CREPE: Rapid Chest X-ray Report Evaluation by Predicting Multi-category Error Counts

Gihun Cho[1,2,3], **Seunghyun Jang**[1,2,3], **Hanbin Ko**[1,2,3], **Inhyeok Baek**[1,2,3], **Chang Min Park**[1,2,3]
[1]Seoul National University
[2]Seoul National University Hospital
[3]Innovative Radiology AI Laboratory (iRAIL)
{gihuncho,jjaa0113,kolucas1219,bih1122,morphius}@snu.ac.kr
https://gihuncho.github.io/crepe

Abstract

We introduce **CREPE** (Rapid **C**hest X-ray **R**eport **E**valuation by **P**redicting Multi-category **E**rror Counts), a rapid, interpretable, and clinically grounded metric for automated chest X-ray report generation. CREPE uses a domain-specific BERT model fine-tuned with a multi-head regression architecture to predict error counts across six clinically meaningful categories. Trained on a large-scale synthetic dataset of 32,000 annotated report pairs, CREPE demonstrates strong generalization and interpretability. On the expert-annotated ReXVal dataset, CREPE achieves a Kendall's τ correlation of 0.786 with radiologist error counts, outperforming traditional and recent metrics. CREPE achieves these results with an inference speed approximately 280 times faster than large language model (LLM)-based approaches, enabling rapid and fine-grained evaluation for scalable development of chest X-ray report generation models.

1 Introduction

The automatic generation of radiology reports from chest X-ray images has seen rapid progress with the advent of advanced generative AI technologies. Such systems hold substantial promise to reduce radiologists' workload and enhance clinical workflow efficiency by enabling the automated production of clinically meaningful reports at scale (Chen et al., 2024; Zambrano Chaves et al., 2025).

A central challenge that remains is the reliable evaluation of these generated reports. Current metrics for evaluating chest X-ray report generation fall into four main categories: overlap-based, embedding-based, named entity recognition (NER)-based, and large language model (LLM)-based approaches. Overlap-based metrics, like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) measure surface-level lexical similarity but miss clinical meaning. Embedding-based metrics, such as BERTScore (Zhang et al., 2020) and SembScore (Smit et al., 2020) capture semantic alignment but not factual or relational accuracy. NER-based metrics, like F1 RadGraph (Jain et al., 2021; Delbrouck et al., 2024) and RaTEScore (Zhao et al., 2024) focus on extracted medical entities and relations, but their reliability depends on the underlying NER system and they can miss nuanced contextual errors. LLM-based metrics, such as GREEN (Ostmeier et al., 2024) and FineRadScore (Huang et al., 2024) achieve strong alignment with human judgment, yet are limited by high computational cost and slow inference. These trade-offs motivate the need for a fast, clinically appropriate, and robust evaluation metric.

In this work, we introduce **CREPE** (Rapid **C**hest X-ray **R**eport **E**valuation by **P**redicting Multi-category **E**rror Counts), a novel evaluation metric designed for both rapid inference and clinically interpretable assessment. CREPE leverages a medical domain-specific BERT model with multiple regression heads to predict continuous error counts across six clinically meaningful categories. By explicitly modeling these error categories, CREPE outputs both a fast, interpretable overall score and a detailed error breakdown, thus providing actionable, clinically relevant feedback beyond conventional summary metrics.

Our primary contributions are as follows:

- **Fast and Fine-Grained Evaluation**: We propose CREPE, a rapid and fine-grained evaluation metric that predicts clinically interpretable multi-category error counts via regression, achieving performance comparable to or exceeding state-of-the-art LLM-based metrics in alignment with radiologist assessments.
- **Large-Scale Synthetic Training Data**: We construct a large-scale synthetic dataset comprising 32,000 report pairs with detailed,

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 21750–21767
November 4-9, 2025

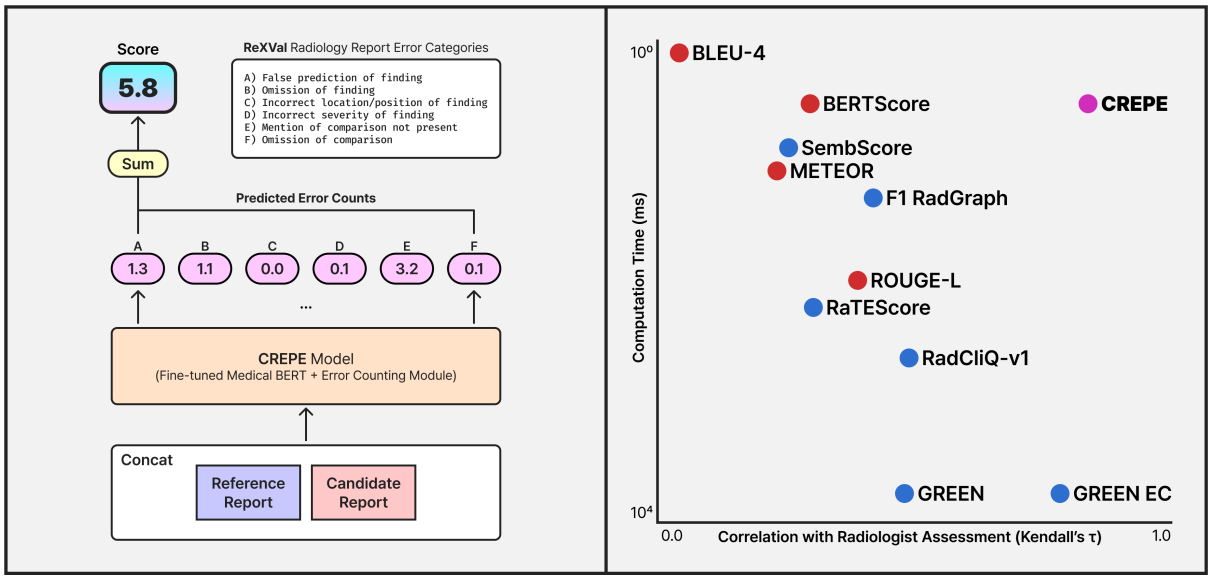


Figure 1: **Overview of CREPE and Comparative Performance.** (Left) Schematic illustration of CREPE's evaluation pipeline: a ground truth and generated chest X-ray report are concatenated and processed by the CREPE model, which predicts continuous error counts across six clinically defined categories; the total error count is used as the final CREPE score. (Right) Comparative analysis of evaluation metrics on the ReXVal dataset, plotting Kendall's τ correlation with radiologist error counts (x-axis) against average computation time per report pair (y-axis, log scale). CREPE achieves superior correlation with expert assessments while maintaining low computational cost, enabling fast and clinically aligned report evaluation.

category-specific error annotations, automatically generated using LLMs under strict data governance.

- **Robust Generalization**: We demonstrate strong generalization and robustness of CREPE across diverse public evaluation benchmarks, including challenging out-of-distribution datasets.
- **Efficiency and Scalability**: CREPE delivers substantial reductions in computational cost and inference time compared to LLM-based evaluation methods, enabling practical and scalable assessment for both research and real-world development pipelines.

An overview of the CREPE pipeline and its comparative performance is illustrated in Figure 1.

2 Related Work

2.1 General Text Evaluation Metrics

Traditional natural language generation (NLG) metrics are widely adopted for evaluating generated text in machine translation and summarization. Overlap-based metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) measure n-gram precision and recall, respectively, providing straightforward and interpretable scores but offering limited insight into the semantic or factual correctness of generated content. Embedding-based metrics like BERTScore (Zhang et al., 2020) address some of these limitations by comparing contextualized token embeddings from pretrained language models, thus better capturing semantic similarity.

