RadEval: A framework for radiology text evaluation

Justin Xu[♠,*] **Xi Zhang**[♢] **Javid Abderezaei**[♡] **Julie Bauml**[♡]

Roger Boodoo[♡] **Fatemeh Haghighi**[♡] **Ali Ganjizadeh**[♡]

Eric Brattain[♡] **Dave Van Veen**[♡] **Zaiqiao Meng**[♢]

David Eyre[♠] **Jean-Benoit Delbrouck**[♡,*]

[♠]**University of Oxford** [♢]**University of Glasgow** [♡]**HOPPR**

justin.xu@ndm.ox.ac.uk, jeanbenoit.delbrouck@hoppr.ai

https://github.com/jbdel/RadEval

🤗 https://huggingface.co/IAMJB/RadEvalModernBERT

Abstract

We introduce RadEval, a unified, open-source framework for evaluating radiology texts. RadEval consolidates a diverse range of metrics – from classic n-gram overlap (BLEU, ROUGE) and contextual measures (BERTScore) to clinical concept-based scores (F1CheXbert, F1RadGraph, RaTEScore, SRR-BERT, TemporalEntityF1) and advanced LLM-based evaluators (GREEN). We refine and standardize implementations, extend GREEN to support multiple imaging modalities with a more lightweight model, and pretrain a domain-specific radiology encoder – demonstrating strong zero-shot retrieval performance. We also release a richly annotated expert dataset with over 450 clinically significant error labels and show how different metrics correlate with radiologist judgment. Finally, RadEval provides statistical testing tools and baseline model evaluations across multiple publicly available datasets, facilitating reproducibility and robust benchmarking in radiology report generation.

1 Introduction

Evaluating automated radiology report generation (RRG) systems remains a fundamental challenge in the development of safe, accurate, and clinically useful medical AI. Unlike general-purpose text generation tasks, RRG demands evaluation methods that can assess not only linguistic fluency but also clinical factuality, domain-specific terminology, uncertainty calibration, and diagnostic relevance. In recent years, the evaluation of radiology report generation has steadily progressed: initial studies relied on classic natural language generation (NLG) metrics such as BLEU and ROUGE (Zhang et al., 2020; Chen et al., 2020); subsequent work emphasized clinical accuracy through disease-classification and natural language inference (NLI)-based metrics (Miura et al., 2021); this was followed by expert-annotated semantic graphs capturing entities and their relations (Delbrouck et al., 2022a); and, most recently, by evaluation approaches that leverage large language models (LLM) (Ostmeier et al., 2024; Bannur et al., 2024a; Huang et al., 2024).

Despite efforts to establish fair benchmarking, such as shared metric codebases released for challenge tracks (Abacha et al., 2021; Delbrouck et al., 2023; Xu et al., 2024b) and a public leaderboard (Zhang et al., 2024b), there is still no open-source repository that reproduces the different factuality-focused metrics, whose scores can vary with implementation choices. For instance, earlier studies have computed BERTScore with different pretrained models and settings – varying the number of layers or whether scores are rescaled with a baseline (Zhang et al., 2019) – or swapped in F1CheXbert embeddings for the calculation (Smit et al., 2020b). Variants of the F1RadGraph metric likewise diverge depending on how they judge the correctness of entities and relations (Delbrouck et al., 2022b). Composite scores such as RadCliQ (Yu et al., 2023b) are similarly challenging to replicate.

`RadEval` brings the following solutions:

- **Unified open-source codebase**: every factuality-oriented metric proposed to date is re-implemented in a single, reproducible repository.
- **Metric refinements**: corrected and improved versions of existing metrics offer more faithful

[*]Equal contributions

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 546–557
November 4-9, 2025

estimates of clinical correctness (Section 5 and Appendix B).

- **Expanded expert test set**: an updated, radiologist-annotated corpus enabling fine-grained studies of how automatic metrics align with human judgments (Section 6).

- **Ready-made baselines**: published predictions from several widely-cited models are included so new systems can be benchmarked out-of-the-box.

- **Built-in statistical testing**: permutation and bootstrap tests (mirroring best practices in image captioning) enable users to determine whether score differences are statistically significant.

2 Existing Radiology Report Metrics

2.1 Lexical Overlap Metrics

Early evaluation methods focused on string-level overlap between generated and reference reports, typically using metrics from the natural language processing (NLP) literature. ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) remain among the most common, measuring n-gram precision and recall. METEOR (Banerjee and Lavie, 2005) and CIDEr (Vedantam et al., 2015) have also been adapted from image captioning literature. These metrics are straightforward to compute and require no domain-specific annotation or models, making them popular baselines.

In addition, BERTScore (Zhang et al., 2019) has been proposed to address limitations of n-gram overlap metrics. Instead of relying on exact token matches, it computes semantic similarity between candidate and reference reports using contextualized embeddings from a pretrained language model (*e.g.*, BERT (Devlin et al., 2019)). Token-level similarity is calculated based on the sum of cosine similarities, offering improved sensitivity to semantic alignment and lexical variation.

