CLEAR: A Clinically Grounded Tabular Framework for Radiology Report Evaluation

Yuyang Jiang[1]*, Chacha Chen[1]†, Shengyuan Wang[2]‡, Feng Li[1]§, Zecong Tang[3]¶,
Benjamin M. Mervak[4]‖, Lydia Chelala[1]§, Christopher M. Straus[1]§, Reve Chahine[4]‖,
Samuel G. Armato III[1]§**, Chenhao Tan[1]†**

[1]University of Chicago   [2]Tsinghua University
[3]Zhejiang University   [4]University of Michigan

Abstract

Existing metrics often lack the granularity and interpretability to capture nuanced clinical differences between candidate and ground-truth radiology reports, resulting in suboptimal evaluation. We introduce a **Cl**inically grounded tabular framework with **E**xpert-curated labels and **A**ttribute-level comparison for **R**adiology report evaluation (**CLEAR**). CLEAR not only examines whether a report can accurately identify the presence or absence of medical conditions, but it also assesses whether the report can precisely describe each positively identified condition across five key attributes: `first occurrence`, `change`, `severity`, `descriptive location`, and `recommendation`. Compared with prior works, CLEAR's multi-dimensional, attribute-level outputs enable a more comprehensive and clinically interpretable evaluation of report quality. Additionally, to measure the clinical alignment of CLEAR, we collaborated with five board-certified radiologists to develop **CLEAR-Bench**, a dataset of 100 chest radiograph reports from MIMIC-CXR, annotated across 6 curated attributes and 13 CheXpert conditions. Our experiments demonstrated that CLEAR achieves high accuracy in extracting clinical attributes and provides automated metrics that are strongly aligned with clinical judgment.

1 Introduction

Evaluation is becoming increasingly challenging in the era of large language models (LLMs). While models continue to rapidly hill-climb on benchmarks rapidly (Maslej et al., 2025; OpenAI, 2025; Anthropic, 2025; Tu et al., 2025; McDuff et al., 2025), it remains unclear whether these reported metrics match task-specific needs (Ganguli et al., 2023; Rauh et al., 2024; Bedi et al., 2025). In the context of radiology, the pursuit of generalist foundation models has achieved promising progress (Bannur et al., 2024; Zambrano Chaves et al., 2025), but do these "appealing" automated metrics truly capture clinically aligned qualities (Paschali et al., 2025)?

In the existing literature, three main types of metrics have been proposed to assess the quality of generated radiology reports, as illustrated in Figure 1: (i) **Lexical metrics** measure surface-level similarity between the generated and ground-truth reports (Papineni et al., 2002; Lin, 2004; Zhang et al., 2020). While straightforward and easy to compute, these metrics struggle to capture nuanced semantics and domain-specific terminology, leading to poor sensitivity for capturing clinically significant errors. (ii) **Clinical efficacy metrics** evaluate the correctness of medical entities and their relationships (Jain et al., 2021; Yu et al., 2023b; Zhao et al., 2024), typically through structured extraction-based comparisons. Although more clinically informed than lexical metrics, they lack the resolution to assess fine-grained attributes such as severity, temporal progression, or treatment recommendations. (iii) **LLM-based metrics** (Ostmeier et al., 2024; Huang et al., 2024; Zambrano Chaves et al., 2025) represent the latest direction, often leveraging the pipeline of LLM-as-a-Judge (Zheng et al., 2023) with pre-defined taxonomies such as the six error categories from the ReXVal dataset (Yu et al., 2023a). While getting closer to expert judgment compared with the previous two types, these methods may still lack comprehensive structured attribution and condition-level interpretability.

Therefore, to address the limitations of existing

---

*Department of Statistics, University of Chicago

†Department of Computer Science, University of Chicago

‡Department of Computer Science and Technology, Tsinghua University

§Department of Radiology, University of Chicago

¶College of Control Science and Engineering, Zhejiang University

‖Department of Radiology, University of Michigan

**Co-senior authorship

Lexical Metrics

BLEU ('02), ROUGE-L ('04), BERTScore ('20)

GT Report
There is evidence of pleural effusion.

Candidate Report 1
Without clear evidence of pleural effusion.

Candidate Report 2
No evidence of pleural effusion.

✗ Fail in capturing nuanced semantics.

Clinical Efficacy Metrics

CheXbert F1 ('20), RadGraph F1 ('21), RaTEScore ('24)

GT Report
There is a left pleural effusion, new since the prior
    Abnormality
exam, associated with atelectasis of the left lower
        Abnormality   Anatomy
lobe. Recommend urgent thoracentesis.

Candidate Report
A right pleural effusion is present likely chronic,
    Abnormality
with associated atelectasis in the lower lobe. No
       Abnormality   Anatomy
intervention is recommended.

✗ Lack the granularity to assess attributes beyond entities and relations.

LLM-based Metrics

FineRadScore ('24), GREEN ('24), CheXprompt ('25)

Based on six error categories (Yu et al., 2023)

1. **False prediction** of finding;
2. **Omission** of finding;
3. Incorrect **location/position** of finding;
4. Incorrect **severity** of finding;
5. Mention of the **comparison that is absent** from the reference impression;
6. **Omission of comparison** describing a change from a previous study.

✗ Lack a structure and does not account for hierarchical or multi-dimensional relationships among errors.

CLEAR (ours)

Tabular Evaluation (13 conditions X 6 attributes)

Condition	Presenc	First Occ.	Change	Severity	Location	Recommend.
Atelectasis	✔	Previous	Worsen	Moderat	Left lower lobe	N/A
Cardiomega	✔	N/A	N/A	N/A	Cardiac silhouette	No immediate action
Consolidati	✔	Current	N/A	Severe	Right middle lobe	N/A
Edema	✔	Previous	Improvir	Mild	N/A	Monitor
Pneumonia	✗	-	-	-	-	-
(...omitted...)						
Support Dev.	✔	-	-	-	PICC line in SVC	Confirm position on X-ray

✔ Provides fine-grained, clinically grounded analysis through a structured, interpretable tabular format that enables easy comparison between reports.

Figure 1: A comparison of existing metrics with CLEAR. Yellow highlights indicate the main evaluation mechanism for each type of metric. Red underlining marks an erroneous term in the candidate report, in contrast to the black underlined term in the ground-truth report, which the designed metric fails to evaluate.

metrics, we introduce **CLEAR** (Section 2), the first clinically-grounded attribute-level evaluation framework that leverages LLMs to map free-text radiology reports to a structured tabular format. Compared with prior work, CLEAR transforms the coarse, single-dimensional taxonomy into a fine-grained, multidimensional structure. Our design not only enables more comprehensive comparisons between candidate and ground-truth reports, but also provides interpretable outputs to assess report quality at the level of condition-attribute pairs. Given the strong adaptability of LLMs across diverse language tasks, they serve as an ideal unified model to operationalize our proposed framework.

Specifically, CLEAR begins with the **Label Extraction Module** (Section 2.1), which evaluates whether the candidate report can precisely identify the presence or absence of specific medical conditions. To ensure robust performance across model scales, this module is enhanced using high-quality, expert-curated labels. Next, for each correctly identified positive condition, the **Description Extraction Module** (Section 2.2) assesses whether the candidate report can accurately describe the condition. Jointly established with one research radiologist and reviewed by one clinical radiologist, five commonly used attributes were defined within a radiology report (first occurrence, change, severity, descriptive location, and recommendation), enabling the first systematic evaluation of these critical facets. Finally, the **Scoring Module** (Section 2.3) compiles and outputs metric scores for each attribute. We carefully designed automated measurements based on the output type from previous modules: accuracy metrics aim at exact matches for single-label outputs while similarity metrics focus on contextual relevance for multi-phrasing outputs.

Additionally, since no existing datasets (Tian et al., 2023; Yu et al., 2023a; Rao et al., 2025) are compatible with CLEAR, we worked closely with radiologists to create **CLEAR-Bench** (Section 3), an expert-curated, attribute-level dataset to assess clinical alignment. CLEAR-Bench consists of 100 studies randomly sampled from MIMIC-CXR-JPG test and validation sets (Johnson et al., 2019, 2024). Each study was annotated and reviewed by at least two radiologists across 6 report attributes and 13 CheXpert conditions[1] (Irvin et al., 2019). CLEAR-Bench includes two components: (i) **Expert ensemble labels** includes ground-truth labels for presence attribute of each condition. These labels were constructed via majority voting among three radiologists, followed by one round of consensus discussion. (ii) **Expert curated attributes** contains the remaining five report attributes for each condition positively identified in the ensemble labels. These attributes were first generated by LLMs, then independently curated by two radiologists, and finalized through one round of discussion and resolution. Additionally, during the curation process, we collected expert Likert scores for each model output, contributing to the assessment of how well the proposed automated metrics align with clinical judgment.

