ReFINE: A Reward-Based Framework for Interpretable and Nuanced Evaluation of Radiology Report Generation

Yunyi Liu[1*], Yingshu Li[1*], Zhanyu Wang[1], Xinyu Liang[4], Lingqiao Liu[3], Lei Wang[2], Luping Zhou[1†]

[1]University of Sydney
[2]University of Wollongong
[3]University of Adelaide
[4]Binzhou Medical University
{yunyi.liu1, yingshu.li, zhanyu.wang, luping.zhou}@sydney.edu.au,
xinyu.liang31@gmail.com, lingqiao.liu@adelaide.edu.au, leiw@uow.edu.au

Abstract

Automated radiology report generation (R2Gen) has advanced significantly, yet evaluation remains challenging due to the complexity of assessing report quality. Traditional metrics often misalign with human judgments, failing to identify specific deficiencies. To address this, we introduce **ReFINE**, a framework for training an Evaluation Model using a novel **margin-based reward enforcement loss**. This approach decomposes report quality into **fine-grained sub-scores** across user-defined criteria, improving interpretability. Leveraging GPT-4, we generate diverse training data with paired accepted and rejected reports to train our model under a reward-based system. The trained **ReFINE Score** provides both **granular sub-scores** and an aggregated quality assessment, enabling **criterion-specific evaluation**. Experiments show that ReFINE achieves stronger correlation with human ratings than traditional metrics, and generalizes well across three expert-annotated datasets—including chest X-rays and multimodal reports spanning nine imaging modalities—under two distinct scoring systems.

Introduction

Automated radiology report generation (R2Gen), which produces free-text descriptions of visual findings in radiographic images, has seen substantial growth (Wang et al. 2023; Li et al. 2024). This complex AI task demands understanding high-level clinical semantics, making both the generation and evaluation of reports highly challenging. Traditional natural language generation (NLG) metrics, such as BLEU (Papineni et al. 2002) and METEOR (Reimers and Gurevych 2019), focus on n-gram matches, often overlooking lexical and structural diversity crucial for capturing diagnostic meaning. While BERTScore (Zhang et al. 2020) leverages contextualized embeddings to better handle paraphrasing, clinically oriented scores like CheXbert (Smit et al. 2020) and RadgraphF1 (Jain et al. 2021) focus on predefined pathological entities but fail to capture broader semantic correlations.

Despite these efforts, existing evaluation metrics often fall short of aligning with human judgment (Liu et al. 2024a). Recent approaches, such as RadCliQ (Yu et al. 2023b), combine multiple metrics and regress combination weights from human-marked scores to better mimic human evaluation. However, RadCliQ's reliance on limited, expensive human annotations restricts its scalability. Meanwhile, advances in large language models (LLMs) like GPT-4 (OpenAI 2023) suggest potential for report evaluation, but directly applying such models raises privacy concerns, demands significant computational resources, and lacks cost-effectiveness for R2Gen tasks.

To address these challenges, we propose ReFINE, an innovative evaluation metric tailored specifically for R2Gen. Unlike previous works, ReFINE utilizes a language model foundation trained with a novel margin-based reward enforcement loss, enabling it to decompose overall report quality into detailed, fine-grained sub-scores across user-specified criteria. This effectively improves the interpretability of the assessment. For example, by combining sub-criteria, we can clearly identify the reasons for a report's poor quality, e.g., whether due to incorrect lesion location, incorrect severity of findings, or omission of findings. Leveraging GPT-4's scoring capabilities, we developed a data generation pipeline that produces evaluation samples mimicking human judgment. These samples are paired as "accepted" and "rejected" reports using our pairing rule and are used to train our evaluation model to assign both individual rewards for each evaluation criterion and a final aggregated score.

Our **ReFINE** metric offers key advantages over existing approaches. Non-trainable metrics, including NLG-based and clinically relevant ones, correlate poorly with human assessments and cannot adapt to customized criteria. Among trainable metrics, RadCliQ combines non-trainable scores linearly, offering limited improvement and flexibility. Some LLM-based metrics (e.g., G_Rad (Chaves et al. 2024b), FineRadScore (Huang et al. 2024), RadFact (Bannur et al. 2024) and CheXprompt (Chaves et al. 2024a)), relying on online LLMs, present privacy concerns. The most relevant methods, MRScore (Liu et al. 2024b) and GREEN (Ostmeier et al. 2024), fall short in different ways. MRScore simplifies training by outputting a single aggregated score, but this sac-

[*]Equal contribution.
[†]Corresponding author.

rifices transparency and interpretability. Its lack of sub-score granularity introduces noise, limits diagnostic utility, and makes error analysis difficult. In contrast, ReFINE provides a structured breakdown of sub-scores, enabling more transparent evaluations and flexible, fine-grained control tailored to specific user needs. GREEN introduces interpretability via free-text analysis but compromises scoring accuracy due to task complexity. Moreover, GREEN's fine-tuning of LLMs lacks a dedicated loss function like ours, limiting its sensitivity to nuanced quality differences. Our approach achieves a higher **Kendall's Tau correlation with human ratings** (0.75 vs. 0.64 for GREEN) while reducing **training costs** (1x NVIDIA A6000 vs. 8x NVIDIA A100) and improving **inference efficiency**. Our contributions are summarized as follows:

(1) We propose a novel approach for training LLMs to perform fine-grained evaluation of radiology reports, using a custom loss to produce human-aligned rewards. ReFINE outperforms existing metrics in expert correlation while remaining offline and practical with moderate computing needs.

(2) Our method provides both overall scores and detailed subscores, enhancing interpretability and enabling users to identify specific weaknesses (according to the criteria) in report quality.

(3) We benchmark ReFINE on three human-annotated datasets—ReXVal, RaTE-Eval, and our custom Rad-100—covering diverse modalities and scoring rubrics. ReFINE consistently outperforms existing metrics in correlation with expert ratings.

