ReXrank: A Public Leaderboard for AI-Powered Radiology Report Generation

Xiaoman Zhang[1], **Hong-Yu Zhou**[1], **Xiaoli Yang**[1], **Oishi Banerjee**[1], **Julián N. Acosta**[1], **Josh Miller**[2], **Ouwen Huang**[2,3], **Pranav Rajpurkar**[1]

[1]Department of Biomedical Informatics, Harvard University, USA
[2]Gradient Health, Durham, NC, USA
[3]Department of Statistical Science, Duke University, Durham, NC, USA

Abstract

AI-driven models have demonstrated significant potential in automating radiology report generation for chest X-rays. However, there is no standardized benchmark for objectively evaluating their performance. To address this, we present **ReXrank**, a public leaderboard and challenge for assessing AI-powered radiology report generation. Our framework incorporates **ReXGradient**, the largest test dataset consisting of **10,000** studies, and three public datasets (MIMIC-CXR, IU-Xray, CheXpert Plus) for report generation assessment. ReXrank employs 8 evaluation metrics and separately assesses models capable of generating only findings sections and those providing both findings and impressions sections. By providing this standardized evaluation framework, ReXrank enables meaningful comparisons of model performance and offers crucial insights into their robustness across diverse clinical settings. Beyond its current focus on chest X-rays, ReXrank's framework sets the stage for comprehensive evaluation of automated reporting across the full spectrum of medical imaging.

Introduction

Writing accurate radiology reports from medical images is a critical but complex task, requiring both deep expertise in medical imaging and the ability to accurately interpret and articulate intricate findings. The demand for such reports has surged with the rapid advancements in imaging technologies, leading to increased workloads for radiologists, risks of information loss, and longer report turnaround times (Bruls and Kwee 2020).

AI-driven solutions have emerged as a potential answer to these challenges, serving as assistive tools to enhance reporting efficiency and ensure access to high-quality, specialty-level interpretations. Medical visual-language models have shown promise in automating the generation of radiology reports from chest X-ray images (Chen et al. 2024; Pellegrini et al. 2023; Chen et al. 2023; Lee et al. 2023; Zhou et al. 2024). However, as the field of AI-assisted medical reporting rapidly evolves, there is a growing need for standardized benchmarks to objectively assess and compare the performance of these models. Existing datasets for chest X-ray report generation, such as MIMIC-CXR (Johnson et al. 2019), are valuable but exhibit limitations that hinder their effectiveness for benchmarking. These datasets frequently suffer from inconsistent data splits and a lack of standardized metrics during evaluation, which impedes reliable comparative analysis across different model architectures. Furthermore, the data distribution in MIMIC-CXR, commonly used in model training, fails to adequately test the models' ability to generalize to new, unseen distributions. To fill this gap, we introduce **ReXrank**, a public leaderboard and challenge specifically designed for evaluating AI-powered radiology report generation from chest X-ray images.

ReXrank offers a comprehensive evaluation framework that sets a standardized benchmark for assessing the effectiveness of different radiology report generation models (Figure 1). To ensure robust and clinically relevant evaluations, it integrates diverse datasets, including MIMIC-CXR (Johnson et al. 2019), IU-Xray (Demner-Fushman et al. 2016), CheXpert Plus (Chambon et al. 2024), and ReXGradient, a large-scale private dataset of 10,000 studies. This broad dataset spectrum allows us to evaluate model performance on data with varying distributions, providing deeper insights into the models' generalization capabilities. Furthermore, ReXrank implements various report evaluation metrics, including BLEU-2 (Papineni et al. 2002), BERTScore (Zhang et al. 2019), SembScore (Smit et al. 2020), RadGraph-F1 (Yu et al. 2023), RadCliQ (Yu et al. 2023), RaTEScore (Zhao et al. 2024), GREEN (Ostmeier et al. 2024), FineRadScore (Huang et al. 2024), etc., to offer a detailed view of each model's strengths and weaknesses. This comprehensive approach enables a more nuanced understanding of model performance and facilitates meaningful comparisons between different radiology report generation systems. Our contributions can be summarized as follows:

- We introduce ReXrank, the first public leaderboard designed for evaluating AI-powered radiology report generation models.
- We curate a new benchmark dataset, ReXGradient, comprising 10,000 studies from 67 US institutions, enabling robust assessments of model generalization and resilience to distributional shifts.
- We establish a comprehensive evaluation framework that incorporates 8 diverse evaluation metrics to assess 16

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

state-of-the-art report generation models and continuously supports model submissions for benchmarking.

Method

In this section, we first introduce the datasets used in ReXrank, comprising three public datasets (MIMIC-CXR, IU-Xray, and CheXpert Plus) and one private dataset ReXGradient, along with our standardized data format for consistent evaluation. We then describe our evaluation framework that employs eight metrics spanning linguistic quality and clinical accuracy and present the 16 state-of-the-art report generation models included in our benchmark, representing diverse architectural approaches from traditional encoder-decoder frameworks to recent medical-specific large language models.