Finally, we evaluated each component of CLEAR using the CLEAR-Bench. Our experimental results (Section 4) demonstrated that: (i) the Label Extraction Module achieved high accuracy compared with expert ensemble labels and significantly outperformed existing labelers across all metrics; (ii) the Description Extraction Module

---

[1]Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax, and Support Devices.
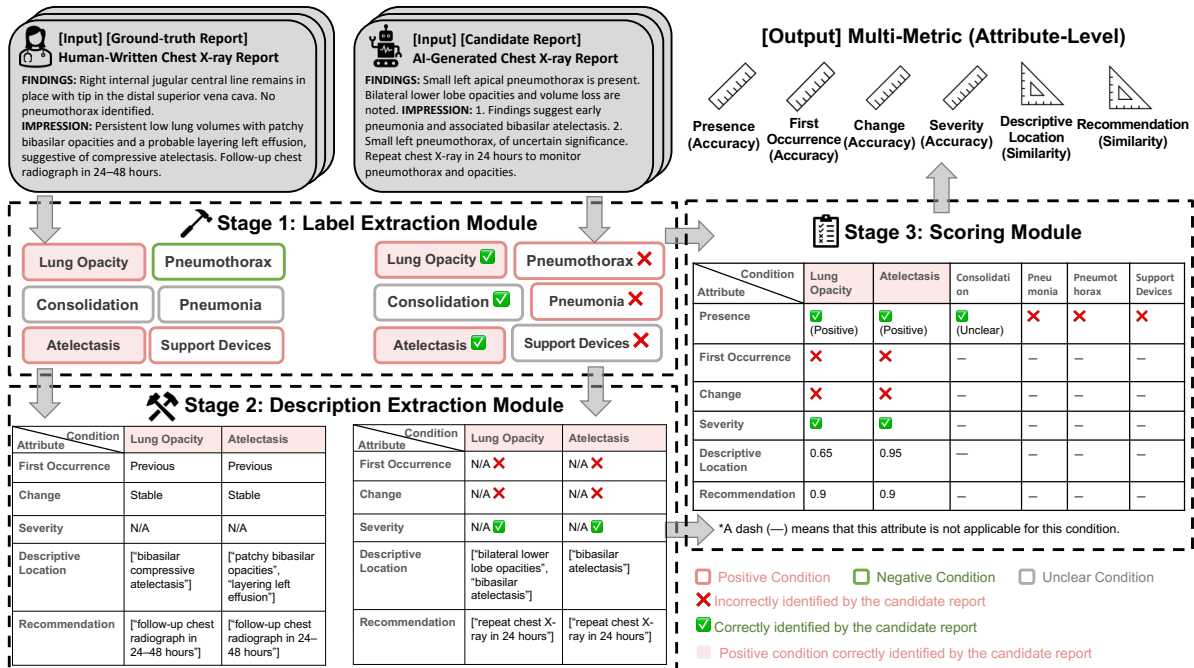
Figure 2: **CLEAR Framework.** Given a pair of ground-truth and candidate reports, CLEAR first assesses whether the candidate report can accurately identify a set of medical observations in the **label extraction module**. For each correctly identified positive condition, the **description extraction module** further evaluates the report's ability to describe the condition across five attributes: `first occurrence`, `change`, `severity`, `descriptive location`, and `recommendation`. Finally, the **scoring module** compiles and outputs the evaluation metrics.

accurately extracted attribute-level information according to clinical assessment; (iii) our proposed automated metrics served as effective proxies for expert scoring.

2 CLEAR Framework

We introduce the CLEAR framework, a hierarchical and fine-grained system for evaluating the clinical accuracy of radiology reports. CLEAR addresses both high-level diagnostic correctness and the descriptive quality of positive findings. As shown in Figure 2, CLEAR includes three sequential stages: label extraction, description extraction, and structured scoring.

Specifically, given a ground-truth and a candidate report pair, CLEAR first identifies whether the candidate correctly recognizes the presence or absence of specific medical conditions (Stage 1). It then examines, for each positively identified condition, whether the ground-truth and candidate reports are aligned across a set of expert-curated descriptive dimensions (Stage 2). Finally, it aggregates these evaluations into standardized, multi-dimensional metrics (Stage 3).

2.1 Stage 1: Label Extraction

This stage determines the presence or absence of 13 pre-defined medical conditions in the candidate re-

port, following the CheXpert structure (Irvin et al., 2019). Since accurately identifying and describing abnormalities is more clinically significant in radiology reporting, we exclude the "No Findings" label and focus on the remaining 13 conditions. Each condition is labeled as `positive`, `unclear`, or `negative` based on report content.

While existing labelers like CheXbert (Smit et al., 2020) and CheXpert (Irvin et al., 2019) are available, our pilot analysis (see Table 2) showed that their performance was limited. Since label extraction involves understanding and interpreting clinical narratives to assign structured labels, we hypothesized that LLMs could offer significant improvements over existing approaches. In particular, LLMs can handle complex linguistic nuances, such as negation, uncertainty, and context-dependent phrasing, more effectively in free-form radiology reports.

Base model variants and training strategies. We support three model scales: small (fine-tuned Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct), medium (Llama-3.3-70B-Instruct and Llama-3.1-70B-Instruct), and large (GPT-4o). For medium and large models, we applied different prompting strategies, including zero-shot (Prompt 1) and five-shot. For small models, we performed full-

Attribute	Value Set	NLP Task	Metric
Presence	$S_1 \in$ {"Positive", "Unclear", "Negative"}	Cls (Prompt 1)	Accuracy
Temporal Assessment			
First Occurrence	$S_2 \in$ {"Previous", "Current", "N/A"}	QA (Prompt 2)	Accuracy
Change	$S_3 \in$ {"Improving", "Stable", "Worsening", "Mixed", "N/A"}	QA (Prompt 3)	Accuracy
Description Assessment			
Severity	$S_4 \in$ {"Severe", "Moderate", "Mild", "Mixed", "N/A"}	QA (Prompt 4)	Accuracy
Descriptive Location	$S_5 = \{Entry_1, ..., Entry_n\}$ (e.g., $Entry_m$ = "left mid lung atelectasis")	IE (Prompt 5)	Similarity
Treatment Assessment			
Recommendation	$S_6 = \{Entry_1, ..., Entry_n\}$ (e.g., $Entry_m$ = "recommend follow-up at 4 weeks")	IE (Prompt 6)	Similarity

* Cls denotes "Classification," QA denotes "Question Answering," and IE denotes "Information Extraction."

Table 1: An overview of our expert-curated fine-grained attributes in CLEAR.

parameter fine-tuning using our curated dataset. To avoid overfitting, we first conducted hyperparameter tuning through 5-fold cross-validation and a grid search over learning rate, gradient accumulation steps, and number of epochs, followed by re-training on the full dataset. Full implementation details are provided in Appendix D.

Expert-in-the-loop label curation. High-quality labeled data is essential for training our label extraction models. To build a gold training dataset, we implemented a multi-stage annotation refinement with expert in the loop. We began with the test set from MIMIC-CXR-JPG (Johnson et al., 2024), which includes a single radiologist's annotations for 13 CheXpert conditions (Irvin et al., 2019). Each condition was originally labeled as positive, negative, unmentioned, or uncertain. In initial discussions with a radiologist, we identified two major issues with the original annotations: labeling errors (e.g., conditions mentioned in the report but left unlabeled) and category ambiguity (e.g., vague distinctions between negative and unmentioned). To address these, we used GPT-4o to pre-screen and re-label the reports, prompting it with the original MIMIC labeling guidelines. We then flagged cases with label mismatches between GPT-4o and the original annotations. We then asked an expert to re-annotate the discrepant cases. To reduce the radiologist's workload, reports with more than five mismatched condition labels were discarded from expert annotation, as such extensive disagreement often signals deeper interpretive ambiguities or quality issues in the original reports. While this introduced potential bias, we prioritized curating a high-quality subset over exhaustively correcting all samples. For the remaining reports, our collaborating radiologist independently re-annotated only

the discrepant conditions, reviewing the original report text without seeing prior labels. During the human annotation process, we observed that the original labeling schema lacked sufficient granularity to reflect the nuanced certainty levels expressed in radiology. In discussion with our expert radiologist, we expanded the label set to:{confidently present, likely present, neutral, likely absent, confidently absent}. In total, we curated 550 studies, each with high-quality labels for all 13 conditions. For consistency with prior work and to simplify downstream modeling, we further merged all labels into three classes {positive, negative, unclear}. A detailed description of the annotation process and instructions are provided in Appendix B.

2.2 Stage 2: Description Extraction

Building on the condition labels from Stage 1, this module extracts fine-grained clinical features that capture essential descriptive information for accurate reporting. The primary motivation is to transform the narrative text of radiology reports into a comprehensive, structured tabular format that distills all clinically significant attributes. In collaboration with two radiologists, we developed five clinically significant dimensions: first occurrence (whether the condition is newly observed), change (progression or improvement from prior studies), severity (the extent or intensity of the condition), descriptive location (specific anatomical site), and recommendation (suggested follow-up actions). These expert-developed attributes were specifically designed to reflect the nuanced but essential information radiologists routinely document when interpreting chest radiographs. By extracting these attributes, our approach enables a more comprehensive evaluation beyond simple condition

detection.

