Beyond Scalar Scores: Exploring LLM-based Metrics for Clinical Significance Evaluation in Radiology Reports

Anonymous ACL submission

Abstract

Automated evaluation of radiology reports is essential for clinical trust, yet clinically significant errors (e.g., missed findings or incorrect descriptions) can critically mislead diagnosis. Traditional metrics fail to distinguish between clinically meaningful and harmless textual variations, collapsing report quality into a single scalar score. Although Large Language Models (LLMs) embody richer medical knowledge, such metrics still lack an explicit boundary of clinical significance. To bridge this gap, we adopt the **ReEvalMed** benchmark as testbed and evaluate metric-level clinical significance from detecting true clinical errors ("*Discrimination*") and tolerating insignificant variations ("*Robustness*"). We design an evaluation framework with one-pass and two-pass settings, and compare 8 LLM evaluators against score-based baselines. Results reveal a *discrimination bias*: LLMs readily detect true clinical errors but struggle to tolerate harmless variations. To mitigate this, we synthesize 4k report pairs and train lightweight interpretable metrics on Qwen3-8B and MedGemma-4B. Our trained metric sharpens the clinical significance boundary, surpassing 32B-scale medical LLMs and remaining competitive with proprietary models. Crucially, the more costly two-pass setting does not consistently improve overall performance; instead, it mainly redistributes errors between discrimination and robustness. These findings suggest one-pass trained metrics as the practical choice for cost-sensitive deployment, with two-pass inference reserved for settings where D–R balance is critical. We will release the dataset and metric to advance future research.

1 Introduction

Radiology reports document imaging findings (e.g., lesion characteristics, anatomical abnormalities, and diagnostic impressions) that directly guide clinical decisions on diagnosis and treatment (Tanno et al., 2025). Vision-language models can generate such reports automatically from medical images (Bannur et al., 2024; Chen et al., 2024), yet reliable and clinically-aligned evaluation remains an open challenge. A clinician-trusted evaluation metric must both identify errors that could significantly affect clinical decisions and remain robust to stylistic variation and clinically insignificant differences. However, as recent study Li et al. (2025a) demonstrate, both traditional lexical metrics and LLM-based medical report metrics that correlate strongly with human judgement (e.g. GREEN (Ostmeier et al., 2024) and RaTEScore (Zhao et al., 2024)) lack the sensitivity to differentiate significant clinical errors from minor ones; such boundary blurring calls their validity as reliable indicators of clinical acceptability into question.

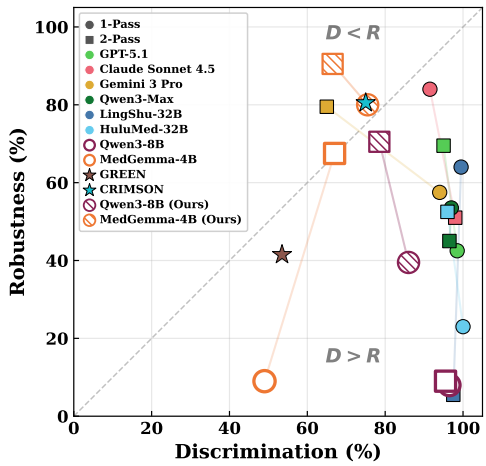

Figure 1: Discrimination–Robustness accuracy of 8 LLM-as-evaluators, 2 medical metrics, and our post-trained models on ReEvalMed. *Discrimination bias*: most LLMs fall below the diagonal ("D > R").

LLMs offer a path toward the interpretable evaluation that previous metrics lack: rather than a scalar, an LLM can identify error spans, classify clinical error aspects, and assess clinical significance, thus providing fine-grained feedback. However, empirical validation of LLMs as evaluators for ra-

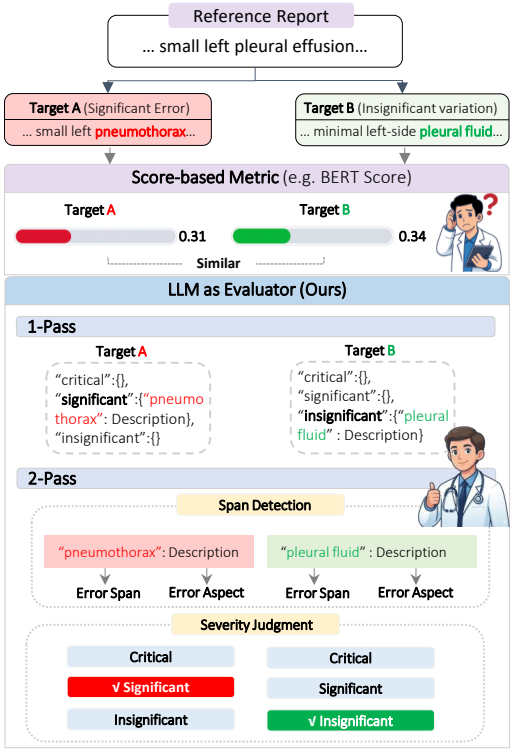

Figure 2: Score-based metrics (e.g. BERTScore) assign comparable scores to clinically significant errors and harmless rephrasings. Our LLM-as-evaluator outputs structured severity judgments, better distinguishing significant errors from insignificant variations.

diology reports remains limited, and it is unclear whether this interpretive capability translates into reliable clinical judgment. To this end, this work investigates the potential of current LLMs as report evaluation metrics from two angles: the zero-shot behaviour of existing open-source and proprietary LLMs, and whether targeted data augmentation and fine-tuning can bridge the performance gap.

Motivated by recent LLM-based evaluators for generated texts (Lu et al., 2024; Kocmi and Federmann, 2023), we design one-pass and two-pass prompts for radiology evaluation (Figure 2), where the former directly outputs structured error annotations and the latter separates ① error span detection from ② clinical significance judgment. As shown in Figure 1, we observe a consistent ***discrimination bias***: *current LLM evaluators struggle to distinguish clinically significant discrepancies from harmless report variations*, resulting in high discrimination but low robustness accuracy.

To mitigate such discrimination bias and better delineate the boundary of clinical significance, we synthesize a balanced dataset containing 2k discrimination and 2k robustness reports across 20 error types using Claude Sonnet, with clinician verification to ensure data quality. To further validate the effectiveness of our synthesized dataset, we train a radiology report evaluation metric based on Qwen3-8B and MedGemma-4B. Our trained metric significantly improve the clinical significance boundary, outperforming 32B-scale medical LLMs and remaining competitive with proprietary LLMs. Another interesting finding is that two-pass inference settings does not eliminate errors, but redistributes them between discrimination and robustness. Our contributions are three-fold:

- **Analysis of LLM-based Evaluators.** We systematically evaluate 11 score-based metrics and 8 LLM-based metrics on ReEvalMed using one-pass and two-pass LLM prompting strategies, analyze the error pattern and revealing a consistent discrimination bias among most LLM metrics.
- **Clinically grounded data synthesis.** We synthesize a balanced dataset of 4k radiology report pairs covering all 20 error categories mentioned in ReEvalMed, annotated at the span level and verified for clinical validity.
- **Lightweight interpretable metric.** Building on the synthesized data, we train Qwen3-8B via supervised fine-tuning (SFT) and reinforcement learning (RL) techniques using both one-pass and two-pass prompting format, achieving 78.5% discrimination and 70.5% robustness accuracy, surpassing medical LLMs at the 32B scale (such as Lingshu-32B (Xu et al., 2025b) and Hulu-Med-32B (Jiang et al., 2025a)).

2 Preliminary

Task formulation A radiology report evaluation metric takes as input a reference report ("**REF**") and a generated candidate ("**TGT**"), and outputs a quality judgment. Traditional metrics, including lexical metrics (BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004)), embedding-based metrics (BERTScore (Zhang et al., 2019)), and clinical NLP metrics (RadGraph (Jain et al., 2021), CheXbert (Smit et al., 2020)), all reduce this judgment to a single **continuous score**. However, a scalar score cannot distinguish clinically significant errors from insignificant ones, as both may incur the same penalty under any scalar metric. In this work, we instead adopt a **structured textual output** that explicitly labels each discrepancy as *significant* or *insignificant*, and identifies its error

Criteria	Error Type	Error Aspect	#
Discrimination	Omission, Fabrication, Inaccuracy	Location, Severity, Description, Comparison/Progression	120
	Inaccuracy Only	Negation, Modality, Size/Distance, Contradiction, Uncertainty, Terminology, Noise, Stylistic Variation	80
Robustness	Omission, Fabrication, Inaccuracy	Location, Severity, Description, Comparison/Progression	120
	Inaccuracy Only	Negation, Modality, Size/Distance, Contradiction, Uncertainty, Terminology, Noise, Stylistic Variation	80

Table 1: ReEvalMed error taxonomy and test set composition. The benchmark defines 20 error categories across two evaluation dimensions, *Discrimination* (detecting clinically significant errors) and *Robustness* (tolerating clinically insignificant variations), each with 200 samples (400 total).

span and clinical aspect, providing fine-grained attribution beyond what a scalar can express.

ReEvalMed benchmark To assess how well a metric aligns with clinical judgment, ReEvalMed (Li et al., 2025a) provides a meta-evaluation benchmark: given report pairs with known clinical labels, it measures whether a metric can correctly identify clinically meaningful discrepancies. Its evaluation criteria are co-developed with clinicians to capture what clinicians actually consider significant in a radiology report. ReEvalMed organises its 20 error categories along two axes, *Error Type* (Omission, Fabrication, Inaccuracy) and *Error Aspect* (e.g., Location, Severity, Negation), as summarised in Table 1. The test set contains 400 report pairs drawn from MIMIC-CXR (Johnson et al., 2019).

Discrimination and Robustness ReEvalMed introduce two metrics along two complementary dimensions to measure the quality of evaluation:

- **Discrimination** (200 test samples): the ability to detect errors that could materially affect diagnosis or treatment (e.g., incorrect severity grade, missed finding).
- **Robustness** (200 test samples): the ability to remain unpenalised by clinically harmless variations (e.g., synonym substitution, phrasing paraphrase).

A clinically significance-sensitive metric should achieve high accuracy on both dimensions simultaneously, i.e., correctly flagging clinically significant errors while tolerating harmless variations.

3 Analysis on LLM-based Metrics

3.1 LLM-as-Evaluator Methodology

Motivated by recent LLM-based evaluators for generated texts (Lu et al., 2024; Kocmi and Federmann, 2023), we design structured prompts that instruct an LLM to compare a REF and TGT pair and output fine-grained error annotations rather than a scalar score. Specifically, the evaluator identifies error spans and assigns each a severity level from three classes: `Critical` (internal contradictions that severely undermine report trust), `Significant` (errors that meaningfully alter clinical decision-making), and `Insignificant` (stylistic variations or clinically harmless deviations). We explore two inference paradigms for this task.