Datasets

- **ReXGradient.** This private test set is provided by Gradient Health, which consists of 10,000 studies collected from 7,004 patients across 67 medical sites in the United States.
- **MIMIC-CXR (Johnson et al. 2019).** This is a large, publicly accessible dataset comprising 377,110 chest X-rays (CXRs) corresponding to 227,835 radiographic studies performed at the Beth Israel Deaconess Medical Center in Boston, MA. For this dataset, we extracted sections of indication, comparison, findings, and impression via keyword matching. We follow the official split of MIMIC-CXR in our experiments and report scores on the test set, which consists of 2,347 studies.
- **IU-Xray (Demner-Fushman et al. 2016).** This is a publicly accessible dataset containing 7,470 pairs of CXRs and radiology reports. Each study in this dataset includes one frontal and one lateral CXR, associated with a single radiology report. We follow the split provided by R2Gen (Chen et al. 2020) and report scores on the test set, which comprises 590 studies.
- **CheXpert Plus (Chambon et al. 2024).** This is a large, publicly accessible consisting of 223,462 unique pairs of radiology reports and chest X-rays. These correspond to 187,711 radiographic studies from 64,725 patients. We follow the official split of CheXpert Plus and report scores on the validation set, which contains 200 studies.

Data format

For each study in our test datasets, the data is organized in a structured format.

- `id`: Unique identifier for the study
- `image_path`: List of paths to all relevant chest X-ray images
- `frontal_lateral`: List indicating the view type of each image
- `key_image_path`: Path to the primary image (typically frontal view)
- `context`: Patient information and clinical context
- `report`: Radiologist's findings and impressions

Evaluation Metrics

- **BLEU-2 (Papineni et al. 2002).** BLEU (Bilingual Evaluation Understudy) is a widely used metric in machine translation and text generation tasks. It evaluates the quality of generated text by comparing n-gram precision between the candidate and reference texts, with scores ranging from 0 to 1. In this work, we specifically use BLEU-2, which focuses on bigram precision to assess the quality of the generated text.
- **BERTScore (Zhang et al. 2019).** BERTScore is a neural metric that uses pre-trained BERT models (Kenton and Toutanova 2019) to evaluate text similarity. It computes cosine similarity between the BERT embeddings of model-generated and groundtruth radiology reports.
- **SembScore (Smit et al. 2020).** SembScore (CheXbert labeler vector similarity) is a domain-specific metric for radiology report evaluation. It computes the cosine similarity between the indicator vectors of 14 pathologies that the CheXbert automatic labeler extracts from model-generated and groundtruth radiology reports
- **RadGraph-F1 (Yu et al. 2023).** RadGraph-F1 is a metric for radiology report evaluation. It computes the overlap in clinical entities and relations that RadGraph (Jain et al. 2021) extracts from candidate and reference reports.
- **1/RadCliQ-v1 (Yu et al. 2023).** RadCliQ is a composite metric designed for evaluating radiology report generation, combining BLEU, BERTScore, SembScore, and RadGraph-F1 to provide a comprehensive assessment of generated reports. For our evaluation, we utilized the official implementation[1], which also includes BLEU, BERTScore, SembScore, RadGraph-F1. While the original RadCliQ metric is designed as lower-is-better, we first calculate the average RadCliQ-v1 score for each model across the dataset, then take its reciprocal (1/RadCliQ-v1) to maintain consistency with other metrics where higher values indicate better performance.
- **RaTEScore (Zhao et al. 2024).** RaTEScore is an entity-aware metric for radiology report evaluation. It emphasizes crucial medical entities like diagnostic outcomes and anatomical details and is robust against complex medical synonyms and sensitive to negation expressions. For our evaluation, we utilized the official package[2].
- **GREEN (Ostmeier et al. 2024).** GREEN (Generative Radiology Report Evaluation and Error Notation) is an LLM-based metric for evaluating radiology report generation. It leverages language models to identify and explain clinically significant errors in quantitative and qualitative candidate reports. We utilized the official implementation[3] for our evaluation.
- **1/FineRadScore (Huang et al. 2024).** FineRadScore is an automated evaluation metric leveraging an LLM to assess the quality of generated chest X-ray reports. It determines the minimum number of line-by-line corrections

[1] https://github.com/rajpurkarlab/CXR-Report-Metric
[2] https://github.com/MAGIC-AI4Med/RaTEScore
[3] https://github.com/Stanford-AIMI/GREEN