Implementation details. We used prompt-based methods to extract each of the five attributes from free-text reports. Each attribute was naturally framed as a standalone language understanding task. To operationalize this, we designed custom prompts tailored to the nature of each attribute: we used a Question Answering (QA) template to prompt the model for `first occurrence` (Prompt 2), `change` (Prompt 3), and `severity` (Prompt 4), and an Information Extraction (IE) template for `descriptive location` (Prompt 5) and `recommendation` (Prompt 6). For QA tasks, the model selects the best answer from multiple-choice options based on its understanding of the report. For IE tasks, it extracts relevant phrases guided by condition-specific example terminologies. Our prompt templates and terminology lists were reviewed by two radiologists and are summarized in Appendix E. We used a single model to process all five prompt types, one prompt per query to extract each attribute from a given report. We evaluated two model scales: a smaller Llama-3.1-8B-Instruct and a larger GPT-4o from OpenAI.

2.3 Stage 3: Scoring and Metrics

In this module, we process outputs from Stage 1 and Stage 2 into numeric metrics for each attribute. Given the i-th pair of ground-truth and candidate attribute sets, denote the attributes extracted from the ground-truth report as $\{S_j^{(i)}\}_{j=1}^6$ and from the candidate report as $\{\hat{S}_j^{(i)}\}_{j=1}^6$. An overview of the attributes is provided in Table 1.

For `presence` $(S_1, \hat{S}_1)$, we evaluated the accuracy of identifying `Positive` and `Negative` conditions. We defined a target class $c \in \{\texttt{Positive}, \texttt{Negative}\}$, treating all other labels as non-target. The corresponding binary F1 score, $F1_c$, was computed for each target class, resulting a positive-F1 and negative-F1. We report these scores at three levels: micro average, Top-5 condition average[2], and across all 13 conditions.

For `first occurrence` $(S_2, \hat{S}_2)$, `change` $(S_3, \hat{S}_3)$, and `severity` $(S_4, \hat{S}_4)$, we assessed the exact match between predictions and ground truth. Considering that these attributes are framed as multiple-choice questions in the prompt, exact match is a natural and appropriate metric. Accu-

racy was calculated as $\text{Acc.}_j = \frac{\sum_i \mathbb{1}[S_j^{(i)} = \hat{S}_j^{(i)}]}{\sum_i 1}$. We report accuracy at the micro level, as well as averaged across reports and the 13 conditions.

For `descriptive location` $(S_5, \hat{S}_5)$ and `recommendation` $(S_6, \hat{S}_6)$, which involve free-text descriptions, we measured phrase-level similarity against clinically meaningful expressions. To evaluate alignment, we first used optimal matching–based metrics with similarity scores such as BLEU-4 (Papineni et al., 2002) and ROUGE-L (Lin, 2004):

$$\text{Score}_j^{(i)} = \frac{1}{|S_j^{(i)}|} \sum_{e \in S_j^{(i)}} \max_{\hat{e} \in \hat{S}_j^{(i)}} \text{Similarity}(e, \hat{e}),$$

where $S_j^{(i)} = \{e_k\}_{k=1}^n$ and $\hat{S}_j^{(i)} = \{\hat{e}_k\}_{k=1}^{n'}$. Additionally, to better approximate clinical judgment from an expert's perspective, we prompted o1-mini (Prompt 8) to directly compare each attribute pair and return a similarity score in the range $[0, 1]$.

3 CLEAR-Bench: Attribute-Level Expert Alignment Dataset

In this section, we introduce CLEAR-Bench, an expert-curated, attribute-level dataset in collaboration with five radiologists. Inspired by recent expert evaluation datasets for chest X-ray reports (Tian et al., 2023; Yu et al., 2023a; Rao et al., 2025), CLEAR-Bench is specifically designed to assess how well automated evaluators such as CLEAR align with radiologist judgments. CLEAR-Bench consists of two annotation subsets: expert ensemble labels and expert-curated attributes. We defer full details of the instruction criteria, interface design, and annotation workflow to Appendix B.

Expert ensemble labels. These provide the ground-truth labels for the `Presence` attribute. We randomly selected 100 studies from the validation and test sets of MIMIC-CXR-JPG (Johnson et al., 2024), excluding any training samples and normal studies. Each report was independently annotated from scratch by three board-certified radiologists. During annotation, the radiologists categorized each of 13 CheXpert conditions (Irvin et al., 2019) into one of five categories (`confidently absent`, `likely absent`, `neutral`, `likely present`, and `confidently present`) based on their best interpretation of the report. After the initial round of annotations, we merged `confidently present` and `likely`

---

[2]Top five conditions in MIMIC-CXR-JPG are Pneumothorax, Pneumonia, Edema, Pleural Effusion, and Consolidation.

Experiments	Pos F1@13	Pos F1@5	Pos F1 (micro)	Neg F1@13	Neg F1@5	Neg F1 (micro)
			LARGE MODELS			
GPT-4o (base)	**0.805**	0.929	0.934	0.476	0.648	0.815
GPT-4o (5-shot)	0.795	**0.940**	**0.934**	0.510	0.723	0.842
			MEDIUM MODELS			
Llama-3.1-70B-Instruct (base)	0.782	0.890	0.924	0.630	0.850	0.920
Llama-3.1-70B-Instruct (5-shot)	0.794	0.916	0.924	**0.744**	0.890	**0.958**
Llama-3.3-70B-Instruct (base)	0.780	0.894	0.925	0.602	0.876	0.926
Llama-3.3-70B-Instruct (5-shot)	0.781	0.907	0.926	0.695	**0.892**	0.953
			SMALL MODELS			
Llama-3.1-8B-Instruct (base)	0.736	0.880	0.910	0.418	0.660	0.714
Llama-3.1-8B-Instruct (550 finetune)	0.729	0.806	0.905	0.482	0.803	0.949
Qwen2.5-7B-Instruct (base)	0.694	0.834	0.880	0.413	0.616	0.736
Qwen2.5-7B-Instruct (550 finetune)	0.727	0.800	0.905	0.511	0.849	0.953
			BASELINES			
CheXbert (Smit et al., 2020)	0.695	0.833	0.897	0.498	0.877	0.952
CheXpert (Irvin et al., 2019)	0.674	0.811	0.888	0.522	0.831	0.948
Δ **Improvement over SOTA**	+15.8%	+12.8%	+4.1%	+42.5%	+1.7%	+0.06%

Table 2: Evaluation of the label extraction module. CLEAR outperforms existing labelers across all metrics in identifying both positive and negative conditions. Specifically, larger models perform better at capturing positive conditions, while techniques such as 5-shot prompting and supervised fine-tuning significantly improve the detection of negative conditions.

present into the single category `positive`, while `likely absent` and `confidently absent` were merged into the single category `negative`. We then assessed agreement across annotators. Remaining disagreements were first resolved by majority vote, followed by a consensus discussion for any unresolved conflicts. The finalized dataset served as the ground truth for evaluating model performance in the Label Extraction Module.

Expert-curated attributes. These cover the remaining five report attributes: `first occurrence`, `change`, `severity`, `descriptive location`, and `recommendation`. We began by preparing two sets of model-generated attributes, one from Llama-3.1-8B-Instruct and the other from GPT-4o, for each positive condition identified in the expert ensemble labels. These two sets were merged and then randomly split into two review sets, each with 50 samples from Llama and 50 samples from GPT-4o. Each set was independently reviewed by separate radiologists. During curation, each radiologist first rated each attribute as `incorrect`, `partially correct`, or `correct`. For non-correct attributes, the radiologist also provided a revised version, which was used to construct the ground-truth attribute set.

4 Experiments

Experimental setup. To evaluate the effectiveness and clinical reliability of our proposed CLEAR framework, we conducted experiments using CLEAR-Bench. For the Label Extraction Module, we compared CLEAR's performance against

two established baselines, the BERT-based labeler CheXbert (Smit et al., 2020) and the rule-based labeler CheXpert (Irvin et al., 2019), using the Expert Ensemble Labels from CLEAR-Bench. We report F1 scores as introduced in Section 2.3. For the Description Extraction Module, we evaluated CLEAR using the Expert-Curated Attributes from CLEAR-Bench. As no prior baselines exist for this task, we report expert evaluation scores directly, along with automated metrics defined in Section 2.3.

