arXiv:2508.15464v1 [cs.CL] 21 Aug 2025

RadReason: Radiology Report Evaluation Metric with Reasons and Sub-Scores

Yingshu Li[1], **Yunyi Liu**[1], **Lingqiao Liu**[2], **Lei Wang**[3], **Luping Zhou**[1]*

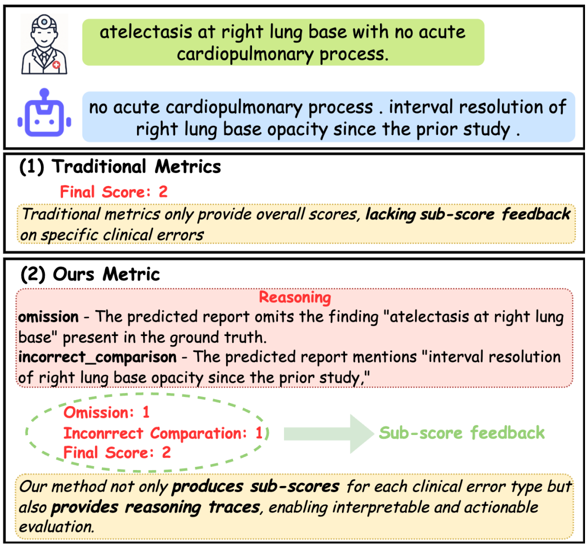
[1]School of Electrical and Computer Engineering, University of Sydney, NSW 2006, Australia
[2]School of Computer Science, University of Adelaide, SA 5005, Australia
[3]School of Computing and Information Technology, University of Wollongong, NSW 2522, Australia

Abstract

Evaluating automatically generated radiology reports remains a fundamental challenge due to the lack of clinically grounded, interpretable, and fine-grained metrics. Existing methods either produce coarse overall scores or rely on opaque black-box models, limiting their usefulness in real-world clinical workflows. We introduce **RadReason**, a novel evaluation framework for radiology reports that not only outputs fine-grained sub-scores across six clinically defined error types, but also produces human-readable justifications that explain the rationale behind each score. Our method builds on Group Relative Policy Optimization and incorporates two key innovations: (1) **Sub-score Dynamic Weighting**, which adaptively prioritizes clinically challenging error types based on live F1 statistics; and (2) **Majority-Guided Advantage Scaling**, which adjusts policy gradient updates based on prompt difficulty derived from sub-score agreement. Together, these components enable more stable optimization and better alignment with expert clinical judgment. Experiments on the ReXVal benchmark show that RadReason surpasses all prior offline metrics and achieves parity with GPT-4-based evaluations, while remaining explainable, cost-efficient, and suitable for clinical deployment. Code will be released upon publication.

1 Introduction

The automatic generation of radiology reports (RRG) from medical images has emerged as a pivotal task in clinical AI, offering the promise of reducing radiologists' workload and improving diagnostic consistency (Huang et al., 2023; Li et al., 2023, 2024; Wang et al., 2024). However, evaluating the quality of generated reports remains a fundamental challenge. Traditional natural language generation (NLG) metrics, such as BLEU, ROUGE, and METEOR (Papineni et al., 2002; Banerjee and Lavie, 2005; Lin, 2004), focus on word overlap and fail to capture clinically meaningful semantic differences, particularly in cases involving paraphrasing, negation, or subtle factual errors. Embedding-based metrics like BERTScore (Zhang et al., 2019) improve semantic alignment but often overlook domain-specific entities critical to clinical interpretation. Structure-aware metrics, including RadGraph F1 (Jain et al., 2021) and CheXbert F1 (Smit et al., 2020), incorporate medical knowledge but lack sensitivity to overall report quality and suffer from limited granularity. RadCliQ (Yu et al., 2023a) takes a step further by learning a regression model over multiple existing metrics to better approximate human judgments. More recently, large language models (LLMs) have been explored for radiology report evaluation (Grattafiori et al., 2024; Yang et al., 2024). MRScore (Liu et al., 2024) proposes a radiology-specific reward model to enable customized scoring, while Green (Ostmeier et al.,


atelectasis at right lung base with no acute cardiopulmonary process.

no acute cardiopulmonary process . interval resolution of right lung base opacity since the prior study .

*Traditional metrics only provide overall scores, **lacking sub-score feedback** on specific clinical errors*

omission - The predicted report omits the finding "atelectasis at right lung base" present in the ground truth.
incorrect_comparison - The predicted report mentions "interval resolution of right lung base opacity since the prior study,"

*Our method not only **produces sub-scores** for each clinical error type but also **provides reasoning traces**, enabling interpretable and actionable evaluation.*

Figure 1: **Comparison of evaluation outputs across metric types.** Traditional metrics yield only overall scores, lacking insight into specific errors. Our method provides both sub-scores and reasoning traces, enabling interpretable and detailed evaluation.

*Corresponding author

2024) evaluates factual correctness through explicit error-type matching, achieving close alignment with expert judgment. RaTEScore (Zhao et al., 2024) improves semantic robustness using entity-aware similarity that handles synonyms and negations. In online GPT-based applications, CheXprompt (Zambrano Chaves et al., 2025) and FineRadScore (Huang et al., 2024) leverage GPT-4 to identify clinical error types and generate detailed corrections through few-shot prompting.