One-pass The LLM receives REF and TGT in a single prompt and directly outputs a JSON object containing three severity buckets (`critical`, `significant`, and `insignificant`), each mapping error spans to their error aspects, together with a free-text explanation.

Two-pass One-pass inference couples span detection and severity judgment into a single output, which may cause compounding errors. To address this, we decouple inference into two passes:

- **Pass 1 (Span Detection)**: the LLM identifies all discrepancies between REF and TGT, outputting a JSON array of error spans with their aspects (e.g., `Location - Omission`, `Negation`).
- **Pass 2 (Severity Judgment)**: for each detected span, a second call outputs exactly one word

Type	Metric	Maximin Threshold			
		D (↑)	R (↑)	Avg (↑)	Gap (↓)
NLP	BLEU	19.0	19.0	19.0 (±4.0)	0.0
	BERTScore	24.0	24.0	24.0 (±4.2)	0.0
	AlignScore	51.0	51.0	51.0 (±5.0)	0.0
Med.	RadGraph	28.5	28.5	28.5 (±4.5)	0.0
	RadBERTScore	26.5	26.5	26.5 (±4.2)	0.0
	RaTEScore	35.5	35.5	35.5 (±4.8)	0.0
	CheXbert	47.0	49.0	48.0 (±5.0)	2.0
LLM	GREEN	53.5	41.5	47.5 (±4.8)	12.0
	RadFact	80.0	56.0	68.0 (±4.5)	24.0
	FineRadScore	**86.5**	68.5	77.5 (±4.0)	18.0
	CRIMSON	75.0	**80.5**	**77.8** (±4.2)	5.5

Table 2: Comparison of score-based metrics on ReEvalMed. D and R scores are in accuracy (%). Maximin threshold is applied to obtain balanced results. Best results are in **bold**. Numbers in parentheses denote 95% bootstrap CIs for Avg.

(`Critical`, `Significant`, or `Insignificant`), conditioned on aspect-specific criteria.

D/R classification rule Note that the three severity levels are *error-level* labels assigned to individual spans, whereas D and R are *metric-level* scores that measure how well a metric's predictions align with clinical ground truth. To bridge the two levels, we aggregate span-level severities into a binary report-level prediction. Denoting a TGT–REF pair as (t, r), let n_c, n_s, n_i be the number of `Critical`, `Significant`, and `Insignificant` spans identified:

$$\text{CLS}(t, r) = \begin{cases} \textit{sig.} & n_c + n_s > 0, \\ \textit{ins.} & n_c = n_s = 0,\ n_i > 0. \end{cases} \quad (1)$$

The **Discrimination score** (D) is the accuracy of predicting *sig.* on the Discrimination subset, where all pairs contain clinically significant errors; the **Robustness score** (R) is the accuracy of predicting *insig.* on the Robustness subset, where all pairs contain only clinically insignificant variations.

3.2 Experimental Setup

Baselines We evaluate 11 score-based metrics that output continuous scores: 3 NLP metrics (BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2019), and AlignScore (Zha et al., 2023)); 4 medical metrics (RadGraph (Jain et al., 2021), RaTEScore (Zhao et al., 2024), CheXbert (Smit et al., 2020)), and RadBERTScore, which replaces the generic BERTScore encoder with a radiology-domain encoder from RadEval (Xu et al., 2025a); and 4 LLM-based metrics (GREEN (Ostmeier et al., 2024), which prompts GPT-4 for finding-level error annotations, and CRIMSON (Baharoon et al., 2026), which fine-tunes MedGemma on 140K report pairs with GPT-5-generated severity labels via LoRA. FineRadScore (Huang et al., 2024) is an LLM-based line-by-line correction metric; RadFact (Bannur et al., 2024) is an LLM-based entailment metric suite for radiology report evaluation, and we use its logical precision/recall scores. We instantiate RadFact and FineRadScore with GPT-5.1 for our experiments.

LLM-as-evaluators We evaluate 3 proprietary LLMs (GPT-5.1, Claude Sonnet 4.5, Gemini 3 Pro) and 5 open-source LLMs (Qwen3-Max (Yang et al., 2025), LingShu-32B (Xu et al., 2025b), Hulu-Med-32B (Jiang et al., 2025a), Qwen3-8B (Yang et al., 2025), and MedGemma-4B (Sellergren et al., 2025)) using our one-pass and two-pass prompts.

Prompt Prompt templates for both one-pass and two-pass paradigms are provided in Appendix I.

Hyperparameters All LLM model inferences use greedy decoding (temperature = 0.0) with a maximum of 1024 output tokens.

Device and platform Proprietary models are accessed via their official APIs. Open-source models (≤32B) are served with vLLM (Kwon et al., 2023) on a single NVIDIA A800 GPU. The same device is used for all training experiments in Section 5.

3.3 Meta-Evaluation

Score-based metrics To align the accuracy scale with LLM-based metrics, score-based metrics require a decision threshold to convert continuous values into binary D/R predictions. We sweep all thresholds and apply the **maximin criterion** $\max_\theta \min(D_\theta, R_\theta)$, selecting the operating point that maximises the worse of the two dimensions[1].

Average and Gap We report **Avg** $= (\text{D} + \text{R})/2$ for overall accuracy and **Gap** $= |\text{D} - \text{R}|$ for D–R imbalance. An ideal evaluator achieves high Avg with low Gap. We report the 95% confidence interval (CI) of Avg using bootstrap resampling over test cases. Detailed results in Appendix B.3.

[1]For FineRadScore (Huang et al., 2024) with severity-level labels, the selected threshold maps severity levels < 2 to insignificant errors and levels ≥ 2 to significant errors.

Type	Model	1-Pass				2-Pass			
		D (↑)	R (↑)	Avg (↑)	Gap (↓)	D (↑)	R (↑)	Avg (↑)	Gap (↓)
Proprietary	GPT-5.1	98.5	42.5	70.5 (±3.5)	56.0	95.0	69.5	**82.3** (±3.5)	25.5
	Claude Sonnet 4.5	91.5	**84.0**	**87.8** (±3.2)	**7.5**	**98.0**	51.0	74.5 (±3.5)	47.0
	Gemini 3 Pro	94.0	57.5	75.8 (±3.8)	36.5	65.0	**79.5**	72.3 (±4.2)	14.5
Open-source	Qwen3-Max	97.0	53.5	75.3 (±3.8)	43.5	96.5	45.0	70.8 (±3.8)	51.5
	LingShu-32B	99.5	64.0	81.8 (±3.3)	35.5	97.5	5.5	51.5 (±2.0)	92.0
	Hulu-Med-32B	**100.0**	23.0	61.5 (±3.0)	77.0	96.0	52.5	74.3 (±3.8)	43.5
	Qwen3-8B	96.5	8.0	52.3 (±2.3)	88.5	95.5	9.0	52.3 (±2.5)	86.5
	MedGemma-4B	49.0	9.0	29.0 (±4.0)	40.0	67.0	67.5	67.3 (±4.8)	**0.5**

Table 3: Comparison of LLM-as-evaluator on ReEvalMed benchmark. D and R scores are in accuracy (%). Numbers in parentheses denote the half-width of the 95% bootstrap CI for Avg. Best results are in **bold**.

3.4 Results

Score-based metrics fail to distinguish significance Score-based metrics require a decision threshold to convert continuous scores into binary significant/insignificant predictions. We select the maximin threshold $\max_\theta \min(D_\theta, R_\theta)$ that maximises the worse of the two dimensions, balancing D and R accuracy (see Appendix G, Figure 7 for full trade-off curves). As shown in Table 2 (left), most non-LLM metrics remain below 55% on both D and R (e.g. BLEU 19.0/19.0, CheXbert 47.0/49.0), indicating they cannot reliably separate significant errors from harmless variations. LLM-based score metrics perform notably better: CRIMSON achieves the best maximin point (75.0/80.5), substantially outperforming GREEN (53.5/41.5) and all non-LLM baselines.

LLM-as-Evaluators are sensitive but poorly calibrated Compared with score-based metrics, LLM-as-Evaluators achieve substantially higher Discrimination: as shown in Table 3, most LLMs reach D $\geq$ 94% under 1-Pass, far exceeding the best score-based result (CRIMSON, 75.0%). However, Robustness varies drastically: R ranges from 8.0% (Qwen3-8B) to 84.0% (Claude Sonnet 4.5), with most models below 60%. Claude Sonnet 4.5 is the only model to achieve strong performance on both dimensions simultaneously (D=91.5%, R=84.0%), suggesting an internal calibration absent from all other tested LLM-based metrics.

1-Pass vs. 2-Pass Switching from 1-Pass to 2-Pass does not uniformly improve calibration; instead, it shifts the bias pattern in model-dependent ways. Some models see their Gap *reduce*: GPT-5.1 narrows from 56.0 to 25.5, and Hulu-Med-32B from 77.0 to 43.5. Others see Gap *amplify*: Claude Sonnet 4.5 increases from 7.5 to 47.0, LingShu-32B from 35.5 to 92.0, and Gemini 3 Pro from 36.5 to 14.5. These results suggest that decoupling span detection from severity judgment redistributes errors between the two passes rather than eliminating them, and the net effect depends on the model's internal calibration.

Discrimination bias Another consistent pattern emerges across both paradigms: nearly all models exhibit large Gap driven by D scores substantially exceeding R, indicating a systematic tendency to classify clinically harmless variations as significant errors. We term this discrimination bias. Without explicit supervision on the significance boundary, models lack a clear understanding of what constitutes a clinically meaningful error and tend to err on the side of caution by flagging variations as significant. Decoupling span detection from severity judgment (2-Pass) partially redistributes this bias, motivating further exploration through data augmentation and fine-tuning in Sections 4–5.

4 Report Synthesis via Clinical Error Injection

The analysis in Section 3 reveals that discrimination bias is pervasive across LLM evaluators, yet no large-scale training resource exists that explicitly supervises the boundary between clinically significant errors and harmless variations. To bridge this gap, we construct 4,000 REF–TGT medical report pairs with specific error annotations covering all 20 error categories with balanced D/R supervision, designed both to mitigate the bias when training lightweight metrics and to serve as a reusable resource, as shown in Figure 3 (left).