However, such metrics perform poorly in RRG settings due to their insensitivity to paraphrasing, semantic equivalence, or clinical correctness. Radiology reports are often sparse, redundant, or variable in linguistic expression, which causes n-gram metrics to underestimate report quality even when the clinical meaning is preserved.

2.2 Clinical Concept-Based Metrics

To improve domain specificity, several works introduced evaluation metrics based on clinical concept extraction and comparison. F1CheXbert (Smit et al., 2020b) utilizes a rule-based labeler that extracts 14 predefined CheXpert disease categories (Irvin et al., 2019) from both reference and generated reports, then computes an F1 score over presence/absence of these labels. SRR-BERT (Delbrouck et al., 2025) expands the label space to 55 labels, and also supports evaluation on structured reports.

F1RadGraph (Delbrouck et al., 2022b) also offers a more expressive alternative by representing reports as structured graphs of anatomical and observational entities and relations. A pretrained graph extraction model is applied to both candidate and reference reports, and graph-level overlap metrics (*e.g.*, precision, recall, F1) are computed. Although F1RadGraph has improved generality and some alignment with radiologist evaluations, it remains limited by the accuracy of the underlying parser and the quality of the training data. While the initial version of F1RadGraph was proposed by Jain et al. (2021), which computed entity and relation overlap separately and reported their average, it did not consider whether entities were matched based on textual spans, semantic types, or shared relations. This simplification may lead to overestimated alignment in complex cases.

RaTEScore (Zhao et al., 2024) identifies medical entities and their types (*e.g.*, anatomy, disease), applies synonym-aware semantic matching, and weights entities by diagnostic importance to better handle terminology variation and negation.

Similar to BERTScore, the CheXbert vector similarity metric (Yu et al., 2022) measures alignment between generated and reference reports by computing the cosine similarity of their embeddings obtained via the CheXbert model (Smit et al., 2020a).

In contrast to these standard clinical metrics, Temporal Entity F1 (Zhang et al., 2025a) is designed to assess temporal information quality in reports. It focuses on capturing the progression or stability of observations (*e.g.*, worsening, improved, or stable) (Bannur et al., 2023), and is particularly useful for detecting temporal hallucinations.

2.3 Composite and Learned Metrics

To improve alignment with human evaluations, ensemble-based or regression-trained metrics such

as RadCliQ (Yu et al., 2023b) have been proposed. RadCliQ uses a linear model trained on radiologist-labeled error counts, combining submetrics like BLEU, BERTScore, CheXbert vector similarity, and F1RadGraph. This strategy improves correlation with expert assessments but introduces interpretability challenges.

2.4 Generative and LLM-Based Evaluation Metrics

More recently, LLM-based generative evaluation has emerged as a promising paradigm for RRG metric design. These approaches leverage foundation models' reasoning and cross-domain generalization to assess report correctness, style, and completeness in a free-text or structured manner.

These LLM-driven evaluators offer high flexibility and often exhibit better alignment with radiologist judgment, especially on nuanced and out-of-distribution findings. For instance, GREEN (Ostmeier et al., 2024) proposed an interpretable and open-source LLM-based evaluation pipeline using a 7B parameter models, and includes a normalized GREEN score, structured error summaries for interpretability, and zero-shot generalization across imaging modalities.

CheXprompt (Zambrano Chaves et al., 2025) is a GPT-based evaluator that detects and categorizes six types of clinically relevant errors: false positive and false negative findings, incorrect location or severity, false positive comparisons, and false negative comparisons.

Similarly, FineRadScore (Huang et al., 2024) evaluates reports line by line, combining clinical severity with the number of incorrect lines. This reflects both the potential clinical risk and the effort required for correction, providing a practical measure of report quality.

RadFact (Bannur et al., 2024a) is also a GPT-based evaluation suite that assesses the factuality of each sentence in a generated report based on the corresponding reference sentences. It supports grounded evaluation and provides interpretable, sentence-level error analysis.

Despite their flexibility and strong alignment with expert judgment, GPT-based evaluation methods face key limitations: high computational cost, deployment barriers due to model size or proprietary APIs, and potential inconsistencies in output. Clinical applications also raise data privacy concerns. RadEval addresses these challenges by standardizing interfaces and supporting lightweight, open-source alternatives like GREEN, enabling local, privacy-preserving, and reproducible evaluation for radiology report generation.

3 RadEval

One of the critical obstacles to building AI systems that can match the accuracy and nuance of expert radiologists is the lack of standardized evaluation metrics. This gap hinders reliable analysis and comparison across different studies, and limits the real-world applicability of research progress.

Despite rapid innovation in metric development, practical barriers remain for adoption and comparison:

- Each metric typically requires separate installation, dependencies, and data pre-processing pipelines.
- Some tools lack public code or require proprietary APIs.
- Evaluation outputs vary in format and interpretability.