More recently, LLM-based evaluators, including Prometheus (Kim et al., 2023) and G-Eval (Liu et al., 2023) have been proposed to deliver holistic and dimension-specific quality assessments through instruction-tuned LLMs. Although these approaches achieve stronger alignment with human judgments, their high computational cost and inference latency limit their practical application in large-scale or rapid evaluation settings.

2.2 Radiology Report Evaluation Metrics

Metrics tailored for clinical text generation aim to address the unique demands of the medical domain, where factual accuracy and clinical interpretability are critical. NER-based metrics such as F1 RadGraph (Jain et al., 2021; Delbrouck et al., 2024) focus on the extraction and matching of medical entities and their relations to evaluate factual consistency, while SembScore (CheXbert vector similarity) (Smit et al., 2020) represents findings as structured vectors for semantic comparison.

Composite metrics such as RadCliQ (Yu et al., 2023a) employ linear regression to combine the outputs of BLEU, BERTScore, SembScore, and F1 RadGraph into a single score. RaTEScore (Zhao et al., 2024) uses medical NER model to emphasize entity-level factual precision. LLM-based metrics such as GREEN (Ostmeier et al., 2024) and FineRadScore (Huang et al., 2024) leverage the reasoning capabilities of language models to provide fine-grained, expert-aligned scoring. While each of these methods advances clinical alignment or interpretability, they often introduce complexity, require specialized domain resources, or incur substantial computational cost, which can hinder scalability and routine deployment.

2.3 Regression-Based Text Evaluation

Regression-based approaches aim to bridge the gap between metric predictions and human judgments by training models to directly regress to quality scores or error counts from text features. Notable examples include COMET (Rei et al., 2020) and BLEURT (Sellam et al., 2020), which use pretrained language model encoders with regression heads to approximate human evaluation. While these methods can offer improved correlation with human ratings, they typically operate as black-box predictors and may not provide detailed, clinically interpretable feedback.

Overall, despite substantial progress in the development of evaluation metrics, a fast, interpretable, and clinically grounded method for chest X-ray report generation remains an open challenge.

3 CREPE

Our proposed metric, **CREPE**, evaluates the quality of chest X-ray reports by directly predicting error counts across clinically relevant categories. By providing structured, category-specific error estimates, CREPE enables both quantitative assessment and interpretable feedback, thereby facilitating robust comparison and targeted diagnostic improvement in automated report generation.

3.1 Problem Formulation

Given a reference report R_{ref} and a generated report R_{cand}, the goal is to predict a vector of error counts,

$$\mathbf{n} = [n_A, n_B, \ldots, n_F],$$

where each element corresponds to one of six clinically defined error categories:

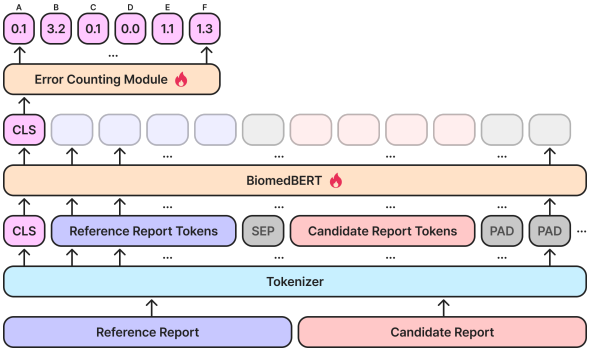


Figure 2: **CREPE Model Architecture.** Tokenized reference and candidate reports are encoded by a medical domain-specific BERT model, and the pooled representation is used by the error counting module to output category-wise error counts.

(A) False prediction of finding

(B) Omission of finding

(C) Incorrect location or position of finding

(D) Incorrect severity of finding

(E) Mention of comparison not present in the reference impression

(F) Omission of comparison describing a change from a previous study

as defined in the annotation protocol used for the ReXVal dataset (Yu et al., 2023b).

The overall error count score, S, is given by the sum of the predicted error counts:

$$S = \sum_{c \in \mathcal{E}} n_c, \tag{1}$$

where $\mathcal{E} = \{A, B, C, D, E, F\}$. Lower values of S indicate higher report quality, with category-level predictions providing actionable insights for further model refinement.

3.2 Model Architecture

The CREPE model is built upon a pretrained BERT model specifically tailored for biomedical text, which we fine-tune for regression-based error prediction. For each error category, the model jointly estimates both the expected error count and the presence of any error, thereby enabling detailed and clinically interpretable evaluation. The overall architecture of the model is shown in Figure 2.

Encoder and Feature Extraction. Let $\mathbf{x}$ denote the tokenized input sequence formed by concatenating R_{ref} and R_{cand}, separated by special tokens. This sequence is processed by the BERT encoder, yielding a pooled representation $\mathbf{h} \in \mathbb{R}^d$:

$$\mathbf{h} = \text{BERT}(\mathbf{x})_{[\text{CLS}]} \tag{2}$$

A dropout layer is applied to $\mathbf{h}$ to prevent overfitting. This pooled output serves as the shared feature for all downstream prediction heads.

Error Regression Heads. For each error category $c \in \mathcal{E}$, an independent regression head predicts the error count:

$$\hat{n}_c = f_c(\mathbf{h}) \tag{3}$$

where $f_c(\cdot)$ is a category-specific feedforward layer outputting a predicted error counts $\hat{n}_c$. The collection of outputs, $\hat{\mathbf{n}} = [\hat{n}_A, \hat{n}_B, \ldots, \hat{n}_F]$, constitutes the predicted error vector.

Error Detection Heads. To address the challenge of class imbalance, where certain error categories are infrequently represented in the training data as shown in Figure 4, the model incorporates auxiliary error detection heads during training. These heads are designed to predict the presence or absence of each error type, and their outputs are used exclusively for loss calculation to enhance learning for rare categories. Specifically, for each category c, the model produces a logit:

$$\hat{p}_c = g_c(\mathbf{h}) \tag{4}$$

where $g_c(\cdot)$ is a category-specific feedforward layer. During training, a sigmoid activation is applied to obtain a probability estimate for error presence, but these predictions are not used at inference time.

Loss Function. As shown in Figure 3, the CREPE model is trained with a dual-objective loss that captures both the count and presence of clinically meaningful errors. For each error category $c \in \mathcal{E}$, the model predicts a continuous error count $\hat{n}_c$ as well as an auxiliary presence logit $\hat{p}_c$. Let $\mathbf{n} = [n_A, \ldots, n_F]$ denote the ground-truth error counts, and $\mathbf{p} = [p_A, \ldots, p_F]$ denote the binary presence indicators, where $p_c = \mathbb{I}[n_c > 0]$.