LLM-based labeler achieves substantial gains over existing labelers. We begin with evaluating the performance of the Label Extraction Module. As shown in Table 2, our text generation-based approach (Prompt 1) significantly outperforms the best BERT-based labeler (Smit et al., 2020) and the top rule-based labeler (Irvin et al., 2019) across all accuracy metrics. In identifying positive conditions, our module achieved a notable improvement in accuracy averaged over all 13 medical conditions (+15.8%), with smaller increase on the Top 5 conditions (+12.8%) and the full label pool (+4.1%). This is likely because text generation models can understand the full sentence and overall report, instead of relying on token-level classification or hard-coded rules. Furthermore, this contextual understanding generalizes across conditions, especially for rare conditions (e.g., fracture) where BERT-based models struggle due to data imbalance, and unseen patterns (e.g., pleural other) where rule-based systems fail to capture beyond their predefined scope. This advantage is even more evident in negative conditions, which

Metric	First Occurrence		Change		Severity		Descriptive Location		Recommendation	
	GPT-4o	Llama 8B	GPT-4o	Llama 8B	GPT-4o	Llama 8B	GPT-4o	Llama 8B	GPT-4o	Llama 8B
EXPERT EVALUATION SCORES										
Experts (condition averaged)	**0.818**	0.685	0.837	0.685	**0.809**	0.565	0.857	0.761	0.933	**0.474**
Experts (report averaged)	0.783	**0.680**	**0.867**	**0.688**	0.771	**0.583**	0.872	**0.763**	**0.940**	0.416
Experts (micro)	0.777	0.662	0.855	0.663	0.777	0.570	**0.867**	0.757	0.936	0.404
ACCURACY METRICS										
Acc. (condition averaged)	0.740	0.688	0.710	0.589	0.682	0.470	–	–	–	–
Acc. (report averaged)	0.755	0.679	0.759	0.596	0.685	0.532	–	–	–	–
Acc. (micro)	0.737	0.665	0.754	0.575	0.671	0.494	–	–	–	–
SIMILARITY METRICS										
o1-mini (micro)	–	–	–	–	–	–	0.785	0.739	0.888	0.361
ROUGE-L (micro)	–	–	–	–	–	–	0.686	0.672	0.887	0.268
BLEU-4 (micro)	–	–	–	–	–	–	0.500	0.402	0.885	0.263
Average (experts)	0.793	0.676	0.853	0.679	0.786	0.573	0.865	0.760	0.936	0.431
Average (all)	0.768	0.677	0.797	0.633	0.733	0.536	0.761	0.682	0.911	0.364
Δ(GPT-4o − Llama)	+0.091		+0.164		+0.197		+0.079		+0.547	

> * A dash (–) indicates the metric is not applicable for this attribute.
> * Bold values highlight the highest scores per metric. Colored cells distinguish GPT-4o (green) from Llama 8B (yellow).
> * The bottom row shows the difference between GPT-4o and Llama 8B for the "Average (all)" metric.

Table 3: Evaluation of the description extraction module. Expert ratings are averaged across all samples (0 = incorrect, 0.5 = partially correct, 1 = correct). According to radiologists' clinical judgment, CLEAR can accurately extract attribute-level information from free-text reports. Additionally, GPT-4o is consistently preferred over Llama-3.1-8B-Instruct, though Llama performs reasonably well, especially on `descriptive location`, and remains a low-cost, open-source option.

require interpreting implicit cues (e.g., "lungs are clear"). Our module achieved a substantial boost (+42.5%) in average accuracy across all conditions, highlighting once again its strength in semantic understanding beyond explicit mentions.

Ablation study of model scales and adaptation. For identifying positive clinical findings, model scale plays a major role, with GPT-4o achieving the highest performance across all accuracy metrics. In contrast, model adaptation strategies, including both few-shot prompting and supervised fine-tuning, have relatively limited impact compared to each base model. This is likely because the base models already encode sufficient clinical knowledge to accurately identify positive findings, and larger model scales are more strongly related with the richness of this knowledge. However, when it comes to negative mentions, model adaptation strategies stand out, with all metrics improving notably across scales. The reason is that these strategies effectively incorporate expert-derived "side" information, which is typically not captured by base models during pre-training, through few-shot examples or supervised training data. Specifically, among different strategies, supervised fine-tuning consistently outperformed few-shot prompting, with average gains of 26.8% for small models from fine-tuning, 7.9% for medium models from few-shot, and 7.3% for large models from few-shot.

LLMs, especially GPT-4o, excel at fine-grained attribute extraction. We next probe our description extraction module to assess how reliably a unified language model can handle all five fine-grained attributes (see Table 3). Overall, GPT-4o showed strong performance across all five attributes, achieving the highest average score of 0.911 (`recommendation` average all) and a minimum of 0.733 (`severity`). When analyzing by task type, GPT-4o performed better on IE tasks (`location` and `recommendation`), with an average score of 0.836, particularly for attributes that involve highly formulaic language (e.g., "follow-up imaging recommended to assess the resolution of opacity" for `recommendation`). In contrast, it achieved a relatively lower score of 0.766 on QA tasks (`first occurrence`, `change`, and `severity`), which typically require deeper clinical contextual understanding. In comparison, Llama-3.1-8B-Instruct (a small-scale model) showed mixed performance across attributes. In QA tasks, it captured temporal information reasonably well, scoring 0.677 for `first Occurrence` average all and 0.633 for `change`, though its interpretation of clinical findings was weaker (0.536 for `severity`). As for IE tasks, hallucinations significantly affected performance, but with a customized terminology list (see Table 9), it achieved 0.682 on `location`, the closest to GPT-4o. How-

Automated Metric	Corr. with Expert Scoring
Accuracy Metrics produced by CLEAR	
Acc. (condition averaged)	0.894
Acc. (report averaged)	0.908
Acc. (micro)	0.915
Similarity Metrics produced by CLEAR	
o1-mini (micro)	0.994
ROUGE-L (micro)	0.977
BLEU-4 (micro)	0.811

Table 4: Pearson correlation between CLEAR and expert scores. All of automated metrics generated by CLEAR show strong alignment with expert evaluations.

ever, unrelated descriptive phrases (e.g., "signs of generalized fluid overload") significantly lowered `recommendation` score to 0.364.

CLEAR aligns well with expert ratings. Generally, all the implementations of CLEAR were highly correlated with expert scoring, as shown in Table 4. However, automated metrics were typically slightly lower than expert scores, as observed in Table 3, likely because similarity metrics based on ROUGE-L and BLEU-4 prioritize exact matches against ground truth, whereas expert scoring includes a `Partially Correct` category, allowing some tolerance for clinically reasonable but not perfectly matched responses. This distinction is further supported by the exceptionally high correlation of o1-mini scores with expert ratings, reaching 0.994. Compared with other lexical metrics, o1-mini can more effectively capture semantic and clinical alignment, making it a closer proxy to expert judgment.

5 Related Work

Lexical metrics. Traditional word-overlap metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004),and METEOR (Banerjee and Lavie, 2005) are commonly used in natural language generation tasks and are therefore also commonly applied to radiology report generation. However, these metrics fail to capture subtle semantic nuances, such as negations or synonyms, which are critical in the clinical domain. Embedding-based metrics such as BERTScore (Zhang et al., 2020) improve on semantic matching but remain inadequate in capturing nuanced semantics and domain-specific medical terms, thereby missing clinically important errors.

Clinical efficacy metrics. To bridge the gap between surface-level fluency and clinical correctness, domain-specific metrics have been introduced.

Label-based metrics such as CheXpert (Irvin et al., 2019) map reports to 14 predefined clinical labels and measure classification accuracy, but their rule-based pipelines propagate annotation noise. CheXbert (Smit et al., 2020) improves semantic understanding over CheXpert by fine-tuning BERT-based classifiers; however, it still lags behind recent LLMs due to the limited capacity of BERT compared with newer and more powerful language models. More recent entity-centric methods such as RadGraph F1 (Jain et al., 2021), Rad-Graph2 (Khanna et al., 2023), MEDCON (Yim et al., 2023) and RaTEScore (Zhao et al., 2024) capture subject–relation–object triples. Although these approaches effectively identify and compare medical entities and their relationships, they often lack the granularity to evaluate specific attributes such as severity, temporal progression, or treatments. To better align automatic metrics with radiologist judgments, RadCliQ (Yu et al., 2023b) combines BLEU, BERTScore, CheXbert similarity, and RadGraph F1 into a weighted score learned from 160 radiologist-annotated report pairs (ReX-Val). These annotations are provided at an aggregate level, quantifying the total number of clinically significant and insignificant errors without distinguishing specific clinical attributes.

LLM-based metrics. More recently, researchers have been using LLMs to assess radiology reports. Several methods, including GREEN and CheXprompt, build on six categories of the clinical-error taxonomy introduced in RadCliQ. GREEN (Ostmeier et al., 2024) tallies the number of errors and matched findings of each type and then aggregates them into a single report-level score, which limits granularity and makes it difficult to isolate specific mistakes. CheXprompt (Zambrano Chaves et al., 2025) uses GPT-4 to quantify clinically significant and insignificant errors in radiology reports, categorizing them into six predefined types. Similarly, it focuses primarily on counting these errors without delving into the nuanced contextual attributes of each error instance. FineRadScore (Huang et al., 2024) takes a different route: it calculates the minimum line-by-line edits required to transform a generated report into a reference report. While this encourages precision, it penalizes semantically equivalent but differently phrased outputs. RadFact (Bannur et al., 2024) decomposes each report into atomic sentences and uses LLM to determine whether each generated sentence is entailed by the

reference report, which does not differentiate different types of clinical errors or severity.

