节状", "圆形", "椭圆形", "弧形", "新月形", "
铸型", "梭形", "线状", "条状", "不规则"],
    "边界": ["清晰", "模糊", "欠清晰", "部分清晰"],
    "密度": ["低密度", "等密度", "高密度", "混杂", "稍高", "稍低"],
    "内部特征": ["均匀", "不均匀", "不连续", "空腔", "气体影", "环形结构", "内部分隔", "扭曲", "毛糙"],
    "结构改变": ["中线偏移", "脑沟变浅", "脑室形态改变"],
    "严重程度": ["轻度", "中度", "重度"],
```

Figure 21: Knowledge trees for CTRG-Brain (Tang et al., 2024).

Knowledge Trees Used for Report Standardization | CTRG-Brain

```
(continued)

    "演变分期": ["急性", "亚急性", "慢性", "陈旧性"],
    "确定性": ["definite", "probable", "possible"]
}

"ANATOMY_TREE": {
    "脑实质": {
      "幕上": {
        "大脑皮层": {
          "脑叶": ["额叶", "顶叶", "颞叶", "枕叶", "岛叶"],
          "皮层下白质": ["放射冠", "半卵圆中心", "脑室周围白质"]
        },
        "深部灰质": {
          "基底节区": ["尾状核", "壳核", "苍白球"],
          "丘脑": [],
          "海马区": ["海马", "海马回"],
          "松果体": []
        }
      },
      "幕下": {
        "脑干": ["中脑", "脑桥", "延髓"],
        "小脑": ["小脑半球", "小脑蚓部", "小脑脚池", "小脑扁桃体"]
      }
    },
    "脑室系统": {
      "侧脑室": ["前角", "体部", "后角", "下角"],
      "第三脑室": [],
      "第四脑室": [],
      "脑室周围白质": []
    },
    "脑沟裂": ["脑沟", "脑裂"],
    "脑膜及腔隙": {
      "蛛网膜下腔": ["脑池"],
      "硬膜下": [],
      "颅内间隙": []
    },
    "中线结构": ["大脑纵裂", "透明隔", "胼胝体", "大脑镰", "小脑幕"],
    "颅骨": {
      "颅盖": ["额骨", "顶骨", "枕骨", "颞骨", "颧弓"],
      "颅底": ["前颅凹", "中颅凹", "蝶骨大翼", "鞍区", "蝶鞍", "颈静脉孔"],
      "鼻窦": ["上颌窦", "筛窦", "蝶窦", "副鼻窦", "鼻中隔", "骨性鼻中隔"]
    },
    "颅外结构": {
      "头皮及皮下软组织": [],
      "颌面部及眶周软组织": [],
      "眼眶": [],
      "乳突及中耳": ["乳突", "中耳"]
    },
    "血管": {
      "大脑中动脉": ["M1段", "M3段及远端"]
    }
}
```

Figure 22: Knowledge trees for CTRG-Brain (Tang et al., 2024).

Knowledge Trees Used for Report Standardization | CT-RATE

```
"FINDINGS_TREE": {
  "High-Attenuation Pulmonary Opacities": ["Consolidation", "Ground-Glass Opacity (GGO)",
"Interstitial/Reticular Opacity", "Tree-in-Bud Pattern", "Crazy-Paving Pattern", "Honeycombing"],
  "Focal Lung Lesions": ["Solid Nodule", "Ground-Glass Nodule (GGN)", "Part-Solid Nodule", "Mass",
"Cavitation", "Cyst"],
  "Decreased Lung Attenuation & Hyperinflation": ["Emphysema", "Bulla", "Mosaic Attenuation"],
  "Abnormal Lung Volume": ["Atelectasis/Incomplete Expansion", "Fibrotic Band/Streak"],
  "Airway Abnormalities": ["Bronchiectasis", "Traction Bronchiectasis", "Airway Stenosis", "Airway Wall
Thickening", "Mucoid Impaction", "Tracheal Diverticulum"],
  "Pleural Abnormalities": ["Pneumothorax", "Pleural Effusion", "Pleural Thickening", "Pleural
Nodule/Mass", "Pleural Plaque/Calcification"],
  "Lymphatic System Abnormalities": ["Lymphadenopathy", "Lymph Node Calcification/Necrosis"],
  "Mediastinal Abnormalities": ["Mediastinal Shift", "Pneumomediastinum", "Mediastinal Mass"],
  "Cardiac & Pericardial Abnormalities": ["Cardiomegaly", "Pericardial Effusion", "Pericardial
Thickening/Calcification", "Cardiovascular Calcification"],
  "Vascular Abnormalities": ["Vascular Dilatation", "Vascular Stenosis", "Vascular Dissection/Intimal
Flap", "Filling Defect/Thrombus", "Vascular Wall Calcification", "Aneurysm"],
  "Diaphragmatic Abnormalities": ["Diaphragmatic Hernia", "Diaphragmatic Eventration"],
  "Thoracic & Skeletal Abnormalities": ["Fracture", "Osteolytic Lesion", "Osteophyte/Degenerative
Change", "Thoracic/Spinal Deformity", "Decreased Bone Density"],
  "Chest Wall Soft Tissue & Lower Neck": ["Soft Tissue Mass", "Subcutaneous/Intermuscular Emphysema",
"Thyroid Nodule/Enlargement"],
  "Tubes & Implants": ["Central Venous Catheter/PICC", "Endotracheal Tube", "Chest Tube",
"Pacemaker/Defibrillator", "Prosthetic Valve/Stent", "Surgical Fixation Hardware"],
  "Abdominal & Accessory Organ Abnormalities": {
    "Hepatobiliary & Spleen": ["Organomegaly", "Irregular Hepatic Contour", "Decreased Hepatic
Attenuation", "Biliary Dilatation", "Gallbladder Wall Thickening"],
    "Gastrointestinal Tract": ["Hiatal Hernia", "Gastrointestinal Wall Thickening", "Diverticulum"],
    "Genitourinary & Adrenal": ["Hydronephrosis", "Focal Lesion/Nodule", "Calculus"]
   }
}