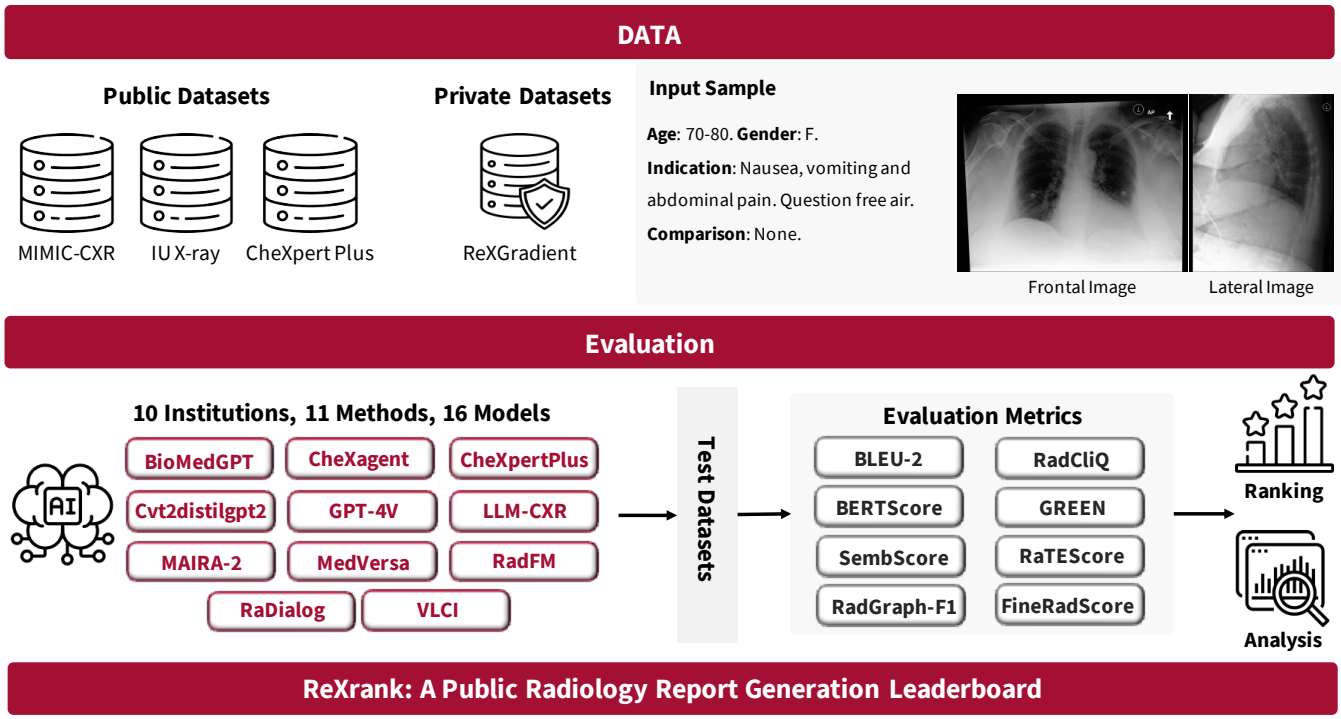

Figure 1: An illustration of ReXrank, a public leaderboard and challenge for AI-powered radiology report generation from chest X-ray images. ReXrank supports model submissions and evaluates them on both public datasets and a large-scale private dataset, providing comprehensive rankings of all submitted models.

needed, with severity ratings from 1 to 4, to transform a candidate report into a ground-truth report. For evaluation, we employ GPT-4o as the LLM and use the official implementation[4]. To obtain a single FineRadScore for each report, we take the maximum clinical severity across all lines. Similar to 1/RadCliQ-v1, we present 1/FineRadScore to maintain consistency where higher values indicate better performance.

Statistical Analysis. For all metrics, we report results as mean ± 95% confidence interval (CI). The CIs are calculated by assuming a normal distribution of the data, using a Z-score of 1.96 multiplied by the standard error of the mean. This approach provides a range that encapsulates the true average value with 95% certainty, enabling reliable comparison of model performance

Participating Models

- **BiomedGPT_IU (Zhang et al. 2024).** BiomedGPT is a lightweight, open-source vision-language foundation model designed for diverse biomedical tasks across modalities. The model was fine-tuned for VQA and image captioning tasks using multiple datasets, including radiology and pathology data. BiomedGPT_IU is fine-tuned on the IU X-ray dataset for image captioning tasks. In our evaluation, we used the publicly available checkpoints trained on the IU-Xray dataset[5].
- **CheXagent (Chen et al. 2024).** CheXagent is an instruction-tuned foundation model specifically designed for chest X-ray interpretation. The model consists of a vision encoder for representing CXR images, and a network to bridge the vision and language modalities. This model is trained on CheXinstruct, a large-scale instruction-tuning dataset curated from 28 publicly-available datasets. For our evaluation, we utilized the publicly available 8 billion parameter checkpoint from Hugging Face[6].
- **CheXpertPlus_CheX (Chambon et al. 2024).** CheXpertPlus_CheX, introduced in the CheXpert Plus paper, utilizes a Swinv2 (Liu et al. 2022) architecture with a two-layer BERT decoder (Kenton and Toutanova 2019) for medical report generation. CheXpertPlus_CheX is trained exclusively on the CheXpert Plus dataset. In our evaluation, we utilized the publicly available Findings Checkpoint[7] and Impression Checkpoint[8]. Our evaluation employs these models sequentially, generating the findings and impression sections separately, and then combining them with appropriate headers to form the complete report.
- **CheXpertPlus_CheX_MIMIC (Chambon et al. 2024).** CheXpertPlus_CheX_MIMIC shares the same architectural design as CheXpertPlus_MIMIC. CheXpert-

[4]https://github.com/rajpurkarlab/FineRadScore
[5]https://github.com/taokz/BiomedGPT
[6]https://huggingface.co/StanfordAIMI/CheXagent-8b
[7]https://huggingface.co/IAMJBchexpert-findings-baseline
[8]https://huggingface.co/IAMJB/chexpert-impression-baseline

Plus_CheX is trained exclusively on the combination of the MIMIC-CXR and CheXpert Plus dataset. In our evaluation, we utilized the publicly available Findings Checkpoint[9] and Impression Checkpoint[10]. Our evaluation employs these models sequentially, generating the findings and impression sections separately, and then combining them with appropriate headers to form the complete report.