To mitigate this fragmentation, we propose a system that consolidates access to a wide range of RRG metrics, spanning from n-gram baselines to modern LLM evaluators. This system is designed to be modular, supporting plug-and-play integration of new metrics. By democratizing access to high-quality RRG evaluation tools, we aim to accelerate research on radiology report generation and encourage more standardized and reproducible benchmarking.

4 Benchmarking

We conducted extensive benchmarking and evaluation of various models on publicly available datasets (Table 3).

4.1 Datasets

For evaluation, we utilized the official test splits of MIMIC-CXR (Johnson et al., 2019b) and ReXGradient-160K (Zhang et al., 2025b), as well as the public validation set of CheXpert Plus (Chambon et al., 2024), as no official test split is available for the latter.

Each study in these datasets may include multiple associated images, all of which were retained for evaluation. Depending on model support, either a single representative image or all available

images were used as input. We focused on specifically evaluating the generation of the "Findings" and "Impression" sections. Reports missing either section were excluded to ensure consistent evaluation across metrics.

MIMIC-CXR A widely used public dataset containing 377,110 chest X-ray images across 227,835 studies, each paired with a radiology report (Johnson et al., 2019a). We use JPEG images from MIMIC-CXR-JPG instead of the original DICOMs. The official test split includes 2,347 studies with "Findings" sections and 2,224 with "Impression" sections.

CheXpert-Plus Comprises 223,462 image–report pairs from 187,711 studies across 64,725 patients (Chambon et al., 2024). We use its validation set, which contains 74 studies with "Findings" and 234 with "Impression" sections.

ReXGradient-160K The largest publicly available chest X-ray dataset to date in terms of patient coverage, including 160,000 image–report pairs from 109,487 patients across 79 U.S. medical sites (Zhang et al., 2025b). Its official test set includes 10,000 studies with both "Findings" and "Impression" sections.

4.2 Baselines

To evaluate the performance of existing RRG systems, we include a representative set of baselines from different institutions and architectures:

CheXpert-Plus (Chambon et al., 2024) Uses a Swinv2-based vision encoder and a two-layer BERT decoder. Two separate models are trained on MIMIC-CXR: one for the *Findings* section and another for the *Impression*.

CheXagent (Chen et al., 2024) An instruction-tuned foundation model for chest X-ray interpretation. It integrates a vision encoder with a cross-modal adapter to align visual and textual information. Training is conducted on CheXinstruct (Chen et al., 2024), a diverse instruction dataset aggregated from 28 open-source medical datasets.

MAIRA-2 (Bannur et al., 2024b) A model designed for grounded radiology report generation, which involves not only producing clinically accurate text but also identifying the spatial locations of findings. It builds on the LLaVA architecture (Liu et al., 2023), incorporating a frozen Rad-DINO-MAIRA-2 vision encoder (Pérez-García et al., 2025), a Vicuna-7B (Zheng et al., 2023) language model, and a four-layer fully connected multilayer perceptron for vision-language alignment.

Libra (Zhang et al., 2025a) A temporally-aware multimodal large language model (MLLM) tailored for generating the *Findings* section in radiology reports. Unlike prior single-image methods, Libra leverages paired chest X-rays to capture disease progression. It integrates a frozen Rad-DINO (Pérez-García et al., 2024) image encoder with the Meditron-7B (Chen et al., 2023) language model via a Temporal Alignment Connector, which combines a Layerwise Feature Extractor and a Temporal Fusion Module to embed multi-scale visual changes over time into the model architecture.

Med-CXRGen (Zhang et al., 2024a) Built on LLaVA-Med (Li et al., 2023), this model uses multi-stage visual instruction tuning and stitches multiple images for unified encoding. Separate models are trained for the "Findings" and "Impression" sections using the RRG24 dataset (Xu et al., 2024a).

5 `RadEvalBERTScore`

In this work, we also introduce a domain-specific radiology language encoder trained using SimCSE (Gao et al., 2022) – a contrastive learning method for sentence embeddings – to capture high-quality representations of radiology report text. We begin with a `ModernBERT-base` architecture and train it on the "Findings" and "Impression" sections from MIMIC-CXR, CheXpert, and ReXGradient-160K. This pretraining setup allows the model to learn clinically meaningful semantic relationships within and across radiology reports.

We demonstrate the effectiveness of our embedding model on a **zero-shot report-to-report retrieval** task. The set-up mirrors a classic text-retrieval scenario:

1. A small set of **query reports** and a larger pool of **candidate reports**, each annotated with one or more radiology labels, are encoded with a frozen text encoder.

2. For every query, we compute the cosine similarity to *all* candidates and obtain a ranked list in descending similarity order.

A candidate is considered relevant to a query if and only if the two reports share at least one label. For each of our experiments, we choose 10 queries and up to 200 positive candidates.