Related Work

Evaluation Metrics for Radiology Reports

Evaluation metrics for radiology reports fall into two categories: language metrics and clinical metrics. Language metrics such as BLEU (Papineni et al. 2002), ROUGE (Lin 2004), METEOR (Banerjee and Lavie 2005), and BERTScore (Zhang et al. 2019) measure textual similarity between generated and reference reports, focusing on n-gram overlap or embedding similarity. However, these metrics often fail to reflect clinical relevance or diagnostic accuracy, especially when evaluating generated texts that may be lexically varied but semantically equivalent. Clinical metrics aim to assess medical correctness. Tools like CheXpert (Irvin et al. 2019) label 14 pathologies as present, absent, or uncertain, with accuracy often evaluated using CheXbert or embedding-based similarity. RadGraph (Jain et al. 2021) extracts clinical entities and relationships. Yet, these extraction-based metrics are limited by fixed entity sets and rigid matching rules, making them less effective for ambiguous or nuanced clinical content. Hybrid methods like RadCliQ (Yu et al. 2023b) and RadEval (Calamida et al. 2023) combine multiple approaches but still struggle to fully capture the richness of clinical descriptions.

Large Language Model for Evaluation

Several recent methods leverage LLMs for radiology report evaluation, including G-Rad (Chaves et al. 2024b), FineRadScore (Huang et al. 2024), RadFact (Bannur et al. 2024), and CheXPrompt (Chaves et al. 2024a), which rely on prompting online APIs (e.g., GPT-4, Claude-3) to score report pairs. While these approaches achieve competitive performance, they raise privacy concerns and depend on costly, non-transparent inference pipelines. On the ReXVal benchmark, their correlation with human ratings is either lower than or at best comparable to that of ReFINE. In contrast, ReFINE operates entirely offline, avoiding potential privacy leakage. MRScore (Liu et al. 2024b) trains an offline evaluator but outputs only a single aggregated score, limiting interpretability and flexibility. GREEN (Ostmeier et al. 2024) also aims to improve explainability but suffers from reduced accuracy and lacks a dedicated loss for fine-grained score alignment. Overall, ReFINE offers a lightweight and interpretable evaluation framework with strong alignment to expert ratings—trained efficiently on a single A6000 GPU and free from online dependencies.

Method

In this section, we propose ReFINE, a metric that aligns well with human assessments by providing both overall and detailed sub-scores for interpretability. Leveraging GPT-4, we generate training samples using a pairing rule (Algorithm 1) and train a reward model with a custom loss function. The model predicts sub-scores, which are summed to form the final overall score (Figure 1).

Scoring Data Generation Pipeline

Recent studies have demonstrated GPT-4's capability in evaluating chest X-ray reports. When prompted with specified criteria, **GPT-4 can generate similarity assessments that correlate well with human evaluations**, as consistently verified in (Chiang and Lee 2023) and (Liu et al. 2024b). For example, in (Chiang and Lee 2023), **GPT-4 achieved Kendall's Tau of 0.735 with radiologists' annotations using RadCliQ scoring system**. In (Liu et al. 2024b), **GPT-4 scored a Kendall's Tau correlation of 0.531 with human ratings using the MRScore scoring system**. The performance gap may come from the difference of the two scoring systems: Unlike RadCliQ, which primarily quantifies errors per subcriterion, MRScore incorporates semantic aspects such as lesion descriptions, report completeness, grammar, and medical terminology accuracy. Building on this observation, we utilize GPT-4 to generate extensive scoring data, including both reports and the corresponding scores, for training purposes. The process is elaborated as follows.

Defining Scoring Criteria. Various assessment criteria have been reported in the literature. In this study, we investigate two scoring systems to demonstrate our model's versatility across different evaluation rules. The RadCliQ scoring system proposed in (Yu et al. 2023b) evaluates both clinically significant and insignificant errors across six error categories: 1) false prediction of a finding, 2) omission of a finding, 3) incorrect location or position of a finding, 4) incorrect severity of a finding, 5) mention of a comparison absent in the reference impression, and 6) omission of a comparison that notes a change from a previous study. The total score is the sum of the error counts, highlighting the importance