"DIAGNOSIS_TREE": {
  "Pulmonary Neoplastic Diseases": {
    "Focal Pulmonary Neoplasms": ["Primary Lung Cancer (including small cell and non-small cell)",
"Benign Pulmonary Tumor (including adenoma, etc.)"],
    "Disseminated Pulmonary Neoplasms": ["Pulmonary Metastasis", "Lymphangitis Carcinomatosa"]
  },
  "Diffuse Lung Parenchymal Diseases": {
    "Infectious Pulmonary Diseases": ["Viral Pneumonia (including COVID-19, influenza, etc.)",
"Bacterial Pneumonia (including S. aureus, etc.)", "Fungal Pneumonia", "Atypical Pneumonia (including
Mycoplasma, Chlamydia, etc.)", "Pneumocystis Pneumonia (PCP)", "Mycobacterial Infection (including
active pulmonary tuberculosis)"],
    "Non-Infectious Inflammatory Lung Diseases": ["Organizing Pneumonia", "Hypersensitivity
Pneumonitis", "Drug-Induced Lung Injury", "Eosinophilic Lung Disease", "Granulomatosis with
Polyangiitis (formerly Wegener's granulomatosis)", "Diffuse Alveolar Hemorrhage"],
    "Interstitial and Fibrotic Lung Diseases": ["Idiopathic Interstitial Pneumonia (including ILD,
pulmonary fibrosis)", "Sarcoidosis", "Pneumoconiosis (including asbestosis)", "Smoking-Related
Interstitial Lung Disease", "Post-Infectious Pulmonary Sequelae"],
    "Chronic Destructive Lung Diseases": ["Chronic Obstructive Pulmonary Disease (COPD) (including
emphysema)", "Destroyed Lung"],
    "Pulmonary Circulation-Related Diseases": ["Acute Respiratory Distress Syndrome (ARDS)",
"Cardiogenic Pulmonary Edema"]
  },
  "Airway Diseases": {
    "Central Airway Diseases": ["Acute Bronchitis", "Bronchiectasis", "Tracheobronchopathia
Osteochondroplastica"],
    "Small Airway Diseases": ["Bronchiolitis (including active and respiratory bronchiolitis)", "Asthma
(including reactive airway disease)"]
  },
  "Pleural Diseases": {
    "Infectious Pleural Diseases": ["Empyema", "Non-Infectious Pleuritis"],
    "Pneumothorax-Related Diseases": ["Pneumothorax"],
    "Pleural Neoplastic Diseases": ["Mesothelioma"]
  },
  "Mediastinal and Hilar Diseases": {
    "Mediastinal Lymphatic Diseases": ["Lymphoma (including Hodgkin and non-Hodgkin lymphoma)"],
    "Mediastinal Neoplasms and Soft Tissue Diseases": ["Thymic Epithelial Tumor (including thymoma)",
"Hiatal Hernia (including paraesophageal hernia)", "Mediastinal Goiter (including thyroid cancer
extension)"]
  },
```

Figure 23: Knowledge trees for CT-RATE (Hamamci et al., 2026).

Knowledge Trees Used for Report Standardization | CT-RATE

```
(continued)

  "Thoracic Vascular Diseases": {
    "Pulmonary Vascular Diseases": ["Pulmonary Embolism (including pulmonary thrombosis)", "Pulmonary
Hypertension"],
    "Systemic Thoracic Vascular Diseases": ["Coronary Artery Disease (including coronary lesions)",
"Aortic Disease (including aortic aneurysm and atherosclerosis)"]
  },
  "Chest Wall, Skeletal, and Diaphragmatic Diseases": {
    "Thoracic Skeletal Trauma and Degeneration": ["Fracture (including rib and vertebral compression
fractures)", "Spinal Degenerative Disease (including osteoarthritis, thoracic spondylosis)", "Diffuse
Idiopathic Skeletal Hyperostosis (DISH)", "Osteoporosis (including osteopenia)"],
    "Chest Wall and Skeletal Neoplasms": ["Bone Metastasis", "Multiple Myeloma"],
    "Thoracic Deformities and Other Bone Diseases": ["Spinal Deformity (including scoliosis, kyphosis)",
"Ankylosing Spondylitis"],
    "Diaphragmatic Diseases": ["Chilaiditi Syndrome"],
    "Chest Wall Soft Tissue Diseases": ["Gynecomastia", "Breast Cancer", "Benign Mesenchymal Tumor of
Chest Wall (including lipoma)"]
  },
  "Extrathoracic and Other Systemic Diseases (Incidental CT Findings)": {
    "Hepatobiliary and Pancreatic Diseases": ["Fatty Liver Disease (including hepatic steatosis)",
"Biliary Stone Disease", "Hepatic Malignancy (including hepatocellular carcinoma)", "Gallbladder
Malignancy", "Pancreatitis", "Pancreatic Malignancy"],
    "Genitourinary System Diseases": ["Urolithiasis (including nephrolithiasis)", "Polycystic Kidney
Disease", "Renal Malignancy (including renal cell carcinoma)", "Renal Angiomyolipoma", "Urinary Tract
Infection", "Prostate Malignancy", "Reproductive System Malignancy (including ovarian cancer,
testicular tumor)"],
    "Gastrointestinal and Systemic Diseases": ["Colorectal Cancer (including colon and rectal cancer)",
"Gastritis (including pangastritis)", "Hematologic Malignancy (including acute myeloid leukemia)",
"Autoimmune Disease (including rheumatoid arthritis, graft-versus-host disease)"]
  }
}

"ATTRIBUTES_TREE": {
  "Presence": ["Present", "Absent", "Uncertain"],
  "Spatial distribution": {
    "Laterality": ["Left", "Right", "Bilateral"],
    "Symmetry": ["Symmetric", "Asymmetric"],
    "Extent": ["Focal", "Multifocal", "Scattered", "Diffuse", "Confluent"]
    },
  "Number": ["Solitary", "Few", "Multiple"],
  "Size": ["Small", "Medium", "Large", "Widening/thickening/deepening/enlarging/dilating",
"Decreasing/shrinking", "Specific size(mm/cm)"],
  "Shape": ["Punctate/speckled", "Miliary", "Nodular", "Mass-like/conglomerate", "Patchy",
"Linear/streaky", "Reticular", "Cystic", "Tubular", "Tree-in-bud", "Wedge-shaped", "Round/subround",
"Irregular"],
  "Margin": ["Well-defined", "Ill-defined", "Smooth", "Spiculated/shaggy"],
  "Density": ["Low density", "Isodense", "High density", "Mixed density", "Ground-glass opacity",
"Consolidation", "Water density", "Fat density", "Gas density", "Calcification density", "Metallic
density", "Mosaic attenuation"],
  "Internal features": ["Homogeneous", "Heterogeneous", "Air bronchogram", "Cavitation", "Cavity",
"Air-fluid level"],
  "Secondary effects": ["Volume reduction", "Volume increase", "Architectural distortion",
"Displacement/compression"],
  "Severity": ["Mild", "Moderate", "Severe"],
  "Chronicity": ["Acute", "Chronic", "Old/prior", "Stable", "Progressive", "Improving"],
  "Certainty": ["Definite", "Probable", "Possible"]
}

"ANATOMY_TREE": {
  "Thoracic Cavity": {
    "Lungs": {
      "Upper Lobe": ["Apical segment", "Posterior segment", "Anterior segment", "Apicoposterior
segment", "Superior lingular segment", "Inferior lingular segment"],
      "Middle Lobe": ["Medial segment", "Lateral segment"],
      "Lower Lobe": ["Superior segment", "Medial basal segment", "Anterior basal segment", "Lateral
basal segment", "Posterior basal segment", "Anteromedial basal segment"]
    },
    "Trachea and Bronchi": {
      "Trachea": ["Cervical trachea", "Thoracic trachea", "Carina"],
```

Figure 24: Knowledge trees for CT-RATE (Hamamci et al., 2026).

Knowledge Trees Used for Report Standardization | CT-RATE

(continued)