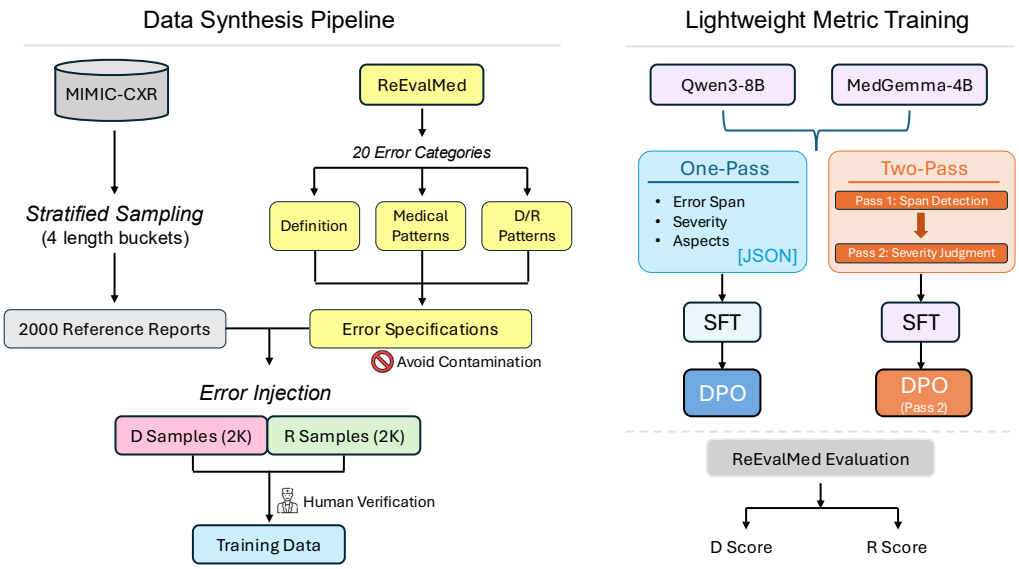

Figure 3: Overview of our data synthesis and training pipeline. **Left:** Reference reports are stratified-sampled from MIMIC-CXR and paired with error specifications for controlled error injection. **Right:** Base models are fine-tuned through SFT and RL under both One-Pass and Two-Pass evaluation paradigms.

4.1 Stratified Sampling

Different error categories naturally occur in reports of varying lengths, e.g. *Stylistic Variation* errors require longer, multi-sentence reports, while *Severity Omission* errors can appear even in short single-finding reports. Inspired by stratified sampling practices in machine translation evaluation (Saldías Fuentes et al., 2022), we sample 2,000 reference reports from the approximately 227K reports in MIMIC-CXR (Johnson et al., 2019) across four character-length buckets (0–200, 200–350, 350–500, >500 characters), combined with category-specific keyword filters derived from the error specifications to guarantee lexical relevance. The resulting length distribution across all 20 error categories is visualised in Appendix E.

4.2 Error Specifications

A key challenge is defining when a textual difference is clinically significant versus harmless. We address this by prompting Claude Sonnet 4.5 to analyse the 20 ReEvalMed test samples per category and distil structured *error specifications*, each containing: (1) a precise clinical error definition, (2) common medical manifestation patterns in CXR reports, (3) discrimination text patterns showing changes that alter clinical meaning (e.g. "left upper lobe nodule" → "pulmonary nodule"), and (4) robustness text patterns showing clinically harmless reformulations (e.g. "hepatic lesion in segment 7" → "right hepatic lobe lesion"). To avoid potential test-set leakage, we deliberately prevent the model from directly reusing ReEvalMed samples during generation; instead, the specifications extract category-level patterns and encourage diverse, novel realisations beyond the original examples. We present the error injection prompt (Appendix I) and representative examples in Appendix H.

4.3 Error Injection and Report Synthesis

Each reference is assigned exactly one Discrimination injection (clinically significant error) and one Robustness injection (clinically harmless variation).[2] Across 20 categories with 100 D + 100 R samples each, the corpus contains **4,000 (REF, TGT) pairs** in total. Claude Sonnet 4.5 generates each TGT conditioned on the error specification and sampled REF; each sample is further annotated with span-level labels (error spans, type, severity) for use in two-pass training.

4.4 Quality Control via Automatic Sanity Checks and Human Evaluation

To guarantee quality, we add automatic sanity checks (e.g. non-empty output, reasonable length ratio) to detect obviously invalid generations and trigger immediate re-generation. In addition, a qualified clinician manually verified 400 randomly

[2] We deliberately limit each reference to a single D and R pair rather than generating multiple variants from the same source, ensuring distributional diversity and preventing the model from overfitting to repeated reference patterns.

sampled pairs (200 D + 200 R) from the 4k synthesized samples to assess clinical validity and label quality.

Although **92.5%** of the samples passed clinician verification, the reviewer flagged 28 out of 400 sampled pairs. After careful inspection, we find the improper generations can be generally summarised into three error patterns: severity mislabelling (the assigned severity disagrees with clinical judgement), injection failure (error injection produces no meaningful change), and aspect mismatch (the injected change is purely stylistic, deviating from the assigned error type). Detailed cases of the representative flagged samples are shown in Appendix F, which we hope will serve as a practical reference when constructing clinically-grounded datasets. We will also release the complete clinician review report and detailed statistical analysis to support transparency and reproducibility.

Model	Setting	D (↑)	R (↑)	Avg (↑)	Gap (↓)
One-Pass					
Qwen3	Base	96.5	8.0	52.3 (±2.3)	88.5
	+ SFT	90.0	19.0	54.5 (±3.5)	71.0
	+ SFT + RL	86.0	39.5	62.8 (±4.2)	46.5
MedGemma	Base	49.0	9.0	29.0 (±4.0)	40.0
	+ SFT	60.5	93.0	76.8 (±4.0)	32.5
	+ SFT + RL	75.5	80.0	**77.8** (±4.2)	4.5
Two-Pass					
Qwen3	Base	95.5	9.0	52.3 (±2.5)	86.5
	+ SFT	84.0	52.0	68.0 (±4.3)	32.0
	+ SFT + RL	78.5	70.5	74.5 (±4.2)	8.0
MedGemma	Base	67.0	67.5	67.3 (±4.8)	**0.5**
	+ SFT	67.0	84.0	75.5 (±4.2)	17.0
	+ SFT + RL	66.5	90.5	**78.5** (±4.0)	24.0

Table 4: Training results on ReEvalMed based on Qwen3-8B (Yang et al., 2025) and MedGemma-4B (Sellergren et al., 2025). D and R scores are in accuracy (%). Best results are in **bold**.

5 Lightweight Metric Training

5.1 Training Stages

To validate the effectiveness of our synthesized 4K corpus in improving evaluation accuracy and reducing discrimination bias, we train a lightweight radiology report metric and evaluate it on ReEvalMed. The 4,000-sample corpus is 1:1 D:R balanced, which prevents model collapse into degenerate always-positive or always-negative prediction. As shown in Figure 3 (right), we adopt a two-stage training pipeline applied to both one-pass and two-pass paradigms: supervised fine-tuning (SFT) to teach the target output format, followed by an RL algorithm to calibrate the significance boundary.

Base models We select two lightweight LLMs of different scales: **Qwen3-8B** (Yang et al., 2025), a general-purpose 8B instruction-tuned LLM, and **MedGemma-4B** (Sellergren et al., 2025), a 4B model pre-trained on biomedical corpora. This allows us to examine whether domain-specific pre-training complements our clinically grounded data.

SFT The SFT stage trains the model to produce the target output format and acquire basic error detection ability. For one-pass, the model learns to produce a structured JSON object containing severity buckets and error spans from a single REF+TGT input. For two-pass, we construct separate training samples for Pass 1 and Pass 2, then fine-tune the same base model on both parts so that it masters error span extraction (Pass 1: a JSON array of error spans with clinical aspects) and severity classification (Pass 2: labelling each span as Critical, Significant, or Insignificant).

RL SFT learns the output format and may improve performance, but does not calibrate the significance boundary. We apply DPO (Rafailov et al., 2023) to address this: preference pairs are constructed from the same 4K corpus by swapping the ground-truth severity label (e.g., Significant → Insignificant) to create *rejected* responses, while the original label serves as the *chosen* response. Note that for two-pass, since the model already captures error spans well after SFT, we only apply DPO to the Pass 2 severity classification.

Training framework All models are trained using LoRA (Hu et al., 2022) via LLaMA-Factory (Zheng et al., 2024). Detailed hyperparameters are provided in Appendix A.

5.2 Results

Table 4 summarises the training results.

Best results After post-training, MedGemma-4B under 1-Pass achieves the best D–R alignment (D=75.5%, R=80.0%, Gap=4.5), while Qwen3-8B under 2-Pass reaches D=78.5%, R=70.5% (Gap=8.0). Both surpass 32B-scale medical LLMs such as LingShu-32B and Hulu-Med-32B (Table 3), and approach Claude Sonnet 4.5 (D=91.5%, R=84.0%) at a fraction of the inference cost.

SFT and RL serve complementary roles SFT teaches models to produce structured outputs but does not calibrate the significance boundary: Qwen3-8B remains heavily D-biased after SFT (1-Pass Gap: 88.5→71.0; 2-Pass Gap: 86.5→32.0), while MedGemma-4B overcorrects toward R-bias under 1-Pass (D=60.5%, R=93.0%). RL then narrows the Gap from both directions: for Qwen3-8B it boosts R while preserving D (2-Pass Gap: 32.0→8.0); for MedGemma-4B it recovers D from 60.5% to 75.5% while maintaining high R (Gap: 32.5→4.5). The two stages are thus complementary—SFT establishes the output format and RL refines the D–R boundary.

1-Pass vs. 2-Pass depends on the model For Qwen3-8B, 2-Pass substantially outperforms 1-Pass (Gap 8.0 vs. 46.5), as decoupling span detection from severity judgment helps the model avoid conflating *where* an error is with *how serious* it is. For MedGemma-4B, however, 1-Pass yields better alignment (Gap 4.5 vs. 24.0): under 2-Pass, R continues to rise (84.0→90.5%) but D stagnates at 66.5%, amplifying R-bias with each training stage. This suggests the optimal pipeline depends on the base model's inherent calibration characteristics.

Effectiveness of data and training Across both models and paradigms, training on our synthesized data consistently reduces the Gap, demonstrating the effectiveness of targeted data augmentation for D–R alignment. Even with lightweight LoRA training on models as small as 4B, the resulting metrics achieve competitive performance with proprietary LLMs, validating both the quality of our synthesized dataset and the training framework.

6 Related Work

Radiology report generation Paired image–report datasets such as CheXpert Plus (Chambon et al., 2024), PadChest (Bustos et al., 2020), and the IU X-Ray collection distributed through Open-i (Demner-Fushman et al., 2016) have enabled radiology report generation models, including recent multimodal LLMs such as μ^2LLM (Li et al., 2025b). Unlike this generation-focused literature, our work evaluates whether automatic metrics can identify clinically significant errors generated by report generation LLMs while tolerating clinically insignificant report variation.