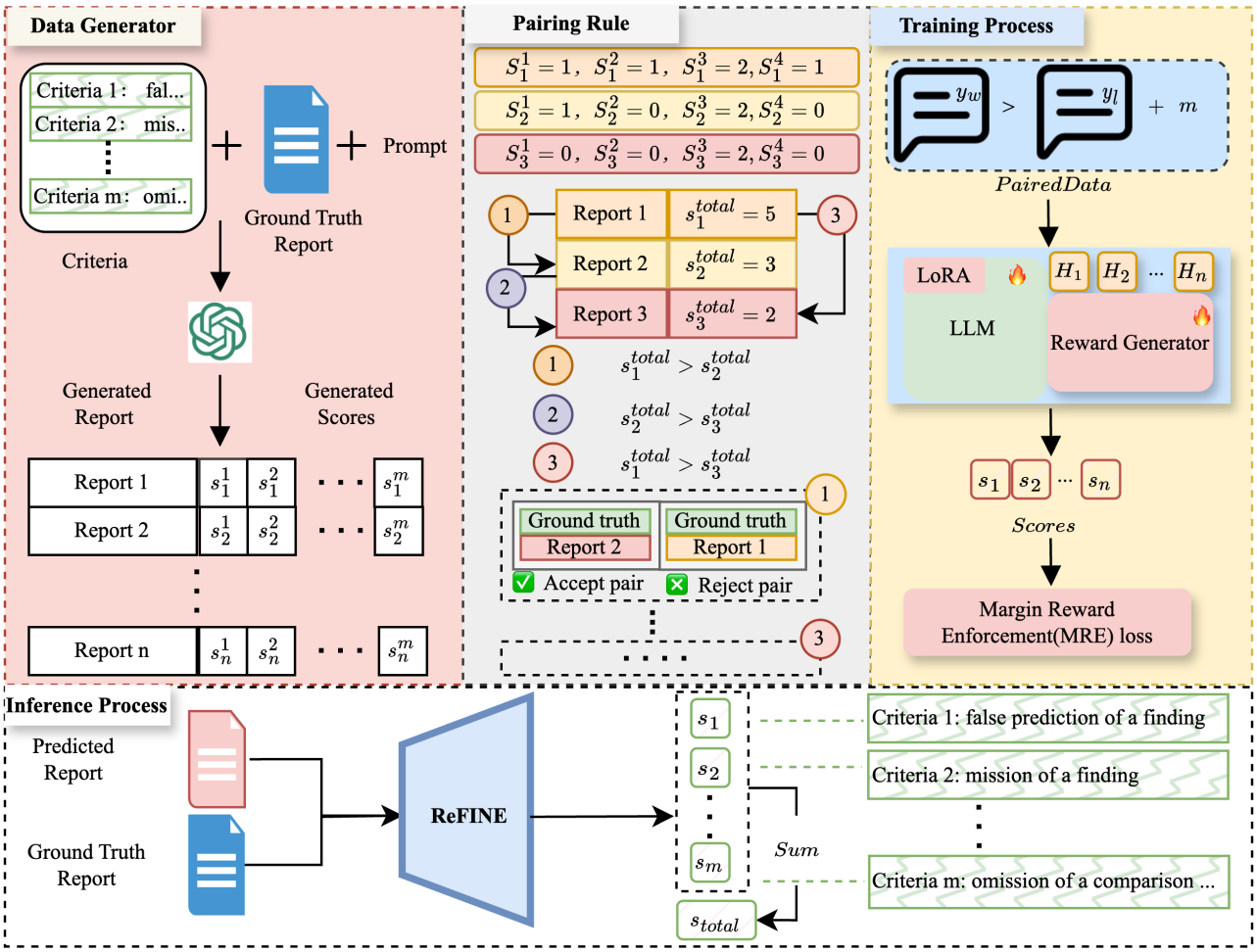

Figure 1: Overview of the framework for our model. The diagram is divided into four main parts: 1) **Data Generator**: Generates reports and corresponding scores based on specified criteria. 2) **Pairing Rule**: Demonstrates the scoring and pairing process using four criteria as an example. Reports are paired into "accepted" and "rejected" categories based on their total scores and margins. 3) **Training Process**: Utilizes paired data to train a reward model through a LoRA-based large language model (LLM) to optimize the MRE loss. 4) **Inference Process**: rate the predicted report by comparing it to the ground truth report, generating both sub-scores and total scores for evaluation.

of clinical findings. Differently, the MRScore scoring system proposed in (Liu et al. 2024b) addresses both clinical findings and linguistic concerns. It involves seven fundamental items from radiologists' expertise and literature review: "impression consistency", "impression organs", "description of lesions", "clinical history", "completeness", "grammar", and "medical terminology", with a detailed explanation. Each item corresponds to an error type with yes/no answers and is assigned a different weight (from $\{30, 20, 20, 10, 10, 5, 5\}$ accordingly) to form individual item scores. The total score is calculated as $\text{Total\_score} = 100 - \sum_{i=1}^{7} S_i \times W_i$, where S_i is error score of the i-th item and W_i is the corresponding weight. With these scoring rules, GPT-4 can be prompted to score reports following these criteria, as elaborated below.

Generating Scoring Training Dataset. With a defined scoring system, we design prompts that encapsulate evaluation criteria, guiding GPT-4 to assess radiology reports in a human-aligned manner (see Appendix for a prompt example). Using the GPT-4 API, we generate reports of varying quality from randomly selected MIMIC-CXR ground-truth reports. For RadCliQ, we sample 8000 ground-truth reports, each used to generate three GPT-4 reports with increasing error levels (0–2, 3–4, 5–6 errors). Each report is annotated with total and sub-criterion error scores. For MRScore, we select 1800 reports, each leading to three GPT-4 reports reflecting different quality tiers (0–40, 40–70, 70–100), with both total and item-level scores. For a sanity check of data quality, we randomly sampled 50 GPT-4–generated training examples and had them scored by two radiologists (10+ years experience). Consensus-based accuracies were: 0.90 (Impression), 0.98 (Impression Organ), 0.86 (Description of Lesion), 0.92 (Clinical History), 0.98 (Completeness), and 1.0 for

Algorithm 1: Report Pairing Rule with Margin Calculation

Input: Ground Truth Report x, Predicted Reports $Y = \{y_1, y_2, ..., y_n\}$, Scoring System S with k criteria
Output: Accepted Reports Y_w, Rejected Reports Y_l, Margins (both total and per criterion)
for $y_i \in Y$ **do**
  Compute total score $S_{\text{tot}}(x, y_i) = \sum_{j=1}^{k} S_j(x, y_i)$
end for
Initialize $Y_w \leftarrow \emptyset$, $Y_l \leftarrow \emptyset$, $M \leftarrow \emptyset$
for $y_i \in Y$ **do**
  for $y_j \in Y$ such that $y_j \neq y_i$ and $S_{\text{tot}}(y_j) < S_{\text{tot}}(y_i)$ **do**
    Compute margin for total score: $M_{\text{tot}} = S_{\text{tot}}(y_i) - S_{\text{tot}}(y_j)$
    Compute margin for each criterion: $M_j = S_j(y_i) - S_j(y_j)$ for all $j \in \{1, 2, ..., k\}$
    if $M_{\text{tot}} > 0$ **then**
      Add (x, y_i) to Y_w
    else
      Add (x, y_i) to Y_l
    end if
    Store M_{tot} and $\{M_1, M_2, ..., M_k\}$ in M
  end for
end for
Return Y_w, Y_l, M

both Grammar and Medical Terminology. Please note that, the GPT-4 generated data were only used for training while our model performance was validated on multiple human-annotated datasets.