Despite recent progress, existing evaluation methods, summarized in Figure 1, face two major limitations: (1) most systems produce only a single overall score, lacking error-type granularity; and (2) few provide explicit reasoning for *why* a particular score was assigned, limiting clinical usability and model transparency. To address these issues, we introduce RadReason, a novel evaluation framework that decomposes report quality into six clinically defined error dimensions (e.g., false prediction, omission, incorrect location) (Yu et al., 2023a), and produces both structured sub-scores and corresponding natural language explanations for each generated report. For example, "the report failed to mention left-sided effusion → omission errors = 1". Technically, RadReason is trained via Group Relative Policy Optimization (GRPO) (Shao et al., 2024; Guo et al., 2025), a reinforcement learning paradigm that models preferences over grouped completions. However, unlike prior work that yields a single scalar score, our framework predicts six distinct sub-scores, each corresponding to a specific error type. This setting introduces two primary challenges: (1) some error types are rare or harder to handle, requiring adaptive prioritization; and (2) report prompts vary in difficulty, with some inducing consistent completions while others exhibit high disagreement. To mitigate these challenges, we incorporate two auxiliary mechanisms: (1) Sub-score Dynamic Weighting, which dynamically adjusts reward weights according to the F1 performance of each error type to target areas of weakness; and (2) Majority-Guided Advantage Scaling, which leverages majority vote statistics to estimate sample difficulty and scale policy gradient updates accordingly. Experiments on the ReXVal benchmark (Yu et al., 2023b) demonstrate that RadReason achieves state-of-the-art correlation with expert ratings, outperforming all prior offline metrics while remaining interpretable. Our key contributions include:

(1) We introduce **RadReason**, one reward-optimization-based evaluation framework for radiology report generation that outputs structured sub-scores and natural language explanations.
(2) We propose two novel training strategies, Sub-score Dynamic Weighting and Majority-Guided Advantage Scaling, to enhance clinical sensitivity.
(3) RadReason achieves state-of-the-art human alignment on ReXVal, while remaining efficient, interpretable, and extensible to new evaluation criteria.

2 Related Works

2.1 Evaluation Metrics for Radiology Reports.

Evaluating radiology report generation (RRG) requires metrics that assess both linguistic fluency and clinical correctness (Yu et al., 2023a). Traditional NLG metrics—such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005), focus on surface-level overlap, failing to capture paraphrasing, negation, or subtle clinical inaccuracies. As a result, semantically accurate reports with alternative phrasing may be penalized unfairly. To address semantic fidelity and clinical content, several domain-aware metrics have been proposed. CheXbert F1 (Smit et al., 2020) leverages a pathology classification model trained on 14 thoracic diseases, while RadGraph F1 (Jain et al., 2021) evaluates factual correctness via structured entity-relation graphs. RadCliQ (Yu et al., 2023a) combines multiple such metrics using a regression model aligned with human annotations, improving correlation but offering limited interpretability. Recent work leverages large language models (LLMs) for radiology report evaluation (Grattafiori et al., 2024; Yang et al., 2024). MRScore (Liu et al., 2024) trains a radiology-specific reward model to define a custom scoring framework. Green (Ostmeier et al., 2024) evaluates factual correctness and clinical significance based on matched findings and identified errors, demonstrating strong alignment with expert assessments. RaTEScore (Zhao et al., 2024) is an entity-aware metric that handles synonyms and negations robustly, further aligning with human judgments. Several methods have also explored prompting commercial LLMs such as GPT-4. CheXprompt (Zambrano Chaves et al., 2025) use GPT-4 (Achiam et al., 2023) to detect six specific error types: false positives, omissions, incorrect location, incorrect severity, irrelevant comparisons, and missing comparative statements. FineRadScore (Huang et al., 2024) applies few-shot prompting to elicit line-by-line corrections and clinical severity ratings for each identified error. However, these methods raise privacy concerns and depend on online access, which limits their practical deployment. Moreover, a common gap remains: most metrics provide only a single overall score, lacking fine-grained sub-aspect feedback or a clear rationale behind score assignments. To address this, we develop an offline evaluation framework capable of producing clinically aligned sub-scores and reasoning explanations per error type, advancing both interpretability and practical applicability.

2.2 Reasoning in Large Language Models.

Recent advances in large language models (LLMs) have shown strong capabilities in mimicking human-like reasoning, particularly by decomposing complex tasks into structured intermediate steps. This paradigm—often referred to as explicit reasoning—enables models to engage in interpretable, step-by-step thinking before arriving at a final output. A variety of techniques have been proposed to follow this approach, including prompting-based strategies such as Chain-of-Thought (CoT) (Wei et al., 2022), planning-oriented methods like Graph-of-Thought and Tree-of-Thought (Besta et al., 2024; Yao et al., 2023). Beyond prompting, supervised fine-tuning (SFT) on datasets annotated with reasoning traces (Kumar et al., 2025) can further enhance reasoning ability, but requires high-quality, labor-intensive annotations that limit scalability. To overcome this, recent work has adopted reinforcement learning (RL) to induce reasoning behaviors without explicit supervision. For example, DeepSeek-R1 (Guo et al., 2025) introduces an RL framework where the model is guided to generate reasoning trajectories followed by answers, and is rewarded based on final correctness, enabling learning from answer-only datasets. Building on this idea, we adopt an RL framework tailored to the evaluation metric. Our goal is to teach the model to both assign fine-grained sub-scores and generate natural language justifications. This enables interpretable, aspect-specific evaluation, critical for clinical auditability and decision support.