Radiology report evaluation Lexical metrics (Papineni et al., 2002; Lin, 2004; Banerjee and Lavie, 2005) remain common despite limitations for clinical text. RadEval (Xu et al., 2025a) unifies lexical, contextual, clinical concept-based, and LLM-based radiology text metrics. CREPE (Cho et al., 2025), CheXbert (Smit et al., 2020), and RadGraph (Jain et al., 2021) provide error counts or structured representations but still reduce quality to scalar scores. GREEN (Ostmeier et al., 2024), RGRG (Tanida et al., 2023), and VERT (Bologna et al., 2026) use LLMs as judges, while CLEAR (Jiang et al., 2025b) evaluates CheXpert conditions and fine-grained attributes; none jointly decompose error types and D–R behavior. RaTEScore (Zhao et al., 2024) and MAIRA (Hyland et al., 2023) target factual grounding but remain scalar. ReEvalMed (Li et al., 2025a) introduces D–R evaluation and reveals failures of existing metrics; we build on it with targeted data construction and post-training.

LLM-as-judge Zheng et al. (2023) established the LLM-as-judge paradigm for general NLP tasks. In the medical domain, Singhal et al. (2023) demonstrated expert-level performance on clinical QA. However, LLMs used as report evaluators tend to over-penalise stylistic variations (Liu et al., 2023a), a failure mode our Robustness dimension systematically quantifies.

LLM post-training Reinforcement learning from human feedback (Ouyang et al., 2022) and DPO (Rafailov et al., 2023) are standard alignment methods, while LoRA (Hu et al., 2022) and LLaMA-Factory (Zheng et al., 2024) make SFT–RL post-training efficient and accessible. Recent work further simplifies preference optimisation by removing the reference model (Hong et al., 2024). We use this stack to train a lightweight D–R calibrated clinical report evaluator.

7 Conclusion

We systematically evaluate LLM-based metrics for clinical significance in radiology reports, revealing a consistent *discrimination bias*: LLMs readily detect true clinical errors but struggle to tolerate harmless variations. To mitigate this, we synthesize 4k report pairs and train lightweight metrics via SFT and RL, sharpening the significance boundary to surpass 32B-scale medical LLMs while remaining competitive with proprietary LLMs. We also find two-pass inference redistributes rather than eliminates errors between discrimination and robustness.

Limitations

The limitations of this study can be summarised as follows:

- **Quality control.** The most rigorous validation of synthesized report pairs would require expert clinician annotation, which is prohibitively expensive at scale. As a practical compromise, we randomly sampled 400 generated pairs for manual verification (Section 4), leaving full-scale clinician review for future work.

- **Training algorithm.** The primary goal of this paper is to explore the proposed LLM-based evaluation framework and demonstrate the effectiveness of our synthesized data; metric training serves as a validation vehicle rather than the core contribution. We therefore adopt DPO as a straightforward offline preference optimisation method. Modern reinforcement learning algorithms (e.g. PPO, GRPO) typically require large-scale rollouts that exceed our current data budget; exploring these and other advanced training pipelines is a promising direction for future work.

- **Domain coverage.** Our models are trained and evaluated on MIMIC-CXR (English, US clinical setting) and Open-i for domain generalization; out-of-distribution generalisation to other languages, imaging modalities, or clinical environments remains to be studied.

- **Inference cost.** The two-pass approach doubles inference cost compared to one-pass; in high-throughput settings, one-pass DPO may provide a better efficiency–calibration trade-off.

Ethics Statement

We take ethical considerations seriously and strictly adhere to the ACL Code of Ethics. All data used in this study are derived from publicly available research datasets. For MIMIC-CXR, we complied with the PhysioNet credentialing process and data-use agreement requirements, including completion of the required human-subjects research training and access under the approved data-use terms. No additional protected patient information was collected or used beyond the de-identified data available through the authorized dataset access. A clinician co-author participated in the human verification of synthesized data (Section 4) to ensure clinical validity. Our proposed approaches and metrics do not include statements that induce the model to generate harmful information; the evaluation framework focuses solely on assessing the clinical significance of textual differences in radiology reports. We ensure that the findings and conclusions of this paper are reported accurately and objectively.

References

Mohammed Baharoon, Thibault Heintz, Siavash Raissi, Mahmoud Alabbad, Mona Alhammad, Hassan AlOmaish, Sung Eun Kim, Oishi Banerjee, and Pranav Rajpurkar. 2026. Crimson: A clinically-grounded llm-based metric for generative radiology report evaluation. *arXiv preprint arXiv:2603.06183*.

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*.

Federica Bologna, Jean-Philippe Corbeil, Matthew Wilkens, and Asma Ben Abacha. 2026. Vert: Reliable llm judges for radiology report evaluation. *arXiv preprint arXiv:2604.03376*.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P. Langlotz. 2024. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *arXiv preprint arXiv:2405.19538*.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, and 1 others. 2024. Chexagent: Towards a foundation model for chest x-ray interpretation. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.

Gihun Cho, Seunghyun Jang, Hanbin Ko, Inhyeok Baek, and Chang Min Park. 2025. Crepe: Rapid chest x-ray report evaluation by predicting multi-category error counts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21749–21766.

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

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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 *Proceedings of the 9th Machine Learning for Healthcare Conference*, volume 252 of *Proceedings of Machine Learning Research*. PMLR.

Stephanie L Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, and 1 others. 2023. Maira-1: A specialised large multimodal model for radiology report generation. *arXiv preprint arXiv:2311.13668*.

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*.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, and 1 others. 2025a. Hulumed: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*.

Yuyang Jiang, Chacha Chen, Shengyuan Wang, Feng Li, Zecong Tang, Benjamin M. Mervak, Lydia Chelala, Christopher M Straus, Reve Chahine, Samuel G. Armato Iii, and Chenhao Tan. 2025b. CLEAR: A clinically grounded tabular framework for radiology report evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15914–15933, Suzhou, China. Association for Computational Linguistics.

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*.

Tom Kocmi and Christian Federmann. 2023. Gemba-mqm: Detecting translation quality error spans with gpt-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775.

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

Ruochen Li, Jun Li, Bailiang Jian, Kun Yuan, and Youxiang Zhu. 2025a. Reevalmed: Rethinking medical report evaluation by aligning metrics with real-world clinical judgment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11823–11837.

Siyou Li, Pengyao Qin, Huanan Wu, Dong Nie, Arun J. Thirunavukarasu, Juntao Yu, and Le Zhang. 2025b. μ^2tokenizer: Differentiable multi-scale multi-modal tokenizer for radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 2511–2522.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Qingyu Lu, Liang Ding, Kanjian Zhang, Jinxia Zhang, and Dacheng Tao. 2025. MQM-APE: Toward high-quality error annotation predictors with automatic post-editing in LLM translation evaluators. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5570–5587, Abu Dhabi, UAE. Association for Computational Linguistics.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. Error analysis prompting enables human-like translation evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8801–8816.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S 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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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.

Belén Saldías Fuentes, George Foster, Markus Freitag, and Qijun Tan. 2022. Toward more effective human evaluation for machine translation. In *Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 76–89, Dublin, Ireland. Association for Computational Linguistics.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

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.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7433–7442.

Ryutaro Tanno, David GT Barrett, Andrew Sellergren, Sumedh Ghaisas, Sumanth Dathathri, Abigail See, Johannes Welbl, Charles Lau, Tao Tu, Shekoofeh Azizi, and 1 others. 2025. Collaboration between clinicians and vision–language models in radiology report generation. *Nature Medicine*, 31(2):599–608.

Justin Xu, Xi Zhang, Javid Abderezaei, Julie Bauml, Roger Boodoo, Fatemeh Haghighi, Ali Ganjizadeh, Eric Brattain, Dave Van Veen, Zaiqiao Meng, David W Eyre, and Jean-Benoit Delbrouck. 2025a. RadEval: A framework for radiology text evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 546–557, Suzhou, China. Association for Computational Linguistics.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, and 1 others. 2025b. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348.

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.

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, and Zheyan Luo. 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)*, pages 400–410.

A Model and Training Hyperparameters

Table 5 lists the model configuration and training hyperparameters used in all experiments. We fine-tune two base models (Qwen3-8B and MedGemma-4B) using LoRA (Hu et al., 2022) via LLaMA-Factory (Zheng et al., 2024). SFT and DPO share the same LoRA and optimiser settings; DPO-specific parameters (preference construction) are listed separately.

Hyperparameter	Value
Model	
Base model 1	Qwen/Qwen3-8B (8B)
Base model 2	google/medgemma-4b-it (4B)
LoRA	
LoRA rank	8
LoRA alpha	16
Training	
Learning rate	1e-4
Epochs	3
Scheduler	cosine
Precision	bfloat16
Gradient accumulation	8
Max sequence length	2048
DPO	
β	0.1
Sampling temperature	1.0
Samples per input	2

Table 5: Model and training hyperparameters for all SFT and DPO experiments. LoRA and optimiser settings are shared across stages; DPO-specific parameters govern preference pair construction via rejection sampling.

B Additional Analysis

B.1 Generalization to Open-i

We further constructed a small Open-i benchmark to test evaluator performance on an external chest X-ray dataset. Following the ReEvalMed pipeline, we sampled 50 cases from Open-i and applied controlled data augmentation to create 25 significant-error samples for discrimination and 25 insignificant-error samples for robustness. To avoid potential same-model bias in the generation of evaluation samples, we used GPT-5.5 for this augmentation process; GPT-5.5 is not included among the evaluator models assessed in this paper. All augmented samples were reviewed by clinicians to ensure that the injected errors were correct, clinically meaningful when intended, and free from duplicates or ambiguous cases.

As shown in Table 6, most methods achieve strong results on Open-i. One likely reason is that Open-i reports are generally simple and contain relatively short sentences. This makes both significant-error detection and robustness to minor wording changes easier than in ReEvalMed. However, the results still reveal differences among evaluators. Qwen3-8B shows strong discrimination but weaker robustness, indicating a tendency to over-detect minor changes as significant errors. After SFT and RL, Qwen3-8B achieves better balance. MedGemma-4B performs well on the base model and reaches perfect performance after SFT and RL.

B.2 Error Aspect Analysis

Beyond the overall D–R gap, we examine *which* error aspects are systematically confused across evaluators. Figure 5 visualizes the aspect-level flows for eight LLM evaluators, with one-pass predictions on the left, ground-truth aspects in the middle, and two-pass predictions on the right. A common pattern is that model predictions collapse toward a small set of broad clinical aspects, especially Description, Location, Severity, and Comparison/Progression, while fine-grained categories such as Modality, Noise, Uncertainty, Terminology, and Stylistic Variation are often under-recognized. For example, GPT-5.1 and Claude Sonnet 4.5 both over-predict Description under two-pass inference (60→113 and 60→96, respectively), whereas Qwen3-8B and MedGemma-4B show even stronger concentration, mapping many cases to Description or Location. Qwen3-Max is comparatively better calibrated, but still shifts substantially toward Location in the two-pass setting (60→99).