- **CheXpertPlus_MIMIC (Chambon et al. 2024).** CheXpertPlus_MIMIC shares the same architectural design as CheXpertPlus_CheX. CheXpertPlus_MIMIC comprises two distinct models trained on MIMIC-CXR: one for findings and another for impressions. In our evaluation, we utilized the publicly available Findings Checkpoint[11] and Impression Checkpoint[12]. Our evaluation employs these models sequentially, generating the findings and impression sections separately, and then combining them with appropriate headers to form the complete report.
- **Cvt2distilgpt2_IU (Nicolson, Dowling, and Koopman 2023).** CvT2DistilGPT2_IU employs a hybrid architecture combining a Convolutional vision Transformer (CvT) (Wu et al. 2021a) pre-trained on ImageNet-21K (Deng et al. 2009) with a DistilGPT2 (Alfarghaly et al. 2021) decoder for chest X-ray report generation. This model leverages the CvT's efficient hierarchical design for image feature extraction and DistilGPT2's natural language generation capabilities. In our evaluation, we utilized the publicly available checkpoint from Github[13] and followed the official evaluation guidelines.
- **Cvt2distilgpt2_MIMIC (Nicolson, Dowling, and Koopman 2023).** Cvt2distilgpt2_MIMIC applies the same Cvt2distilgpt2 architecture but is trained on the MIMIC-CXR dataset. Our evaluation utilized the publicly available checkpoints from Github.
- **RGRG (Tanida et al. 2023).** RGRG (Region-Guided Radiology Report Generation) employs object detection to extract localized visual features from 29 anatomical regions in chest X-rays. It uses binary classifiers to select salient features and encode abnormalities, followed by a language model generating sentences for each selected region. RGRG was trained on the Chest ImaGenome dataset (Wu et al. 2021b). In our evaluation, we utilized the publicly available checkpoint from github[14] and followed the official evaluation guidelines.
- **RaDialog (Pellegrini et al. 2023).** RaDialog is a large vision-language model for radiology report generation and interactive dialogue. It integrates visual image features and structured pathology findings with a large language model (LLM), adapted to radiology using parameter-efficient fine-tuning. RaDialog was trained on the MIMIC-CXR for radiology report generation tasks. In our evaluation, we utilized the publicly available LLaVA version checkpoint from huggingface[15] and followed the official evaluation guidelines.
- **GPT-4V (Yang et al. 2023).** GPT-4V (GPT-4 with vision) is a multimodal LLM released by OpenAI, which enables users to instruct GPT-4 to analyze image inputs provided by the user. In our evaluation, we used the API of model "gpt4o05132024" and followed the official evaluation protocols to assess its performance. The prompt we used is "You are a helpful assistant. Please generate a report for the given images, including both findings and impressions. Return the report in the following format: Findings: {} Impression: {}. ".
- **LLM-CXR (Lee et al. 2023).** LLM-CXR is a multimodal large language model that utilizes VQ-GAN to tokenize images, integrating both image and text tokens as input to its base LLM architecture. This model enables CXR-to-report generation, report-to-CXR generation, and CXR-related visual question answering (VQA). For our evaluation, we used the publicly available checkpoints[16] and followed the official evaluation guidelines.
- **MAIRA-2 (Bannur et al. 2024).** MAIRA-2 is a large multimodal model that combines a radiology-specific image encoder with a Large Language Model (LLM), trained for grounded report generation from chest X-rays. For input, the model accepts X-ray images along with indication, comparison, and technique information. For our evaluation, we used the publicly available checkpoints[17] and followed the official evaluation guidelines. For studies containing both frontal and lateral views, we input the technique that "PA and lateral views of the chest were obtained.". For studies with only frontal views, we use "PA view of the chest was obtained.".
- **MedVersa (Zhou et al. 2024).** MedVersa is a compound medical AI system that can coordinate multimodal inputs, orchestrate models and tools for varying tasks, and generate multimodal outputs. MedVersa was trained on the MIMIC-CXR training and validation dataset for medical report generation tasks. In our evaluation, we utilized the publicly available checkpoint from huggingface[18] and followed the official evaluation guidelines. The standard prompt structure we employed for report generation was "Can you provide a report of {input_image_token} with findings and impression?", where {input_image_token} represents the placeholder for the input image.
- **RadFM (Wu et al. 2023).** RadFM is a versatile radiology foundation model trained on large-scale multimodal datasets. It supports both 2D and 3D scans, multi-image input, and visual-language interleaving cases. The