The regression loss encourages accurate estimation of error counts and is computed using mean squared error (MSE):

$$\mathcal{L}_{\text{reg}} = \frac{1}{|\mathcal{E}|} \sum_{c \in \mathcal{E}} \text{MSE}(\hat{n}_c, n_c) \tag{5}$$

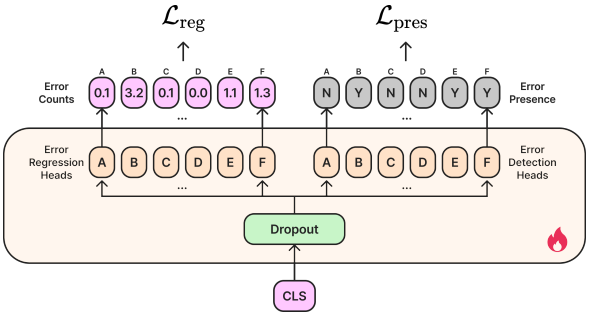


Figure 3: **Prediction Heads Architecture.** The pooled [CLS] embedding is passed through a dropout layer and then fed into two parallel sets of output heads for each error category: error regression heads (left) predict continuous error counts, while auxiliary error detection heads (right) predict the presence or absence of errors during training.

The presence loss penalizes incorrect predictions of whether an error of a given type is present and is computed using binary cross-entropy (BCE):

$$\mathcal{L}_{\text{pres}} = \frac{1}{|\mathcal{E}|} \sum_{c \in \mathcal{E}} \text{BCE}(\hat{p}_c, p_c) \tag{6}$$

The total loss used for training is the average of the regression and presence losses:

$$\mathcal{L} = \frac{\mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{pres}}}{2} \tag{7}$$

This combined objective allows the model to learn both precise error count estimation and improved sensitivity to rare or underrepresented error categories, resulting in more accurate and interpretable evaluation of chest X-ray reports.

Inference and Scoring. As shown in Figure 1, the CREPE model outputs predicted error counts $\hat{n}_c$ for each of the six error categories given a pair of reference and generated reports. The CREPE metric is defined as the sum of predicted error counts:

$$\text{CREPE} = \sum_{c \in \mathcal{E}} \hat{n}_c \tag{8}$$

A lower CREPE score indicates higher report quality, and this aggregate value serves as the primary metric for evaluating overall performance through correlation-based analysis. In addition, category-level error predictions can be analyzed individually, for example using mean absolute error (MAE), to provide more granular diagnostic feedback or to target specific clinical priorities.

Metric	ReXVal		ReFiSco-v0		RadEvalX		RaTE-Eval		RaTE-Eval†	
	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ
BLEU-4	0.383	0.516	0.489	0.616	0.074	0.096	0.197	0.247	0.134	0.166
ROUGE-L	0.570	0.748	0.524	0.662	0.257	0.356	0.200	0.281	0.220	0.302
METEOR	0.484	0.653	0.468	0.617	0.201	0.284	0.174	0.245	0.248	0.338
BERTScore	0.521	0.694	0.541	0.689	0.326	0.452	0.224	0.315	0.256	0.351
F1 RadGraph	0.585	0.765	0.475	0.609	0.171	0.243	0.306	0.393	0.258	0.328
SembScore	0.495	0.666	0.461	0.605	0.318	0.434	0.198	0.280	0.245	0.336
RaTEScore	0.520	0.697	0.433	0.571	0.316	0.438	0.339	**0.460**	0.310	0.419
RadCliQ-v1	0.623	0.809	0.510	0.656	0.326	0.449	0.299	0.415	0.304	0.414
GREEN	0.626	0.798	0.592	0.709	0.411	0.539	**0.374**	0.457	0.409	0.494
GREEN EC	0.775	0.899	**0.723**	0.811	0.448	0.577	0.252	0.315	**0.432**	0.517
CREPE	**0.786**	**0.933**	0.697	**0.825**	**0.580**	**0.745**	0.267	0.375	0.407	**0.541**

Table 1: **Correlation with Radiologist Error Counts Across Datasets.** Kendall's τ and Spearman's ρ correlation coefficients for each evaluation metric on five benchmark datasets. Bold indicates the best and underline the second-best value for each metric and dataset. 95% confidence intervals are reported in Table 5.

3.3 Synthetic Training Data Generation

Obtaining large-scale, expert-annotated error counts for chest X-ray reports is logistically and financially challenging. To address this, we constructed a synthetic training dataset through an automated pipeline, designed to maximize clinical relevance while adhering strictly to responsible data use policies for sensitive medical data.

Report Pair Sampling. We randomly sampled 32,000 image-report pairs from the MIMIC-CXR (Johnson et al., 2019) training set to serve as the basis for synthetic data generation.

Synthetic Report Generation. For each selected image, a candidate report was generated using CheXagent (Chen et al., 2024), a vision-language foundation model specialized for chest X-ray interpretation.

Error Analysis. To identify and classify errors between reference and candidate reports, we used Gemini 2.5 Pro 03-25 preview version deployed via Vertex AI on Google Cloud (Google, 2025). This ensured that all analysis was performed in a secure environment, fully compliant with the MIMIC-CXR data use agreement, which prohibits sending data to external services such as public LLM APIs.

Label Extraction. We adapted the prompt from GREEN (Ostmeier et al., 2024) to obtain, for each reference–candidate report pair, error counts across six predefined categories, separately for clinically significant and clinically insignificant errors. For each category, we summed the significant and insignificant error counts to create the final label vector $[n_A, n_B, \ldots, n_F]$ used for CREPE model training. The distribution of error counts for each category is illustrated in Figure 4.

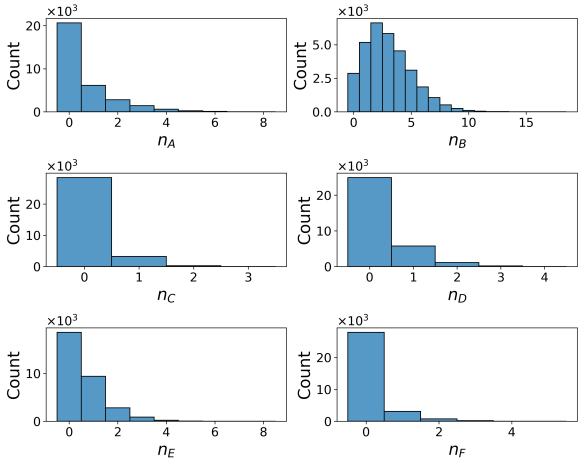


Figure 4: **Distribution of per-category error counts in the synthetic training dataset.** The highly skewed distributions highlight substantial class imbalance, motivating the use of a dual-objective loss to improve learning of rare error categories.

This data generation pipeline enabled us to create a diverse and representative training dataset, supporting robust model learning across a wide range of clinically meaningful error types while strictly adhering to data governance requirements. For full details on data generation, including prompts and hyperparameters used for the generation models, see Appendix A.2.

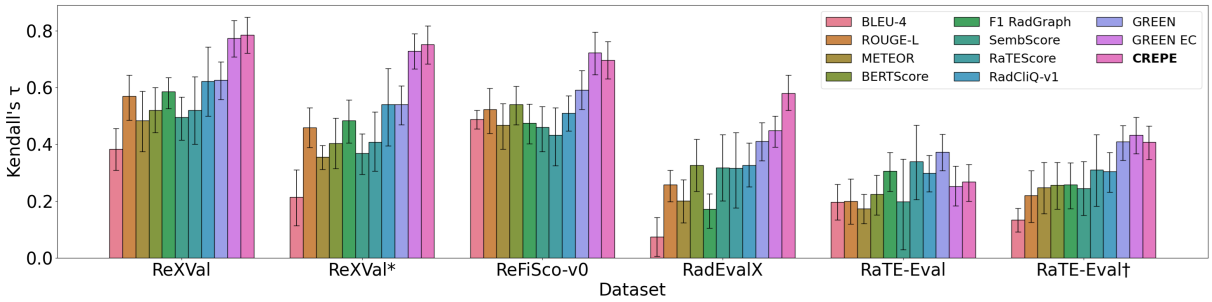


Figure 5: **Kendall's τ Correlation with Radiologist Error Counts Across Datasets.** Bar plots show Kendall's τ values (with 95% confidence intervals) for each evaluation metric on multiple public benchmarks. CREPE consistently achieves the highest or near-highest correlation with expert error counts across all datasets.