Reward Model

ReFINE is our innovative evaluation metric designed to be versatile across various evaluation frameworks. This LLM-based reward model leverages a pretrained language model, such as Llama3 (Touvron et al. 2023), fine-tuning it to align with human evaluations using pairs of reports guided by our novel reward system. The core of ReFINE is its training process, which involves pairs of reports generated from the same ground-truth report but with different qualities. This pairing mechanism is essential for calibrating the model to distinguish between different quality levels effectively. During training, the model learns to assign higher rewards to high-quality reports while simultaneously generating multiple individual criterion scores. These criterion scores are critical as they provide detailed insights into specific aspects of the report's quality. At the inference stage, the model predicts rewards for each individual criterion. These rewards are then summed to generate the final ReFINE. To ensure precise differentiation, we also introduce a scoring margin for each criterion and the overall score. This margin enables the model to recognize and learn subtle differences in report quality, enhancing its evaluative capability.

Model Input Our model requires paired reports and their score margins as input. Each pair consists of an "accepted" report and a "rejected" report, both derived from the same ground-truth report, with the "accepted" report having a higher score than the "rejected" one. Algorithm in 1 illustrates the pairing rule, showing the selection process for accepted and rejected reports and the calculation of their respective margins.

Algorithm 1 outlines the Report Pairing Rule, which selects the optimal predicted report for a given ground truth. It calculates total scores for each predicted report based on multiple criteria, sorts them in descending order, and compares scores to determine acceptance or rejection. Margins, representing the score differences between accepted and rejected pairs, are computed for both total and individual criteria. The algorithm outputs accepted/rejected pairs and their margins, ensuring effective and interpretable report selection.

Multi-Reward Generator Our reward model, based on the LLaMA3 (Meta 2024) backbone, incorporates a multi-reward head to generate the ReFINE score. LLaMA3 was selected for its exceptional language comprehension with just 6.8M trainable parameters over 7 billion in total. The multi-reward head is a linear projection layer mapping LLaMA3's last layer feature map to an $N \times 1$ vector, where N is the total number of sub-scores. This model is fine-tuned using Low-Rank Adaptation (LoRA) (Hu et al. 2022), allowing effective fine-tuning with minimal parameter changes. Training pairs of "accepted" and "rejected" reports calibrate the model for reward prediction. During training, the model learns to distinguish report qualities by adhering to a scoring margin reflecting quality differences. Sub-scores discern quality differences per report aspect, with their summation producing the overall assessment for generated reports.

Multi-Reward Learning. Our multi-reward model aims to mimic human judgement via GPT-4 by optimizing a function based on the GPT-4 rankings of radiology reports. It discerns and predicts the preferred report within each pair, capturing subtle differences that distinguish superior reports. Instead of rewarding based merely on the whole report, our objective function is devised to learn also the preference per individual criterion. The objective function is elaborated in MRE loss section. Through our objective function, we can effectively utilize total margin to control the overall quality of the report and also respect each sub-score's margin to manage the differences in sub-scores across varied overall quality levels. By adjusting the size of the margin, corresponding penalties are applied, thus training the model to produce appropriate rewards.

Margin Reward Enforcement(MRE) Loss Function

Considering a pair of generated reports $< y_w^i, y_l^i >$[1] corresponding to the same i-th ground truth report x^i, the accepted report y_w^i receives a higher ground truth score (denoted as s_w^i), while the rejected report y_l^i receives a lower ground truth score (s_l^i). Let $s_w^{i,j}$ and $s_l^{i,j}$ denote the j-th sub-score of s_w^i and s_l^i, respectively, where $j = 1, \ldots, N$, and N is the number of sub-scores under a specific scoring system. Note that although the total score s_w^i is greater than s_l^i, an individual sub-score $s_w^{i,j}$ may not necessarily be greater than $s_l^{i,j}$. Our goal is to train the model to distinguish ranking relationships at both the individual sub-score and total score levels, ensuring alignment with the ground truth ranking structure, formulated as follows.

[1]Here "w" stands for "win", indicating the accepted report, and "l" for "lose", indicating the rejected report.

Individual Reward Loss $\mathcal{L}_{\text{ind}}$ This loss focuses on correctly ranking the individual sub-scores for each scoring criterion.

$$\mathcal{L}_{\text{ind}}(y_{\text{w}}^i, y_{\text{l}}^i) = \frac{1}{N}\sum_{j=1}^{N}\Big[\mathbb{1}_j \ \text{ReLU}\left(-t_w \Delta r_j + t_w m^{i,j}\right) + (1-\mathbb{1}_j)\ \text{ReLU}\left(|\Delta r_j| - c\right)\Big], \tag{1}$$

where $\mathbb{1}_j = 1\Big(s_w^{i,j} \neq s_l^{i,j}\Big)$ and $\Delta r_j = r_w^{i,j} - r_l^{i,j}$.

- $t_w = 1$ if $m^{i,j} > 0$, otherwise $t_w = -1$.
- $m^{i,j} = s_w^{i,j} - s_l^{i,j}$ is the margin between sub-scores for the j-th criterion.
- $1(s_w^{i,j} \neq s_l^{i,j})$ ensures this term applies only when the sub-scores differ.

Here $r_w^{i,j}$ and $r_l^{i,j}$ denote the j-th individual rewards assigned to the reports y_w^i and y_l^i, respectively. The margin between the total scores s_w^i and s_l^i is denoted by $m^i = s_w^i - s_l^i$, where $m^i > 0$. The individual "margin" $m^{i,j} = s_w^{i,j} - s_l^{i,j}$ is not necessarily positive. The variable t_w acts as a flag: $t_w = 1$ if $m^{i,j} > 0$, otherwise $t_w = -1$. The function $1(\cdot)$ is an indicator function, returning 1 when the event occurs and 0 otherwise. K is the total number of report pairs.