Models	CheXpert 5×200					HOPPR 8×200				
	P@5	P@10	P@50	P@100	mAP	P@5	P@10	P@50	P@100	mAP
Devlin et al. (2019)	46.6 ± 5.5	44.8 ± 4.8	37.0 ± 2.8	33.4 ± 2.1	30.1 ± 1.6	28.1 ± 2.1	24.6 ± 2.0	19.1 ± 1.0	17.3 ± 0.8	15.6 ± 0.4
Warner et al. (2024)	39.4 ± 4.1	35.7 ± 4.1	30.5 ± 2.2	28.3 ± 1.7	26.5 ± 1.2	23.0 ± 2.1	20.3 ± 1.6	16.8 ± 0.5	15.8 ± 0.5	14.6 ± 0.2
Sounack et al. (2025)	38.6 ± 4.6	36.6 ± 3.9	30.6 ± 2.3	28.2 ± 1.6	25.6 ± 0.9	23.4 ± 2.3	21.4 ± 1.6	17.4 ± 1.0	15.9 ± 0.5	14.7 ± 0.3
RadEvalBERT	60.3 ± 3.1	56.4 ± 2.6	46.4 ± 2.0	41.4 ± 1.7	36.4 ± 1.2	38.6 ± 2.0	34.8 ± 2.2	26.9 ± 1.0	23.7 ± 1.0	20.1 ± 0.6

Models	CheXbert Test Set				
	P@5	P@10	nDCG@5	nDCG@10	mAP
Devlin et al. (2019)	64.3 ± 2.0	59.2 ± 1.8	54.5 ± 1.3	48.7 ± 1.0	50.1 ± 1.1
Warner et al. (2024)	62.0 ± 2.1	57.2 ± 1.3	51.8 ± 1.5	46.2 ± 0.9	48.7 ± 1.0
Sounack et al. (2025)	61.8 ± 1.0	56.8 ± 1.0	51.2 ± 1.0	45.8 ± 0.8	48.3 ± 0.9
RadEvalBERT	70.2 ± 2.2	64.6 ± 1.5	58.9 ± 1.6	53.3 ± 1.1	53.4 ± 1.0

Table 1: Zero-shot report-to-report retrieval performance of RadEvalBERTScore Evaluation. We evaluate on three datasets: CheXpert 5×200 (five single-label categories), HOPPR 8×200 (eight out-of-domain single-label categories), and the CheXbert multi-label test set. For CheXpert and HOPPR, we report Precision@$\{5, 10, 50, 100\}$ and mean Average Precision (mAP); for CheXbert we report Precision@$\{5, 10\}$, normalized Discounted Cumulative Gain (nDCG@$\{5, 10\}$), and mAP. Values are presented as mean ± standard deviation over 10 random seeds.

5.1 Metrics

Precision@K

Fraction of the first K retrieved reports that are relevant. A hit is counted as soon as *one* label overlaps with the query.

mean Average Precision (mAP)

For each query, we compute *Average Precision* (*i.e.*, the mean of the precision values at every rank where a relevant report occurs. The scan stops at the last relevant rank, so items appearing afterwards cannot influence the score. mAP is the mean AP over all queries and reflects how early, on average, relevant reports are surfaced.

nDCG@K

A graded variant that rewards richer matches. The gain between a query and a candidate is the number of shared labels. DCG is accumulated over the top K positions with a logarithmic discount $1/\log_2(\text{rank}+1)$; nDCG normalises by the ideal DCG, yielding a score in $[0, 1]$ where 1 means the system ranks the strongest overlaps highest.

5.2 Datasets

CheXpert 5×200

Five single-label categories (Atelectasis, Cardiomegaly, Edema, Consolidation, Pleural Effusion).

→We report Precision@$\{5, 10, 50, 100\}$ and mAP.

CheXbert Test Set

Multi-label reports with 14 chexpert possible findings (*e.g.*, Airspace Opacity, Pneumonia, Support Devices).

→Because at most 20 positives exist per query, we report Precision@$\{5, 10\}$ and nDCG@$\{5, 10\}$ together with mAP.

HOPPR 8×200

This is an out-of-domain, single-label dataset used to evaluate generalization performance. The label categories include: acute rib fracture, air space opacity, cardiomegaly, lung nodule or mass, non acute rib fracture, pleural fluid, pneumothorax, and pulmonary artery enlargement.

→We report Precision@$\{5, 10, 50, 100\}$ and mAP.

This protocol provides complementary views of retrieval quality: Precision@K for top-K exact-hit rate, mAP for overall early-ranking performance, and nDCG@K for sensitivity to *how many* labels overlap.