3 Methods

We propose a reinforcement learning framework for radiology report evaluation that produces interpretable sub-scores and explanations across clinical error types, as shown in Figure 2. Our method integrates three key components: (1) sub-score prediction with natural language reasoning; (2) **Dynamic Sub-score Weighting**, which emphasizes clinically challenging aspects by adapting to per-dimension performance; and (3) **Majority-Guided Advantage Scaling**, which modulates policy updates based on prompt difficulty to reward rare but valuable completions.

3.1 Background

Group Relative Policy Optimization (GRPO) (Guo et al., 2025) is a reinforcement learning algorithm designed for optimizing language models using group-wise preference signals. Unlike pairwise methods like DPO (Rafailov et al., 2023), GRPO compares multiple completions within a group and computes relative advantages to guide policy updates. For each prompt, GRPO samples a set of completions and assigns rewards via a reward model. Let $\mathbf{r} = \{r_1, r_2, \cdots, r_G\}$ denote the set of rewards for a group of G completions for question q. The advantage of each completion is normalized within its group:

$$\hat{A}_{i,t} = \tilde{r}_i = \frac{r_i - \mathrm{mean}(\mathbf{r})}{\mathrm{std}(\mathbf{r})}. \tag{1}$$

The GRPO loss function can be defined as:

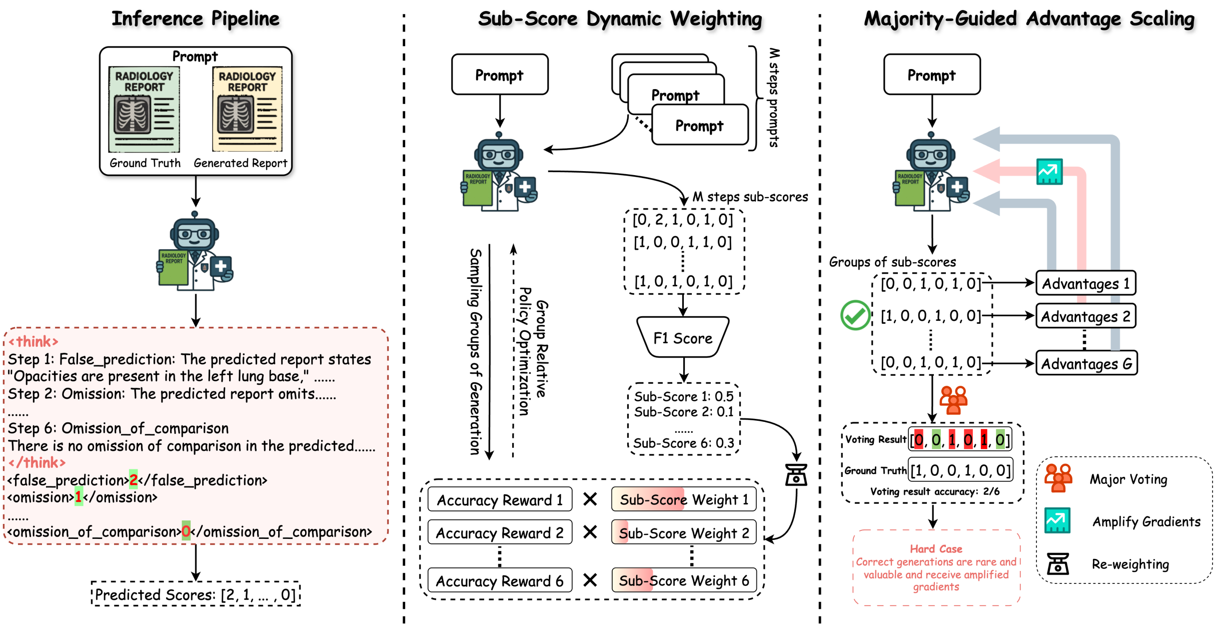
$$\mathcal{L}_{\mathrm{GRPO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\Bigg[\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\mathrm{old}}}(o_{i,t} \mid q, o_{i,<t})}\hat{A}_{i,t} - \beta\, D_{\mathrm{KL}}\left(\pi_\theta \,\|\, \pi_{\theta_{\mathrm{ref}}}\right)\Bigg], \tag{2}$$

where o_i is the set of sampled completions for the i-th prompt, and $\pi_\theta(o_{i,t} \mid q, o_{i,<t})$ denotes the token-level likelihood under the current policy.

3.2 Rewards Construction

To guide GRPO training, we design three reward functions that capture distinct aspects of evaluation quality: reasoning completeness, formatting correctness, and sub-score accuracy. For each sampled completion, we compute rewards as follows:

Structured Reasoning Reward. This reward promotes interpretability by verifying whether the model explicitly discusses all six clinical error types proposed in (Yu et al., 2023a): false prediction, omission, incorrect location, incorrect severity, incorrect comparison, and omission of comparison. We use regular expressions to detect whether each aspect is addressed with appropriate structural cues (e.g., “Step 1: false prediction”).



```
<think>
Step 1: False_prediction: The predicted report states
"Opacities are present in the left lung base," ......
Step 2: Omission: The predicted report omits......
......
Step 6: Omission_of_comparison
There is no omission of comparison in the predicted......
</think>
<false_prediction>2</false_prediction>
<omission>1</omission>
......
<omission_of_comparison>0</omission_of_comparison>
```

Figure 2: Overview of our training framework. **Left:** The model produces detailed sub-scores and explanations across six clinically defined error types. **Middle:** During training, we apply Dynamic Sub-score Weighting by periodically computing F1 gaps and reweighting each dimension's reward accordingly. **Right:** We introduce Majority-Guided Advantage Scaling, which estimates prompt difficulty by comparing majority-voted sub-scores across completions with the ground truth. Our method amplifies gradients for correct completions on hard prompts, while penalizing incorrect completions more heavily on easy ones.