4 Experiments

4.1 Experimental Setup

We designed a comprehensive experimental framework to evaluate CREPE's performance across multiple aspects of chest X-ray report evaluation. Below, we describe the datasets used for benchmarking, the baseline metrics for comparison, and the main implementation and optimization details for reproducibility.

Evaluation Datasets. We evaluated CREPE on four publicly available datasets containing radiologist-annotated errors: ReXVal (Yu et al., 2023b), ReFiSco-v0 (Tian et al., 2023), RadEvalX (Calamida et al., 2024), and RaTE-Eval (Zhao et al., 2024). ReXVal consists of expert-labeled report-level errors from the MIMIC-CXR dataset, and we additionally use a filtered variant, ReXVal*, with identical report pairs removed to reduce class imbalance. ReFiSco-v0 provides line-level severity annotations, which we map to binary error labels and aggregate at the report level. RadEvalX comprises 100 IU-Xray (Pavlopoulos et al., 2019) reports with expert annotations for six standard error categories, plus two uncertainty-related categories. For RaTE-Eval, we use only the sentence-level human rating task, which covers nine imaging modalities and 22 anatomies and includes normalization of error counts by potential error opportunities. We also report raw, unnormalized counts for comparison and denote the raw-count variant as RaTE-Eval†. Detailed descriptions of dataset construction, annotation protocols, and pre-processing are provided in Appendix A.3.

Baselines. To contextualize CREPE's performance, we compared it against general text evaluation metrics, including BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang et al., 2020), as well as radiology report evaluation metrics, such as F1 RadGraph (Jain et al., 2021), SembScore (Smit et al., 2020), RaTEScore (Zhao et al., 2024), RadCliQ (Yu et al., 2023a), GREEN (Ostmeier et al., 2024), and GREEN error count (GREEN EC).

Implementation Details. Our experiments were conducted using `BiomedBERT` (Gu et al., 2021), a medical domain-specific BERT model, as the foundation for the CREPE model. The entire BERT encoder and all regression and presence heads were fine-tuned on the synthetic dataset for 10 epochs, with a validation split of 0.1 and a batch size of 64. Sequences were truncated or padded to a maximum length of 512 tokens, selected based on corpus statistics. The longest sequence observed was 502 tokens in the fMIMIC-CXR training set and 451 tokens in the synthetic corpus, making 512 sufficient to cover all cases without introducing excessive padding. We employed mixed-precision training (FP16) to accelerate computation. All experiments were performed on a single NVIDIA A6000 GPU.

Hyperparameter Optimization. Key hyperparameters, including learning rate, weight decay, and warmup ratio, were optimized via automated search using Optuna (Akiba et al., 2019). The optimal configuration was selected based on validation set performance, as measured by Kendall's τ correlation.

4.2 Results

We comprehensively evaluate CREPE across four key dimensions: correlation with human expert judgments (Sec. 4.3), robustness to class imbalance (Sec. 4.4), computational efficiency (Sec. 4.5), and absolute error-prediction accuracy (Sec. 4.6).

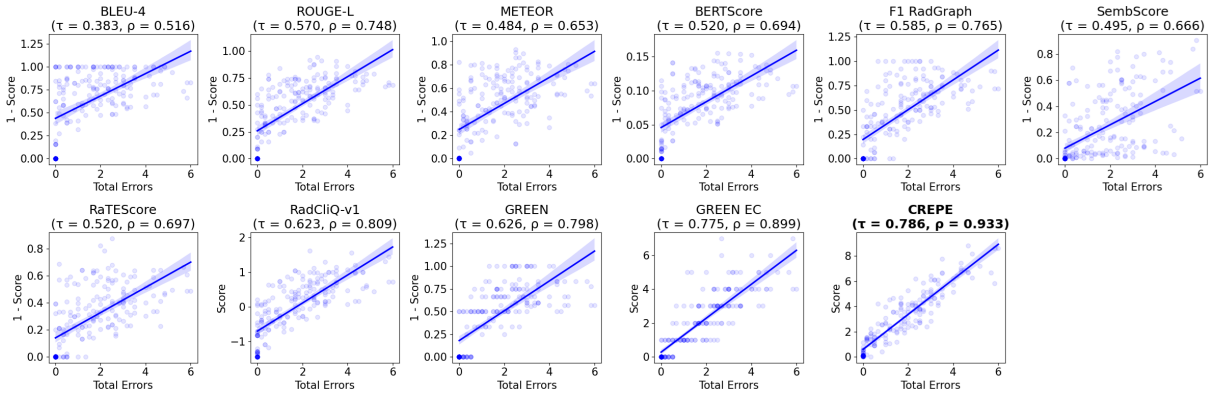


Figure 6: **Correlation Between Metric Scores and Radiologist-Identified Errors (ReXVal).** Scatter plots show the relationship between each evaluation metric's score and the total number of radiologist-identified errors on the ReXVal dataset. Each subplot includes a regression line and 95% confidence intervals. CREPE achieves the highest correlation with expert error counts. Kendall's τ and Spearman's ρ are shown in the subplot titles.

4.3 Correlation with Human Judgments

To assess clinical validity, we evaluate the correlation between each metric's predictions and radiologist-annotated error counts across five public benchmarks with varying annotation protocols and levels of difficulty. Table 1 reports Kendall's τ and Spearman's ρ for all metrics. CREPE achieves the highest or second-highest correlation on most datasets, demonstrating strong alignment with human expert ratings. On ReXVal, CREPE attains a Kendall's τ of 0.786 and a Spearman's ρ of 0.933, while maintaining competitive performance on line-level (ReFiSco-v0), category-extended (RadEvalX), and out-of-distribution multi-modality (RaTE-Eval) datasets.

Figure 5 summarizes Kendall's τ (with 95% confidence intervals) across all benchmarks, illustrating that CREPE is consistently among the top-performing metrics in terms of agreement with radiologist judgments.

For a more detailed view, Figure 6 visualizes the relationship between metric scores and total radiologist-identified errors on ReXVal. The scatter plots show that CREPE exhibits a monotonic association with expert error counts, in line with its correlation statistics.

Taken together, these results suggest that CREPE provides reliable alignment with expert judgment across a range of evaluation settings and dataset characteristics.

4.4 Robustness to Class Imbalance

To examine the impact of class imbalance, particularly the prevalence of zero-error pairs in ReXVal, we constructed a filtered version, ReXVal*, by removing report pairs where the reference and candidate reports were identical. As shown in Table 2, this filtering results in a notable decline in the performance of conventional metrics; the average decrease in Kendall's τ is approximately 0.12. In contrast, CREPE's correlation with human judgments decreases by only 0.033, corresponding to a 4.2 percent relative drop. These results suggest that CREPE remains robust and continues to reliably differentiate clinically meaningful errors even in more challenging evaluation settings.