6 Conclusion

We present CLEAR, the first clinically grounded, attribute-level evaluation framework that leverages LLMs to convert free-text radiology reports into a structured tabular format. CLEAR consists of three components: (1) a label extraction module to assess the accurate identification of medical conditions; (2) a description extraction module to evaluate the precision of condition descriptions; and (3) a scoring module to compile multi-metric evaluation results. We also introduce CLEAR-Bench, an expert-curated alignment dataset covering 6 report attributes and 13 medical conditions. Our experiments demonstrated that CLEAR can effectively identify clinical conditions, faithfully extract attribute-level information consistent with clinical validation, and provide automated metrics that serve as reliable proxies for expert scoring.

Limitations

While CLEAR provides a clinically grounded framework and demonstrates strong alignment with expert clinical assessment, it has several limitations. First, like all existing evaluation metrics, CLEAR relies solely on ground-truth reports without incorporating image information, overlooking the fact that reference reports may not fully capture all relevant findings present in the image. Future work could explore integrating image-based evaluation to better reflect clinical completeness. Second, CLEAR is built on the CheXpert label structure, which is limited in both granularity and anatomical coverage. Extending the framework to include additional specialties such as breast imaging, cardiology, and gastroenterology in the future could enhance its generalizability. Lastly, although we prioritize high-quality annotations, both the training and evaluation datasets remain relatively small due to the common tradeoff between annotation quality and dataset scale.

Acknowledgements

References

Anthropic. 2025. Claude 3.7 sonnet system card. Accessed: 2025-05-19.

Satanjeev Banerjee and Alon Lavie. 2005. 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*, pages 65–72.

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando P'erez-Garc'ia, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Prasanna Ranjit, Shaury Srivastav, Julia Gong, Fabian Falck, Ozan Oktay, Anja Thieme, Matthew P. Lungren, Maria T. A. Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. 2024. Maira-2: Grounded radiology report generation. *arXiv*, abs/2406.04449.

Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A. Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, Hyo Jung Hong, Mehr Kashyap, Akash R. Chaurasia, Nirav R. Shah, Karandeep Singh, Troy Tazbaz, Arnold Milstein, Michael A. Pfeffer, and Nigam H. Shah. 2025. Testing and evaluation of health care applications of large language models: A systematic review. *JAMA*, 333(4):319–328.

Deep Ganguli, Nicholas Schiefer, Marina Favaro, and Jack Clark. 2023. Challenges in evaluating AI systems.

Alyssa Huang, Oishi Banerjee, Kay Wu, Eduardo Pontes Reis, and Pranav Rajpurkar. 2024. Fineradscore: A radiology report line-by-line evaluation technique generating corrections with severity scores. In *Machine Learning for Healthcare Conference*. PMLR.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, and 1 others. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew Lungren, Andrew Ng, Curtis Langlotz, Pranav Rajpurkar, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Alistair Johnson, Matthew Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. 2024. Mimic-cxr-jpg - chest radiographs with structured labels (version 2.1.0). https://doi.org/10.13026/jsn5-t979.

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. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Sameer Khanna, Adam Dejl, Kibo Yoon, Steven QH Truong, Hanh Duong, Agustina Saenz, and Pranav Rajpurkar. 2023. Radgraph2: Modeling disease progression in radiology reports via hierarchical information extraction. In *Machine Learning for Healthcare Conference*, pages 381–402. PMLR.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Curtis P. Langlotz. 2015. *The Radiology Report: A Guide to Thoughtful Communication for Radiologists and Other Medical Professionals*. CreateSpace Independent Publishing Platform, North Charleston, SC.

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

Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Tobi Walsh, Armin Hamrah, Lapo Santarlasci, and 4 others. 2025. The ai index 2025 annual report. Technical report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA. https://hai.stanford.edu/ai-index/2025-ai-index-report.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, and 1 others. 2025. Towards accurate differential diagnosis with large language models. *Nature*, pages 1–7.

OpenAI. 2025. Introducing gpt-4.5. Accessed: 2025-05-19.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Md, Michael Moseley, Curtis Langlotz, Akshay Chaudhari, and 1 others. 2024. Green: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 374–390.

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

Magdalini Paschali, Zhihong Chen, Louis Blankemeier, Maya Varma, Alaa Youssef, Christian Bluethgen, Curtis Langlotz, Sergios Gatidis, and Akshay Chaudhari. 2025. Foundation models in radiology: What, how, why, and why not. *Radiology*, 314(2):e240597.

V. Rao, S. Zhang, J. Acosta, S. Adithan, and P. Rajpurkar. 2025. Rexerr-v1: Clinically meaningful chest x-ray report errors derived from mimic-cxr. PhysioNet.

Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Ramona Comanescu, Canfer Akbulut, Tom Stepleton, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, William Isaac, and Laura Weidinger. 2024. Gaps in the safety evaluation of generative ai. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):1200–1217.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. 2020. 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)*, pages 1500–1519.

Kevin Tian, S. J. Hartung, A. A. Li, J. Jeong, F. Behzadi, J. Calle-Toro, S. Adithan, M. Pohlen, D. Osayande, and P. Rajpurkar. 2023. Refisco: Report fix and score dataset for radiology report generation. PhysioNet.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, and 7 others. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, pages 1–9.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific data*, 10(1):586.

F. Yu, M. Endo, R. Krishnan, I. Pan, A. Tsai, E. P. Reis, E. Kaiser Ururahy Nunes Fonseca, H. Lee, Z. Shakeri, A. Ng, C. Langlotz, V. K. Venugopal, and P. Rajpurkar. 2023a. Radiology report expert evaluation (rexval) dataset (version 1.0.0).

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, and 1 others. 2023b. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9).

15923

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, and 8 others. 2025. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(1):3108.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhenxuan Zhang, Kinhei Lee, Weihang Deng, Huichi Zhou, Zihao Jin, Jiahao Huang, Zhifan Gao, Dominic C Marshall, Yingying Fang, and Guang Yang. 2025. Gema-score: Granular explainable multi-agent score for radiology report evaluation. *arXiv preprint arXiv:2503.05347*.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. RaTEScore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15004–15019, Miami, Florida, USA. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Appendix

A   Open-sourced Artifacts

We release the CLEAR codebase at `https://github.com/ChicagoHAI/CLEAR-evaluator` . The current version supports both open-source models via the vLLM backend and closed-source models through the Azure OpenAI API.

Our ground-truth dataset, CLEAR-Bench, along with comprehensive documentation, is publicly available on PhysioNet (`https://physionet.org/`) to facilitate future research. Licensing and citation guidelines for using the dataset are detailed in the accompanying documentation.

B   Data Annotation and Curation

We accessed MIMIC-CXR-JPG data by following the required steps on `https://physionet.org/content/mimic-cxr-jpg/2.1.0/`. We first registered and applied to be a credentialed user, and then completed the required training of CITI Data or Specimens Only Research. Data license can be found at `https://physionet.org/content/mimic-cxr-jpg/view-license/2.1.0/`.

During each human annotation process, we followed a traditional paradigm: initial pilot rounds were conducted to gather user feedback, followed by formal, independent large-scale annotation, data analysis for quality control, and final resolution via consensus discussion. Our annotation platform is built upon an open source data labeling tool, Label Studio (Tkachenko et al., 2020-2022).

B.1   Label Structure Refinement

MIMIC-CXR-JPG Labeling Criteria

Positive (1.0): The label is positively mentioned in the report and present in one or more associated images.
Example: "A large pleural effusion"

Negative (0.0): The label is negatively mentioned in the report and should not be present in any associated image.
Example: "No pneumothorax."

Uncertain (-1.0): The label is either: (1) mentioned with uncertainty, so presence in the image is unclear; or (2) described ambiguously, with uncertain existence.
Explicit uncertainty: "The cardiac size cannot be evaluated."
Ambiguous language: "The cardiac contours are stable."

Unmentioned (Missing): The label is not mentioned in the report at all.

Figure 3: 4-type labeling criteria in MIMIC.

During the interaction of pilot training, we closely worked with all involved radiologists and collect a lot of valuable feedback regarding the user experience with designed interfaces and task instruction.

After summarizing input feedback, we recognized some shared and repeatedly mentioned issues in the 4-type label structure of MIMIC-CXR-JPG (see Figure 3): (1) The "unmentioned" category has a high degree of overlap with other categories, particularly with "negative" labels. This is because radiologists often do not explicitly state negative findings in the report. However, indirect phrases

such as "Lungs are clear" can implicitly negate a wide range of lung-related abnormalities. (2) Additionally, different radiologists have varying tendencies in labeling conditions. More conservative radiologists may lean toward assigning "uncertain" rather than "positive" labels, even when the evidence suggests a likely presence. This inconsistency introduced label noise and ambiguity, particularly when these labels are used for supervised training or evaluation purposes.