Format Reward. We enforce output by requiring a single <think>...</think> block containing the reasoning text, and exactly one valid numerical value within each sub-score tag (e.g., <omission>1.0</omission>). This promotes format consistency and ensures reliable sub-score extraction.

Accuracy Reward. This component measures how closely the predicted sub-scores align with ground-truth annotations. For each generated report, we extract predicted sub-scores via regular expressions and compare them against the corresponding ground-truth values. Unlike previous GRPO-based approaches that use binary (0/1) reward signals, penalizing any deviation from the ground truth with a zero reward, such rigid schemes fail to differentiate between near-correct and completely incorrect predictions, leading to sparse and unstable learning signals. To address this, we introduce a smooth gaussian reward function that penalizes prediction errors based on the squared distance from the ground truth, thereby providing a more stable and informative learning signal. The reward for sub-score j is computed as:

$$r^{(j)} = \exp\left(-\frac{(\text{pred}^{(j)} - \text{gt}^{(j)})^2}{2\sigma^2}\right), \tag{3}$$

with $\sigma = 0.5$ as the standard deviation controlling the tolerance to prediction error.

The sub-score rewards are then averaged across all $K = 6$ dimensions:

$$r_{\text{sub}} = \frac{1}{K}\sum_{j=1}^{K} r^{(j)}. \tag{4}$$

In addition to evaluating each sub-score individually, we incorporate a total-score alignment term that encourages the sum of predicted sub-scores to match the ground-truth sum:

$$r_{\text{total}} = \exp\left(-\frac{(\hat{s}_{\text{total}} - s_{\text{total}})^2}{2\sigma^2}\right), \tag{5}$$

where $\hat{s}_{\text{total}}$ and s_{total} denote the predicted and ground-truth total scores respectively.

The final accuracy reward is a weighted combination of the two components:

$$r_{\text{acc}} = r_{\text{sub}} + r_{\text{total}}. \tag{6}$$

The final reward for one output is:

$$r = r_{reasoning} + r_{format} + r_{acc} \tag{7}$$

3.3 Sub-score Dynamic Weighting

While the accuracy reward evaluates each sub-score independently, it implicitly assumes equal importance across all error types. In practice, these aspects differ in frequency, ambiguity, and clinical impact. For example, *omission* and *false prediction* errors occur more frequently but tend to be ambiguous, while incorrect comparison is rarer yet often carries critical clinical implications. Consequently, uniformly averaging sub-score rewards can lead to biased optimization, with frequent and easier-to-learn aspects disproportionately influencing the gradient updates.

To address this imbalance, we propose a Sub-score Dynamic Weighting **(SDW)** strategy, which dynamically emphasizes sub-scores where the model exhibits weaker performance. Concretely, every M steps, we compute the F1 score $F1^{(j)}$ for each aspect $j \in \{1, ..., K\}$, and define the relative performance gap as: $\Delta_j = \bar{F}1 - F1^{(j)}$, where $\bar{F}1 = \frac{1}{K}\sum_{j=1}^{K} F1^{(j)}$.

We then convert these difficulty scores into a normalized set of weights using a softmax function:

$$w_j = 1 + \frac{\exp(\alpha \cdot \Delta_j)}{\sum_{k=1}^{K} \exp(\alpha \cdot \Delta_k)}, \tag{8}$$

where α is a temperature hyperparameter controlling the sharpness of the focus on under-performing dimensions.

We then compute the final sub-score reward as a weighted average across dimensions:

$$r_{\text{sub}}^{\text{dyn}} = \frac{1}{K}\sum_{j=1}^{K} w_j \cdot r^{(j)}. \tag{9}$$

This strategy allows the model to continuously reallocate supervision toward clinically challenging or underperforming aspects. Unlike static reward averaging, SDW encourages more balanced learning and improves robustness across diverse error types.

Algorithm 1 Sub-score Dynamic Weighting

```
Require: F1 scores, update interval $M$, temperature $\alpha$
  Initialize $w_j \leftarrow 1$ for $j = 1, \ldots, K$
  for each training step $t = 1$ to $T$ do
      if $t \mod M = 0$ then
          Calculate $F1^{(j)}$ score for each aspect $j$
          Compute average F1: $\bar{F}1 \leftarrow \frac{1}{K}\sum_{j=1}^{K} F1^{(j)}$
          for each aspect $j = 1$ to $K$ do
              Compute F1 gap: $\Delta_j \leftarrow \bar{F}1 - F1^{(j)}$
          end for
          Update weights with softmax:
              $w_j \leftarrow 1 + \frac{\exp(\alpha \cdot \Delta_j)}{\sum_{k=1}^{K} \exp(\alpha \cdot \Delta_k)}$
      end if
  end for
```

3.4 Majority-Guided Advantage Scaling

While GRPO models relative preferences between completions for a given prompt, it treats all training prompts equally, regardless of their difficulty. However, not all samples offer equal learning utility. Some prompts are difficult, as evidenced by consistently poor completions; in such cases, high-quality generations are both rare and informative. Conversely, other prompts are comparatively easy, where most generations perform well, and errors on such prompts may indicate critical model failures.