Metric	ReXVal	ReXVal*	$\Delta\tau$
BLEU-4	0.383	0.215	−0.168
ROUGE-L	0.570	0.459	−0.111
METEOR	0.484	0.355	−0.129
BERTScore	0.521	0.404	−0.117
F1 RadGraph	0.585	0.484	−0.101
SembScore	0.495	0.368	−0.127
RaTEScore	0.520	0.408	−0.113
RadCliQ-v1	0.623	0.500	−0.123
GREEN	0.626	0.500	−0.126
GREEN EC	0.775	0.729	−0.046
CREPE	**0.786**	**0.753**	**−0.033**

Table 2: **Robustness to Class Imbalance.** Kendall's τ for each metric on the original ReXVal dataset and the filtered set ReXVal*, after removing 26 identical report pairs. $\Delta\tau$ shows the change in correlation.

4.5 Computational Efficiency

We compare the inference time per sample for all metrics using the ReXVal dataset. As shown in Figure 1 and Table 3, CREPE processes a report in 9.5 milliseconds, which is comparable to the fastest

Metric	Time (ms)
BLEU-4	4.29
ROUGE-L	125.99
METEOR	24.83
BERTScore	9.65
F1 RadGraph	36.59
SembScore	18.86
RaTEScore	160.28
RadCliQ-v1	370.58
GREEN	2642.37
GREEN EC	2642.37
CREPE	**9.53**

Table 3: **Speed Comparison.** Average computation time per sample (milliseconds) for each evaluation metric on the ReXVal dataset.

Head Counts	τ
1	0.723
2	0.730
6 (ours)	**0.753**
Loss Function	τ
MSE only	0.718
Poisson NLL only	0.738
MSE + BCE (ours)	**0.753**
Encoder Backbone	τ
`ClinicalBERT`	0.693
`Bio_ClinicalBERT`	0.716
`BiomedBERT` (ours)	**0.753**

Table 4: **Ablation on Model Components.** Impact of regression head granularity, loss function, and encoder backbone on Kendall's τ for ReXVal*.

neural baselines such as BERTScore. In contrast, LLM-based methods such as GREEN require over 2,600 milliseconds per sample, making them approximately 280 times slower. This substantial speed advantage demonstrates that CREPE is well-suited for rapid and large-scale evaluation, without sacrificing alignment with expert judgment.

4.6 Error Score Accuracy

Among public benchmarks, the ReXVal and RaTE-Eval datasets provide category-level error counts, enabling direct assessment of absolute prediction accuracy. On the ReXVal dataset, CREPE achieves a mean absolute error (MAE) of 1.252 ± 1.022, representing a 25.2% reduction compared to GREEN EC (1.674 ± 1.241). On the RaTE-Eval dataset, CREPE achieves an MAE of 0.739 ± 0.627, which is 33% lower than the MAE of GREEN EC (1.102 ± 0.632). Both metrics offer category-specific predictions, but CREPE's consistently lower error across datasets demonstrates the effectiveness of regression-based modeling when detailed supervision is available.

4.7 Qualitative Case Studies

To complement the quantitative evaluation, we also provide qualitative case studies that highlight both typical behaviors and failure modes of CREPE. The examples drawn from the ReXVal dataset are presented in Appendix A.1.

5 Ablation Studies

We conducted ablation studies on the balanced ReXVal* dataset to quantify the contributions of key architectural and modeling choices. Unless otherwise specified, all variants share the same hyperparameters as the full model (six regression heads, dual-objective MSE + BCE loss, and `BiomedBERT` backbone).

5.1 Number of Regression Heads

To assess the impact of output granularity, we compared three regression head configurations: (1) a single head predicting the total error count, (2) two heads for clinically significant and insignificant errors, and (3) six heads for category-specific errors, which is the default setting. As shown in Table 4, increasing the number of regression heads leads to consistently higher agreement with radiologist judgments, with the six-head model achieving the best Kendall's τ. This shows that dividing the task into error-specific heads is more effective than relying on a single head, as each head only needs to capture one type of clinical error.

5.2 Loss Function

We evaluated three loss formulations: mean squared error (MSE) only, Poisson negative log likelihood (NLL) only, and a dual-objective MSE+BCE loss that incorporates an auxiliary error presence term. As shown in Table 4, the dual-objective loss yields the highest correlation with expert annotations. This indicates that predicting both the error count and the presence of each error type helps the model handle rare categories more effectively, complementing the multi-head structure where each head specializes in a single error type.

5.3 Encoder Backbone

To determine the importance of domain-specific pre-training, we replaced `BiomedBERT` with two clinical BERT models: `ClinicalBERT` (Wang et al., 2023; Liu et al., 2025) and `Bio_ClinicalBERT` (Alsentzer et al., 2019), keeping all other settings fixed. As reported in Table 4, `BiomedBERT` yields the highest Kendall's τ. The reason for `BiomedBERT`'s superior performance may be related to its pre-training on a broader biomedical literature corpus, potentially offering better coverage of radiology terminology than models trained solely on clinical notes. Details about encoder backbones are provided in Appendix A.4.

6 Conclusion

We presented **CREPE**, an efficient evaluation metric for automated chest X-ray report generation that predicts clinically meaningful error counts using a domain-specific BERT model with a multi-head regression architecture. CREPE provides both an overall score and interpretable category-level feedback, demonstrating strong correlation with expert judgments, robustness to class imbalance, and fast inference compared to existing evaluation methods. Future work includes extending this approach to additional medical imaging domains and further investigating its generalizability.

Limitations

CREPE has several limitations. The model requires a GPU for efficient inference, which may restrict use in low-resource environments compared to lightweight rule-based metrics. Although inference is fast, the creation of synthetic training data involves substantial computation and large LLMs, which can limit reproducibility. Performance on multiple public benchmarks is strong, yet accuracy on out-of-distribution modalities and reporting styles is not guaranteed.

Another limitation is the reliance on LLM-based synthetic annotations rather than large-scale expert curation, which may introduce biases; radiologist validation is a key next step. In terms of scoring, CREPE currently aggregates errors with an unweighted sum, which does not reflect clinical priorities, though per-category outputs allow custom weighting. Finally, while all corpora used here fit within 512 tokens, longer reports in other domains may require chunking or longer-context encoders.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. RS-2024-00463651) and partly by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00454977).

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2623–2631, New York, NY, USA. Association for Computing Machinery.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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.

Amos Rubin Calamida, Farhad Nooralahzadeh, Morteza Rohanian, Mizuho Nishio, Koji Fujimoto, and Michael Krauthammer. 2024. Radiology Report Generation Models Evaluation Dataset For Chest X-rays (RadEvalX).

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. *arXiv preprint*. ArXiv:2401.12208 [cs].

Jean-Benoit Delbrouck, Pierre Chambon, Zhihong Chen, Maya Varma, Andrew Johnston, Louis Blankemeier, Dave Van Veen, Tan Bui, Steven Truong, and

Curtis Langlotz. 2024. RadGraph-XL: A Large-Scale Expert-Annotated Dataset for Entity and Relation Extraction from Radiology Reports. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12902–12915, Bangkok, Thailand. Association for Computational Linguistics.

Google. 2025. Gemini 2.5: Our most intelligent AI model.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare*, 3(1):2:1–2:23.

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 [cs].

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong N. Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew Lungren, Andrew Ng, Curtis Langlotz, and Pranav Rajpurkar. 2021. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1.