Our Refined Labeling Criteria

Confidently Absent: The condition is clearly stated as not present in the report.
Example: "No pneumothorax."

Likely Absent: The report implies the condition is likely absent, but the language is ambiguous or uncertain.
Example: "Heart size is normal though increased."

Neutral: The report does not clearly indicate presence or absence.
Explicit uncertainty: "The cardiac size cannot be evaluated."
Ambiguous language: "The cardiac contours are stable."

Likely Present: The report suggests the condition may be present, but uses uncertain or ambiguous language.
Example: "Likely reflecting compressive atelectasis."

Confidently Present: The condition is clearly stated as present in the report.
Example: "A small right pleural effusion."

Figure 4: Our refined 5-type labeling criteria during expert annotation.

Therefore, we refined the original MIMIC label structure into a "5+1" annotation framework. The "5" refers to an extension of MIMIC's original "Positive," "Negative," and "Uncertain" categories into five more nuanced types, as shown in Figure 4. The "+1" refers to retaining the "Unmentioned" label as a separate flag. Specifically, radiologists were asked to select one of the five labels for each condition and additionally indicate whether this label was explicitly mentioned in the report or not.

After collecting radiologist responses, we mapped the five types into a final three-type scheme for downstream use: "Confidently Present" and "Likely Present" were merged into "Positive," "Confidently Absent" and "Likely Absent" were merged into "Negative," and "Neutral" was renamed as "Unclear." We then proceeded with inter-rater alignment checks for quality control. Notably, the "mentioned" flag was not incorporated into the
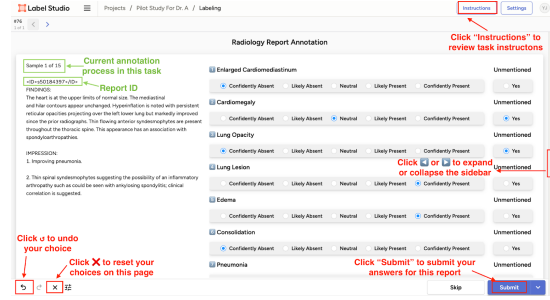

Figure 5: Interface for Label Annotation.

final label itself but served as a supporting indicator for data managers to differentiate between labeling disagreements due to quality issues versus differences in individual clinical interpretation. This overall process enabled us to accommodate variability in radiologist judgment while maintaining high annotation quality.

B.2 Expert-in-the-loop Dataset Curation

We first excluded two cases without any "FINDINGS" or "IMPRESSION" sections in the radiology report and 30 cases labeled as "No Finding" in the radiologist annotation dataset from MIMIC-CXR-JPG (containing 687 studies in total). Then, we randomly selected 20 cases to serve as a pilot set for initial review and refinement of the process.

We then prompted GPT-4o to generate condition labels following the same guidelines used in the original MIMIC documentation for the remaining studies excluding the 20 pilot cases. After identifying discrepancies between the model-generated labels and the original dataset annotations, we isolated the suspected noisy labels for further review.

For each case, we extracted only the relevant report sections (FINDINGS and IMPRESSION), with no images involved, and presented them to a board-certified radiologist. The radiologist independently re-annotated the report from scratch based on their clinical judgment.

During the curation, we discarded five cases due to GPT-4o generation failures. To manage the annotation workload, we limited each review to reports with one to five mismatched conditions per case.

The full curation process took approximately one month, resulting in 550 finalized reports, each annotated with 13 condition labels.

Task instruction can be checked in Figure 8 and the interface can be checked in Figure 5.

B.3 CLEAR-Bench: Expert Ensemble

After excluding "No Finding" cases and those already annotated in the curation stage, we selected
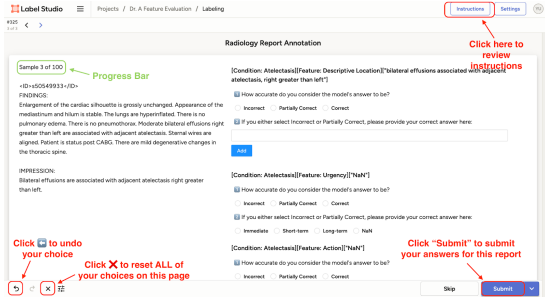
Figure 6: Interface for Attribute Curation.

five cases for pilot training and randomly sampled 100 reports from the test and validation sets of MIMIC-CXR-JPG to construct our final evaluation dataset.

Following a brief onboarding process using 5 pilot cases, we collected independent annotations from three radiologists, each labeling the 100 reports from scratch. After an initial round of majority voting, 25 reports with 32 individual condition labels in total remained unresolved. These were finalized through a single round of discussion and consensus among the experts.

The full expert ensemble workflow was completed over the course of three months, resulting in 100 fully annotated reports, each with 13 condition labels.

Task instruction can be checked in Figure 8 and interface can be checked in Figure 5.

B.4 CLEAR-Bench: Attribute Curation

The blueprint for attribute design was initially inspired by the concept of an "Attribute-Value Format" proposed by Dr. Langlotz in his practical guide to writing radiology reports (Langlotz, 2015, 207). Driven by this concept, we generated a list of commonly used report attributes with the assistance of GPT-4o, and refined it through discussion with our collaborating research radiologist, who is also a co-author. Together, we determined which attributes to include, revise, or remove. During this process, we not only developed a concise yet comprehensive attribute structure and also collected useful example phrases and sentences for each attribute. These examples were later incorporated into the prompts used in the Description Extraction Module (see Appendix E). The final version of the prompt set and word list was also reviewed and approved by a clinical radiologist.

We curated attributes using the same 100 studies described earlier, excluding two cases that lacked any positively identified conditions in expert ensemble labels. Following a round of pilot training,

Metric	Correlation
BERTScore (Zhang et al., 2020)	0.27
METEOR (Banerjee and Lavie, 2005)	0.49
GREEN (Ostmeier et al., 2024)	0.63
GEMA-Score (Zhang et al., 2025)	0.70
CLEAR (GPT-4o, o1-mini)	0.70

Table 5: Pearson correlation with clinical judgment on ReXVal. All baseline results are reproduced from (Zhang et al., 2025).

the formal curation process proceeded as detailed in Section 3. After collecting radiologist responses, we conducted a second round of quality control to finalize the ground-truth attributes. The full human curation process took approximately one month.

Task instructions are shown in Figure 9, and the annotation interface is illustrated in Figure 6.

C Comparison

C.1 CLEAR vs. Existing LLM-based Metrics

Notably, most existing LLM-based metrics yield a single holistic score for a (candidate, reference) pair and do not evaluate the structured, fine-grained *attribute-level* information captured by our expert annotations (e.g., *location* of pneumonia), which makes a like-for-like comparison between CLEAR and existing metrics impossible in Table 4.

However, for the consideration of completeness, in this section, we constructed a naive scalarization of CLEAR by averaging its condition–attribute outputs into a single score, and evaluated it on ReXVal (Yu et al., 2023a) following the latest GEMA protocol (Zhang et al., 2025). Correlation results are reported in Table 5. Under this setup, the scalarized CLEAR performs on par with GEMA-Score (Zhang et al., 2025) and outperforms other recent semantic or clinical-efficacy metrics (e.g., GREEN (Ostmeier et al., 2024)), while still providing interpretable, condition-specific feedback that report-level metrics lack.

We additionally present a brief case study comparing CLEAR with GEMA-Score in Figure 7.

Methodologically, GEMA-Score and other LLM-based radiology metrics (e.g., GREEN (Ostmeier et al., 2024), FineRadScore (Huang et al., 2024)) follow a common paradigm: identify errors using a fixed error-category taxonomy, aggregate them into a single report-level score (e.g., F1), and assess alignment with radiologist judgments on expert-annotated sets. This paradigm is inherently constrained to single-score outputs compatible with datasets such as ReXVal (Zhang et al., 2025). In

Aspect	ReXVal	CLEAR-Bench
Report Section Used	Impression only	Full report (Findings + Impression)
Annotation Scope	6 coarse error categories (e.g., false positive, omission, comparison)	13 conditions $\times$ 6 expert-curated clinical attributes
Error Distribution	88.7% of annotations had 0 errors; location & change underrepresented	All reports selected for at least one positive finding; balanced coverage of all 6 attributes

Table 6: Comparison between ReXVal and CLEAR-Bench.

contrast, CLEAR is designed around a clinically interpretable, two-dimensional *condition–attribute* framework. Rather than collapsing quality into one number, CLEAR produces a multi-metric "examination sheet" that explicitly indicates which conditions, attributes, and condition–attribute pairs are present or missing in generated reports.