6 RadEval Expert Dataset

Dataset. We release an updated RadEval-expert dataset with board-certified radiologists annotating clinically significant and insignificant errors across different error categories. Building on ReXVal (Yu et al., 2023a), we annotate false predictions of findings, omissions of findings, incorrect locations/positions, incorrect severities, spurious comparisons, omissions of changes from prior studies, as well

Table 2: Top-3 metrics by Kendall's τ_b (more negative is better; higher metric ⇒ fewer errors). "✓ aligned" = 95% CI < 0; "✗ misaligned" = 95% CI > 0; "ns" = CI overlaps 0. **Scope:** pooled rows show `(pairs, n)`; blocked rows show `(blocks, pairs)`. Each study has K=3 candidates (Findings: 148 studies; Impression: 60).

Endpoint	Metric	τ_b [95% CI]	Sig	Scope
Overall (pooled) vs. total significant errors				
	green	$-0.183\,[-0.246, -0.122]$	✓ aligned	(pairs 194,376, n 624)
	srr_bert	$-0.133\,[-0.193, -0.071]$	✓ aligned	(pairs 194,376, n 624)
	radcliq	$-0.107\,[-0.160, -0.052]$	✓ aligned	(pairs 194,376, n 624)
	radevalbertscore	$-0.076\,[-0.131, -0.017]$	✓ aligned	(pairs 194,376, n 624)
	rouge	$-0.038\,[-0.091, 0.017]$	ns	(pairs 194,376, n 624)
	bertscore	$-0.027\,[-0.085, 0.032]$	ns	(pairs 194,376, n 624)
	radgraph	$-0.011\,[-0.068, 0.048]$	ns	(pairs 194,376, n 624)
	bleu	$0.034\,[-0.020, 0.086]$	ns	(pairs 194,376, n 624)
	chexbert	$0.074\,[0.012, 0.137]$	✗ misaligned	(pairs 194,376, n 624)
ALL (blocked) vs. total significant errors				
	green	$-0.195\,[-0.295, -0.087]$	✓ aligned	(blocks 159, pairs 477)
	bertscore	$-0.160\,[-0.260, -0.059]$	✓ aligned	(blocks 188, pairs 564)
	bleu	$-0.153\,[-0.273, -0.029]$	✓ aligned	(blocks 89, pairs 267)
ALL (blocked) vs. total insignificant errors				
	radcliq	$-0.039\,[-0.147, 0.076]$	ns	(blocks 143, pairs 429)
	radevalbertscore	$-0.009\,[-0.126, 0.103]$	ns	(blocks 133, pairs 399)
	bertscore	$-0.008\,[-0.107, 0.094]$	ns	(blocks 143, pairs 429)
Impression only (blocked) vs. total significant errors				
	bertscore	$-0.225\,[-0.399, -0.049]$	✓ aligned	(blocks 57, pairs 171)
	rouge	$-0.215\,[-0.383, -0.042]$	✓ aligned	(blocks 57, pairs 171)
	radevalbertscore	$-0.206\,[-0.399, -0.010]$	✓ aligned	(blocks 53, pairs 159)
Findings only (blocked) vs. total significant errors				
	green	$-0.196\,[-0.314, -0.075]$	✓ aligned	(blocks 113, pairs 339)
	bleu	$-0.168\,[-0.313, -0.020]$	✓ aligned	(blocks 68, pairs 204)
	bertscore	$-0.132\,[-0.246, -0.017]$	✓ aligned	(blocks 131, pairs 393)
Per-category (blocked, significant-error endpoints)				
significant: false prediction of finding	green	$-0.082\,[-0.198, 0.030]$	ns	(blocks 134, pairs 400)
	radcliq	$-0.052\,[-0.157, 0.052]$	ns	(blocks 162, pairs 482)
	bertscore	$-0.049\,[-0.158, 0.061]$	ns	(blocks 162, pairs 482)
significant: omission of finding	srr_bert	$-0.512\,[-0.625, -0.390]$	✓ aligned	(blocks 91, pairs 271)
	chexbert	$-0.503\,[-0.626, -0.366]$	✓ aligned	(blocks 89, pairs 267)
	green	$-0.250\,[-0.374, -0.126]$	✓ aligned	(blocks 113, pairs 337)
significant: incorrect location/ position of finding	bertscore	$-0.068\,[-0.213, 0.079]$	ns	(blocks 80, pairs 238)
	radevalbertscore	$-0.040\,[-0.201, 0.123]$	ns	(blocks 76, pairs 226)
	rouge	$-0.018\,[-0.174, 0.139]$	ns	(blocks 80, pairs 238)
significant: incorrect severity of finding	radevalbertscore	$-0.033\,[-0.223, 0.160]$	ns	(blocks 56, pairs 168)
	bleu	$-0.001\,[-0.235, 0.219]$	ns	(blocks 35, pairs 105)
	rouge	$0.007\,[-0.170, 0.191]$	ns	(blocks 61, pairs 181)
significant: spurious comparison (not in reference)	bertscore	$-0.153\,[-0.300, 0.001]$	ns	(blocks 81, pairs 241)
	radcliq	$-0.125\,[-0.263, 0.014]$	ns	(blocks 81, pairs 241)
	radevalbertscore	$-0.103\,[-0.247, 0.063]$	ns	(blocks 77, pairs 229)
significant: omission of change from previous study	bleu	$-0.127\,[-0.335, 0.099]$	ns	(blocks 37, pairs 111)
	green	$-0.066\,[-0.241, 0.113]$	ns	(blocks 65, pairs 195)
	radgraph	$-0.027\,[-0.199, 0.137]$	ns	(blocks 61, pairs 183)
significant: inarticulate report (grammar/readability)	radcliq	$-0.350\,[-0.560, -0.140]$	✓ aligned	(blocks 35, pairs 105)
	radevalbertscore	$-0.266\,[-0.476, -0.046]$	✓ aligned	(blocks 34, pairs 102)
	bertscore	$-0.251\,[-0.480, -0.013]$	✓ aligned	(blocks 35, pairs 105)