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

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2023. Prometheus: Inducing Fine-Grained Evaluation Capability in Language Models. In *12th International Conference on Learning Representations (ICLR 2024)*.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xiaohong Liu, Hao Liu, Guoxing Yang, Zeyu Jiang, Shuguang Cui, Zhaoze Zhang, Huan Wang, Liyuan Tao, Yongchang Sun, Zhu Song, Tianpei Hong, Jin Yang, Tianrun Gao, Jiangjiang Zhang, Xiaohu Li, Jing Zhang, Ye Sang, Zhao Yang, Kanmin Xue, and 5 others. 2025. A generalist medical language model for disease diagnosis assistance. *Nature Medicine*, 31(3):932–942. Publisher: Nature Publishing Group.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. 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.

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.

John Pavlopoulos, Vasiliki Kougia, and Ion Androutsopoulos. 2019. A survey on biomedical image captioning. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 26–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew 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, Online. Association for Computational Linguistics.

Katherine Tian, Sina J Hartung, Andrew A Li, Jaehwan Jeong, Fardad Behzadi, Juan Calle-Toro, Subathra Adithan, Michael Pohlen, David Osayande, and Pranav Rajpurkar. 2023. ReFiSco: Report Fix and Score Dataset for Radiology Report Generation.

Guangyu Wang, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, Hongmei Yan, Yuxing Lu, Yuanxu Gao, Kanmin Xue, Xiaoying Li, and Ying Chen. 2023. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*, 29(10):2633–2642. Publisher: Nature Publishing Group.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint*. ArXiv:1910.03771 [cs].

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. 2023a. Evaluating progress in automatic chest X-ray radiology report generation. *Patterns*, 4(9):100802.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Lee, Zahra Shakeri, Andrew Ng, Curtis Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. 2023b. Radiology Report Expert Evaluation (ReXVal) Dataset.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations (ICLR 2020)*.

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.

A Appendix

A.1 Qualitative Case Studies

To complement the quantitative evaluation, we present several qualitative examples from the ReXVal dataset that illustrate how CREPE behaves in practice (Figure 8). Each panel compares the reference and candidate reports and the corresponding distribution of errors over categories A–F.

For clarity, the six error categories are as follows:

(A) False prediction of finding

(B) Omission of finding

(C) Incorrect location or position of finding

(D) Incorrect severity of finding

(E) Mention of comparison not present in the reference impression

(F) Omission of comparison describing a change from a previous study

Successful Case 1 (post-lobectomy). The candidate omits the post–left upper lobectomy context present in the reference. Both the radiologist and CREPE assign most of the penalty to category B, indicating a clinically meaningful omission. CREPE adds small mass in categories E and F, reflecting minor surface or phrasing differences, but the dominant signal aligns with the expert assessment.

Successful Case 2 (esophageal findings and aspiration). The reference documents a dilated distal esophagus and directs the reader to prior CT details, while the candidate emphasizes "no aspiration," describes intraluminal air, and proposes a barium swallow. Both the radiologist and CREPE assign high error counts in A and B, capturing that the candidate both under-reports salient findings and introduces unsupported content. The tails in C–F are small and closely track each other, showing that CREPE follows the expert's overall error profile in a mixed-error setting.

Failure Case 1 (resolving pneumonia). Both reports agree on improved aeration of the left lower lobe; the candidate additionally states "no pleural effusion," which is not discussed in the reference. The radiologist marks minimal penalties, primarily in A and B. CREPE, however, amplifies the errors, assigning a stronger penalty in B. This overprediction reflects that the model is sensitive to seemingly innocuous additions and tends to over-penalize content which does not materially alter the clinical interpretation.

Failure Case 2 (pneumothorax with communication details). The key clinical finding, "a small right apical pneumothorax", is consistent across reports; the remaining differences are templated metadata about communication (roles, times). The radiologist assigns only a small penalty (mainly in A), recognizing that clinical content is preserved. CREPE, in contrast, spreads penalties across multiple categories (A–E), overestimating errors in what are essentially administrative details. This represents another failure case, where the model overpredicts categories in response to non-clinical discrepancies. Addressing such cases may require masking boilerplate fields.

Across these cases, CREPE generally matches the radiologist on the clinically central errors (notably omissions and contradictions). Its main weaknesses appear in failure cases, where the model overpredicts certain categories, either due to sensitivity to benign additions or to metadata that carries no clinical relevance. These examples illustrate both the strengths of CREPE in capturing clinically aware discrepancies and the areas where refinement is needed to further align with expert judgment.

A.2 Details for Synthetic Data Generation

We provide here the details and prompts used to generate the synthetic training data for the CREPE model.

A.2.1 Candidate Report Generation

To generate candidate reports, we utilized `StanfordAIMI/CheXagent-8b` from the Hugging Face Hub (Wolf et al., 2020) with default inference settings. From the MIMIC-CXR training split, we randomly sampled 32,000 examples. For each sample, the 'FINDINGS' section was used as the reference report; if this section was unavailable, we instead used the 'IMPRESSION' section. The candidate report was generated by providing the associated chest X-ray image along with a standardized prompt.

```
{image} Provide a radiological report for the following image. ASSISTANT:
```

A.2.2 Error Count Data Generation

To obtain error counts, we used `gemini-2.5-pro-preview-03-25` accessed via Vertex AI on Google Cloud, again with default parameters (temperature 1.0, topP 0.95, candidateCount 1). For each reference–candidate pair, the model was prompted to compare the candidate report against the reference, following explicit instructions to assess both clinically significant and clinically insignificant errors across six predefined categories. The prompt also required the model to return a structured output, detailing error counts and explanations for each category, as well as matched findings between the reports.

```
Objective:
Evaluate the accuracy of a candidate radiology report in comparison to a reference radiology report
composed by expert radiologists.

Process Overview:
You will be presented with:
1. The criteria for making a judgment.
2. The reference radiology report.
3. The candidate radiology report.
4. The desired format for your assessment.

1. Criteria for Judgment:
For each candidate report, determine:
    - The count of clinically significant errors.
    - The count of clinically insignificant errors.

Errors can fall into one of these categories:
    a) False report of a finding in the candidate.
    b) Missing a finding present in the reference.
    c) Misidentification of a finding's anatomic location/position.
    d) Misassessment of the severity of a finding.
    e) Mentioning a comparison that isn't in the reference.
    f) Omitting a comparison detailing a change from a prior study.

Note: Concentrate on the clinical findings rather than the report's writing style.
Evaluate only the findings that appear in both reports.

2. Reference Report:
{reference_report}

3. Candidate Report:
{candidate_report}

4. Reporting Your Assessment:
Follow this specific format for your output, even if no errors are found:
Do NOT abbreviate the name of the error type.
```
```
[Explanation]:
<Explanation>

[Clinically Significant Errors]:
(a) <Error Type>: <The number of errors>. <Error 1>; <Error 2>; ...; <Error n>
....
(f) <Error Type>: <The number of errors>. <Error 1>; <Error 2>; ...; <Error n>

[Clinically Insignificant Errors]:
(a) <Error Type>: <The number of errors>. <Error 1>; <Error 2>; ...; <Error n>
....
(f) <Error Type>: <The number of errors>. <Error 1>; <Error 2>; ...; <Error n>

[Matched Findings]:
<The number of matched findings>. <Finding 1>; <Finding 2>; ...; <Finding n>
```

A.3 Dataset Details and Processing

We provide detailed descriptions of each benchmark used in our evaluation. Figure 7 visualizes the distribution of total error counts across all datasets.