The Sankey diagrams also show that two-pass inference does not simply "fix" one-pass errors; rather, it redistributes them. In several models, the second pass reduces some over-predicted categories but introduces new failure modes, including large Unclassified flows for Gemini 3 Pro and LingShu-32B, suggesting that span extraction and severity judging can propagate formatting or classification errors. Across models, explicit factual aspects such as Location, Severity, and Negation are more frequently detected, while subtle imaging- or wording-specific aspects such as Modality, Noise, and Terminology remain difficult. This aspect-level evidence supports our central finding that current LLM evaluators tend to over-detect discrepancies but remain under-calibrated in distinguishing clinically meaningful errors from harmless variations.

B.3 Statistical Analysis

Confidence intervals. We report 95% bootstrap confidence intervals (CIs) for all evaluation scores in Table 7, Table 8 and Table 9. For each metric or model, we resample test cases with replacement and recompute Discrimination (D), Robustness (R),

Type	Model	D (↑)	R (↑)	Avg (↑)	Gap (↓)
Baseline	RadFact	**100.0**	**100.0**	**100.0**	**0.0**
	FineRadScore	**100.0**	**100.0**	**100.0**	**0.0**
	CRIMSON	92.0	**100.0**	96.0	8.0
Proprietary	Claude Sonnet 4.5	**100.0**	**100.0**	**100.0**	**0.0**
Open-source	Qwen3-8B	**100.0**	76.0	88.0	24.0
	Qwen3-8B + SFT + RL	**100.0**	96.0	98.0	4.0
	MedGemma-4B	**100.0**	92.0	96.0	8.0
	MedGemma-4B + SFT + RL	**100.0**	**100.0**	**100.0**	**0.0**

Table 6: Comparison of evaluators on the Open-i 50-sample benchmark. D and R scores are in accuracy (%). Avg is the mean of D and R, and Gap is the absolute difference between D and R.

Avg, and Gap over 10,000 bootstrap replicates. We report the half-width of the 95% percentile bootstrap CI in parentheses. All CIs are reported in percentage points.

Across these tables, the 95% bootstrap CIs are generally moderate, with Avg half-widths mostly around 3–5 percentage points. The large gaps between conventional similarity metrics and the best LLM-based evaluators are therefore unlikely to be explained by sampling noise alone. By contrast, several top-performing LLM-based metrics and models have overlapping CIs, so their small numerical differences should be interpreted cautiously rather than as definitive rankings.

Pairwise Significance Test. We further conduct pairwise paired significance tests to compare metrics and evaluators on Avg in Figure 4. For each pair of methods, we compute the per-example correctness difference on the same discrimination and robustness test cases, and test whether the paired Avg difference is significantly different from zero using a paired randomization test. Since all methods are evaluated on the same test instances, this paired test controls for example-level difficulty and is more appropriate than an independent-sample comparison.

The significance matrix shows that broad differences between weak similarity-based metrics and stronger LLM-based evaluators are statistically significant after Holm correction. However, several top-performing methods have no significant differences from each other, including CRIMSON, FineRadScore, and our MedGemma SFT+RL evaluator. This suggests that their small numerical Avg differences should be interpreted cautiously, while the improvements from trained evaluators over their base counterparts are statistically reliable.

C Discussion

C.1 Main Contribution and Practical Value

Contribution. Our contribution goes beyond applying ReEvalMed as a benchmark. While ReEvalMed reports D and R mainly through averaged metric scores, we reformulate clinical-significance evaluation as a classification problem, which *reduces the influence of outlier scores and makes the D–R behavior of LLM-as-judge metrics more explicit*. This redefinition allows the discrimination bias of LLM evaluators to be measured more directly: a reliable evaluator should both identify clinically significant errors and tolerate insignificant variations. A second contribution is showing that *significance-oriented data augmentation can improve this balance*. By synthesizing report pairs with controlled significance levels and applying a standard SFT–DPO post-training pipeline, we can calibrate lightweight LLM evaluators on the D–R boundary without designing a complex task-specific training algorithm.

C.2 Same-Model Bias on Claude

Same-model bias is a potential concern in LLM-as-judge studies (Zheng et al., 2023; Liu et al., 2023b), especially when model-generated outputs are evaluated by the same or closely related model. However, this concern does not apply to our main evaluation setting. All main results, including Tables 3 and 4, are evaluated on ReEvalMed, a clinically annotated benchmark that **is not generated by Claude. Thus, our evaluation does not form a closed loop in which Claude-generated data are judged again by Claude.** Claude is used only to construct supervision signals for data augmentation, which

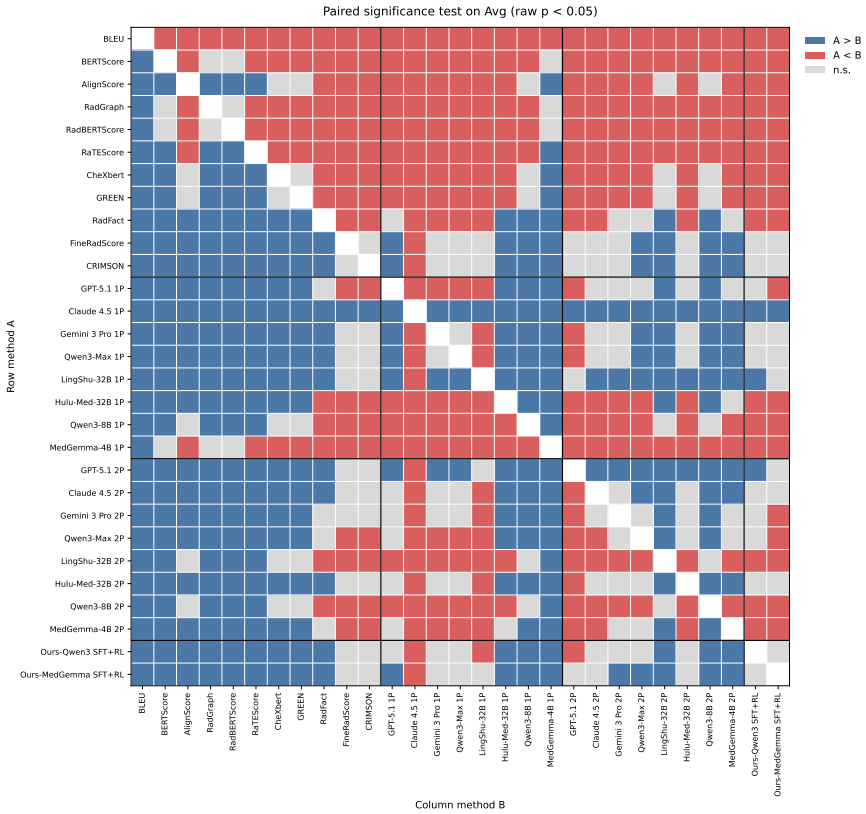

Figure 4: Pairwise paired significance tests on Avg across metrics and evaluators. Each cell compares the row method against the column method using a paired randomization test over matched test cases. Blue indicates that the row method significantly outperforms the column method, red indicates that it performs significantly worse, and gray indicates no significant difference after Holm correction ($p \geq 0.05$).

are then used to post-train lightweight Qwen and MedGemma models, not to optimize Claude itself. Using a stronger model to generate supervision and transfer this capability to smaller models is a common distillation practice, and should not be interpreted as same-model bias in the evaluation.

To further reduce the possibility that model-generated errors are later judged by the same model, we used GPT-5.5 rather than Claude when constructing the Open-i evaluation set. GPT-5.5 is not included among the evaluator models in this paper. This design minimizes potential same-model bias in the external evaluation and further supports that our trained metrics do not merely imitate Claude's judgments, but improve their ability to distinguish clinically significant errors from insignificant variations.

C.3 Potential Data Leakage

Potential data leakage from MIMIC-CXR is unlikely to directly affect the main conclusions of this work. Both the ReEvalMed test set and the samples constructed in our study are based on MIMIC-CXR reports with controlled injected clinical errors. Even if an LLM had been exposed to the original ground-truth MIMIC-CXR reports during pretraining, such exposure would not imply an ability to correctly identify diverse injected errors or determine their clinical significance, which is the central focus of our evaluation.

Moreover, ReEvalMed was released only at the end of 2025, whereas the evaluated LLMs were trained before that release, making direct benchmark leakage unlikely. To further reduce concerns about dataset-specific leakage, we additionally introduce an external Open-i evaluation set. Since Open-i is independent of MIMIC-CXR, this held-out evaluation provides complementary evidence that the observed behavior is not solely driven by memorization of MIMIC-CXR reports.

Although limited domain coverage remains a limitation, as discussed in the Limitation section, the added Open-i analysis helps assess whether the conclusions generalize beyond the primary MIMIC-CXR-derived setting.