C.2 CLEAR-Bench vs. Existing Expert Evaluation Datasets

Commonly used datasets such as ReXVal (Yu et al., 2023a) are not directly compatible with CLEAR's condition-by-attribute evaluation design. ReXVal focuses on the *impression* section, lacks structured attribute annotations, and is highly sparse—over 88% of annotator ratings indicate zero errors, with most labeled errors concentrated in a single attribute ("presence" of disease). These characteristics make it challenging to evaluate attribute-level metrics like CLEAR.

To facilitate direct comparison where possible, we also report CLEAR's performance on ReXVal (Yu et al., 2023a) using the naive scalarization described above. Nevertheless, while ReXVal remains useful for coarse, report-level scoring, CLEAR-Bench is purpose-built to evaluate condition- and attribute-level fidelity and thus provides a more rigorous foundation for assessing the full capabilities of CLEAR.

Detailed comparison can be found in Table 6.

D CLEAR: Implementation Details

Base Model	GAS	LR	Epochs
Llama3.1-8B-Instruct	1	7.0×10^{-6}	4
Qwen2.5-7B-Instruct	1	9.0×10^{-6}	5

Table 7: Hyperparameter search results. GAS denotes the number of gradient-accumulation steps, LR the learning rate, and Epochs the total training epochs.

Supervised finetuning details. All fine-tuned models were obtained through supervised fine-tuning with LLaMA-Factory (Zheng et al., 2024).

To identify an optimal configuration, we developed an automated hyperparameter optimization (HPO) framework that combines five-fold cross-validation with a grid search. Learning rate, number of epoch, and gradient accumulation steps were three objects to be optimized. For learning rate, the searching space wad $[3.0e^{-6}, 3.0e^{-5}]$, with an interval of $2.0e^{-6}$. For epoch, searching space is $\{2, 3, 4, 5\}$. For gradient accumulation steps, the searching target was $\{1, 2, 4\}$. We conducted extensive experiments to assess each hyperparameter's influence. A total of 360 models were finetuned for one base model to determine the best hyperparameter settings. The best-performing settings, summarized in Table 7, were used for all experiments reported in Table 2. Hyperparameter optimization and model training were performed on NVIDIA A100 80G and NVIDIA H100 94G GPUs. The HPO stage took 93 h 51 m 20 s on four A100s and 14 h 39 m 36 s on four H100s.

Model	Standard Pricing (per 1M Tokens)
GPT-4o-2024-1120 (Global)	Input: $2.50 Cached: $1.25 Output: $10.00
o1-mini-2024-09-12 (Global)	Input: $1.10 Cached: $0.55 Output: $4.40

Table 8: Standard API pricing per 1M tokens for GPT-4o and o1-mini models, based on Azure OpenAI pricing: https://azure.microsoft.com/en-us/pricing/details/cognitive-services/openai-service/#pricing.

Inference details for local models. We served the models locally with vLLM (0.8.5.post1) (Kwon et al., 2023). Inference runs were made with a temperature of 1e-5 and a max_tokens of 4,096; all other sampling parameters remained at their default settings. A single NVIDIA A100 80G was sufficient for inference under this setting.

API Details We access OpenAI's GPT-4o (2024-11-20) and o1-mini (2024-09-12) via Microsoft's Azure. Pricing details can be checked in Table 8.

Groundtruth Report

Findings: Right internal jugular central line remains in place with tip in the distal superior vena cava. No pneumothorax identified.

Impression: Persistent low lung volumes with patchy bibasilar opacities and a probable layering left effusion, suggestive of compressive atelectasis. Follow-up chest radiograph in 24--48 hours.

Candidate Report

Findings: Small left apical pneumothorax is present. Bilateral lower lobe opacities and volume loss are noted.

Impression: (1) Findings suggest early pneumonia and associated bibasilar atelectasis. (2) Small left pneumothorax, of uncertain significance. Repeat chest X-ray in 24 hours to monitor pneumothorax and opacities.

1. GEMA-Score: 0.547

Objective Score: 0.5
- False prediction of finding: 1
- Omission of finding: 1
- Incorrect location/position: 2
- Omission of location: 0
- Incorrect severity: 1
- Omission of severity: 1
- Mention of uncertainty not in reference: 1
- Omission of uncertainty in reference: 1

Subjective Score:
- completeness_score: 0.6
- readability_score: 0.8
- clinical_utility_score: 0.8

2. CLEAR (GPT-4o, o1-mini): (0.51 + 0.57 + 0 + 0 + 0)/5 = **0.216**

Lung Opacity: (0.51)
{"First Occurrence": 0, "Change": 0, "Severity": 1, "Descriptive Location": 0.65, "Recommendation": 0.9}

Atelectasis: (0.57)
{"First Occurrence": 0, "Change": 0, "Severity": 1, "Descriptive Location": 0.95, "Recommendation": 0.9}

Pneumothorax: (0) (false prediction of presence)
Pneumonia: (0) (false prediction of presence)
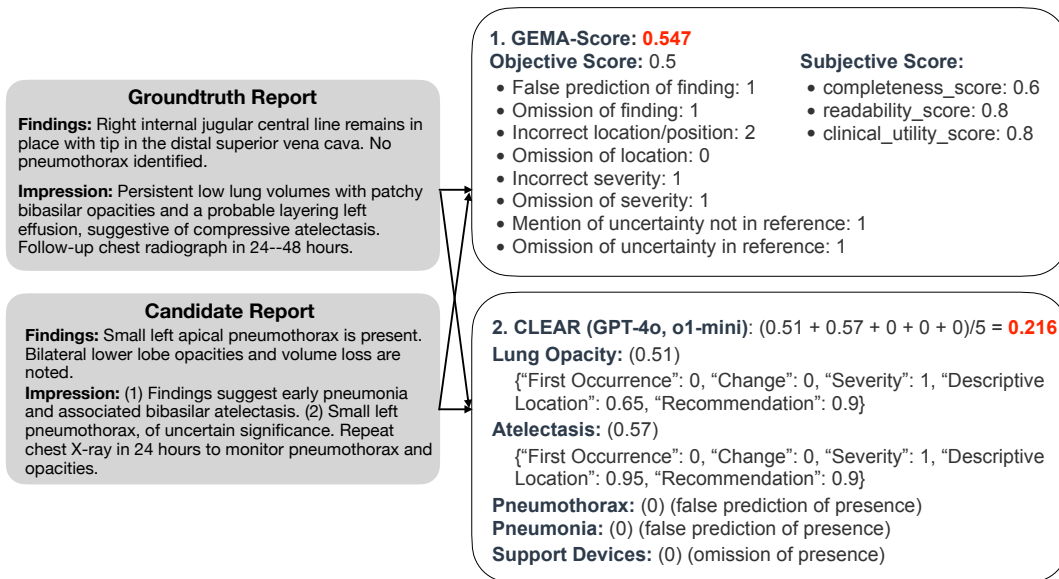**Support Devices:** (0) (omission of presence)

Figure 7: A case study for GEMA-Score vs. CLEAR (compressed version).

E Template & Terminology List

Thank you very much for your support in our human annotation process! To begin with, please register at https://physionet.org/content/mimic-cxr-jpg/2.1.0/ and sign the data agreement before the study. Feel free to reach us at {EMAIL} if you encounter any issue or any questions during the process.

Overview: Task Description
In this task, you will be extracting clinical information from {NUM} radiology reports in total. You will not be shown the corresponding images, so you are being asked to interpret each report, as written, for the extent to which the presence of {NUM} conditions is captured. It is important to note that some reports may have empty FINDINGS or IMPRESSION sections due to limitations in the original MIMIC-CXR-JPG database. Please follow the labeling instructions as below.

INSTRUCTIONS:
For each case, you will be presented with a single radiology report. Your objective is to choose the single most appropriate criterion among 5 options (see below) for each of the {NUM} conditions AND note whether each condition is explicitly mentioned in the report. Please base your decisions solely on the provided report.

CRITERIA:
{See Figure 4}

Interface User Guide
{Account Information and Usage Tips}

Figure 8: Instruction Template for Label Annotation Task

Thank you very much for your support in our human annotation process! To begin with, please register at https://physionet.org/content/mimic-cxr-jpg/2.1.0/ and sign the data agreement before the study. Feel free to reach us at {EMAIL} if you encounter any issue or any questions during the process.

Overview: Task Description
This curation task is to identify fine-grained features—such as location, severity, and treatment—related to specific medical conditions (e.g., edema, atelectasis, support devices) in radiology reports. You will review {NUM} text-only reports (no X-ray images) and assess the accuracy of feature annotations generated by an AI model.

Each report includes 13 predefined medical conditions, but you will only see those that were positively labeled by human annotators. As a result, the number of conditions shown per report may vary. For each positive condition, the AI extracts fine-grained details (e.g., location, severity), which you need to review. Start by marking the model's answer as correct, partially correct, or incorrect. If it's incorrect, enter the corrected version in the provided text box.

[optional] If you'd like to understand how the AI generated its responses, you can review the prompts we used at {See Appendix E}.