```
    "Bronchial Branches": ["Main bronchus", "Bronchus intermedius", "Lobar bronchus", "Segmental
bronchus", "Subsegmental bronchus", "Bronchiole"]
    },
    "Hilar Region": ["Pulmonary hilum"]
  },
  "Pleura and Pleural Spaces": {
    "Pleura and Fissures": ["Visceral pleura", "Parietal pleura", "Mediastinal pleura", "Diaphragmatic
pleura", "Horizontal fissure", "Oblique fissure", "Accessory fissure"],
    "Spaces and Angles": ["Pleural cavity", "Costophrenic angle", "Cardiophrenic angle"]
    },
  "Mediastinum": {
    "Heart": ["Atrium", "Ventricle", "Myocardium", "Endocardium", "Atrial appendage", "Aortic valve",
"Mitral valve", "Tricuspid valve", "Pulmonary valve"],
    "Mediastinal Spaces": ["Anterior mediastinum", "Middle mediastinum", "Posterior mediastinum",
"Prevascular space", "Paratracheal space", "Aortopulmonary window", "Subcarinal space", "Paraesophageal
space", "Paravertebral space"],
    "Mediastinal Ancillary Structures": ["Thymus", "Thoracic esophagus", "Thoracic duct", "Pericardium",
"Pericardial cavity", "Pericardial fat pad"]
    },
  "Lymph Nodes": {
    "Upper Mediastinal and Cervical Zones": ["Supraclavicular lymph node", "Upper paratracheal lymph
node", "Prevascular lymph node", "Retrotracheal lymph node", "Lower paratracheal lymph node"],
    "Aortic and Lower Mediastinal Zones": ["Aortopulmonary window lymph node", "Para-aortic lymph node",
"Subcarinal lymph node", "Paraesophageal lymph node", "Pulmonary ligament lymph node"],
    "Hilar and Intrapulmonary Zones": ["Hilar lymph node", "Interlobar lymph node", "Lobar lymph node",
"Segmental lymph node", "Subsegmental lymph node"],
    "Chest Wall and Extrathoracic Zones": ["Axillary lymph node", "Internal mammary lymph node",
"Cardiophrenic angle lymph node"]
  },
  "Blood Vessels": {
    "Pulmonary Vessels": ["Pulmonary trunk", "Pulmonary conus", "Main pulmonary artery", "Interlobar
artery", "Lobar artery", "Segmental artery", "Superior pulmonary vein", "Inferior pulmonary vein",
"Lobar vein"],
    "Aorta and Branches": ["Aortic root", "Aortic sinus", "Sinotubular junction", "Ascending aorta",
"Aortic arch", "Thoracic aorta", "Descending aorta", "Brachiocephalic trunk", "Common carotid artery",
"Subclavian artery"],
    "Other Great Vessels": ["Superior vena cava", "Inferior vena cava", "Brachiocephalic vein", "Azygos
vein", "Hemiazygos vein", "Coronary artery"]
  },
  "Chest Wall and Skeletal System": {
    "Skeletal System": {
      "Spine": ["Cervical vertebra", "Thoracic vertebra", "Lumbar vertebra", "Vertebral body",
"Intervertebral disc", "Pedicle", "Posterior elements"],
      "Thoracic Cage and Periphery": ["Sternum", "Rib", "Costal cartilage", "Clavicle", "Scapula",
"Humerus"]
      },
    "Body Wall Soft Tissues": ["Pectoralis major", "Pectoralis minor", "Intercostal muscle", "Serratus
anterior", "Erector spinae", "Breast", "Subcutaneous fat", "Skin"]
  },
    "Lower Neck": {
      "Cervical Organs": ["Thyroid lobe", "Thyroid isthmus", "Larynx", "Glottic region"]
    },
    "Upper Abdomen and Diaphragm": {
      "Upper Gastrointestinal Tract": ["Esophagogastric junction", "Gastric fundus", "Gastric body",
"Gastric antrum", "Duodenum"],
      "Hepatobiliary, Pancreas, and Spleen": ["Hepatic lobe", "Caudate lobe", "Hepatic segment",
"Gallbladder", "Common bile duct", "Spleen", "Pancreas"],
      "Genitourinary and Endocrine": ["Renal cortex", "Renal medulla", "Renal pelvis", "Ureter",
"Adrenal body", "Medial limb of adrenal gland", "Lateral limb of adrenal gland"],
      "Diaphragmatic Structures": ["Diaphragm", "Diaphragmatic crus", "Esophageal hiatus"]
      },
    "Mid-Lower Abdomen and Pelvic Edges": {
      "Mid-Lower Gastrointestinal Tract": ["Colon", "Rectum"],
      "Pelvic Organs": ["Urinary bladder", "Uterus", "Ovary", "Prostate", "Testis"],
      "Abdominopelvic Supporting Structures": ["Inguinal canal", "Peritoneal cavity", "Retroperitoneal
space", "Omentum", "Mesentery"]
    },
    "Nerves": {
      "Autonomic and Peripheral Nerves": ["Vagus nerve", "Recurrent laryngeal nerve", "Phrenic nerve",
"Thoracic sympathetic trunk"]}}
```


Figure 25: Knowledge trees for CT-RATE (Hamamci et al., 2026).

Knowledge Trees Used for Report Standardization | AMOS-MM