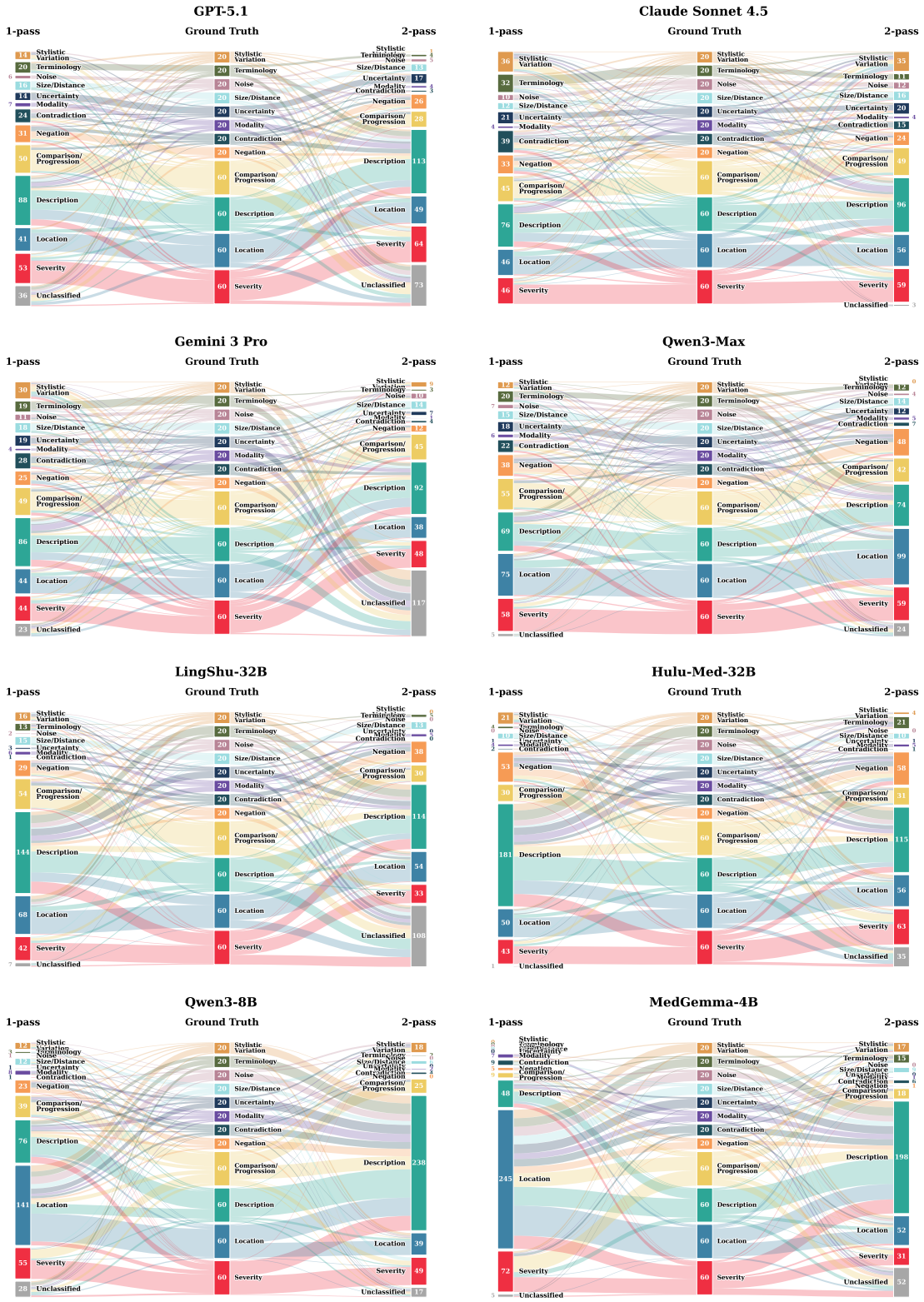

Figure 5: Aspect-flow Sankey diagrams across eight LLM evaluators. Each diagram shows one-pass predicted aspects on the left, ground-truth aspects in the middle, and two-pass predicted aspects on the right. Row totals are displayed on all axes.

Type	Metric	Maximin Threshold			
		D (↑)	R (↑)	Avg (↑)	Gap (↓)
NLP	BLEU	19.0 (±5.5)	19.0 (±5.5)	19.0 (±4.0)	0.0 (±8.0)
	BERTScore	24.0 (±6.0)	24.0 (±6.0)	24.0 (±4.2)	0.0 (±8.5)
	AlignScore	51.0 (±7.0)	51.0 (±7.0)	51.0 (±5.0)	0.0 (±10.5)
Med.	RadGraph	28.5 (±6.5)	28.5 (±6.0)	28.5 (±4.5)	0.0 (±10.0)
	RadBERTScore	26.5 (±6.0)	26.5 (±6.5)	26.5 (±4.2)	0.0 (±10.0)
	RaTEScore	35.5 (±7.0)	35.5 (±7.0)	35.5 (±4.8)	0.0 (±10.0)
	CheXbert	47.0 (±7.0)	49.0 (±7.0)	48.0 (±5.0)	2.0 (±10.0)
LLM	GREEN	53.5 (±7.0)	41.5 (±7.0)	47.5 (±4.8)	12.0 (±10.0)
	RadFact	80.0 (±5.5)	56.0 (±7.0)	68.0 (±4.5)	24.0 (±9.5)
	FineRadScore	**86.5** (±5.0)	68.5 (±6.5)	77.5 (±4.0)	18.0 (±8.0)
	CRIMSON	75.0 (±6.0)	**80.5** (±5.5)	**77.8** (±4.2)	5.5 (±8.0)

Table 7: Comparison of score-based metrics on ReEvalMed. With Numbers in parentheses denote the half-width of 95% bootstrap CIs.

Type	Model	1-Pass				2-Pass			
		D (↑)	R (↑)	Avg (↑)	Gap (↓)	D (↑)	R (↑)	Avg (↑)	Gap (↓)
Proprietary	GPT-5.1	98.5 (±2.0)	42.5 (±7.0)	70.5 (±3.5)	56.0 (±7.0)	95.0 (±3.0)	69.5 (±6.5)	**82.3** (±3.5)	25.5 (±7.0)
	Claude Sonnet 4.5	91.5 (±4.0)	**84.0** (±5.5)	**87.8** (±3.2)	**7.5** (±6.5)	**98.0** (±2.0)	51.0 (±7.0)	74.5 (±3.5)	47.0 (±7.0)
	Gemini 3 Pro	94.0 (±3.5)	57.5 (±6.5)	75.8 (±3.8)	36.5 (±7.5)	65.0 (±6.5)	**79.5** (±6.0)	72.3 (±4.2)	14.5 (±9.0)
Open-source	Qwen3-Max	97.0 (±2.5)	53.5 (±7.0)	75.3 (±3.8)	43.5 (±7.5)	96.5 (±2.5)	45.0 (±7.0)	70.8 (±3.8)	51.5 (±7.5)
	LingShu-32B	99.5 (±1.0)	64.0 (±6.5)	81.8 (±3.3)	35.5 (±6.5)	97.5 (±2.5)	5.5 (±3.5)	51.5 (±2.0)	92.0 (±4.0)
	Hulu-Med-32B	**100.0** (±0.0)	23.0 (±6.0)	61.5 (±3.0)	77.0 (±6.0)	96.0 (±3.0)	52.5 (±7.0)	74.3 (±3.8)	43.5 (±7.5)
	Qwen3-8B	96.5 (±2.5)	8.0 (±4.0)	52.3 (±2.3)	88.5 (±5.0)	95.5 (±3.0)	9.0 (±4.0)	52.3 (±2.5)	86.5 (±5.0)
	MedGemma-4B	49.0 (±7.0)	9.0 (±4.0)	29.0 (±4.0)	40.0 (±8.0)	67.0 (±6.5)	67.5 (±6.5)	67.3 (±4.8)	**0.5** (±10.0)

Table 8: Comparison of LLM-as-evaluator on ReEvalMed benchmark. D and R scores are in accuracy (%). Numbers in parentheses denote the half-width of 95% bootstrap CIs. Best results are in **bold**.

C.4 CRIMSON and MedGemma Fine-tuning

Compared with CRIMSON, our MedGemma-4B + SFT + RL evaluator changes the error profile from a more balanced score-based evaluator to a more robustness-oriented evaluator. CRIMSON achieves stronger discrimination accuracy (75.0 vs. 66.5), whereas our model achieves substantially higher robustness accuracy (90.5 vs. 80.5). As a result, the overall Avg is numerically similar, with our model slightly higher (78.5 vs. 77.8), but the paired significance test indicates that this Avg difference is not statistically significant. This suggests that the main gain of our approach is not a large increase in aggregate Avg over CRIMSON, but a shift toward better preservation of clinically insignificant variations.

The improvement is most visible on robustness cases, where our model correctly accepts more benign perturbations than CRIMSON. In particular, robustness accuracy improves substantially for fabrication-type cases (50.0 to 100.0), severity-related cases (73.3 to 96.7), contradiction cases (30.0 to 80.0), and terminology/style-related cases. This indicates that SFT + RL makes the evaluator less likely to over-penalize clinically harmless edits. However, this comes with a trade-off in discrimination: CRIMSON remains better at detecting significant errors overall, especially for size/distance, terminology, uncertainty, and severity-related significant errors. Thus, our method improves robustness and reduces false alarms on insignificant changes, while CRIMSON remains more sensitive to certain clinically significant discrepancies.

C.5 Critical and Significant Severity Levels

Our evaluation maps both *critical* and *significant* errors to the clinically significant class. This design follows the practical goal of identifying report

Model	Setting	D (↑)	R (↑)	Avg (↑)	Gap (↓)
One-Pass					
Qwen3	Base	96.5 (±3.0)	8.0 (±4.0)	52.3 (±2.3)	88.5 (±4.5)
	+ SFT	90.0 (±4.5)	19.0 (±5.5)	54.5 (±3.5)	71.0 (±7.0)
	+ SFT + RL	86.0 (±5.0)	39.5 (±7.0)	62.8 (±4.2)	46.5 (±8.5)
MedGemma	Base	49.0 (±7.0)	9.0 (±4.0)	29.0 (±4.0)	40.0 (±8.0)
	+ SFT	60.5 (±7.0)	93.0 (±3.5)	76.8 (±4.0)	32.5 (±7.5)
	+ SFT + RL	75.5 (±6.0)	80.0 (±5.5)	**77.8** (±4.2)	4.5 (±8.0)
Two-Pass					
Qwen3	Base	95.5 (±3.0)	9.0 (±4.0)	52.3 (±2.5)	86.5 (±5.0)
	+ SFT	84.0 (±5.0)	52.0 (±7.0)	68.0 (±4.3)	32.0 (±8.5)
	+ SFT + RL	78.5 (±6.0)	70.5 (±6.5)	74.5 (±4.2)	8.0 (±8.5)
MedGemma	Base	67.0 (±6.5)	67.5 (±6.5)	67.3 (±4.8)	**0.5** (±10.0)
	+ SFT	67.0 (±6.5)	84.0 (±5.0)	75.5 (±4.2)	17.0 (±8.5)
	+ SFT + RL	66.5 (±6.5)	90.5 (±4.5)	**78.5** (±4.0)	24.0 (±7.5)

Table 9: Training results on ReEvalMed based on Qwen3-8B (Yang et al., 2025) and MedGemma-4B (Sellergren et al., 2025). D and R scores are in accuracy (%). Numbers in parentheses denote the half-width of 95% bootstrap CIs. Best results are in **bold**.

differences that may affect clinical interpretation or downstream decisions. However, the boundary between critical and significant can itself be subjective and context-dependent. For this reason, we treat the central task as distinguishing clinically meaningful from clinically insignificant differences, while leaving finer-grained severity calibration as an important direction for future work.

C.6 Potential Error Propagation

Our current two-pass design is itself motivated by common practice in machine translation evaluation. As pointed out in related work such as MQM-APE (Lu et al., 2025), in error span detection, completely missing an error span is relatively rare, while producing too many redundant spans is often a more common issue. Based on this observation, pass2 does not handle spans missed by pass1, but it can suppress redundant or unreasonable spans. We will clarify this design motivation and add the corresponding citation more explicitly in the revised version.

C.7 Why Use a Standard Post-training Design

We use a standard SFT followed by DPO training design because our goal is to test whether targeted supervision can sharpen the clinical-significance boundary, rather than to introduce a new optimization algorithm. SFT teaches the model the structured output format and exposes it to controlled clinical-error patterns. DPO then directly optimizes preference pairs that reflect the desired D–R behavior. This simple and reproducible design makes the effect of the data and supervision signal easier to interpret.

The same principle also applies to our prompting design. Our prompts are intended to provide a clear and consistent evaluation protocol for testing whether the proposed clinical-significance formulation is effective, rather than to exhaustively optimize prompt engineering strategies. A systematic exploration of alternative prompts, prompt ensembles, and model-specific prompting designs is therefore left for future work.