[9]https://huggingface.co/IAMJB/chexpert-mimic-cxr-findings-baseline

[10]https://huggingface.co/IAMJB/chexpert-mimic-cxr-impression-baseline

[11]https://huggingface.co/IAMJB/mimic-cxr-findings-baseline

[12]https://huggingface.co/IAMJB/mimic-cxr-impression-baseline

[13]https://github.com/aehrc/cvt2distilgpt2

[14]https://github.com/ttanida/rgrg

[15]https://huggingface.co/ChantalPellegrini/RaDialog-interactive-radiology-report-generation

[16]https://github.com/hyn2028/llm-cxr

[17]https://huggingface.co/microsoft/maira-2

[18]https://huggingface.co/hyzhou/MedVersa

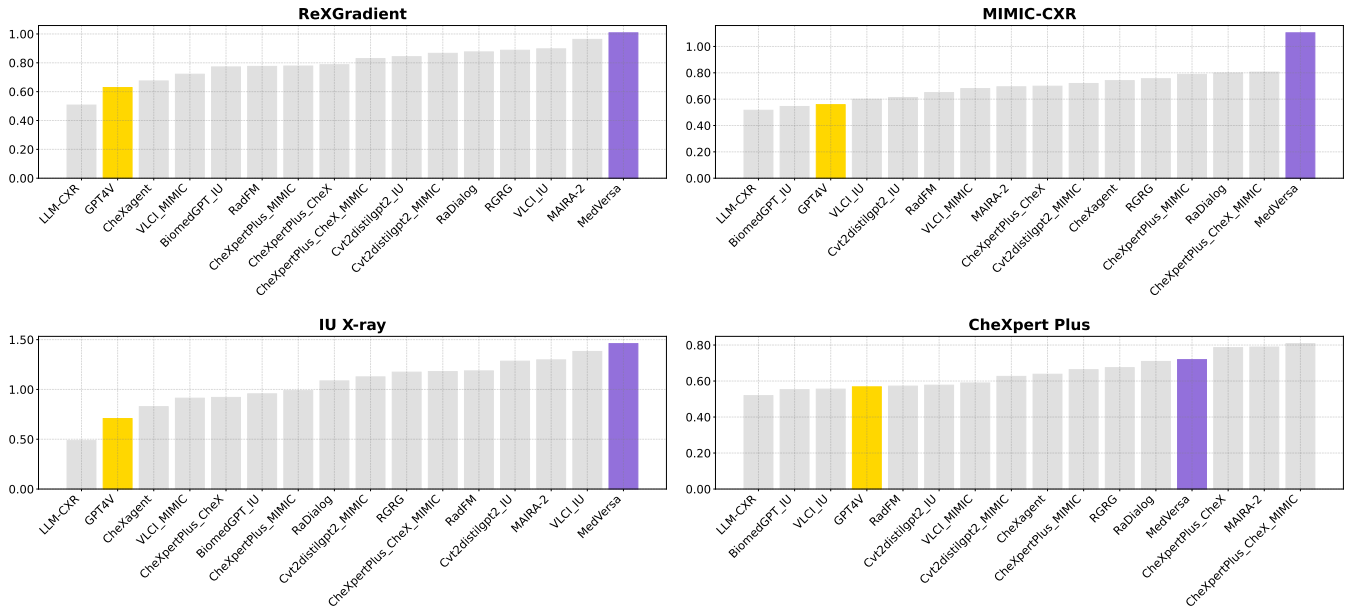

Figure 2: Comprehensive performance evaluation and ranking of report generation models based on the **1/RadCliQ-v1** metric across four distinct datasets: ReXGradient, MIMIC-CXR, IU X-ray, and CheXpert Plus. MedVersa (highlighted in purple) demonstrates consistently superior performance across all datasets, achieving significantly higher scores compared to other models, including GPT4V (highlighted in yellow).

model's training included the MIMIC-CXR dataset. For our evaluation, we utilized the publicly available checkpoint from huggingface[19] and followed the official evaluation guidelines. The prompt we employed for report generation was "Can you provide a radiology report for this medical image?".

- **VLCI_IU (Chen et al. 2023).** VLCI (Visual-Linguistic Causal Intervention) combines Visual linguistic pre-training using a multiway transformer for cross-modal alignment with Visual-linguistic causal intervention, integrating a pre-trained transformer and Visual and linguistic de-confounding Modules to mitigate cross-modal bias through local and global visual sampling and linguistic estimation using a vocabulary dictionary and visual features. In our evaluation, we used the publicly available checkpoints trained on the IU-Xray dataset[20].
- **VLCI_MIMIC (Chen et al. 2023).** VLCI_MIMIC applies the same VLCI architecture but is trained on the MIMIC-CXR dataset. Our evaluation utilized the publicly available MIMIC-CXR-trained checkpoints.