- If $s_w^{i,j} > s_l^{i,j}$ (i.e., $m^{i,j} > 0$): A penalty is incurred if $r_l^{i,j} > r_w^{i,j} - m^{i,j}$, indicating that the "losing" report's reward exceeds the acceptable range relative to the "winning" report.
- If $s_w^{i,j} < s_l^{i,j}$ (i.e., $m^{i,j} < 0$): A penalty is incurred if $r_l^{i,j} < r_w^{i,j} - m^{i,j}$, meaning the "losing" report's reward falls below the acceptable range.
- If $s_w^{i,j} = s_l^{i,j}$ (i.e., $m^{i,j} = 0$): A penalty is incurred if $|r_w^{i,j} - r_l^{i,j}| > c$, ensuring that rewards remain sufficiently close when scores are equal.

Total Reward Loss $\mathcal{L}_{\text{tot}}$ This loss enforces a clear margin between the total scores of the winning and losing reports.

$$\mathcal{L}_{\text{tot}}(y_{\text{w}}^i, y_{\text{l}}^i) = \text{ReLU}\left(-\left(\sum_{j=1}^{N} r_w^{i,j} - \sum_{j=1}^{N} r_l^{i,j}\right) + m^i\right), \tag{2}$$

where $m^i = s_w^i - s_l^i$ is the margin between total scores s_w^i and s_l^i.

- If $\sum_j r_l^{i,j} \leq \sum_j r_w^{i,j} - m^i$: The total reward of the rejected report y_l^i is within the acceptable range relative to the winning report y_w^i, and no penalty is applied.
- If $\sum_j r_l^{i,j} > \sum_j r_w^{i,j} - m^i$: A penalty is incurred, indicating that the rejected report y_l^i has a total reward exceeding the acceptable range defined by the winning report y_w^i and the margin m^i.

Margin Reward Enforcement (MRE) Loss The overall loss,$\mathcal{L}_{\text{MRE}}$, combines these two terms: the individual reward loss $\mathcal{L}_{\text{ind}}$ and the total reward loss $\mathcal{L}_{\text{tot}}$, balanced by the hyperparameter λ, an ablation of hyperparameter λ shows in Table 3.This combined loss ensures that the model learns to rank individual sub-scores correctly through $\mathcal{L}_{\text{ind}}$, while maintaining a sufficient margin between total scores via $\mathcal{L}_{\text{tot}}$. By minimizing $\mathcal{L}_{\text{overall}}$, the model effectively balances individual and total reward losses, enabling it to provide nuanced insights into the assessment results by optimizing both individual sub-scores and the total score.

$$\mathcal{L}_{\text{MRE}} = \sum_{i=1}^{K}\Big(\mathcal{L}_{\text{ind}}(y_{\text{w}}^i, y_{\text{l}}^i) + \lambda \mathcal{L}_{\text{tot}}(y_{\text{w}}^i, y_{\text{l}}^i)\Big), \tag{3}$$

The loss function consists of multiple components, including ReLU and absolute functions, whose gradients can be derived using their subgradient properties:

$$\frac{\partial \mathcal{L}_{\text{ind}}}{\partial r_w^{i,j}} = \begin{cases} -t_w, & \text{if } s_w^{i,j} \neq s_l^{i,j} \text{ and } r_w^{i,j} - r_l^{i,j} < 0, \\ 1, & \text{if } s_w^{i,j} = s_l^{i,j} \text{ and } |r_w^{i,j} - r_l^{i,j}| > c, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

$$\frac{\partial \mathcal{L}_{\text{tot}}}{\partial r_w^{i,j}} = \begin{cases} -1, & \text{if } \sum_{j=1}^{N} r_w^{i,j} - \sum_{j=1}^{N} r_l^{i,j} < m^i, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Experiments and Result

Datasets

We evaluated ReFINE on three expert-annotated datasets: ReXVal (Yu et al. 2023a), RaTE-Eval (Zhao et al. 2024), and Rad-100. ReXVal and RaTE-Eval share the RadCliQ scoring framework but differ in scope. ReXVal includes chest X-rays (CXRs) only, while RaTE-Eval spans 9 imaging modalities and 22 anatomical regions. In contrast, Rad-100 focuses solely on CXR data but uses the MRScore rubric, allowing us to evaluate ReFINE's robustness across scoring schemes.

ReXVal (Yu et al. 2023a) contains 200 pairs from 50 MIMIC-CXR studies (4 per study), annotated by six board-certified radiologists using RadCliQ's six-category error taxonomy. It serves as a public benchmark for metric–human alignment in radiology report evaluation.

RaTE-Eval (Zhao et al. 2024) is a large-scale, public dataset of 2,215 sentence-level report pairs from MIMIC-IV. Reports are annotated using RadCliQ criteria by two radiologists with over five years of experience. It includes diverse imaging modalities (CT, MRI, Ultrasound, etc.) and supports metric training and validation through an 8:2 train/test split.

Rad-100 is a CXR dataset we curated using the MRScore criteria (Section 3.1). It comprises 100 generated reports (from R2Gen) paired with randomly sampled MIMIC-CXR ground-truth references. Each report was independently scored by three senior radiologists (10+ years' experience), with majority voting to finalize scores for individual items. The total score is obtained by summing subscores. As Rad-100 is not used to train ReFINE [2], it provides an independent benchmark for validating model performance to a scoring system different than RadCliQ.

Performance on ReXVal Dataset

Correlation Analysis of Sub-criteria. Table 1 shows ReFINE's performance on the ReXVal dataset using the RadCliQ Scoring System. ReFINE demonstrates strong alignment with expert judgments across error categories, with

[2]Rad-100 is entirely separate from ReFINE's training data.