To address this, we introduce a Majority-Guided Advantage Scaling **(MGAS)** mechanism, which adjusts the advantage magnitude based on the inferred difficulty of each prompt. Specifically, for each prompt group, we aggregate the predicted sub-scores from all G completions and perform majority voting for each of the $K = 6$ sub-score dimensions. For each sub-score dimension, we aggregate all predicted values across the group and compute the majority vote. This majority-voted value is then compared against the corresponding ground-truth label. If the majority prediction fails to match the ground truth under a defined agreement threshold, we consider it a hard case.

We average the correctness across all six aspects to compute a majority-selected score γ. This score reflects how easy the prompt is—i.e., how often the majority prediction per aspect matches the ground truth. Formally, let $P^{(j)} = [p_1^{(j)}, p_2^{(j)}, ..., p_G^{(j)}]$ denote the group of predicted values for the j-th sub-score aspects across G completions, and let $y^{(j)}$ be the ground truth value. We define the majority-selected score γ as:

$$\gamma = \frac{1}{K}\sum_{j=1}^{K} \mathbf{1}\left[\text{mode}\left(P^{(j)}\right) = y^{(j)}\right], \tag{10}$$

where $\text{mode}(\cdot)$ returns the most frequent value in the group. The score $\gamma \in [0, 1]$ reflects the proportion of sub-score aspects where the majority prediction matches the ground truth. The scaling method is defined as:

$$s_i(\gamma) = \phi_- + (\phi_+ - \phi_-) \cdot (1 + (\psi_i(\gamma) - c))^{-\beta}. \tag{11}$$

Here, ϕ_- and ϕ_+ are the lower and upper bounds of the scaling. c is a difficulty threshold, and β controls the sharpness of the modulation. The function $\psi(\gamma)$ is defined as:

$$\psi_i(\gamma) = \begin{cases} \gamma, & \hat{A}_{i,t} > 0 \\ 1 - \gamma, & \hat{A}_{i,t} < 0 \end{cases} \tag{12}$$

The final updated advantages can be defined as:

$$\hat{A}'_{i,t} = s_i(\gamma) \cdot \hat{A}_{i,t}. \tag{13}$$

Algorithm 2 Majority-Guided Advantage Scaling

```
Input: A Group of Predicted sub-scores $P$, ground truth $y$, original advantages $\hat{A}_{i,t}$, scaling parameters $(\phi_-, \phi_+, c)$
Output: Scaled advantages $\hat{A}'_{i,t}$
for each aspect $j = 1, \dots, K$ do
    Collect predictions $P^{(j)} = [p_1^{(j)}, p_2^{(j)}, ..., p_G^{(j)}]$
    Compute majority prediction: $m^{(j)} \leftarrow \mathrm{mode}(P^{(j)})$
    $\gamma^{(j)} \leftarrow \mathbf{1}[m^{(j)} = y^{(j)}]$
end for
Compute agreement: $\gamma \leftarrow \frac{1}{K} \sum_{j=1}^{K} \gamma^{(j)}$
for each completion $i = 1, \dots, G$ do
    if $\hat{A}_{i,t} > 0$ then
        $\psi_i \leftarrow \gamma$
    else
        $\psi_i \leftarrow 1 - \gamma$
    end if
    Compute scaling factor:
    $s_i \leftarrow \phi_- + (\phi_+ - \phi_-) \cdot (1 + (\psi_i(\gamma) - c))^{-\beta}$ .
    Update advantages:
    $\hat{A}'_{i,t} \leftarrow s_i \cdot \hat{A}_{i,t}$
end for
```

In summary, we introduce two strategies to enhance optimization in GRPO-based training. **Sub-score Dynamic Weighting** focuses learning on difficult error types by adjusting weights based on F1 performance. **Majority-Guided Advantage Scaling** amplifies gradients for correct completions on hard prompts and downweights errors on easy ones. Together, these mechanisms guide learning toward clinically meaningful and robust behavior.

4 Experiments and Results

4.1 Datasets

Training Data Generation. We sample 1,000 radiology reports from the MIMIC-CXR dataset to serve as ground-truth anchors. For each case, we prompt GPT-4, previously shown to align closely with expert radiologists in evaluation tasks (Liu et al., 2024), to generate synthetic diagnostic reports exhibiting varied error profiles. We control report quality by systematically injecting clinical errors (e.g., omissions, false findings, localization mistakes) to produce reports of varying fidelity. Specifically, GPT-4 is instructed to generate: (1) high-quality reports containing 0–1 errors; (2) medium-quality reports containing 2–3 errors; (3) low-quality reports containing 4 or more errors.

This error-count–based prompting enables fine-grained control over semantic fidelity, ensuring comprehensive coverage across clinically plausible quality levels. In total, we collect 3,968 labeled report completions, each paired with its corresponding ground-truth anchor. Detailed prompt is in Appendix A.

ReXVal (Yu et al., 2023b) is a publicly available benchmark designed to evaluate alignment between automated metrics and expert human judgments in radiology report assessment. It consists of 200 candidate–reference report pairs sampled from 50 MIMIC-CXR studies, with four candidate reports per study. Each pair is annotated by six board-certified radiologists using the six-category RadCliQ error taxonomy, which provides fine-grained sub-score annotations based on the number of errors in each category.

4.2 Experimental Settings

We evaluate our proposed metric against a diverse set of established baselines, including traditional NLG metrics (BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005)), semantic similarity metrics (BERTScore (Zhang et al., 2019)), and clinically informed or structure-aware metrics such as RadGraph F1 (Jain et al., 2021), Semb score, and RadCliQ-v1 (Yu et al., 2023a). Following prior work (Yu et al., 2023b), we assess alignment with human judgment by computing Kendall's Tau and Spearman correlation between metric outputs and expert annotations on the ReXVal benchmark.