as a new category: **inarticulate report/grammar**. All error categories are labeled as either significant or insignificant. We also extend beyond the "Impression" to also cover the "Findings"' section. The corpus comprises **208 studies** (**148 findings** and **60 impressions**), and **each study has exactly** $K{=}3$ **annotated candidate reports per ground truth**. Ground-truth reports come from MIMIC-CXR, CheXpert-Plus, and ReXGradient-160K, and candidate reports are generated by CheXagent, the CheXpert-Plus model, and MAIRA-2.

Methods. We measure agreement between automatic metrics and radiologists' judgments using Kendall's τ_b (Kendall, 1945) against radiologist error counts. We report: (1) a **pooled** (overall) τ_b versus the **total number of significant errors** (treating each candidate independently; ignores study grouping), and (2) a **blocked** (within-study, tie-aware) τ_b with 95% block-bootstrap confidence intervals obtained by resampling study blocks (preserving section type). Blocked analyses cover: totals of significant and insignificant errors across all sections ("ALL"), the significant-error total within each section (Impression, Findings), and each individual significant-error category. We label ✓ aligned when the 95% CI lies entirely below 0 (higher metric ⇒ fewer errors), ✗ misaligned when the CI lies entirely above 0 (higher scores ⇒ more errors), and **ns** otherwise.

Results. Overall (pooled vs. total significant errors), *green*, *srr_bert*, *radcliq*, and *radeval-bertscore* show meaningful negative correlations (✓aligned), while *rouge*, *bertscore*, *radgraph*, and *bleu* are not significant; *chexbert* is ✗misaligned (higher scores with more errors). Within studies (blocked, ALL), *green*, *bertscore*, and *bleu* are top for total significant errors (all ✓aligned), and no metric tracks total *insignificant* errors (all ns).

Table 2 reports correlations by section and across different error category types. From these results, *green* emerges as the most reliable single metric for tracking clinically significant errors, followed by *srr_bert*.

7 Conclusion

We introduced RadEval, a unified framework for evaluating RRG. By consolidating and standardizing a diverse suite of evaluation metrics, including lexical overlap, clinical concept extraction, structured graph comparison, and LLM-based scoring, RadEval addresses longstanding reproducibility and benchmarking challenges in the RRG domain. We refined existing metrics, released a high-fidelity expert-annotated test set, and benchmarked state-of-the-art report generation systems across multiple publicly available datasets. In addition, we demonstrated the utility of a new domain-specific sentence encoder through a zero-shot retrieval task and introduced an updated lightweight version of the GREEN metric capable of cross-modality evaluation. RadEval's modular architecture will help facilitate robust, reproducible, and clinically grounded evaluation – ultimately helping accelerate the safe deployment of radiology AI systems.

Limitations

While RadEval already unifies a broad set of automated radiology text evaluation metrics, recently proposed LLM-based metrics (*e.g.*, CheXprompt, FineRadScore, and RadFact) are not currently implemented due to their reliance on LLM APIs or lack of standardization – though they remain valuable future additions. Additionally, some metrics depend on upstream parsers that may introduce noise. Current evaluations also focus primarily on chest X-ray radiology reports written in English from institutions in the U.S., limiting generalizability across languages and geographical regions. Finally, we aim to continue to expand the expert-labeled dataset with additional reports and clinical annotators, as well as to compute detailed correlation analyses across all automated metrics and annotations to better assess metric alignment with clinical judgment.

References

Asma Ben Abacha, Yassine Mrabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. 2021. Overview of the mediqa 2021 shared task on summarization in the medical domain. *NAACL-HLT 2021*, page 74.

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, Ann Arbor, Michigan. Association for Computational Linguistics.

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Sal-

vatelli, Harshita Sharma, and 1 others. 2024a. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*.