```
"FINDINGS_TREE": {
  "Chest Imaging Abnormalities": {
    "Density and Parenchymal Abnormalities": {
      "High-Density Opacities": ["Consolidation", "Ground-Glass Opacity", "Linear Opacities", "Patchy
Opacities", "Calcification"],
      "Low-Density and Air-Filled Spaces": ["Lucency", "Cystic Lesions", "Cavitation"],
      "Space-Occupying Lesions": ["Pulmonary Nodule", "Pulmonary Mass"]
    },
    "Morphological and Volumetric Abnormalities": ["Cardiomegaly", "Hilar Enlargement", "Atelectasis",
"Lung Volume Loss", "Mediastinal Shift", "Mediastinal Widening"],
    "Luminal and Airway Abnormalities": ["Bronchiectasis", "Bronchial Stenosis or Occlusion",
"Bronchial Wall Thickening", "Air Bronchogram", "Tree-in-Bud Pattern"],
    "Interstitial Abnormalities": ["Interlobular Septal Thickening", "Reticular Opacities", "Increased
Lung Markings", "Honeycomb Pattern"],
    "Serosal Cavity and Space Abnormalities": ["Pleural Effusion", "Pericardial Effusion",
"Pneumothorax", "Pleural Thickening", "Pericardial Thickening", "Widened Interlobar Fissure",
"Hemothorax/Hematoma"],
    "Chest Wall and Skeletal Abnormalities": ["Fracture Line", "Bone Destruction", "Hyperostosis",
"Cortical Disruption"],
    "Chest Lymph Node Abnormalities": ["Mediastinal Lymphadenopathy", "Hilar Lymphadenopathy", "Lymph
Node Calcification"]
  },
  "Abdominal Imaging Abnormalities": {
    "Density and Enhancement Abnormalities": {
      "High-Density Features": ["Calculus", "Calcification", "Contrast or Lipiodol Deposition",
"Implant or Device"],
      "Low/Mixed-Density Features": ["Fat-Density Lesion", "Fluid/Cystic-Density Lesion", "Mixed-
Density Lesion"],
      "Enhancement Abnormalities": ["Avidly Enhancing Lesion", "Mildly Enhancing Lesion",
"Heterogeneous Enhancement", "Ring Enhancement", "Abnormal Mucosal Enhancement"]
    },
    "Parenchymal and Space-Occupying Abnormalities": ["Parenchymal Nodule", "Parenchymal Mass",
"Parenchymal Cyst"],
    "Organ Morphological and Volumetric Changes": ["Solid Organ Enlargement", "Solid Organ Atrophy",
"Irregular Organ Contour", "Focal Organ Bulge"],
    "Abdominal Luminal and Wall Abnormalities": ["Bowel Dilatation", "Biliary Dilatation", "Luminal
Stenosis or Occlusion", "Bowel Wall Thickening", "Gallbladder Wall Thickening", "Intraluminal Filling
Defect", "Diverticulum", "Fistula/Sinus Tract"],
    "Vascular Abnormalities": ["Vascular Dilatation or Aneurysm", "Vascular Wall Thickening", "Vascular
False Lumen"],
    "Peritoneal Cavity and Space Abnormalities": {
      "Abnormal Space Collections": ["Ascites", "Loculated Fluid Collection", "Air-Fluid Level",
"Pneumoperitoneum", "Localized Gas Collection", "Hemoperitoneum/Hematoma"],
      "Fat and Fascial Changes": ["Fat Stranding", "Fascial Thickening", "Exudative Opacities"]
    },
    "Abdominal Lymph Node Abnormalities": ["Abdominal Lymphadenopathy"]
  },
  "Pelvic Region Imaging Abnormalities": {
    "Pelvic Density and Space-Occupying Abnormalities": {
      "Pelvic Region High-Density Opacities": [],
      "Pelvic Region Low/Mixed-Density Opacities": ["Fat-Density Lesion", "Cystic/Fluid-Density
Lesion"],
      "Space-Occupying Lesions": ["Pelvic Nodule", "Pelvic Mass", "Adnexal Space-Occupying Lesion"]
    },
    "Pelvic Region Organ Morphological Changes": {
      "Reproductive System Abnormalities": ["Prostatic Enlargement", "Uterine Enlargement", "Cervical
Thickening"],
      "Urinary System Abnormalities": ["Bladder Wall Thickening"]
    },
    "Pelvic Region Skeletal Abnormalities": ["Bone Destruction", "Cortical Thinning", "Osteosclerosis",
"Cortical Disruption", "Fracture Line"],
    "Pelvic Cavity Content Abnormalities": ["Pelvic Fluid Collection", "Hematocele/Hematoma", "Pelvic
Free Air", "Pelvic Implant/Device"],
    "Pelvic Lymphatic and Vascular Abnormalities": {
      "Lymph Node Abnormalities": ["Pelvic Lymphadenopathy", "Pelvic Lymph Node Calcification"],
      "Vascular Abnormalities": ["Pelvic Vein Engorgement and Tortuosity"]
    }
  }
}
```


Figure 26: Knowledge trees for AMOS-MM (Ji et al., 2022).

Knowledge Trees Used for Report Standardization | AMOS-MM

```
"DIAGNOSIS_TREE": {
  "Thoracic Diseases": {
    "Pulmonary Parenchymal and Interstitial Diseases": ["Pneumonia", "Emphysema", "Pulmonary
Tuberculosis", "Lung Abscess", "Lung Cancer", "Pulmonary Metastasis", "Pulmonary Lymphoma",
"Interstitial Lung Disease", "Pulmonary Bulla"],
    "Tracheal and Bronchial Diseases": ["Bronchitis", "Bronchiectasis", "Tracheal Malignancy",
"Bronchial Malignancy"],
    "Pleural Diseases": ["Pleurisy", "Pneumothorax", "Pleural Mesothelioma", "Empyema"],
    "Mediastinal Diseases": ["Mediastinal Lymphoma", "Mediastinal Teratoma", "Mediastinal Goiter",
"Nodular Goiter", "Foregut Cyst", "Thymoma"],
    "Cardiovascular Diseases": ["Atherosclerosis", "Aortic Dissection", "Aneurysm", "Hypertension",
"Pulmonary Embolism", "Coronary Atherosclerotic Heart Disease", "Pericarditis"],
    "Chest Wall and Breast Diseases": ["Breast Cancer", "Chest Wall Fracture", "Chest Wall Tumor"]
  },
  "Abdominal Diseases": {
    "Solid Abdominal Organ Diseases": {
      "Hepatic Diseases": ["Fatty Liver", "Liver Cirrhosis", "Hepatic Hemangioma", "Focal Nodular
Hyperplasia of Liver", "Hepatocellular Carcinoma", "Hepatocellular Adenoma", "Dysplastic Nodule of
Liver", "Hemochromatosis", "Hepatic Cyst", "Hepatic Abscess"],
      "Splenic Diseases": ["Splenomegaly", "Accessory Spleen", "Splenic Cyst", "Splenic Infarction",
"Splenic Rupture"],
      "Pancreatic Diseases": ["Pancreatitis", "Ectopic Pancreas", "Pancreatic Cancer", "Intraductal
Papillary Mucinous Neoplasm", "Pancreatic Neuroendocrine Tumor"],
      "Adrenal Diseases": ["Adrenal Adenoma", "Pheochromocytoma", "Adrenal Myelolipoma",
"Adrenocortical Hyperplasia"],
      "Renal and Ureteral Diseases": ["Polycystic Kidney Disease", "Hydronephrosis", "Renal Calculus",
"Clear Cell Carcinoma", "Renal Cell Carcinoma", "Renal Cyst", "Renal Angiomyolipoma"]
    },
    "Gastrointestinal Diseases": {
      "Esophageal/Gastric/Duodenal Diseases": ["Gastric Cancer", "Gastrointestinal Stromal Tumor",
"Esophageal Cancer", "Peptic Ulcer", "Hiatal Hernia"],
      "Small Intestinal Diseases": ["Intestinal Obstruction", "Intussusception", "Intestinal
Perforation", "Volvulus", "Ischemic Bowel Disease"],
      "Colorectal/Appendiceal Diseases": ["Appendicitis", "Colon Cancer", "Colorectal Cancer", "Rectal
Cancer", "Ulcerative Colitis", "Diverticulum", "Polyp", "Appendicolith"]
    },
    "Biliary Tract Diseases": ["Cholecystitis", "Cholangitis", "Cholestasis", "Cholangiocarcinoma",
"Adenomyomatosis of Gallbladder", "Gallstone", "Choledocholithiasis", "Gallbladder Polyp", "Gallbladder
Carcinoma"],
    "Abdominal Vascular Diseases": ["Portal Hypertension", "Thrombosis", "Varices", "Abdominal Aortic
Aneurysm", "Mesenteric Artery Embolism"],
    "Peritoneal and Retroperitoneal Diseases": ["Peritonitis", "Hernia"],
    "Abdominal Wall Diseases": ["Lipoma", "Abdominal Wall Hernia"]
  },
  "Pelvic Diseases": {
    "Pelvic Organ Diseases": {
      "Bladder and Distal Urethral Diseases": ["Bladder Cancer", "Bladder Calculus", "Cystitis"],
      "Female Reproductive System Diseases": ["Uterine Fibroids", "Uterine Leiomyoma", "Uterine
Leiomyosarcoma", "Chocolate Cyst", "Hydrosalpinx", "Ovarian Cyst", "Ovarian Cancer", "Cervical Cancer",
"Endometrial Cancer", "Pelvic Inflammatory Disease"],
      "Male Reproductive System Diseases": ["Benign Prostatic Hyperplasia", "Hydrocele", "Prostate
Cancer"]
    }
  },
  "Skeletal System Diseases": {
    "Spinal and Rib Diseases": ["Vertebral Compression Fracture", "Rib Fracture", "Bone Metastasis",
"Multiple Myeloma", "Ankylosing Spondylitis"]
  }
}
```

Figure 27: Knowledge trees for AMOS-MM (Ji et al., 2022).

Knowledge Trees Used for Report Standardization | AMOS-MM