We adopt Qwen2.5-7B[1] (Team, 2024) as the base language model. Fine-tuning is conducted using LoRA (Hu et al., 2022) with a rank of 16 and a scaling factor of $\alpha = 16$. We set the learning rate to 1e$-$5, batch size to 4, and train for 2 epochs using 2×NVIDIA RTX A6000 (48GB) GPUs. For Majority-Guided Advantage Scaling, we use $\phi_- = 0.8$, $\phi_+ = 1.2$, and difficulty threshold $c = 0.5$.

4.3 Main Results

Comparison with Existing Metrics. Table 1 presents a comparative evaluation of our proposed metric against a comprehensive set of baseline methods on the ReXVal dataset, using Kendall's Tau and Spearman correlation coefficients to assess alignment with expert ratings. Our method achieves the highest overall performance among all non-LLM approaches, with a Kendall's Tau of 0.730 and Spearman correlation of 0.871. It

[1] https://huggingface.co/Qwen/Qwen2.5-7B

Ground truth report: post left upper lobectomy changes, with no superimposed acute intrathoracic process detected.

Predicted report: small left pleural effusion. no superimposed acute intrathoracic process.

Reasoning	Ours	Human
Step 1: false_prediction - The predicted report mentions "small left pleural effusion" which is not mentioned in the ground truth. This is an additional finding not present in the ground truth report.	**1.0**	**1.0**
Step 2: omission - The ground truth mentions "post left upper lobectomy changes" which is omitted in the predicted report.	**1.0**	**1.0**
Step 3: incorrect_location - The predicted report does not misidentify the location of any finding.	0.0	0.0
Step 4: incorrect_severity - The predicted report does not misassess the severity of any finding.	0.0	0.0
Step 5: incorrect_comparison - No comparison is made in either report.	0.0	0.0
Step 6: omission_of_comparison - No comparison is mentioned in either report.	0.0	0.0

Ground truth report: ap chest compared to ___: previous mild but asymmetric pulmonary edema continues to improve. the residual opacification in the right upper lobe raises concern for pneumonia. heart size is normal. there is no pleural effusion.

Predicted report: compared to chest radiographs ___ . pulmonary vascular congestion which developed between ___ and ___ persists . very mild interstitial pulmonary edema has improved , but small right pleural effusion has accumulated . the absence of the vasculature in the right upper lobe is usually an indication of emphysema . heart size is normal .

Reasoning	Ours	Human
Step 1: false_prediction - The predicted report mentions "pulmonary vascular congestion" and "small right pleural effusion" which are not mentioned in the ground truth. It also incorrectly states that the absence of vasculature in the right upper lobe indicates emphysema instead of pneumonia.	**2.0**	**1.667**
Step 2: omission - The predicted report omits the continuation of improvement in asymmetric pulmonary edema and the residual opacification in the right upper lobe that raises concern for pneumonia.	**1.0**	**0.667**
Step 3: incorrect_location - Not applicable in this case as the locations of the findings are not misidentified.	0.0	0.0
Step 4: incorrect_severity - The predicted report downgrades the severity of pulmonary edema from "mild but asymmetric" to "very mild."	**1.0**	**0.667**
Step 5: incorrect_comparison - Not applicable in this case.	0.0	0.333
Step 6: omission_of_comparison - Not applicable in this case.	0.0	0.333

Figure 3: Case studies comparing our model's sub-score predictions with human annotations, with step-by-step reasoning shown for each of the six clinical error types. Color highlights mark reasoning findings.

consistently outperforms traditional NLG metrics (BLEU-4, ROUGE-L, METEOR), semantic similarity scores (BERTScore, Semb Score), and structure-aware clinical metrics (RadGraph F1, RadCliQ-v1). Recent advanced models, such as GREEN (Kendall: 0.640) and RaTEScore (0.527), lack fine-grained interpretability and do not provide sub-scores, limiting their utility in detailed evaluation scenarios. Online methods such as CheXprompt and FineRadScore achieve comparable Kendall scores (0.750 and 0.737, respectively) but rely on commercial LLM APIs (e.g., GPT-4), which introduce concerns related to privacy, reproducibility, and operational cost.

4.4 Ablation Study

To evaluate the contribution of each component, we perform an ablation study summarized in Table 2. Beginning with a baseline model trained via supervised fine-tuning, we incrementally introduce GRPO, Sub-score Dynamic Weighting (SDW), and Majority-Guided Advantage Scaling (MGAS).

Introducing GRPO substantially improves alignment with expert annotations (Kendall: 0.495 → 0.690; Spearman: 0.634 → 0.832), confirming the effectiveness of preference-based learning in evaluation radiology reports. We attribute part of this gain to GRPO's ability to better capture intermediate reasoning signals during generation. Adding SDW yields further improvements in sub-scores such as *Omission of finding* and *Absence of comparison*, highlighting its utility in dynamically emphasizing clinically underrepresented yet critical error types. Finally, incorporating MGAS leads to the best overall performance (Kendall: **0.730**, Spearman: **0.871**). This validates our intuition

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
BertScore (Zhang et al., 2019)	0.507	0.677
RadGraphF1 (Jain et al., 2021)	0.516	0.702
Semb_score (Yu et al., 2023a)	0.494	0.665
RadCliQ-v1 (Yu et al., 2023a)	0.631	0.816
GREEN (Ostmeier et al., 2024)	0.640	–
RaTEScore (Zhao et al., 2024)	0.527	–
Ours	**0.730**	**0.871**
Results below are not strictly comparable because they are using an online model (e.g., GPT4).		
RadFact (Bannur et al., 2024)	0.590	–
CheXprompt (Zambrano Chaves et al., 2025)	0.750	–
FineRadScore (Huang et al., 2024)	0.737	–

Table 1: Human Correlation Comparison of Evaluation Metrics on ReXVal Dataset.