Criteria	Kendall's Tau↑	Spearman↑
Score 1: False prediction of a finding	0.680	0.842
Score 2: Omission of a finding	0.507	0.673
Score 3: Incorrect location or position of a finding	0.246	0.327
Score 4: Incorrect severity of a finding	0.443	0.569
Score 5: Mention of a comparison absent in the reference impression	0.433	0.545
Score 6: Omission of a comparison that notes a change from a previous study	0.267	0.345
Total	0.751	0.910

Table 1: Human Correlations of ReFINE Subscores on ReXVal Dataset

$\mathcal{L}_{\text{tot}}$	$\mathcal{L}_{\text{ind}}$	Spearman (↑)	Kendall's Tau (↑)
		0.319	0.215
✓		0.899	0.740
	✓	0.899	0.738
✓	✓	0.910	0.751

Table 2: Ablation study on different loss terms

overall Kendall's Tau of 0.751 and Spearman correlation of 0.910. High scores in categories such as "False prediction of a finding" (Kendall: 0.680, Spearman: 0.842) and "Omission of a finding" (0.507, 0.673) indicate effective recognition of common radiological errors. Lower scores for "Incorrect location or position" (0.246, 0.327) suggest difficulty capturing subtle spatial details. Unlike single-score metrics (Yu et al. 2023b; Zhang et al. 2019; Jain et al. 2021), ReFINE's sub-criteria output helps pinpoint specific model weaknesses, guiding targeted improvements.

Comparison with other metrics. Table 5 compares the performance of different metrics using Kendall's Tau and Spearman correlation on ReXVal Dataset. The comparison is based on the total score. Unlike ReFINE, *the existing metrics cannot be customized to user-specific sub-criteria*, making subscore comparison impossible[3]. We include traditional NLG metrics (BLEU-4 (Papineni et al. 2002), ROUGE-L (Lin 2004), METEOR (Banerjee and Lavie 2005), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015)), clinical metrics (BERTScore (Zhang et al. 2019), RadGraphF1 (Jain et al. 2021), and Semb_score (Yu et al. 2023b)), hybrid metrics (RadCliQ (Yu et al. 2023b)), and both offline (GREEN (Ostmeier et al. 2024), RaTEScore (Zhao et al. 2024)) and online LLM-based models (RadFact (Bannur et al. 2024), CheXPrompt (Chaves et al. 2024a), G-Rad (Chaves et al. 2024b), FineRadScore (Huang et al. 2024)). ReFINE achieves the strongest overall correlations (Kendall: 0.751, Spearman: 0.910). CheXPrompt and G-Rad report comparable Kendall scores (0.750) but depend on proprietary GPT-based APIs, limiting privacy and reproducibility. Offline methods such as GREEN (0.640, 0.816) and RadCliQ-v1 (0.631, 0.816) lag behind, with GREEN requiring substantially heavier compute (8 A100 GPUs for 12 epochs vs. ReFINE's 1 A6000 for 4 epochs). In addition, ReFINE supports customizable sub-criteria evaluation, offering stronger alignment, interpretability, and efficiency than prior metrics.

[3]GREEN only provides error counts for each subcategory without human correlation, making direct comparison of subscore correlations unfeasible. Additionally, its overall correlation with human assessments is significantly lower than ours.

λ	0.5	0.8	1.0	1.2	2.0	3.0
Spearman	0.904	0.906	0.910	0.900	0.895	0.893
Kendall	0.743	0.746	0.751	0.740	0.735	0.729

Table 3: Ablation study about varying λ values

Performance on RaTE-Eval Dataset

We train and test on the RaTE-Eval dataset, which includes 9 different imaging modalities. We compare the results directly with (Zhao et al. 2024), which only reports Pearson correlation without subscore-level breakdown. As shown in Table 6, our Pearson correlation 0.61 surpasses all existing metrics. Beyond the total score, subscore breakdowns on the RaTE-Eval dataset are shown in Table 7.

Performance on Rad-100 Dataset

Since the scoring system used by Rad-100 is a binary format where the presence of an error is marked as 1 and the absence as 0 (check Appendix for detail), the results are multiplied by pre-defined weights before forming the final score. Accordingly, we evaluate the accuracy of binary classification for each sub-criterion, as reported in Table 4. The comparison of overall scores is presented in Table 8. ReFINE again attains superior performance, achieving a Kendall's Tau of 0.230 and a Spearman correlation of 0.293 - significantly higher than those of other metrics. The corresponding p-values ($p = 0.003$ for both correlations) confirm the statistical significance of ReFINE's human correlations.

Ablation Study

Loss terms. The loss we proposed comprises two terms: the individual reward loss L_{ind} and the total reward loss L_{tot}. An ablation of the loss functions is given in Table 2. As shown, if we train L_{tot} alone for predicting sub-scores, Kendall's Tau will drop from 0.751 to 0.740 for the total score, a sum of the sub-scores. If we train L_{ind} alone, Kendall's Tau will drop from 0.751 to 0.738, demonstrating the effectiveness of the regularization from L_{tot}. **Hyperparameters.** Our loss function involves two hyper-parameters: the hyperparameter c is a small positive rounding number when judging whether r_w equals r_l, which we set to 1e-2. The hyperparameter λ balances the two loss terms L_{ind} and L_{tot} and we examined its effect through the ablation study shown in Table 3. As seen, our model is insensitive to λ. When it varies in a reasonable range, our model produces better human correlations than the existing metrics.

LLM backbones. Table 9 presents a performance comparison of various LLM backbones. Llama3 demonstrates superior performance with a medium size of trainable parameters. To ensure the scoring system is easily deployable, we focused on models with 7 billion parameters in total or fewer.

Sub-criteria	Imp. Cons.	Imp. Org.	Desc. Les.	Clin. Hist.	Comp.	Gram.	Med. Term.
Accuracy	0.589	0.730	0.770	0.410	0.380	0.980	0.720

Table 4: Accuracy of Different Sub-scores in Rad-100 test dataset. Here, 'Imp. Cons.' stands for Impression Consistency, 'Imp. Org.' for Impression Organ, 'Desc. Les.' for Description of Lesion, 'Clin. Hist.' for Clinical History, 'Comp.' for Completeness, 'Gram.' for Grammar, and 'Med. Term.' for Medical Terminology.