```
"ATTRIBUTES_TREE": {
  "Presence": ["Present", "Absent", "Uncertain"],
  "Spatial distribution": {
    "Laterality": ["Left", "Right", "Bilateral"],
    "Symmetry": ["Symmetric", "Asymmetric"],
    "Extent": ["Focal", "Multifocal", "Scattered", "Diffuse", "Confluent"]
  },
  "Number": ["Solitary", "Few", "Multiple"],
  "Size": ["Small", "Medium", "Large", "Widening/thickening/deepening/enlarging/dilating",
"Decreasing/shrinking", "Specific size(mm/cm)"],
  "Shape": ["Punctate/Speckled", "Miliary", "Nodular", "Mass-like/Conglomerate", "Patchy",
"Linear/Streaky", "Tubular", "Reticular", "Cystic", "Round/Oval", "Annular/Arcuate", "Wedge-shaped",
"Tree-in-bud", "Lobulated", "Irregular"],
  "Margin": ["Well-defined", "Ill-defined", "Smooth", "Spiculated/shaggy"],
  "Density": ["Hypodense", "Isodense", "Hyperdense", "Mixed Density", "Ground-Glass Opacity", "Soft-
Tissue Density", "Fluid Density", "Fat Density", "Gas Density", "Calcific Density", "Metallic Density"],
  "Internal features": ["Homogeneous", "Heterogeneous", "Calcification", "Gas", "Fluid/Cystic", "Air-
fluid level", "Necrosis/Liquefaction", "Septation/Loculation", "Internal nodules", "Vascular",
"Hemorrhagic", "Fat-containing", "Capsular"],
  "Secondary effects": ["Volume reduction", "Volume increase", "Architectural distortion",
"Displacement/compression"],
  "Severity": ["Mild", "Moderate", "Severe"],
  "Chronicity": ["Acute", "Chronic", "Old/prior", "Stable", "Progressive", "Improving"],
  "Certainty": ["Definite", "Probable", "Possible"]
  }

"ANATOMY_TREE": {
  "Neck": {
    "Neck Organs and Airway": ["Thyroid Gland", "Larynx", "Hypopharynx", "Cervical Trachea", "Cervical
Esophagus"],
    "Cervical Vascular System": ["Common Carotid Artery", "Internal Carotid Artery", "External Carotid
Artery", "Internal Jugular Vein", "External Jugular Vein", "Cervical Vascular Bundle"]
  },
  "Chest": {
    "Airway and Lung Parenchyma": {
      "Trachea and Bronchial Tree": ["Trachea", "Carina", "Main Bronchus", "Bronchus Intermedius",
"Lobar Bronchus", "Segmental Bronchus"],
      "Lung Lobes": ["Upper Lobe", "Middle Lobe", "Lower Lobe"],
      "Lung Segments": ["Apical Segment", "Posterior Segment", "Anterior Segment", "Lateral Segment",
"Medial Segment", "Superior Segment", "Medial Basal Segment", "Anterior Basal Segment", "Lateral Basal
Segment", "Posterior Basal Segment", "Superior Lingular Segment", "Inferior Lingular Segment",
"Apicoposterior Segment", "Anteromedial Basal Segment"],
      "Pulmonary Interstitium and Microstructures": ["Bronchovascular Bundle", "Centrilobular Core"],
      "Pulmonary Fissures": ["Oblique Fissure", "Horizontal Fissure"]
    },
    "Pleura and Pleural Cavity": {
      "Pleura": ["Pleura"],
      "Pleural Cavity and Recesses": ["Free Pleural Space", "Costodiaphragmatic Recess",
"Costomediastinal Recess", "Phrenicomediastinal Recess"]
    },
    "Mediastinum and Cardiovascular System": {
      "Mediastinal Compartments": ["Superior Mediastinum", "Anterior Mediastinum", "Middle Mediastinum",
"Posterior Mediastinum"],
      "Cardiac Chambers and Structures": ["Atrium", "Ventricle", "Atrial Appendage", "Mitral Valve",
"Tricuspid Valve", "Aortic Valve", "Pulmonary Valve", "Myocardium", "Base of Pericardium"],
      "Pericardial System": ["Pericardium", "Pericardial Cavity"],
      "Thoracic Arterial System": ["Aortic Root", "Ascending Aorta", "Aortic Arch", "Descending
Thoracic Aorta", "Brachiocephalic Trunk", "Subclavian Artery", "Main Pulmonary Artery", "Main Pulmonary
Artery Branches", "Coronary Artery"],
      "Thoracic Venous System": ["Superior Vena Cava", "Brachiocephalic Vein", "Azygos Vein",
"Hemiazygos Vein", "Thoracic Inferior Vena Cava", "Pulmonary Vein", "Coronary Sinus"],
      "Other Mediastinal Organs": ["Thoracic Esophagus", "Thymus"]
    },
    "Chest Wall and Diaphragm": {
      "Chest Wall Soft Tissues": ["Subcutaneous Fat of Chest Wall", "Breast Glandular Tissue",
"Retromammary Space", "Axillary Fat Pad", "Pectoralis Major", "Pectoralis Minor", "Intercostal Muscle",
"Latissimus Dorsi", "Erector Spinae"],
      "Thoracic Skeleton": ["Clavicle", "Scapula", "Manubrium Sterni", "Body of Sternum", "Xiphoid
Process", "Rib Bone", "Costal Cartilage", "Thoracic Vertebral Body", "Posterior Elements of Thoracic
Vertebra"],
```

Figure 28: Knowledge trees for AMOS-MM (Ji et al., 2022).

Knowledge Trees Used for Report Standardization | AMOS-MM