D Error Category Descriptions

Table 10 provides detailed descriptions of each of the 20 error categories in ReEvalMed, organised by Error Type and Error Aspect.

E Length Distribution of Synthesized Data

Figure 6 shows the length distribution of our synthesized data compared to the ReEvalMed test set, broken down by error category and stratified across four length buckets. The generated corpus covers all buckets in every category, confirming that stratified sampling produces diverse report lengths.

F Human Validation on Report Synthesis

To verify the clinical fidelity of our synthesized data, we randomly sample 100 pairs (50 D + 50 R) covering all 20 error categories and have a clinician

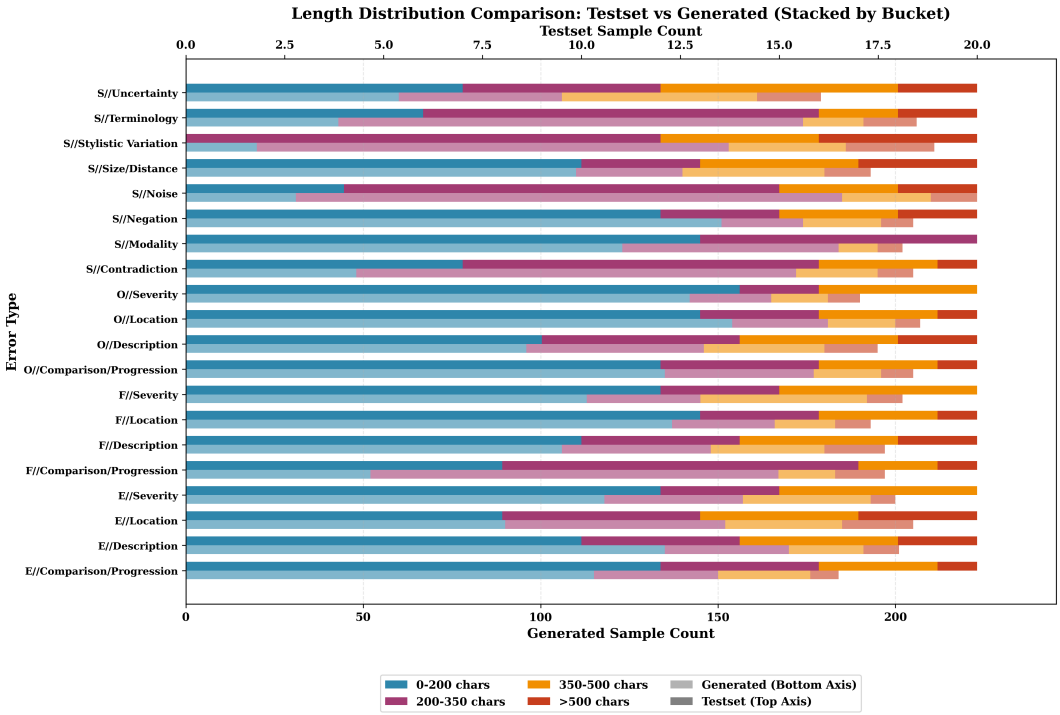

Figure 6: Length distribution comparison between the ReEvalMed test set (top axis) and our generated training corpus (bottom axis), stacked by length bucket across all 20 error categories.

manually review each pair. For each sample, the clinician assesses whether: (1) the injected error is clinically plausible and correctly categorised, (2) the severity label (significant vs. insignificant) is appropriate, and (3) the generated report reads naturally without obvious artefacts.

Case studies Table 11 presents representative examples from the human validation, illustrating both successful and problematic generations.

G Score-based Metric Results

Unlike LLM-based metrics that output binary significant/insignificant predictions, score-based metrics produce continuous scores requiring a decision threshold to convert to D/R accuracy. In the main paper (Table 3), we report results at the **maximin threshold** $\max_\theta \min(D_\theta, R_\theta)$, which selects the operating point that maximises the worse of the two dimensions.

Figure 7 shows the full D–R trade-off curves for all 8 score-based metrics. Each curve traces the (D, R) accuracy pair as the decision threshold varies; the filled circle marks the maximin operating point reported in Table 3. CRIMSON achieves the best maximin point (75.0, 80.5), substantially outperforming all other score-based metrics. NLP metrics (BLEU, BERTScore, AlignScore) cluster in the low-accuracy region, while medical metrics (RadGraph, RaTEScore, CheXbert) offer moderate improvements.

H Error Specification Examples

We provide two representative error specifications generated by Claude Sonnet 4.5 (Section 4). Each specification contains a clinical definition, medical patterns, and contrastive text patterns that distinguish significant errors from harmless variations.

Error Specification: Location – Omission

Definition: Omission of anatomical location information present in the reference report, including missing specific anatomical sites, laterality (left/right), spatial relationships, or regional descriptors.

Medical Patterns:

- Missing laterality specification (left/right, bilateral)
- Missing specific anatomical region or lobe identification
- Missing spatial descriptors (upper/lower, anterior/posterior)

Discrimination Text Patterns (signifi-

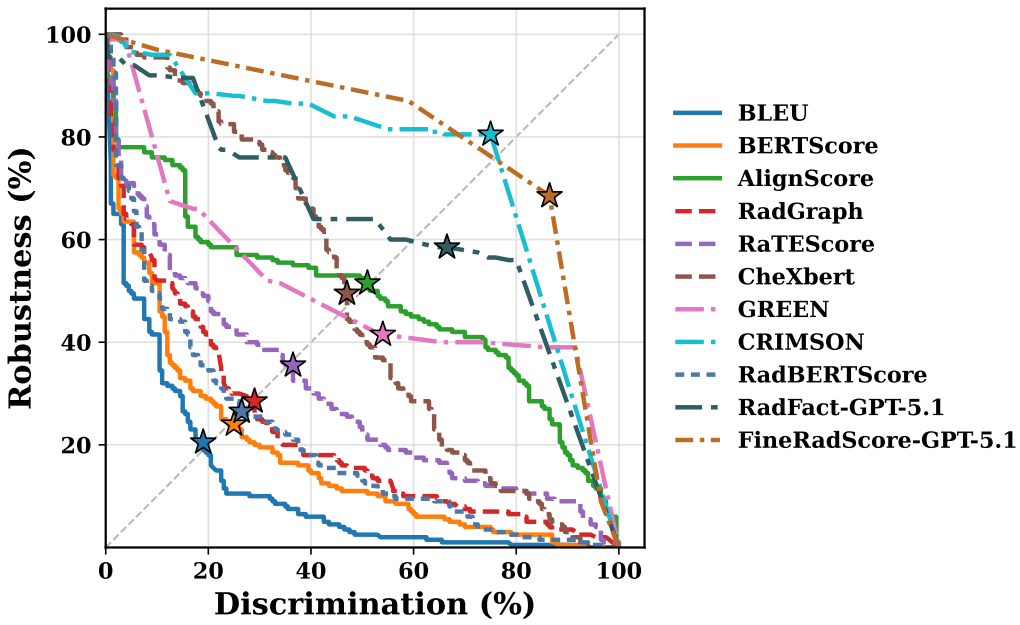

Figure 7: D–R trade-off curves for all score-based metrics. Each curve traces (D, R) accuracy across decision thresholds; filled circles mark the maximin operating point $\max_\theta \min(D_\theta, R_\theta)$ reported in Table 3.

cant):

- "left upper lobe nodule" → "pulmonary nodule" (missing critical laterality and lobe)
- "bilateral pleural effusions" → "pleural effusion" (missing bilateral nature)
- "fracture of L3 vertebral body" → "lumbar fracture" (missing specific vertebral level)

Robustness Text Patterns (insignificant):

- "hepatic lesion in segment 7" → "right hepatic lobe lesion" (alternative valid descriptor)
- "distal esophagus" → "lower esophagus" (synonymous location terms)
- "left lung base" → "left lower lung" (equivalent regional description)

Error Specification: Description – Fabrication

Definition: Adding findings, observations, or anatomical descriptions in TGT that do not exist in REF, including introducing new pathological findings, normal structures not mentioned, or additional descriptive details.

Medical Patterns:

- Adding new pathological findings not present in REF (e.g. pleural effusion, nodules)
- Adding quantitative details or measurements that do not exist in REF
- Introducing temporal or comparative information not in REF

Discrimination Text Patterns (significant):

- REF: "clear lungs" → TGT adds "bilateral pleural effusions"
- REF: "normal heart" → TGT adds "cardiomegaly"
- REF: "opacity present" → TGT adds "large 5cm mass"

Robustness Text Patterns (insignificant):

- REF: "no acute findings" → TGT: "heart, lungs, and mediastinum show no acute findings"
- REF: "unremarkable" → TGT lists "bones intact, soft tissues normal"
- REF gives findings → TGT adds "adequate inspiration" as contextual information

I Prompt Templates

We provide the full prompt templates used for inference (Section 3) and data synthesis (Section 4). In all templates, {REF} and {TGT} are replaced with the ground-truth and target reports, respectively.

One-Pass Evaluation Prompt

SYSTEM:
You are a clinical evaluation agent for chest X-ray reports. You will be given: 1) a ground-truth report written by professional radiologists, and 2) a target report to be evaluated.
Your task is to:

- Identify all clinically significant and insignificant errors in the target report.
- Classify each error into one of the predefined error categories.
- Determine the error type (Omission / Fabrication / Inaccuracy) when applicable.
- Provide a concise explanation of the errors and their classifications.

Follow all definitions and constraints provided in subsequent instructions.

USER:
You must evaluate the target report against the ground-truth report and produce a structured JSON output.
Your output must be a JSON object with the following structure:

- "critical": a dictionary mapping error spans to error aspects for critical errors
- "significant": a dictionary mapping error spans to error aspects for significant errors
- "insignificant": a dictionary mapping error spans to error aspects for insignificant errors
- "explanation": a concise explanation of the errors identified and the rationale for the categorization

Definition of significance:

- **Critical**: Refer to internal inconsistencies or logical contradictions, such as describing the presence of a structure previously stated to be absent, that can severely undermine clinician trust in the report and are considered critical failures.
- **Significant**: Errors that meaningfully alter, mislead, or distort clinical decision-making.
- **Insignificant**: Errors that represent stylistic variation, minor wording changes, or clinically harmless deviations.

Error Aspects:
Below are the 12 error categories and their descriptions:
{error_aspect_info}
Every error must be assigned exactly one of these aspects.
Error Types:
For the "Comparison/Progression", "Description", "Location", "Severity" aspects, you must further specify the error type using one of:
{error_type_info}
For all other error aspects, you must NOT include an error type; only the aspect label should be provided.
Example:

- "rib fractures": "Location - Omission"
- "no pneumothorax": "Negation" (no error type needed)

If no errors are identified, the "significant" and "insignificant" fields should be empty dictionaries.