Metric	Kendall's Tau↑	Spearman↑
BLEU-4 (Papineni et al. 2002)	0.345	0.475
ROUGE-L (Lin 2004)	0.491	0.663
METEOR (Banerjee and Lavie 2005)	0.464	0.627
CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015)	0.499	0.664
BertScore (Zhang et al. 2019)	0.507	0.677
RadGraphF1 (Jain et al. 2021)	0.516	0.702
Semb_score (Yu et al. 2023b)	0.494	0.665
RadCliQ-v1 (Yu et al. 2023b)	0.631	0.816
GREEN (Ostmeier et al. 2024)	0.640	-
RaTEScore (Zhao et al. 2024)	0.527	-
ReFINE (Ours)	**0.751**	**0.910**
*Results below are not strictly comparable because they are **online** model(e.g., GPT4).*		
RadFact (online) (Bannur et al. 2024)	0.590	-
CheXprompt (online) (Chaves et al. 2024a)	0.750	-
G-Rad (online) (Chaves et al. 2024b)	0.750	-
FineRadScore (online) (Huang et al. 2024)	0.737	-

Table 5: Human Correlation Comparison of Evaluation Metrics on ReXVal Dataset.All reported correlations are statistically significant with $p < 0.01$. *Gray part indicates online methods which directly use online model API.*

Metric	Pearson↑	Spearman↑	Kendall's Tau↑
BLEU	0.27†	0.23	0.16
ROUGE	0.34†	0.21	0.15
METEOR	0.39†	0.33	0.24
CIDEr	0.25†	0.28	0.20
BERTScore	0.40†	0.35	0.25
RadGraph	0.44†	0.42	0.31
RadCliQ	0.46†	0.41	0.30
RaTEScore	0.54†	–	–
ReFINE (Ours)	**0.61**	**0.59**	**0.45**

Table 6: Human Correlation on the Multimodal RaTE-Eval Dataset (Sentence-level correlation). † indicates values directly cited from the RaTEScore paper; others are reproduced. All correlations are statistically significant with $p < 0.01$.

Qualitative Analysis

A visual example is provided in Figure 2, demonstrating how the ReFINE score correlates with human ratings using the RadCliQ scoring system. As shown, the generated report inaccurately describes the severity of the "left pleural effusion" (highlighted in red), resulting in a high ReFINE score for "incorrect severity of a finding", which aligns with the human rating. Additionally, the report erroneously mentions a "right pleural effusion", leading to an "incorrect location/position of a finding", again perceived similarly by both the ReFINE score and human ratings. Lastly, the generated report fails to mention the "left retrocardiac opacification", leading to a score of '1.0' for "false prediction of a finding" from both the ReFINE score and the human rating.

Conclusions

ReFINE provides a human-aligned, interpretable metric for radiology report evaluation by mapping each item to its corresponding sub-score. Leveraging GPT-4 to generate tailored

Score	Spearman ↑	Kendall ↑	Pearson ↑
Score 1	0.41 (3.85e-19)	0.30 (5.99e-18)	0.41 (1.51e-18)
Score 2	0.15 (3.94e-2)	0.12 (3.79e-2)	0.15 (1.79e-2)
Score 3	0.45 (9.66e-25)	0.34 (7.03e-23)	0.44 (2.20e-22)
Score 4	0.32 (1.18e-11)	0.27 (1.54e-11)	0.37 (1.42e-14)
Score 5	0.25 (1.71e-10)	0.20 (3.08e-10)	0.23 (1.33e-9)
Score 6	0.11 (2.61e-2)	0.09 (2.66e-2)	0.10 (7.33e-2)

Table 7: Human Correlation (with p-values) on RaTE-Eval (Sentence-level, 9 modalities)

Metric	Kendall's Tau↑	Spearman↑
BLEU-4	0.07	0.05
ROUGE-L	0.16	0.12
METEOR	0.11	0.08
CIDEr	0.04	0.03
BERTScore	0.13	0.09
RadGraphF1	0.09	0.06
semb_score	0.01	0.01
RadCliQ-v1	0.08	0.06
ReFINE (Ours)	**0.23**	**0.29**

Table 8: Human Correlation on Rad-100 Dataset.

Model	Trainable Params (%)	Kendall	Spearman
Llama3 (Meta 2024)	6.8M (0.090)	0.751	0.910
Vicuna-7b (Chiang et al. 2023)	8.4M (0.127)	0.738	0.901
Meditron (Chen et al. 2023)	8.4M (0.127)	0.709	0.880
Gemma-7b (Gemma Team et al. 2024)	6.4M (0.075)	0.707	0.876
Qwen1.5-7b(Bai et al. 2023)	8.4M (0.110)	0.684	0.858
Phi-2 (Li et al. 2023)	5.3M (0.196)	0.591	0.784

Table 9: Ablation of LLM Backbones on ReXVal Dataset

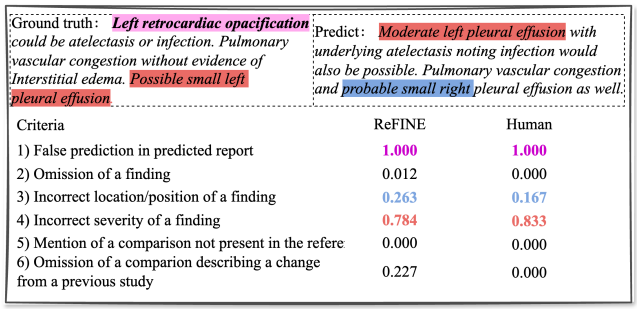

Figure 2: A visual example of ReFINE on ReXVal, where highlighted sentences and scores share matching colors.

training samples, we fine-tune LLMs with a reward-based system adaptable to diverse scoring criteria. **Limitations**. ReFINE's explainability could be improved with richer textual rationales, which are not yet included. Moreover, the cost of human evaluation restricts the size of our test sets, though their scale remains comparable to prior work.