Criteria	**Baseline**		**+ GRPO**		**+ SDW**		**+ MGAS (Ours)**	
	Kendall	Spearman	Kendall	Spearman	Kendall	Spearman	Kendall	Spearman
False prediction	0.507	0.581	0.608	0.704	0.614	0.704	**0.645**	**0.742**
Omission of finding	0.323	0.366	0.576	0.662	0.583	0.682	**0.596**	**0.706**
Incorrect location	0.375	0.401	0.461	0.482	**0.533**	**0.571**	0.473	0.506
Incorrect severity	0.430	0.460	0.571	0.614	0.450	0.482	**0.570**	**0.611**
Absence of comparison	0.106	0.112	0.170	0.182	0.176	0.189	**0.186**	**0.196**
Omission of comparison	0.160	0.168	0.194	0.204	**0.317**	**0.333**	0.238	0.252
Total	0.495	0.634	0.690	0.832	0.698	0.838	**0.730**	**0.871**

Table 2: Ablation study of reward design. Our full model combines GRPO training with Sub-score Dynamic Weighting (SDW) and Majority-Guided Advantage Scaling (MGAS), achieving the highest correlation in bold.

that difficult prompts deserve greater reward signal amplification when answered correctly, while easy prompts should be penalized more harshly if mistakes occur. MGAS helps stabilize updates by aligning learning signals with case difficulty.

Overall, these results demonstrate the complementary benefits of SDW and MGAS in enhancing the optimization dynamics of GRPO, achieving high correlation across diverse error categories.

4.5 Qualitative Analysis

Figure 3 provides visual examples of how our model's sub-score predictions align with human annotations under the RadCliQ framework. In the first case, the generated report incorrectly introduces a new finding "small left pleural effusion" and omits the key phrase "post left upper lobectomy changes" (both highlighted), resulting in scores of 1.0 for both "false prediction" and "omission of finding," fully aligned with expert ratings. In the second case, the model hallucinates "pulmonary vascular congestion" and "small right pleural effusion," while omitting the clinically important improvement in "pulmonary edema" and "right upper lobe opacification." These errors lead to elevated scores for "false prediction," "omission," and "incorrect severity," closely mirroring human assessments. These examples illustrate the model's ability to perform fine-grained error identification consistent with expert judgment.

5 Conclusion

We introduce **RadReason**, an interpretable evaluation framework for radiology reports that produces structured sub-scores and explicit reasoning across clinically meaningful error categories. By embedding *Sub-score Dynamic Weighting* and *Majority-Guided Advantage Scaling*, our method adaptively focuses on harder sub-aspects and calibrates learning based on prompt difficulty. Empirical results on the ReXVal benchmark demonstrate that RadReason not only surpasses prior metrics.

Limitations

Our evaluation is conducted on ReXVal (Yu et al., 2023b), the standard benchmark for radiology report assessment aligned with human judgment. While moderate in size due to the cost of expert annotation, it enables fair and meaningful comparisons across methods. We adopt the six clinically grounded error categories defined in RadCliQ (Yu et al., 2023a); though fixed, our framework is modular and readily extensible to alternative or hierarchical taxonomies. While our experiments focus on chest X-ray reports from MIMIC-CXR, the reward-based approach is modality-agnostic and can generalize to structured diagnostic outputs from CT, MRI, or multimodal reports. Looking forward, the sub-score-based reward formulation may also inspire evaluation methods for other clinical generation tasks such as medical VQA.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, and 1 others. 2024. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and 1 others. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

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.

Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. 2023. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19809–19818.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, and 1 others. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.

Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip HS Torr, Fahad Shahbaz Khan, and Salman Khan. 2025. Llm post-training: A deep dive into reasoning large language models. *arXiv preprint arXiv:2502.21321*.

Yingshu Li, Yunyi Liu, Zhanyu Wang, Xinyu Liang, Lingqiao Liu, Lei Wang, Leyang Cui, Zhaopeng Tu, Longyue Wang, and Luping Zhou. 2023. A comprehensive study of gpt-4v's multimodal capabilities in medical imaging. *medRxiv*, pages 2023–11.

Yingshu Li, Zhanyu Wang, Yunyi Liu, Lei Wang, Lingqiao Liu, and Luping Zhou. 2024. Kargen: Knowledge-enhanced automated radiology report generation using large language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 382–392. Springer.

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

Yunyi Liu, Zhanyu Wang, Yingshu Li, Xinyu Liang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2024. Mrscore: Evaluating medical report with llm-based reward system. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 283–292. Springer.