```
(continued)

 "Diaphragmatic Parts": ["Diaphragmatic Crus", "Central Tendon of Diaphragm", "Costal Part of
Diaphragm", "Sternal Part of Diaphragm"]
    },
    "Thoracic Lymph Node Stations": ["Supraclavicular Lymph Node", "Upper Paratracheal Lymph Node",
"Pretracheal Lymph Node", "Retrotracheal Lymph Node", "Aortopulmonary Window Lymph Node", "Para-aortic
Lymph Node", "Subcarinal Lymph Node", "Paraesophageal Lymph Node", "Pulmonary Ligament Lymph Node",
"Hilar Lymph Node", "Interlobar Lymph Node", "Lobar Lymph Node", "Segmental Lymph Node", "Axillary
Lymph Node", "Internal Mammary Lymph Node"]
  },
  "Abdomen": {
    "Hepatobiliary System": {
      "Liver Surface and Landmarks": ["Liver Dome", "Falciform Ligament", "Fissure for Ligamentum
Venosum", "Fissure for Ligamentum Teres", "Porta Hepatis"],
      "Hepatic Segments": ["Hepatic Segment S1", "Hepatic Segment S2", "Hepatic Segment S3", "Hepatic
Segment S4", "Hepatic Segment S5", "Hepatic Segment S6", "Hepatic Segment S7", "Hepatic Segment S8"],
      "Biliary System": ["Intrahepatic Bile Duct", "Common Hepatic Duct", "Cystic Duct", "Common Bile
Duct", "Major Duodenal Papilla"],
      "Gallbladder Structures": ["Gallbladder Fundus", "Gallbladder Body", "Gallbladder Neck",
"Gallbladder Wall", "Gallbladder Lumen"]
    },
    "Pancreaticosplenic System": {
      "Pancreatic Structures": ["Pancreatic Head", "Pancreatic Neck", "Pancreatic Body", "Pancreatic
Tail", "Uncinate Process", "Main Pancreatic Duct"],
      "Splenic Structures": ["Upper Pole of Spleen", "Splenic Parenchyma", "Splenic Hilar Vascular
Sheath", "Accessory Spleen"]
    },
    "Gastrointestinal System": {
      "Gastric Divisions": ["Cardia", "Gastric Fundus", "Gastric Body", "Gastric Antrum", "Pylorus",
"Greater Curvature of Stomach", "Lesser Curvature of Stomach"],
      "Duodenum and Small Intestine": ["Duodenal Bulb", "Descending Duodenum", "Horizontal Duodenum",
"Ascending Duodenum", "Jejunum", "Ileum", "Terminal Ileum"],
      "Colorectum and Appendix": ["Cecum", "Ileocecal Valve", "Appendix", "Ascending Colon", "Hepatic
Flexure", "Transverse Colon", "Splenic Flexure", "Descending Colon", "Mid-upper Sigmoid Colon",
"Rectosigmoid Junction"]
    },
    "Urinary System and Adrenal Glands": {
      "Renal Parenchyma and Collecting System": ["Upper Pole of Kidney", "Renal Cortex", "Renal
Medulla", "Renal Column", "Minor Calyx", "Major Calyx", "Renal Pelvis", "Renal Hilum"],
      "Ureters and Adrenal Glands": ["Abdominal Ureter", "Adrenal Cortex", "Adrenal Medulla"]
    },
    "Major Abdominal Vessels": {
      "Abdominal Arterial Trunks and Branches": ["Abdominal Aorta", "Celiac Trunk", "Left Gastric
Artery", "Common Hepatic Artery", "Splenic Artery", "Superior Mesenteric Artery", "Inferior Mesenteric
Artery", "Renal Artery", "Lumbar Artery"],
      "Abdominal Venous Trunks and Branches": ["Main Portal Vein", "Portal Vein Branches", "Splenic
Vein", "Superior Mesenteric Vein", "Inferior Mesenteric Vein", "Abdominal Inferior Vena Cava", "Hepatic
Vein", "Renal Vein", "Gonadal Vein"]
    },
    "Peritoneal Cavity and Retroperitoneal Space": {
      "Omentum and Mesentery": ["Greater Omentum", "Lesser Omentum", "Small Bowel Mesentery",
"Transverse Mesocolon", "Sigmoid Mesocolon", "Lower Abdominal Greater Omentum"],
      "Peritoneal Recesses and Spaces": ["Subphrenic Space", "Hepatorenal Recess", "Suprahepatic Space",
"Subhepatic Space", "Lesser Sac", "Paracolic Gutter", "Mesenteric Sinus"],
      "Retroperitoneal Compartments": ["Anterior Pararenal Space", "Perirenal Space", "Posterior
Pararenal Space", "Paravascular Space"]
    },
    "Abdominal Wall and Skeletal Muscle Group": {
      "Abdominal Wall Musculature": ["Rectus Abdominis", "External Oblique Muscle", "Internal Oblique
Muscle", "Transversus Abdominis", "Psoas Major", "Quadratus Lumborum", "Iliacus Muscle", "Latissimus
Dorsi", "Erector Spinae"],
      "Abdominal Skeleton": ["Lower Lumbar Spine", "Lumbar Vertebral Body", "Posterior Elements of
Lumbar Vertebra", "Lower Ribs"]
    },
    "Abdominal Lymph Node Stations": {
      "Abdominal and Retroperitoneal Lymph Nodes": ["Retroperitoneal Lymph Node", "Celiac Lymph Node",
"Perigastric Lymph Node", "Porta Hepatis Lymph Node", "Splenic Hilar Lymph Node", "Peripancreatic Lymph
Node", "Mesenteric Lymph Node", "Paracolic Lymph Node", "Para-aortic Lymph Node", "Paracaval Lymph
Node", "Renal Hilar Lymph Node"]
    }
  },
```

Figure 29: Knowledge trees for AMOS-MM (Ji et al., 2022).

Knowledge Trees Used for Report Standardization | AMOS-MM

(continued)