Results

Overall Model Performance. Table 1, 2, 3 and 4 summarize the performance of various medical report generation models across four datasets: ReXGradient, MIMIC-CXR, IU X-ray, and CheXpert Plus. For better visualization of model performance comparisons, we present the rankings based on 1/RadCliQ-v1 in Figure 1. MedVersa demonstrates superior performance, achieving the best 1/RadCliQ-v1 scores on ReXGradient (1.01 ± 0.01), MIMIC-CXR (1.10 ± 0.02), and IU X-ray (1.46 ± 0.03) on the findings section. This consistent top performance across different datasets indicates MedVersa's robust generalization capabilities and effectiveness in medical report generation. Notably, compared to the generalist model GPT-4V, medical-specific report generation models consistently achieve higher performance. For instance, MedVersa outperforms GPT-4V by significant margins across all datasets (60.3% on ReXGradient, 96.4% on MIMIC-CXR, and 105.6% on IU X-ray), highlights the importance of domain-specific training. Additionally, when compared between findings-only and findings + impression tasks on ReXGradient, MedVersa shows slight performance degradation when generating both findings and impressions (1/RadCliQ-v1 decreasing from 1.01 ± 0.01 to 0.98 ± 0.05), while CheXpertPlus_CheX_MIMIC shows improvement (from 0.83 ± 0.01 to 0.85 ± 0.01). This may be due to CheXpertPlus_CheX_MIMIC using separate models for findings and impression generation, while MedVersa only uses a single model architecture. The separate models may allow CheXpertPlus_CheX_MIMIC to better specialize in each subtask.

Impact of Training Strategies. Models trained on multiple datasets (e.g., CheXpertPlus_CheX_MIMIC) tend to outperform those trained on individual datasets, suggesting that a multi-dataset training approach helps bridge the distributional gap and enhances generalization. Models perform better when evaluated on the same distribution seen in training, for instance, VLCI_IU achieves superior performance on IU X-ray (RadCliQ-v1: 1.38 ± 0.04) compared

[19]https://huggingface.co/chaoyi-wu/RadFM
[20]https://github.com/WissingChen/VLCI

Table 1: Comprehensive evaluation of medical report generation models on ReXGradient. Models are ranked by 1/RadCliQ-v1. Model evaluation results with 95% confidence intervals (mean ± CI) under normality assumption. The best results for each metric are shown in **bold**.