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Noel C. F. Codella, Fabian Falck, Ozan Oktay, Matthew P. Lungren, Maria Teodora Wetscherek, and 2 others. 2024b. Maira-2: Grounded radiology report generation. *Preprint*, arXiv:2406.04449.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. 2023. Learning to exploit temporal structure for biomedical vision-language processing. *Preprint*, arXiv:2301.04558.

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. *Preprint*, arXiv:2405.19538.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449.

Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Andrew Johnston, Alaa Youssef, Louis Blankemeier, Christian Bluethgen, Stephan Altmayer, Jeya Maria Jose Valanarasu, Mohamed Siddig Eltayeb Muneer, Eduardo Pontes Reis, Joseph Paul Cohen, Cameron Olsen, Tanishq Mathew Abraham, Emily B. Tsai, Christopher F. Beaulieu, Jenia Jitsev, and 4 others. 2024. A vision-language foundation model to enhance efficiency of chest x-ray interpretation. *Preprint*, arXiv:2401.12208.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022a. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022b. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jean-Benoit Delbrouck, Maya Varma, Pierre Chambon, and Curtis Langlotz. 2023. Overview of the radsum23 shared task on multi-modal and multi-anatomical radiology report summarization. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 478–482.

Jean-Benoit Delbrouck, Justin Xu, Johannes Moll, Alois Thomas, Zhihong Chen, Sophie Ostmeier, Asfandyar Azhar, Kelvin Zhenghao Li, Andrew Johnston, Christian Bluethgen, Eduardo Reis, Mohamed Muneer, Maya Varma, and Curtis Langlotz. 2025. Automated structured radiology report generation. *Preprint*, arXiv:2505.24223.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2022. Simcse: Simple contrastive learning of sentence embeddings. *Preprint*, arXiv:2104.08821.

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. *arXiv preprint arXiv:2405.20613*.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Preprint*, arXiv:1901.07031.

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

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

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

M. G. Kendall. 1945. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Preprint*, arXiv:2306.00890.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and Jean-Benoit Delbrouck. 2024. GREEN: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 374–390, Miami, Florida, USA. Association for Computational Linguistics.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, Maria Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. 2024. Raddino: Exploring scalable medical image encoders beyond text supervision. *Preprint*, arXiv:2401.10815.

Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, Maria Teodora Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. 2025. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7(1):119–130.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 2020a. 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, Online. Association for Computational Linguistics.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. 2020b. 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.

Thomas Sounack, Joshua Davis, Brigitte Durieux, Antoine Chaffin, Tom J Pollard, Eric Lehman, Alistair EW Johnson, Matthew McDermott, Tristan Naumann, and Charlotta Lindvall. 2025. Bioclinical modernbert: A state-of-the-art long-context encoder for biomedical and clinical nlp. *arXiv preprint arXiv:2506.10896*.

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

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J. Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, Curtis Langlotz, and Jean-Benoit Delbrouck. 2024a. Overview of the first shared task on clinical text generation: RRG24 and "discharge me!". In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 85–98, Bangkok, Thailand. Association for Computational Linguistics.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, and 1 others. 2024b. Overview of the first shared task on clinical text generation: Rrg24 and

"discharge me!". In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 85–98.

F. Yu, M. Endo, R. Krishnan, I. Pan, A. Tsai, E. P. Reis, E. Kaiser Ururahy Nunes Fonseca, H. Lee, Z. Shakeri, A. Ng, C. Langlotz, V. K. Venugopal, and P. Rajpurkar. 2023a. Radiology report expert evaluation (rexval) dataset (version 1.0.0). RRID:SCR_007345.

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, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. 2022. Evaluating progress in automatic chest x-ray radiology report generation. *medRxiv*.

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

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*.

Xi Zhang, Zaiqiao Meng, Jake Lever, and Edmond S. L. Ho. 2025a. Libra: Leveraging temporal images for biomedical radiology analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17275–17303, Vienna, Austria. Association for Computational Linguistics.

Xi Zhang, Zaiqiao Meng, Jake Lever, and Edmond S.L. Ho. 2024a. Gla-AI4BioMed at RRG24: Visual instruction-tuned adaptation for radiology report generation. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 624–634, Bangkok, Thailand. Association for Computational Linguistics.

Xiaoman Zhang, Julián N. Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. 2025b. Rexgradient-160k: A large-scale publicly available dataset of chest radiographs with free-text reports. *Preprint*, arXiv:2505.00228.

Xiaoman Zhang, Hong-Yu Zhou, Xiaoli Yang, Oishi Banerjee, Julián N Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. 2024b. Rexrank: A public leaderboard for ai-powered radiology report generation. *arXiv preprint arXiv:2411.15122*.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D Manning, and Curtis Langlotz. 2020. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Score table

Table 3: Benchmarking results across multiple models on standard datasets under default system prompts.