```
  "Pelvis": {
    "Pelvic Digestive System": {
      "Rectum and Anal Canal": ["Middle Rectum", "Lower Rectum", "Anal Canal", "Anal Sphincter Complex"]
    },
    "Pelvic Urinary System": {
      "Lower Urinary Tract Structures": ["Pelvic Ureter", "Bladder Dome", "Bladder Body", "Bladder
Base", "Bladder Neck", "Urethra"]
    },
    "Pelvic Reproductive System": {
      "Female Reproductive Organs": ["Cervix", "Uterine Body", "Uterine Fundus", "Endometrium",
"Myometrium", "Adnexal Region", "Ovary", "Vaginal Fornix", "Vaginal Wall"],
      "Male Reproductive Organs": ["Prostate", "Seminal Vesicle", "Spermatic Cord", "Scrotal Wall",
"Testis", "Epididymis", "Penis"]
    },
    "Pelvic Floor Support and Surrounding Spaces": {
      "Pelvic Floor Musculature": ["Levator Ani", "Puborectalis", "Obturator Internus", "Piriformis"],
      "Pelvic Spaces": ["Rectouterine Pouch", "Vesicouterine Pouch", "Rectovesical Pouch", "Retropubic
Space", "Retrorectal Space", "Ischioanal Fossa"]
    },
    "Pelvic Vascular System": {
      "Pelvic Arteries": ["Common Iliac Artery", "External Iliac Artery", "Internal Iliac Artery",
"Obturator Artery", "Uterine Artery", "Internal Pudendal Artery", "Pelvic Vessels"],
      "Pelvic Veins": ["Common Iliac Vein", "External Iliac Vein", "Internal Iliac Vein", "Prostatic
Venous Plexus", "Uterine Venous Plexus"]
    },
    "Pelvic Skeleton and Joints": {
      "Bony Pelvis": ["Sacrum", "Coccyx", "Ilium", "Iliac Wing", "Ischium", "Pubis"],
      "Pelvic Joints": ["Sacroiliac Joint", "Pubic Symphysis", "Hip Joint"]
    }
  },
  "Thigh and Superficial Structures": {
    "Proximal Thigh Structures": ["Femoral Head", "Femoral Neck", "Upper Thigh Musculature"],
    "Superficial and External Genitalia": ["Superficial Tissues of External Genitalia"]
  }
}
```

Figure 30: Knowledge trees for AMOS-MM (Ji et al., 2022).

Knowledge Trees Used for Report Standardization | MIMIC-CXR

```
"FINDINGS_TREE": {
  "Increased Pulmonary Parenchymal Opacity": ["Ground-Glass Opacity", "Air-Space Opacity",
"Interstitial Opacity", "Linear Or Reticular Opacity", "Focal Opacity"],
  "Focal Pulmonary Lesion": ["Nodule", "Mass", "Cavity", "Cystic Lesion", "Granuloma"],
  "Increased Pulmonary Lucency": ["Diffuse Increased Lucency", "Emphysematous Change", "Bulla"],
  "Abnormal Lung Volume": ["Atelectasis", "Low Lung Volume"],
  "Airway Abnormality": ["Bronchiectasis", "Airway Narrowing", "Airway Wall Thickening", "Mucus
Plugging", "Tracheal Deviation"],
  "Pleural Abnormality": ["Pneumothorax", "Hydropneumothorax", "Pleural Effusion", "Pleural Thickening",
"Pleural Plaque", "Pleural Calcification"],
  "Mediastinal Abnormality": ["Mediastinal Widening", "Mediastinal Shift", "Pneumomediastinum",
"Mediastinal Mass"],
      "Hilar Abnormality": ["Hilar Enlargement", "Hilar Indistinctness", "Lymphadenopathy", "Lymph Node
Calcification"],
      "Cardiac Silhouette Abnormality": ["Cardiomegaly"],
      "Great Vessel Abnormality": ["Aortic Tortuosity Or Unfolding", "Aneurysmal Dilatation",
"Prominent Pulmonary Artery", "Vascular Calcification"],
      "Pulmonary Vascular Abnormality": ["Pulmonary Vascular Congestion", "Prominent Pulmonary
Vasculature", "Pulmonary Vascular Redistribution", "Increased Pulmonary Vascular Markings"],
      "Diaphragmatic Abnormality": ["Elevated Hemidiaphragm", "Flattened Diaphragm", "Diaphragmatic
Eventration"],
      "Thoracic Cage And Osseous Abnormality": ["Chest Wall Deformity", "Scoliosis", "Kyphotic
Deformity", "Non-Compression Fracture", "Compression Fracture", "Osteopenia", "Osseous Destruction",
"Cortical Irregularity", "Osteophyte Or Ossification"],
      "Chest Wall Soft Tissue Abnormality": ["Chest Wall Soft Tissue Mass", "Non-Device Foreign Body",
"Subcutaneous Emphysema", "Deep Neck Emphysema"],
      "Tubal Or Catheter Abnormality": ["Central Venous Catheter", "PICC Line", "Endotracheal Tube",
"Chest Tube", "Enteric Tube", "Dialysis Catheter", "Pulmonary Artery Catheter", "Intra-Aortic Balloon
Pump", "Tracheostomy Tube", "Malpositioned Catheter"],
  "Implanted Device Abnormality": ["Stent", "Pacemaker Or Defibrillator", "Pacing Lead", "Prosthetic
Valve", "Other Implanted Device"],
  "Postoperative Change": ["Sternal Wire", "Surgical Clip Or Suture", "Orthopedic Or Spinal Fixation
Hardware"],
  "Subdiaphragmatic Free Air": []
}

"DIAGNOSIS_TREE": {
  "Airspace Diseases": {
    "Infectious Lung Diseases": ["Pneumonia", "Pulmonary Tuberculosis", "Lung Abscess",
"Lobar/Bronchopneumonia"],
    "Non-Infectious Airspace Diseases": ["Aspiration and Aspiration Pneumonia", "Pulmonary Edema",
"Pulmonary Hemorrhage", "Acute Respiratory Distress Syndrome (ARDS)"]
  },
  "Interstitial and Diffuse Lung Parenchymal Diseases": ["Interstitial Lung Disease (ILD)",
"Cryptogenic Organizing Pneumonia (COP)", "Pulmonary Fibrosis", "Sarcoidosis", "Granulomatous Disease",
"Drug-Induced Lung Injury", "Lymphangioleiomyomatosis (LAM)"],
  "Airway and Chronic Obstructive Pulmonary Diseases": ["Chronic Obstructive Pulmonary Disease
(COPD/Emphysematous Changes)", "Bronchiectasis", "Emphysema and Hyperinflation"],
  "Pulmonary Neoplasms and Nodules": ["Pulmonary Nodule", "Mass/Space-Occupying Lesion", "Granulomas
and Calcifications", "Lung Cancer/Adenocarcinoma", "Pulmonary Metastases", "Lymphangitic
Carcinomatosis"],
  "Pleural Diseases": ["Empyema", "Hemothorax", "Pleural Plaques and Thickening", "Chronic Pleural
Scarring"],
  "Air Leak Syndromes": ["Pneumothorax", "Hydropneumothorax", "Pneumomediastinum", "Subcutaneous/Deep
Cervical Emphysema"],
  "Cardiac and Pericardial Diseases": ["Heart Failure", "Cardiac Decompensation",
"Cardiomyopathy/Dilated Cardiomyopathy", "Valve Replacement and Stenosis", "Pericardial Effusion"],
  "Great Vessel Diseases": ["Aneurysm", "Aortic Dissection", "Atherosclerosis"],
  "Mediastinal and Esophageal Diseases": {
    "Mediastinal Diseases": ["Mediastinal Tumor", "Mediastinal Lymphadenopathy"],
    "Esophageal and Thyroid Diseases": ["Hiatal Hernia", "Esophageal Tumor", "Esophagitis", "Achalasia",
"Enlarged Thyroid / Goiter"]
  },
  "Pulmonary Vascular and Volume Overload Diseases": ["Pulmonary Embolism", "Pulmonary Infarction",
"Pulmonary Vascular Congestion and Prominence"],
  "Diaphragmatic Diseases": ["Diaphragmatic Paralysis", "Subdiaphragmatic Free Air / Pneumoperitoneum"],
  "Chest Wall and Skeletal Diseases": ["Sternal/Rib Fractures", "Chest Contusion and Laceration",
"Osteoarthritis/Degenerative Changes", "Scoliosis/Deformity", "Ankylosing Spondylitis", "Syndesmophyte",
"Osseous Metastases", "Lytic Destruction", "Osteosclerosis"],
  "Postoperative and Device-Related Chest Abnormalities": ["Post-Thoracotomy/Sternotomy", "Post-
Lobectomy/Pneumonectomy", "Disrupted Sternal Wires/Dehiscence", "Medical Device Malposition and
Complications"]}
```

Figure 31: Knowledge trees for MIMIC-CXR (Johnson et al., 2019a).

Knowledge Trees Used for Report Standardization | MIMIC-CXR