Reports to evaluate:
Ground-Truth Report: {REF}
Target Report: {TGT}

Please output only a single valid JSON object.
IMPORTANT:

- Do NOT wrap the JSON in Markdown code fences.
- Do NOT include any language labels such as "json" before the JSON.
- The output must start directly with {{ and end with }}.
- Do NOT include any extra text, comments, or explanations outside the JSON.

Two-Pass: Pass 1 (Span Detection)

SYSTEM:
You are a clinical evaluator for chest X-ray reports. Given a ground-truth report (REF) and a target report (TGT), identify all differences.
For each difference, output a JSON object with:

- "span": the exact text span that changed (from TGT for Fabrication/Inaccuracy/Style; from REF for Omission)
- "aspect": one of the 20 error aspect codes

Valid aspect codes: Location - Omission, Location - Fabrication, Location - Inaccuracy, Severity - Omission, Severity - Fabrication, Severity - Inaccuracy, Description - Omission, Description - Fabrication, Description - Inaccuracy, Comparison/Progression - Omission, Comparison/Progression - Fabrication, Comparison/Progression - Inaccuracy, Negation, Modality, Size/Distance, Contradiction, Uncertainty, Terminology, Noise, Stylistic Variation
Output a JSON array. If no differences, output []. Do NOT wrap in markdown. Start directly with [and end with].

USER:
Ground-Truth Report:
{REF}
Target Report:
{TGT}
Identify all differences and output the JSON array of {span, aspect} objects.

Two-Pass: Pass 2 (Severity Judgment)

SYSTEM:
You are a clinical evaluator for chest X-ray reports. You will be given a REF/TGT report pair, a specific error span, its aspect, and judgment criteria. Determine whether the error is SIGNIFICANT or INSIGNIFICANT.
Output a single JSON object: {"severity": "significant"/"insignificant"/"critical", "explanation": "..."}
Do NOT wrap in markdown. Start with { and end with }.

USER:
Ground-Truth Report:
{REF}
Target Report:
{TGT}

Identified Error Span: "{span}"
Error Aspect: {aspect}
Error Type: {etype_label}

Judgment Criteria for [{aspect}]:
{criteria}

Is this error SIGNIFICANT or INSIGNIFICANT? Output only the JSON.

Error Injection Prompt (Data Synthesis)

SYSTEM:
You are a medical imaging report error generation system specialized in radiology reports.

USER:
Error Type Definition:
Error Category: {error_type} ({error_category_name})
Error Aspect: {error_aspect}
{error_definition}

Reference Medical Context:
The following patterns show how this error type manifests in real medical reports:
{medical_patterns}

Severity Level: {severity_description}
Example Text Patterns:
{selected_text_patterns}

Task:
{severity_instruction}
Reference Report: {REF}
Instructions:

- Introduce the specified error type into the reference report
- Follow the text patterns shown above as guidance
- Keep the rest of the report unchanged
- Maintain medical plausibility and report structure

- Preserve the overall report length ($\pm 20\%$)
- Ensure the severity matches the specified level

Output ONLY the modified target report text, without any explanations or formatting.

Type	Aspect	Description	Significant Example	Insignificant Example
O / F / E	Location	Errors in the precise anatomical site of a finding (e.g., laterality, lobe, region).	**REF**: *left-sided rib fractures* **TGT**: *right rib fractures*	**REF**: *left retrocardiac opacity* **TGT**: *opacity behind the heart on the left side*
O / F / E	Severity	Errors in describing the extent or clinical seriousness of a finding.	**REF**: *Heart is mildly enlarged* **TGT**: *Severely enlarged heart*	**REF**: *Severe cardiomegaly* **TGT**: *Moderate-to-severe cardiomegaly*
O / F / E	Description	Errors in morphological characteristics such as shape, margins, or appearance.	**REF**: *irregular mass with spiculated margins* **TGT**: *round, smooth mass*	**REF**: *patchy ill-defined opacities* **TGT**: *faint and poorly marginated opacities*
O / F / E	Comp./Prog.	Errors in describing interval changes compared to prior imaging.	**REF**: *Pulmonary edema has improved* **TGT**: *Pulmonary edema has worsened*	**REF**: *No interval change in pleural effusion* **TGT**: *Pleural effusion is essentially unchanged*
S	Negation	Incorrect presence or absence of a finding.	**REF**: *No evidence of pneumothorax* **TGT**: *Pneumothorax is present*	**REF**: *No pleural effusion is seen* **TGT**: *There is no definite pleural effusion*
S	Modality	Conflicts with the imaging modality.	**REF**: *Refer to prior CT torso for details* **TGT**: *Refer to prior abdominal ultrasound*	**REF**: *Consider chest CT for further evaluation* **TGT**: *CT can be considered for further assessment*
S	Size/Distance	Errors in quantitative measurements.	**REF**: *3-cm mass in the lingula* **TGT**: *8-cm mass in the lingula*	**REF**: *ET tube within 1 cm of the carina* **TGT**: *ET tube within 0.9 cm of the carina*
S	Contradiction	Internal logical inconsistencies within the same report.	**REF**: *The lungs are clear.* **TGT**: *The lungs are clear. There is consolidation in the right base*	**REF**: *No evidence of larger pleural effusions* **TGT**: *No evidence of larger pleural effusions. Minimal effusions may exist*
S	Uncertainty	Incorrect use of hedging terms conveying diagnostic uncertainty.	**REF**: *Whether this is pneumonia is radiographically indeterminate* **TGT**: *Pneumonia exists*	**REF**: *A possible infiltrate is suggested* **TGT**: *An infiltrate is likely present*
S	Terminology	Inaccurate or unclear medical terminology.	**REF**: *A cavitary lesion, suggesting tuberculosis* **TGT**: *A hole, suggesting infection*	**REF**: *A 3-cm mass* **TGT**: *A 3-cm lesion*
S	Noise	Grammatical mistakes, typographical errors, or other linguistic noise.	**REF**: *3-cm mass in the lingula has grown* **TGT**: *3-cm lingula margins has been growing irregularly*	**REF**: *subtle opacity may represent atelectasis* **TGT**: *subtble opaciti may represent atelectasi*
S	Stylistic Var.	Variations in phrasing that do not alter clinical meaning.	**REF**: *Bilateral left greater than right pleural effusion* **TGT**: *Fluid accumulation on both sides of the chest, more on the right*	**REF**: *lung… pulmonary edema… pleural effusions* **TGT**: *pleural effusions… lung… pulmonary edema*

Table 10: Descriptions and examples of the 20 error categories in ReEvalMed (Li et al., 2025a). The first four aspects combine with three error types (Omission (O), Fabrication (F), and Inaccuracy (E)), yielding $4 \times 3 = 12$ categories. The remaining eight aspects apply to Inaccuracy Only (S), yielding 8 categories. Each category appears in both the Discrimination and Robustness test sets. Examples are shown as REF–TGT pairs; significant examples represent clinically meaningful errors, while insignificant examples represent harmless variations.

Case 1 Severity Mislabel	*Significant / Noise*
REF: ... There is no intraperitoneal free air. The lungs are clear without effusion or pneumothorax. ... IMPRESSION: Normal chest. **TGT:** ... Severe motion artifact and noise significantly limit evaluation; intraperitoneal free air cannot be excluded. ... Recommend repeat imaging to exclude free air.	
Error Span: "no intraperitoneal free air" → "cannot be excluded" **Clinician:** Retake imaging only; no serious clinical consequence. Label should be *insignificant*.	
Case 2 Severity Mislabel	*Significant / Noise*
REF: ... Frontal and lateral radiographs of the chest demonstrate normal heart size, mediastinal and hilar contours. Clear lungs. No pneumothorax or pleural effusion. IMPRESSION: Normal chest radiograph. **TGT:** ... Severe motion artifact and excessive noise degrade image quality, limiting diagnostic evaluation. ... Pneumothorax or pleural effusion cannot be excluded. IMPRESSION: Nondiagnostic chest radiograph ... Recommend repeat imaging ...	
Error Span: "normal heart size ... Clear lungs" → "Severe motion artifact ... cannot be reliably assessed" **Clinician:** Same pattern as Case 1; normal study changed to nondiagnostic with repeat imaging recommendation. No actual pathology introduced. Label should be *insignificant*.	
Case 3 Injection Failure	*Significant / Location-Omission*
REF: CHEST ON __ HISTORY: Pneumonitis. REFERENCE EXAM: __. Compared to the prior exam, there is no significant interval change. **TGT:** CHEST ON __ HISTORY: Pneumonitis. REFERENCE EXAM: __. Compared to the prior exam, there is no significant interval change in the pneumonitis.	
Error Span: *null* → "in the pneumonitis" **Clinician:** Appears to critique the REF rather than inject an error into the TGT; the added location specificity is arguably an improvement rather than an omission.	
Case 4 Injection Failure	*Significant / Location-Omission*
REF: HISTORY: Hickman catheter placement. FINDINGS: The catheter tip lies at the level of the mid portion of the SVC. No evidence of acute pneumonia or vascular congestion. **TGT:** *(identical to REF)*	
Error Span: *null* → *null* **Clinician:** Error injection failed entirely; REF and TGT are identical. The model's explanation reveals an aborted attempt: "*This is actually appropriate positioning. Let me generate a proper SEVERE omission error.*"	
Case 5 Aspect Mismatch	*Insignificant / Negation*
REF: ... IMPRESSION: In comparison with the study of __ **TGT:** ... IMPRESSION: Compared with the study of __	
Error Span: "In comparison with" → "Compared with" **Clinician:** Purely stylistic; the model failed to inject the assigned *Negation* error.	
Case 6 Aspect Mismatch	*Insignificant / Contradiction*
REF: ... There is no large pleural effusion or pneumothorax. ... **TGT:** ... No large pleural effusion or pneumothorax is identified. ...	
Error Span: "There is no large pleural effusion or pneumothorax." → "No large pleural effusion or pneumothorax is identified." **Clinician:** Purely stylistic; the model failed to inject the assigned *Contradiction* error.	
Case 7 Aspect Mismatch	*Insignificant / Size-Distance*
REF: ... IMPRESSION: AP and lateral chest compared to __: Normal heart, lungs, hila, mediastinum, and pleural surfaces. **TGT:** ... IMPRESSION: AP and lateral chest compared to __: Normal heart size, lungs, hila, mediastinum, and pleural surfaces.	
Error Span: "Normal heart" → "Normal heart size" **Clinician:** Purely stylistic; the model failed to inject the assigned *Size/Distance* error.	

Table 11: Representative cases from human validation of synthesized data. Cases are categorised by failure type: Severity Mislabel (severity label disagrees with clinical judgement), Injection Failure (error injection produces no meaningful change), and Aspect Mismatch (injected change is purely stylistic, deviating from the assigned error type).