Models	GREEN Llama	ROUGE L	F1 RG	RG ER	RG $\overline{ER}$	BLEU	BERT Score	SRR BERT	F1cXb 5	F1cXb 14	RaTE Score	Rad CLiQ v1	Temp. Entity F1	Rad Eval BERT
CheXpert-Plus Validation Findings														
CheXagent	26.2	21.4	27.4	24.8	17.6	4.1	33.9	50.5	55.8	58.1	51.1	67.9	40.0	8.2
CheXpert-Plus	25.2	21.7	25.8	22.9	17.2	4.7	47.1	46.9	48.7	52.8	51.0	77.3	34.4	12.7
Med-CXRGen-F	25.5	22.1	23.7	21.2	15.5	6.4	48.0	44.5	41.2	47.6	50.1	75.3	32.6	14.2
Libra-v1.0-3B	21.9	18.8	18.6	15.9	11.0	2.8	42.9	47.1	50.9	50.7	47.0	65.1	19.1	10.9
Libra-v1.0-7B	22.0	18.0	17.9	16.2	11.0	2.5	43.8	44.1	47.8	49.1	47.3	65.6	26.0	11.5
MAIRA-2	15.5	16.4	14.6	13.3	9.2	1.9	41.8	41.0	44.0	47.1	44.0	63.6	29.1	10.7
CheXpert-Plus Validation Impressions														
CheXpert-Plus	23.7	25.3	3.4	3.1	8.0	2.9	51.2	44.5	45.7	49.0	47.7	48.8	35.6	14.8
Med-CXRGen-I	26.6	23.6	18.7	16.7	12.6	6.5	50.8	44.4	48.2	48.3	45.5	64.6	35.3	20.0
MIMIC-CXR Test Findings														
CheXagent	29.6	23.1	26.8	24.2	17.4	4.9	39.0	49.7	60.1	55.1	56.0	71.6	22.5	37.2
CheXpert-Plus	30.6	25.7	24.2	22.1	17.0	5.7	54.2	48.2	54.1	47.4	54.2	84.6	22.5	46.8
Med-CXRGen-F	28.1	22.7	20.9	18.6	13.8	5.9	50.4	44.6	53.3	45.2	52.4	75.0	16.8	44.1
Libra-v1.0-3B	29.2	21.7	20.3	18.0	12.9	5.1	49.4	46.4	60.0	52.5	53.5	76.0	14.5	46.4
Libra-v1.0-7B	28.0	20.9	19.9	17.5	12.3	4.6	49.5	45.8	61.5	53.7	52.7	75.1	18.0	46.5
MAIRA-2	22.4	18.5	17.3	15.3	10.9	3.0	47.6	42.8	59.1	50.7	51.0	71.1	16.4	40.2
MIMIC-CXR Test Impressions														
CheXpert-Plus	25.0	23.6	22.6	20.1	16.8	2.9	46.2	36.8	50.4	45.3	46.4	84.6	37.7	29.9
Med-CXRGen-I	20.8	14.6	13.5	11.7	8.5	2.3	39.3	35.4	49.1	41.5	43.4	61.4	16.7	9.7
ReXGradient-160K Test Findings														
CheXagent	44.5	21.7	25.9	24.5	21.0	3.3	39.2	41.0	4.5	5.5	59.2	69.7	57.1	36.4
CheXpert-Plus	33.6	23.5	22.0	20.7	16.8	3.5	50.9	45.4	16.5	21.0	55.0	76.7	45.8	36.5
Med-CXRGen-F	29.6	14.7	10.1	9.2	4.4	1.6	42.0	39.7	0.7	13.0	43.6	65.8	62.1	33.1
Libra-v1.0-7B	44.8	21.9	20.5	18.5	13.1	3.8	51.0	45.7	7.4	15.7	56.2	78.7	40.5	40.6
MAIRA-2	36.5	21.4	20.1	18.3	14.3	2.3	48.9	43.0	4.1	14.0	52.9	77.9	60.8	35.5
ReXGradient-160K Test Impressions														
CheXpert-Plus	3.8	7.6	2.5	1.9	1.2	0.2	30.3	37.8	16.3	9.4	30.1	51.2	21.2	18.7
Med-CXRGen-I	22.3	8.2	3.9	3.6	2.4	0.4	31.5	40.6	2.6	24.4	31.4	51.7	55.6	10.5

B Refined GREEN Metric

To extend the capabilities of the GREEN metric, we finetuned a compact Gemma-2B model on the original GREEN dataset along with an additional 50,000 annotated radiology report pairs spanning multiple imaging modalities, including CT, MRI, and ultrasound. This represents an evolution beyond the original implementation, which focused exclusively on chest X-rays. By leveraging a smaller, more lightweight language model, we achieve substantial improvements in computational efficiency – averaging inference times of 2–3 seconds per report – while maintaining performance comparable to larger models such as Llama and Phi. This reduced resource footprint makes the updated GREEN model more practical for large-scale or real-time deployment settings. Our work underscores the potential for continued improvement by incorporating more diverse imaging contexts and exploring even lighter architectures without sacrificing clinical alignment or interpretability.