Model	1/RadCliQ-v1 ↑	BLEU-2 ↑	BertScore ↑	SembScore ↑	RadGraph ↑	RaTEScore ↑	GREEN ↑	1/FineRadScore ↑
Findings + Impression								
MedVersa	**0.98 ± 0.01**	0.17 ± 0.01	**0.44 ± 0.01**	**0.48 ± 0.02**	**0.19 ± 0.01**	**0.53 ± 0.00**	**0.52 ± 0.01**	**0.47 ± 0.02**
CheXpertPlus_CheX_MIMIC	0.85 ± 0.01	**0.20 ± 0.00**	0.39 ± 0.00	0.43 ± 0.00	0.17 ± 0.00	0.50 ± 0.00	0.51 ± 0.01	0.47 ± 0.02
CheXpertPlus_MIMIC	0.80 ± 0.01	0.18 ± 0.00	0.36 ± 0.00	0.43 ± 0.00	0.14 ± 0.00	0.48 ± 0.00	0.52 ± 0.01	0.47 ± 0.02
CheXpertPlus_CheX	0.76 ± 0.01	0.17 ± 0.00	0.33 ± 0.00	0.40 ± 0.00	0.15 ± 0.00	0.50 ± 0.00	0.47 ± 0.01	0.42 ± 0.02
RadFM	0.74 ± 0.01	0.13 ± 0.00	0.34 ± 0.00	0.38 ± 0.00	0.13 ± 0.00	0.47 ± 0.00	0.41 ± 0.01	0.43 ± 0.02
GPT4V	0.66 ± 0.01	0.07 ± 0.00	0.21 ± 0.00	0.36 ± 0.00	0.17 ± 0.00	0.46 ± 0.00	0.36 ± 0.01	0.42 ± 0.02
Findings								
MedVersa	**1.01 ± 0.01**	0.21 ± 0.00	0.43 ± 0.00	**0.50 ± 0.00**	0.20 ± 0.00	0.53 ± 0.00	0.53 ± 0.01	**0.47 ± 0.02**
MAIRA-2	0.96 ± 0.01	0.20 ± 0.00	**0.44 ± 0.00**	0.46 ± 0.00	0.19 ± 0.00	0.56 ± 0.00	0.53 ± 0.01	**0.47 ± 0.02**
VLCI_IU	0.90 ± 0.01	**0.21 ± 0.00**	0.36 ± 0.00	0.47 ± 0.00	**0.21 ± 0.00**	**0.57 ± 0.00**	**0.54 ± 0.01**	0.45 ± 0.02
RGRG	0.89 ± 0.01	0.19 ± 0.00	0.39 ± 0.00	0.47 ± 0.00	0.17 ± 0.00	0.54 ± 0.00	0.49 ± 0.01	0.46 ± 0.02
RaDialog	0.88 ± 0.01	0.19 ± 0.00	0.40 ± 0.00	0.45 ± 0.00	0.16 ± 0.00	0.52 ± 0.00	0.43 ± 0.01	0.46 ± 0.02
Cvt2distilgpt2_MIMIC	0.87 ± 0.01	0.19 ± 0.00	0.37 ± 0.00	0.46 ± 0.00	0.18 ± 0.00	0.52 ± 0.00	0.51 ± 0.01	0.47 ± 0.02
Cvt2distilgpt2_IU	0.84 ± 0.01	0.18 ± 0.00	0.40 ± 0.00	0.41 ± 0.00	0.17 ± 0.00	0.52 ± 0.00	0.47 ± 0.01	0.46 ± 0.02
CheXpertPlus_CheX_MIMIC	0.83 ± 0.01	0.17 ± 0.00	0.37 ± 0.00	0.44 ± 0.00	0.15 ± 0.00	0.52 ± 0.00	0.49 ± 0.01	0.47 ± 0.02
CheXpertPlus_CheX	0.79 ± 0.01	0.14 ± 0.00	0.36 ± 0.00	0.43 ± 0.00	0.12 ± 0.00	0.48 ± 0.00	0.41 ± 0.01	0.41 ± 0.02
CheXpertPlus_MIMIC	0.78 ± 0.01	0.15 ± 0.00	0.34 ± 0.00	0.44 ± 0.00	0.13 ± 0.00	0.50 ± 0.00	0.52 ± 0.01	0.47 ± 0.02
RadFM	0.78 ± 0.01	0.16 ± 0.00	0.36 ± 0.00	0.39 ± 0.00	0.14 ± 0.00	0.50 ± 0.00	0.41 ± 0.01	0.44 ± 0.02
BiomedGPT_IU	0.77 ± 0.01	0.10 ± 0.00	0.32 ± 0.00	0.44 ± 0.00	0.16 ± 0.00	0.47 ± 0.00	0.39 ± 0.01	0.45 ± 0.02
VLCI_MIMIC	0.72 ± 0.01	0.16 ± 0.00	0.31 ± 0.00	0.40 ± 0.00	0.12 ± 0.00	0.49 ± 0.00	0.48 ± 0.01	0.46 ± 0.02
CheXagent	0.67 ± 0.01	0.09 ± 0.00	0.30 ± 0.00	0.37 ± 0.00	0.08 ± 0.00	0.43 ± 0.00	0.24 ± 0.01	0.46 ± 0.02
GPT4V	0.63 ± 0.01	0.07 ± 0.00	0.21 ± 0.00	0.34 ± 0.00	0.14 ± 0.00	0.47 ± 0.00	0.50 ± 0.01	0.43 ± 0.02
LLM-CXR	0.51 ± 0.01	0.04 ± 0.00	0.18 ± 0.00	0.14 ± 0.00	0.03 ± 0.00	0.32 ± 0.00	0.04 ± 0.00	0.33 ± 0.02

Table 2: Comprehensive evaluation of medical report generation models on the MIMIC-CXR datasets. Models are ranked by 1/RadCliQ-v1. Model evaluation results with 95% confidence intervals (mean ± CI) under normality assumption. The best results for each metric are shown in **bold**.