```
"ATTRIBUTES_TREE": {
  "Presence": ["Present", "Absent", "Uncertain"],
  "Spatial distribution": {
    "Laterality": ["Left", "Right", "Bilateral", "Midline"],
    "Symmetry": ["Symmetrical", "Asymmetrical"],
    "Extent": ["Focal", "Multifocal/Multiple", "Scattered", "Diffuse/Extensive", "Confluent"]
  },
  "Number": ["Solitary", "Few", "Multiple"],
  "Size": ["Small", "Moderate", "Large", "Widening/thickening/deepening/enlarging/dilating",
"Decreasing/shrinking", "Specific dimension (mm/cm)"],
  "Shape": ["Punctate/Nodular", "Mass-like", "Patchy", "Linear/Band-like", "Reticular", "Wedge-
shaped/Triangular", "Flat/Discoid", "Round/Oval", "Tortuous/Coiled", "Loculated/Encapsulated",
"Stratified/Laminated", "Irregular"],
  "Margin": ["Clear/Well-defined", "Ill-defined/Obscured", "Irregular/Spiculated", "Blunted"],
  "Density": ["Low density/Radiolucent", "High density/Radiopaque", "Mixed density", "Ground-
glass/Hazy", "Calcific density", "Metallic/Foreign body density"],
  "Internal features": ["Homogeneous", "Heterogeneous", "Air bronchogram", "Cavity", "Cyst/Air space",
"Air-fluid level", "Calcification"],
  "Secondary effects": ["Volume loss/Collapse", "Hyperinflation/Volume expansion", "Architectural
distortion/Crowding", "Mass effect/Displacement", "Traction/Retraction", "Silhouette sign/Obscuration"],
  "Severity": ["Mild", "Moderate", "Severe"],
  "Chronicity": ["Acute/New", "Chronic/Old", "Stable/Unchanged", "Progressive/Worsening",
"Resolving/Improving", "Postoperative/Post-treatment"],
  "Certainty": ["Definite", "Probable", "Possible"]
}

"ANATOMY_TREE": {
  "Airway and Pulmonary System": {
    "Tracheobronchial Region": ["Trachea", "Carina", "Bronchus"],
    "Lung Parenchyma (Anatomical)": ["Upper lobe parenchyma", "Middle lobe parenchyma", "Lower lobe
parenchyma", "Lingular parenchyma", "Lung apex", "Lung base"],
    "Lung Zones (2D Projection)": ["Upper lung zone", "Middle lung zone", "Lower lung zone", "Hilar
region", "Retrocardiac region", "Paramediastinal region"],
    "Interstitial and Parenchymal Tissues": ["Airspaces", "Pulmonary interstitium"]
  },
  "Cardiovascular System": {
    "Cardiac Chambers and Walls": ["Atrium", "Ventricle", "Myocardial contour"],
    "Cardiac Valves": ["Aortic valve", "Mitral valve", "Tricuspid valve"],
    "Aorta and Major Branches": ["Aortic sinus/root", "Ascending aorta", "Aortic arch", "Descending
aorta", "Aortic knob"],
    "Major Veins (Catheter locations)": ["Superior vena cava (SVC)", "Inferior vena cava (IVC)",
"Cavoatrial junction (CAJ)", "Azygos vein", "Brachiocephalic vein", "Internal jugular vein",
"Subclavian vein"],
    "Pulmonary Arterial System": ["Main pulmonary trunk", "Main pulmonary artery branches",
"Aortopulmonary window"],
    "Peripheral Arterial System": ["Coronary artery", "Subclavian artery", "Splenic artery"]
  },
  "Pleural and Diaphragmatic System": {
    "Pleural Surfaces and Spaces": ["Pleural cavity", "Visceral/Parietal pleura", "Costophrenic
angle/sinus", "Cardiophrenic angle/sinus"],
    "Pleural Fissures": ["Oblique fissure", "Horizontal fissure"],
    "Diaphragm": ["Diaphragmatic dome"]
  },
  "Skeletal and Body Wall System": {
    "Axial Skeleton": ["Cervical vertebra", "Thoracic vertebra", "Lumbar vertebra", "Vertebral body",
"Manubrium", "Sternal body", "Rib", "Costal cartilage"],
    "Appendicular Skeleton": ["Clavicle", "Scapula", "Proximal humerus", "Shoulder joint",
"Acromioclavicular joint"],
    "Soft Tissues": ["Cervical soft tissues", "Chest wall soft tissues", "Subcutaneous fat", "Axillary
region", "Pectoral musculature"]
  },
  "Marginal Organs": {
    "Cervical Organs": ["Thyroid gland"],
    "Upper Gastrointestinal System": ["Esophagus", "Gastroesophageal (GE) junction", "Stomach",
"Duodenum"],
    "Abdominal Solid and Hollow Organs": ["Spleen", "Biliary system", "Peritoneal cavity",
"Bowel/Intestines"]
  },
  "Medical Devices and Implants (High-frequency Non-anatomical Entities)": {
    "Tubes and Lines": ["Central venous catheter (CVC)", "Peripherally inserted central catheter
(PICC)", "Nasogastric/Orogastric (NG/OG) tube", "Chest tube / Pleural drain"],
    "Metals and Implants": ["Pacemaker", "ECG leads", "Vascular stent", "Sternotomy wires"]}}
```

Figure 32: Knowledge trees for MIMIC-CXR (Johnson et al., 2019a).