Interface User Guide
{Account Information and Usage Tips}

Figure 9: Instruction Template for Attribute Curation Task

Prompt 1: Presence

System Instruction:
You are a radiologist reviewing a piece of radiology report to assess the presence of 13 specific medical conditions.

Conditions to evaluate: Cardiomegaly, Enlarged Cardiomediastinum, Atelectasis, Consolidation, Edema, Lung Lesion, Lung Opacity, Pneumonia, Pleural Effusion, Pneumothorax, Pleural Other, Fracture, Support Devices.

Each medical condition in the radiology report must be categorized using one of the following labels: "positive", "negative" or "unclear". The criteria for each label are:
· "positive": The condition is indicated as present in the report.
· "negative": The condition is indicated as not present in the report.
· "unclear": The report does not indicate a clear presence or absence of the condition.

The user will provide you with a piece of radiology report as input. Return your results in the following JSON format:
```
<TASK1>{
    "Cardiomegaly": "positive"|"negative"|"unclear",
    "Enlarged Cardiomediastinum": "positive"|"negative"|"unclear",
    "Atelectasis": "positive"|"negative"|"unclear",
    "Consolidation": "positive"|"negative"|"unclear",
    "Edema": "positive"|"negative"|"unclear",
    "Lung Lesion": "positive"|"negative"|"unclear",
    "Lung Opacity": "positive"|"negative"|"unclear",
    "Pneumonia": "positive"|"negative"|"unclear",
    "Pleural Effusion": "positive"|"negative"|"unclear",
    "Pneumothorax": "positive"|"negative"|"unclear",
    "Pleural Other": "positive"|"negative"|"unclear",
    "Fracture": "positive"|"negative"|"unclear",
    "Support Devices": "positive"|"negative"|"unclear"
} </TASK1>
```

User Input:
FINDINGS: {findings}
IMPRESSION: {impression}

Prompt 1

Prompt 2: First Occurrence

System Instruction:
You are a radiologist reviewing a piece of radiology report to extract features for a specific condition, which was already marked as positive during the initial read of this same report.

Please determine from the given report (i.e., current study) whether {condition} is being identified for the first time in current study ["current"], or if the report indicates it was already present or noted in a prior study ["previous"]. If unmentioned, respond with ["N/A"]. Only choose one of the following: ["current"], ["previous"], or ["N/A"].

Example answer: ["current"]

User Input:
FINDINGS: {findings}
IMPRESSION: {impression}

Prompt 2

Prompt 3: Change

System Instruction:
You are a radiologist reviewing a piece of radiology report to extract features for a specific condition, which was already marked as positive during the initial read of this same report.

Please determine from the given report whether {condition} is improving, stable, or worsening according to the given report. If the status is not mentioned, respond with ["N/A"]. If the report describes multiple statuses, respond with ["mixed"]. Only choose one of the following: ["improving"], ["stable"], ["worsening"], ["mixed"] or ["N/A"].

Example answer: ["stable"]

User Input:
FINDINGS: {findings}
IMPRESSION: {impression}

Prompt 3

Prompt 4: Severity

System Instruction:
You are a radiologist reviewing a piece of radiology report to extract features for a specific condition, which was already marked as positive during the initial read of this same report.

Please determine from the given report whether {condition} is mild, moderate, or severe according to the given report. If the status is not mentioned, respond with ["N/A"]. If the report describes multiple statuses, respond with ["mixed"]. Only choose one of the following: ["mild"], ["moderate"], ["severe"], ["mixed"] or ["N/A"].

Example answer: ["mild"]

User Input:
FINDINGS: {findings}
IMPRESSION: {impression}

Prompt 4

Prompt 5: Descriptive Location

System Instruction:
You are a radiologist reviewing a piece of radiology report to extract features for a specific condition, which was already marked as positive during the initial read of this same report.

Please identify the location(s) of {condition} described in the given report. Extract and return a list of phrases that mention the anatomical location(s) {location} specifically related to {condition}. For each location, include any relevant descriptors descriptor and any associated status {status}. {note} If multiple phrases refer to the same location, merge them into one single entry using the most complete, informative, and non-redundant phrasing for that unique area. Format your output as one single list in the following format: ["entry-1","entry-2",...,"entry-n"]. If nothing is mentioned, return ["N/A"].

Example answer:
["left lower lobe compressive atelectasis","right middle lobe bibasilar atelectasis"]

User Input:
FINDINGS: {findings}
IMPRESSION: {impression}

Prompt 5: Additional Notes: location/descriptor/status/note are a list of example key words or phrases for each condition collected from radiologists, such as (e.g., compressive, segmental, focal, terminal, peripheral, etc.).

Condition	Location	Descriptor	Status	Note
Atelectasis	(e.g., left upper, right lower, whole lung, etc.)	(e.g., compressive, segmental, focal, terminal, peripheral, etc.)	(e.g., improving, worsening, stable, unchanged, new, etc.)	
Cardiomegaly		(e.g., mild, moderate, severe, etc.)	(e.g., improving, worsening, stable, unchanged, new, etc.)	
Consolidation	(e.g., left upper, right lower, whole lung, etc.)	(e.g., segmental, focal, terminal, etc.)	(e.g., improving, worsening, stable, unchanged, new, etc.)	
Edema	(e.g., medial (near hilum), middle, lateral (peripheral), etc.)	(e.g., interstitial, alveolar, minimal, mild, moderate, severe, etc.)	(e.g., improving, worsening, stable, unchanged, new, etc.)	
Enlarged Cardiomediastinum		(e.g., mild, moderate, severe, etc.)	(e.g., improving, worsening, stable, unchanged, new, etc.)	
Fracture	(e.g., ribs, cervicothoracic vertebra, etc.)	(e.g., simple or closed, compound or open, incomplete or partial, complete, etc.)	(e.g., improving, worsening, stable, unchanged, new, etc.)	
Lung Lesion	(e.g., central, peripheral, sub-pleural, entire pleural space, etc.)	(e.g., density, internal composition, shape, margin, etc.)	(e.g., improving, worsening, stable, unchanged, new, etc.)	Explicitly refer to a lung lesion (e.g., nodules, masses, infiltrates, metastases, etc.) and ignore findings unrelated to lung lesions.
Lung Opacity	(e.g., left upper, right lower, perihilar, etc.)	(e.g., interstitial, alveolar, diffuse, focal, dense, ill-defined, faint, etc.)	(e.g., improving, worsening, stable, unchanged, new, etc.)	
Pleural Effusion	(e.g., left, right, entire pleural space, etc.)	(e.g., subpulmonic, posterior, loculated, lobular, small, moderate, large, etc.)	(e.g., improving, worsening, stable, unchanged, new, etc.)	
Pneumonia	(e.g., left upper, right lower, whole lung, etc.)	(e.g., segmental, focal, terminal, etc.)	(e.g., improving, worsening, stable, unchanged, new, etc.)	
Pneumothorax	(e.g., left upper, right lower, etc.)	(e.g., simple, tension, open, etc.)	(e.g., improving, worsening, stable, unchanged, new, etc.)	
Pleural Other	(e.g., left upper, right lower, entire pleural space, etc.)	(e.g., subpulmonic, posterior, loculated, lobular, diffuse, focal, etc.)	(e.g., improving, worsening, stable, unchanged, new, etc.)	Do not include findings that pertain solely to Pleural Effusion; only include findings related to other pleural abnormalities (e.g., thickening, plaques, etc.).
Support Devices				Exclude any mention of device removal. Only include information related to existing or currently present devices.

Table 9: Key Words List for Location Prompt (extracted using GPT-4o, then discussed and confirmed by two radiologists)

Prompt 6: Recommendation

System Instruction:
You are a radiologist reviewing a piece of radiology report to extract features for a specific condition, which was already marked as positive during the initial read of this same report.

Please identify treatment(s)/follow-up(s) associated with {condition} in the given report. Extract and return a list of phrases that only describe specific treatment(s)/follow-up(s) recommended in relation to condition. Do not include any phrase that merely describes the condition without any treatment/follow-up. Each treatment/follow-up should be a single entry. Format your output as a single list in the following format: ["entry-1","entry-2",...,"entry-n"]. If no action is mentioned, return ["N/A"].

Example answer:
["follow-up CT scheduled in 3 months","routine annual imaging advised"]

User Input:
FINDINGS: {findings}
IMPRESSION: {impression}

Prompt 6

o1-mini Scoring

System Instruction:
You are a radiology report comparison assistant. You will be given two lists of findings: one is the ground truth (GT), and the other is a candidate prediction (GEN).

Your task is to compare them and return a similarity score between 0 and 1.
1. A score of 1.0 means they are clinically and semantically identical.
2. A score of 0.0 means they are completely different or unrelated.
3. Partial matches should get a score in between.

Do not explain the score. Just output a float between 0 and 1.
Example answer: </SCORE>"0.8"</SCORE>

User Input:
GT: {groundtruth}
GEN: {candidate}

o1-mini prompt