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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. 2020. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Xiao Wang, Fuling Wang, Yuehang Li, Qingchuan Ma, Shiao Wang, Bo Jiang, Chuanfu Li, and Jin Tang. 2024. Cxpmrg-bench: Pre-training and benchmarking for x-ray medical report generation on chexpert plus dataset. *arXiv preprint arXiv:2410.00379*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

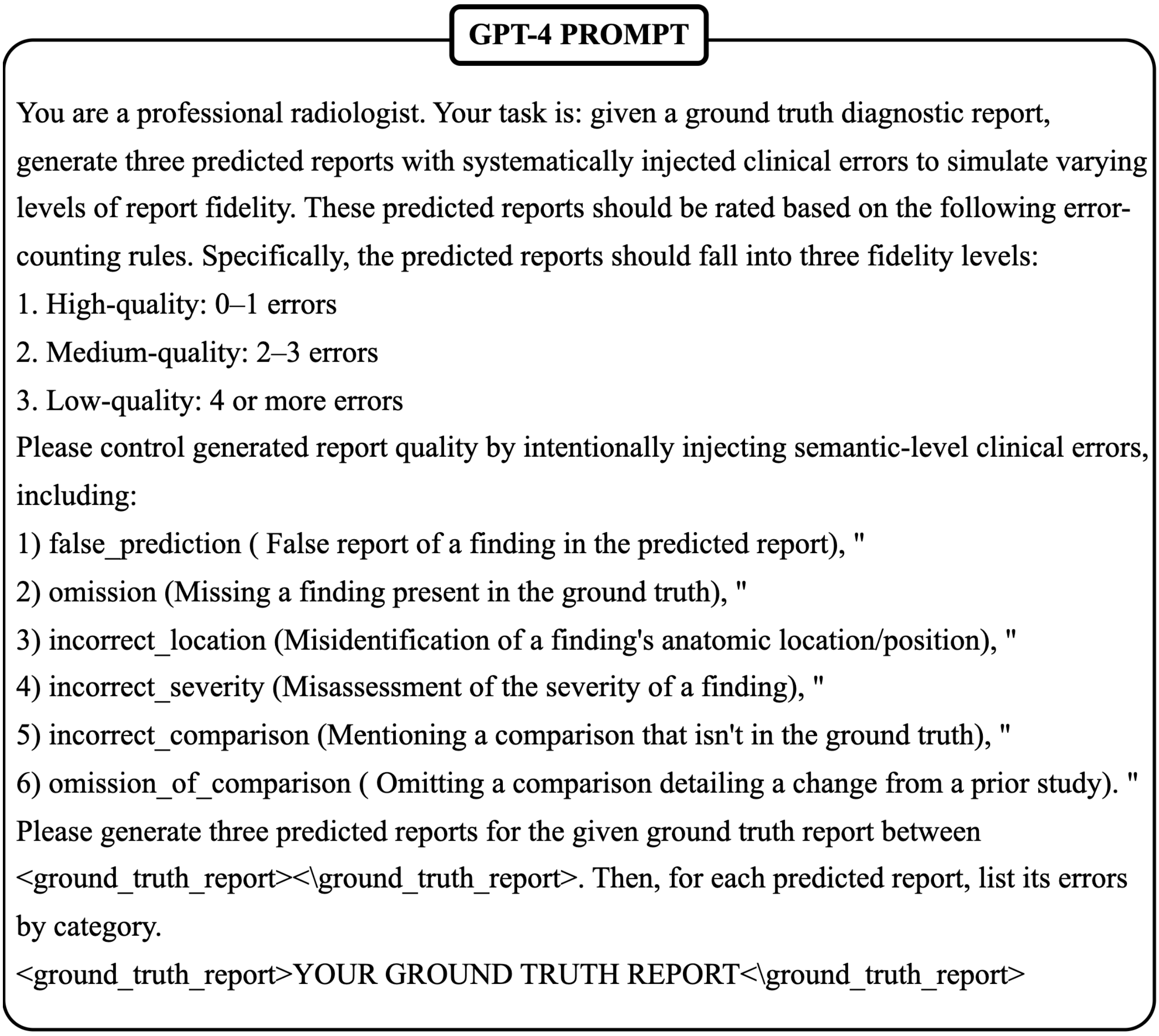
Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, and 1 others. 2023a. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9).

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, EKU Fonseca, Henrique Lee, Zahra Shakeri, Andrew Ng, and 1 others. 2023b. Radiology report expert evaluation (rexval) dataset.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, and 1 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. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

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.



GPT-4 PROMPT

You are a professional radiologist. Your task is: given a ground truth diagnostic report, generate three predicted reports with systematically injected clinical errors to simulate varying levels of report fidelity. These predicted reports should be rated based on the following error-counting rules. Specifically, the predicted reports should fall into three fidelity levels:

1. High-quality: 0–1 errors
2. Medium-quality: 2–3 errors
3. Low-quality: 4 or more errors

Please control generated report quality by intentionally injecting semantic-level clinical errors, including:

1) false_prediction (False report of a finding in the predicted report), "
2) omission (Missing a finding present in the ground truth), "
3) incorrect_location (Misidentification of a finding's anatomic location/position), "
4) incorrect_severity (Misassessment of the severity of a finding), "
5) incorrect_comparison (Mentioning a comparison that isn't in the ground truth), "
6) omission_of_comparison (Omitting a comparison detailing a change from a prior study). "

Please generate three predicted reports for the given ground truth report between <ground_truth_report><\ground_truth_report>. Then, for each predicted report, list its errors by category.

<ground_truth_report>YOUR GROUND TRUTH REPORT<\ground_truth_report>

Figure 4: GPT-4 Prompt Example.

A Appendix

A.1 GPT-4 Prompt Template

The prompt in Figure 4 was used in GPT-4 to generate the training data.