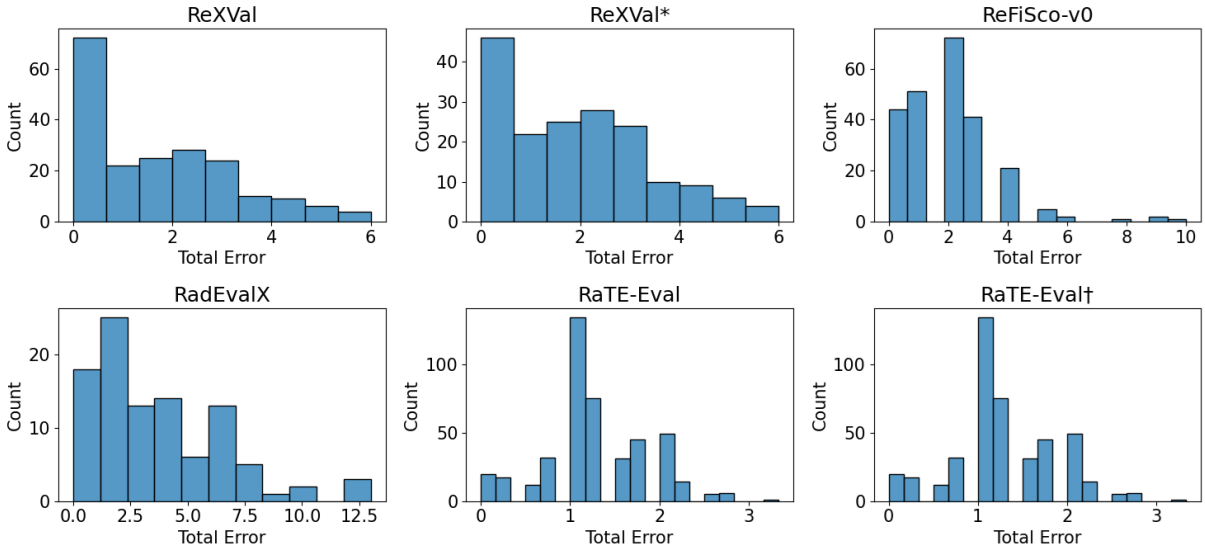


Figure 7: **Total error count distributions for all evaluation datasets.**

A.3.1 ReXVal

ReXVal contains radiologist-annotated errors for generated reports compared to ground-truth reports from the MIMIC-CXR dataset. Six radiologists independently evaluated both clinically significant and insignificant errors, assigning counts for each of six predefined error categories. For each ground-truth report, four candidate reports were generated by different automated methods. To address class imbalance, we also evaluate on ReXVal*, a variant where pairs with identical reference and candidate reports are removed.

A.3.2 ReFiSco-v0

ReFiSco-v0 provides line-level radiologist error annotations for MIMIC-CXR reports, categorizing each line as 'No error', 'Not actionable', 'Actionable nonurgent error', 'Urgent error', or 'Emergent error.' We map 'No error' to zero and all other categories to one, then aggregate the binary error labels across lines to produce report-level error counts.

A.3.3 RadEvalX

RadEvalX consists of 100 reports sampled from the IU-Xray dataset, selected to balance normal and abnormal findings. Each report and its generated counterpart were annotated by experts for six standard error categories, consistent with ReXVal, as well as two additional categories related to uncertainty (mention or omission of uncertainty). For our evaluation, we sum error counts across all categories.

A.3.4 RaTE-Eval

For RaTE-Eval, we use the sentence-level human rating benchmark, which spans nine imaging modalities and 22 anatomical regions, representing a multi-modality and out-of-distribution test case. Annotators counted errors per sentence, and scores are normalized by the number of potential error opportunities. For consistency, we also report results using the raw, unnormalized error counts, denoted as RaTE-Eval†

A.3.5 Additional Notes

All datasets were used in accordance with their respective data use agreements and ethical guidelines. Where necessary, we standardized error category definitions across datasets for consistency in evaluation.

A.4 Details for BERT Encoder Backbones

We summarize here the key characteristics and pretraining corpora for each encoder backbone used in our experiments.

A.4.1 `medicalai/ClinicalBERT`

`ClinicalBERT` is a domain-adapted BERT model initially trained on general English text and then further pre-trained on a large corpus of de-identified clinical notes. Its training corpus encompasses approximately 1.2 billion words of clinical narratives, including diverse disease phenotypes and a wide range of free-text observations from electronic health records. The model is designed for masked language modeling and fine-tuned for downstream clinical NLP tasks such as information extraction, symptom detection, and temporal representation of patient trajectories. `ClinicalBERT` has demonstrated high performance in extracting clinically relevant information from free-text notes, with an average F1 score of 94.5% in symptom extraction tasks on annotated samples.

A.4.2 `emilyalsentzer/Bio_ClinicalBERT`

`Bio_ClinicalBERT` is based on the BERT architecture, initialized from BioBERT, and further pre-trained on approximately two million clinical notes from the MIMIC-III database. This corpus includes a comprehensive range of note types, including both discharge summaries and general clinical narratives. The pre-training process follows standard masked language modeling objectives and leverages the full MIMIC corpus for greater domain coverage. `Bio_ClinicalBERT` embeddings are intended as a community resource for downstream medical NLP tasks. Pre-training required significant computational resources, estimated at 17–18 days on a high-end GPU.

A.4.3 `microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract`

`BiomedBERT`, previously named `PubMedBERT`, used as the primary encoder backbone for CREPE, is pre-trained on large-scale biomedical text corpora, including PubMed abstracts and clinical literature. The model employs the standard BERT-base architecture and is optimized for masked language modeling to capture domain-specific biomedical terminology and semantics. For our experiments, `BiomedBERT` was further fine-tuned on the CREPE training set for the specific task of clinical error count regression.