References

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; Hui, B.; Ji, L.; Li, M.; Lin, J.; Lin, R.; Liu, D.; Liu, G.; Lu, C.; Lu, K.; Ma, J.; Men, R.; Ren, X.; Ren, X.; Tan, C.; Tan, S.; Tu, J.; Wang, P.; Wang, S.; Wang, W.; Wu, S.; Xu, B.; Xu, J.; Yang, A.; Yang, H.; Yang, J.; Yang, S.; Yao, Y.; Yu, B.; Yuan, H.; Yuan, Z.; Zhang, J.; Zhang, X.; Zhang, Y.; Zhang, Z.; Zhou, C.; Zhou, J.; Zhou, X.; and Zhu, T. 2023. Qwen Technical Report. *arXiv preprint arXiv:2309.16609*.

Banerjee, S.; and Lavie, A. 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*, 65–72.

Bannur, S.; Bouzid, K.; Castro, D. C.; Schwaighofer, A.; Thieme, A.; Bond-Taylor, S.; Ilse, M.; Pérez-García, F.; Salvatelli, V.; Sharma, H.; et al. 2024. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*.

Calamida, A.; Nooralahzadeh, F.; Rohanian, M.; Fujimoto, K.; Nishio, M.; and Krauthammer, M. 2023. Radiology-Aware Model-Based Evaluation Metric for Report Generation. *arXiv preprint arXiv:2311.16764*.

Chaves, J. M. Z.; Huang, S.-C.; Xu, Y.; Xu, H.; Usuyama, N.; Zhang, S.; Wang, F.; Xie, Y.; Khademi, M.; Yang, Z.; et al. 2024a. Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation. *arXiv preprint arXiv:2403.08002*.

Chaves, J. M. Z.; Huang, S.-C.; Xu, Y.; Xu, H.; Usuyama, N.; Zhang, S.; Wang, F.; Xie, Y.; Khademi, M.; Yang, Z.; et al. 2024b. Training small multimodal models to bridge biomedical competency gap: A case study in radiology imaging. *CoRR*.

Chen, Z.; Cano, A. H.; Romanou, A.; Bonnet, A.; Matoba, K.; Salvi, F.; Pagliardini, M.; Fan, S.; Köpf, A.; Mohtashami, A.; et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Chiang, C.-H.; and Lee, H.-y. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. *URL https://lmsys. org/blog/2023-03-30-vicuna*, 3(5).

Gemma Team, T. M.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; Tafti, P.; Hussenot, L.; and et al. 2024. Gemma.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Huang, A.; Banerjee, O.; Wu, K.; Reis, E. P.; and Rajpurkar, P. 2024. FineRadScore: A Radiology Report Line-by-Line Evaluation Technique Generating Corrections with Severity Scores. *arXiv preprint arXiv:2405.20613*.

Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R. L.; Shpanskaya, K. S.; Seekins, J.; Mong, D. A.; Halabi, S. S.; Sandberg, J. K.; Jones, R.; Larson, D. B.; Langlotz, C. P.; Patel, B. N.; Lungren, M. P.; and Ng, A. Y. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Hawaii, USA, January 27 - February 1, 2019*, 590–597. AAAI Press.

Jain, S.; Agrawal, A.; Saporta, A.; Truong, S. Q.; Duong, D. N.; Bui, T.; Chambon, P.; Zhang, Y.; Lungren, M. P.; Ng, A. Y.; et al. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.

Li, Y.; Bubeck, S.; Eldan, R.; Giorno, A. D.; Gunasekar, S.; and Lee, Y. T. 2023. Textbooks Are All You Need II: phi-1.5 technical report. arXiv:2309.05463.

Li, Y.; Wang, Z.; Liu, Y.; Wang, L.; Liu, L.; and Zhou, L. 2024. KARGEN: Knowledge-Enhanced Automated Radiology Report Generation Using Large Language Models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 382–392. Springer.

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

Liu, Y.; Li, Y.; Wang, Z.; Liang, X.; Liu, L.; Wang, L.; Cui, L.; Tu, Z.; Wang, L.; and Zhou, L. 2024a. A systematic evaluation of gpt-4v's multimodal capability for chest x-ray image analysis. *Meta-Radiology*, 100099.

Liu, Y.; Wang, Z.; Li, Y.; Liang, X.; Liu, L.; Wang, L.; and Zhou, L. 2024b. MRScore: Evaluating Radiology Report Generation with LLM-based Reward System. *arXiv preprint arXiv:2404.17778*.

Meta. 2024. Introducing Meta Llama 3: The Most Capable Openly Available LLM to Date. https://ai.meta.com/blog/meta-llama-3/. Accessed: 2024-05-20.

OpenAI. 2023. GPT-4 Technical Report. *ArXiv*, abs/2303.08774.

Ostmeier, S.; Xu, J.; Chen, Z.; Varma, M.; Blankemeier, L.; Bluethgen, C.; Michalson, A. E.; Moseley, M.; Langlotz, C.; Chaudhari, A. S.; et al. 2024. GREEN: Generative Radiology Report Evaluation and Error Notation. *arXiv preprint arXiv:2405.03595*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 3980–3990. Association for Computational Linguistics.

Smit, A.; Jain, S.; Rajpurkar, P.; Pareek, A.; Ng, A. Y.; and Lungren, M. P. 2020. CheXbert: combining automatic la-

belers and expert annotations for accurate radiology report labeling using BERT. *arXiv preprint arXiv:2004.09167*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Wang, Z.; Liu, L.; Wang, L.; and Zhou, L. 2023. R2GenGPT: Radiology Report Generation with Frozen LLMs. *arXiv preprint arXiv:2309.09812*.

Yu, F.; Endo, M.; Krishnan, R.; Pan, I.; Tsai, A.; Reis, E. P.; Fonseca, E.; Lee, H.; Shakeri, Z.; Ng, A.; et al. 2023a. Radiology Report Expert Evaluation (ReXVal) Dataset.

Yu, F.; Endo, M.; Krishnan, R.; Pan, I.; Tsai, A.; Reis, E. P.; Fonseca, E. K. U. N.; Lee, H. M. H.; Abad, Z. S. H.; Ng, A. Y.; et al. 2023b. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9).

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 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.

Zhao, W.; Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2024. Ratescore: A metric for radiology report generation. *arXiv preprint arXiv:2406.16845*.