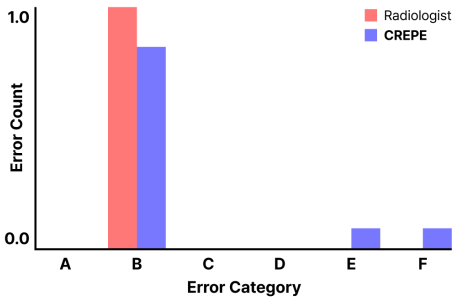
Metric	ReXVal		ReXVal*		ReFiSco-v0	
	τ	ρ	τ	ρ	τ	ρ
BLEU-4	0.383 [0.274, 0.482]	0.516 [0.371, 0.633]	0.215 [0.104, 0.333]	0.294 [0.144, 0.447]	0.489 [0.408, 0.561]	0.616 [0.520, 0.692]
ROUGE-L	0.570 [0.490, 0.636]	0.748 [0.661, 0.809]	0.459 [0.373, 0.537]	0.633 [0.523, 0.716]	0.524 [0.443, 0.602]	0.662 [0.569, 0.742]
METEOR	0.484 [0.400, 0.558]	0.653 [0.546, 0.738]	0.355 [0.260, 0.444]	0.497 [0.369, 0.606]	0.468 [0.388, 0.538]	0.617 [0.514, 0.694]
BERTScore	0.521 [0.441, 0.598]	0.694 [0.597, 0.776]	0.404 [0.313, 0.497]	0.558 [0.438, 0.668]	0.541 [0.467, 0.606]	0.689 [0.604, 0.759]
F1 RadGraph	0.585 [0.512, 0.650]	0.765 [0.686, 0.823]	0.484 [0.401, 0.563]	0.661 [0.562, 0.742]	0.475 [0.392, 0.547]	0.609 [0.509, 0.690]
SembScore	0.495 [0.416, 0.574]	0.666 [0.571, 0.751]	0.368 [0.271, 0.455]	0.513 [0.388, 0.621]	0.461 [0.387, 0.538]	0.605 [0.512, 0.691]
RaTEScore	0.520 [0.439, 0.589]	0.697 [0.601, 0.770]	0.408 [0.319, 0.495]	0.564 [0.448, 0.667]	0.433 [0.356, 0.504]	0.571 [0.472, 0.654]
RadCliQ-v1	0.623 [0.566, 0.676]	0.809 [0.749, 0.855]	0.540 [0.475, 0.602]	0.730 [0.654, 0.791]	0.510 [0.433, 0.576]	0.656 [0.564, 0.729]
GREEN	0.626 [0.555, 0.685]	0.798 [0.729, 0.843]	0.541 [0.459, 0.614]	0.713 [0.617, 0.789]	0.592 [0.518, 0.663]	0.709 [0.632, 0.781]
GREEN EC	0.775 [0.728, 0.814]	0.899 [0.861, 0.924]	0.729 [0.667, 0.776]	0.864 [0.807, 0.900]	**0.723** [0.660, 0.780]	0.811 [0.744, 0.866]
CREPE	**0.786** [0.749, 0.816]	**0.933** [0.907, 0.949]	**0.753** [0.703, 0.794]	**0.911** [0.871, 0.937]	0.697 [0.640, 0.747]	**0.825** [0.761, 0.873]
Metric	**RadEvalX**		**RaTE-Eval**		**RaTE-Eval†**	
	τ	ρ	τ	ρ	τ	ρ
BLEU-4	0.074 [-0.092, 0.231]	0.096 [-0.120, 0.301]	0.197 [0.119, 0.270]	0.247 [0.151, 0.339]	0.134 [0.060, 0.208]	0.166 [0.075, 0.255]
ROUGE-L	0.257 [0.111, 0.382]	0.356 [0.155, 0.519]	0.200 [0.136, 0.260]	0.281 [0.192, 0.362]	0.220 [0.146, 0.284]	0.302 [0.201, 0.387]
METEOR	0.201 [0.065, 0.334]	0.284 [0.100, 0.458]	0.174 [0.105, 0.244]	0.245 [0.150, 0.340]	0.248 [0.183, 0.310]	0.338 [0.251, 0.420]
BERTScore	0.326 [0.195, 0.442]	0.452 [0.274, 0.596]	0.224 [0.160, 0.284]	0.315 [0.226, 0.395]	0.256 [0.196, 0.316]	0.351 [0.270, 0.432]
F1 RadGraph	0.171 [0.053, 0.294]	0.243 [0.070, 0.409]	0.306 [0.235, 0.374]	0.393 [0.304, 0.476]	0.258 [0.185, 0.329]	0.328 [0.235, 0.418]
SembScore	0.318 [0.165, 0.447]	0.434 [0.228, 0.593]	0.198 [0.134, 0.258]	0.280 [0.190, 0.360]	0.245 [0.186, 0.308]	0.336 [0.256, 0.421]
RaTEScore	0.316 [0.192, 0.435]	0.438 [0.270, 0.583]	0.339 [0.280, 0.396]	**0.460** [0.379, 0.534]	0.310 [0.250, 0.369]	0.419 [0.340, 0.494]
RadCliQ-v1	0.326 [0.206, 0.459]	0.449 [0.288, 0.609]	0.299 [0.238, 0.357]	0.415 [0.332, 0.492]	0.304 [0.247, 0.363]	0.414 [0.338, 0.488]
GREEN	0.411 [0.308, 0.517]	0.539 [0.408, 0.661]	**0.374** [0.303, 0.439]	0.457 [0.371, 0.532]	0.409 [0.343, 0.470]	0.494 [0.419, 0.562]
GREEN EC	0.448 [0.341, 0.546]	0.577 [0.441, 0.686]	0.252 [0.184, 0.320]	0.315 [0.229, 0.396]	**0.432** [0.369, 0.495]	0.517 [0.443, 0.585]
CREPE	**0.580** [0.477, 0.669]	**0.745** [0.625, 0.826]	0.267 [0.202, 0.331]	0.375 [0.288, 0.457]	0.407 [0.349, 0.466]	**0.541** [0.468, 0.611]

Table 5: **Correlation with Radiologist Error Counts and Confidence Intervals.** Kendall's τ and Spearman's ρ correlation coefficients, with 95% confidence intervals, for each evaluation metric on six datasets. Bold indicates the best and underline the second-best values for each metric and dataset.

Successful Case 1

Reference Report
Post left upper lobectomy changes, with no superimposed acute intrathoracic process detected.

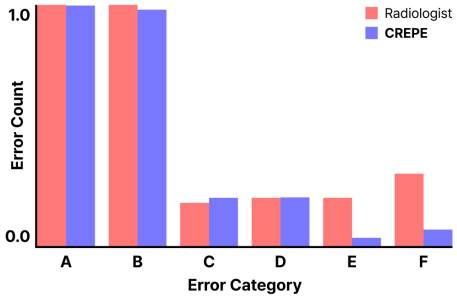
Candidate Report
No acute intrathoracic process detected.



Successful Case 2

Reference Report
Dilated distal esophagus as seen previously containing ingested food contents. No signs of aspiration. Please refer to prior CT torso for full descriptive details of esophageal abnormalities.

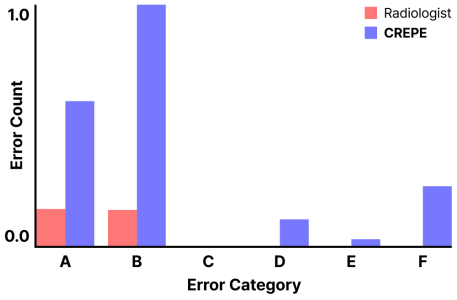
Candidate Report
No aspiration. Air in the esophagus in multiple locations. Further assessment with a barium swallow may be considered.



Failure Case 1

Reference Report
Improved aeration of the left lower lobe suggesting resolving pneumonia.

Candidate Report
Improved aeration of the left lower lobe with no pleural effusion.



Failure Case 2

Reference Report
Small right apical pneumothorax. Findings were discussed with Dr. ___ by Dr. ___ by telephone on ___ at 10:40 a.m., time of discovery 10:35 a.m.

Candidate Report
Small right apical pneumothorax. These findings were discussed with ___, NP by Dr. ___ via telephone on ___ at 12:22 p.m., at time of discovery.

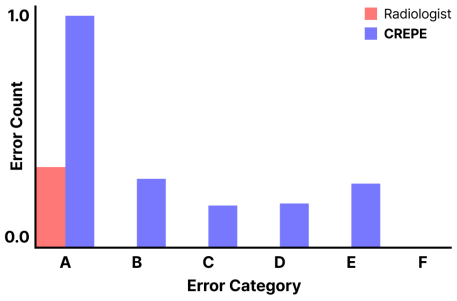


Figure 8: **Qualitative Case Studies.** Examples derived from ReXVal. The reference (ground truth) and candidate reports, and the error counts across categories A–F from a board-certified radiologist and from CREPE.