Model	1/RadCliQ-v1 ↑	BLEU-2 ↑	BertScore ↑	SembScore ↑	RadGraph ↑	RaTEScore ↑	GREEN ↑	1/FineRadScore ↑
Findings + Impression								
MedVersa	**0.92 ± 0.02**	**0.19 ± 0.00**	**0.43 ± 0.01**	0.32 ± 0.01	**0.27 ± 0.01**	**0.55 ± 0.01**	**0.42 ± 0.01**	0.36 ± 0.03
CheXpertPlus_CheX_MIMIC	0.83 ± 0.02	0.17 ± 0.00	0.36 ± 0.00	**0.39 ± 0.01**	0.20 ± 0.00	0.52 ± 0.00	0.37 ± 0.01	**0.36 ± 0.03**
CheXpertPlus_MIMIC	0.80 ± 0.02	0.17 ± 0.00	0.35 ± 0.00	0.38 ± 0.01	0.19 ± 0.00	0.51 ± 0.00	0.38 ± 0.01	**0.36 ± 0.03**
CheXpertPlus_CheX	0.71 ± 0.01	0.13 ± 0.00	0.30 ± 0.00	0.34 ± 0.01	0.17 ± 0.00	0.51 ± 0.00	0.30 ± 0.01	0.35 ± 0.03
RadFM	0.62 ± 0.02	0.08 ± 0.00	0.28 ± 0.00	0.24 ± 0.01	0.11 ± 0.00	0.45 ± 0.00	0.21 ± 0.01	0.35 ± 0.03
GPT4V	0.55 ± 0.01	0.07 ± 0.00	0.20 ± 0.00	0.19 ± 0.01	0.09 ± 0.00	0.43 ± 0.00	0.13 ± 0.01	0.33 ± 0.03
Findings								
MedVersa	**1.10 ± 0.02**	**0.21 ± 0.00**	**0.45 ± 0.01**	**0.47 ± 0.01**	**0.27 ± 0.01**	**0.55 ± 0.01**	**0.37 ± 0.01**	**0.36 ± 0.03**
CheXpertPlus_CheX_MIMIC	0.81 ± 0.02	0.14 ± 0.00	0.37 ± 0.00	0.38 ± 0.01	0.18 ± 0.01	0.49 ± 0.01	0.30 ± 0.01	0.36 ± 0.03
RaDialog	0.80 ± 0.02	0.13 ± 0.00	0.36 ± 0.00	0.39 ± 0.01	0.17 ± 0.00	0.48 ± 0.00	0.27 ± 0.01	0.36 ± 0.03
CheXpertPlus_MIMIC	0.79 ± 0.02	0.14 ± 0.00	0.36 ± 0.00	0.38 ± 0.01	0.17 ± 0.00	0.48 ± 0.01	0.31 ± 0.01	0.36 ± 0.03
RGRG	0.76 ± 0.02	0.13 ± 0.00	0.35 ± 0.00	0.34 ± 0.01	0.17 ± 0.00	0.49 ± 0.00	0.27 ± 0.01	0.35 ± 0.03
CheXagent	0.74 ± 0.02	0.11 ± 0.00	0.35 ± 0.01	0.35 ± 0.01	0.15 ± 0.00	0.47 ± 0.01	0.26 ± 0.01	0.35 ± 0.03
Cvt2distilgpt2_MIMIC	0.72 ± 0.02	0.13 ± 0.00	0.33 ± 0.01	0.33 ± 0.01	0.15 ± 0.01	0.43 ± 0.01	0.27 ± 0.01	0.36 ± 0.03
CheXpertPlus_CheX	0.70 ± 0.01	0.08 ± 0.00	0.31 ± 0.00	0.33 ± 0.01	0.14 ± 0.00	0.47 ± 0.00	0.23 ± 0.01	0.35 ± 0.03
MAIRA-2	0.69 ± 0.02	0.09 ± 0.00	0.31 ± 0.00	0.34 ± 0.01	0.13 ± 0.00	0.52 ± 0.00	0.22 ± 0.01	0.36 ± 0.03
VLCI_MIMIC	0.68 ± 0.02	0.14 ± 0.00	0.30 ± 0.00	0.30 ± 0.01	0.14 ± 0.00	0.45 ± 0.01	0.26 ± 0.01	0.36 ± 0.03
RadFM	0.65 ± 0.02	0.09 ± 0.00	0.31 ± 0.00	0.26 ± 0.01	0.11 ± 0.00	0.45 ± 0.01	0.18 ± 0.01	0.35 ± 0.03
Cvt2distilgpt2_IU	0.61 ± 0.02	0.06 ± 0.00	0.30 ± 0.00	0.19 ± 0.01	0.10 ± 0.00	0.45 ± 0.01	0.16 ± 0.01	0.35 ± 0.03
VLCI_IU	0.60 ± 0.02	0.07 ± 0.00	0.26 ± 0.00	0.21 ± 0.01	0.11 ± 0.00	0.45 ± 0.01	0.21 ± 0.01	0.35 ± 0.03
GPT4V	0.56 ± 0.01	0.07 ± 0.00	0.21 ± 0.00	0.21 ± 0.01	0.08 ± 0.00	0.42 ± 0.00	0.16 ± 0.01	0.34 ± 0.03
BiomedGPT_IU	0.54 ± 0.01	0.02 ± 0.00	0.19 ± 0.00	0.22 ± 0.01	0.06 ± 0.00	0.36 ± 0.00	0.12 ± 0.01	0.34 ± 0.03
LLM-CXR	0.52 ± 0.01	0.04 ± 0.00	0.18 ± 0.00	0.16 ± 0.01	0.05 ± 0.00	0.34 ± 0.00	0.04 ± 0.00	0.31 ± 0.03

Table 3: Comprehensive evaluation of medical report generation models on the IU X-ray datasets. Models are ranked by 1/RadCliQ-v1. Model evaluation results with 95% confidence intervals (mean ± CI) under normality assumption. The best results for each metric are shown in **bold**.

Model	1/RadCliQ-v1 ↑	BLEU ↑	BertScore ↑	SembScore ↑	RadGraph ↑	RaTEScore ↑	GREEN ↑	1/FineRadScore ↑
Findings + Impression								
MedVersa	**1.45 ± 0.04**	0.20 ± 0.01	**0.52 ± 0.01**	**0.60 ± 0.02**	**0.24 ± 0.01**	**0.63 ± 0.01**	0.66 ± 0.02	0.58 ± 0.07
CheXpertPlus_CheX_MIMIC	1.28 ± 0.04	**0.24 ± 0.01**	0.48 ± 0.01	0.60 ± 0.02	0.23 ± 0.01	0.61 ± 0.01	0.69 ± 0.02	0.59 ± 0.07
RadFM	1.23 ± 0.05	0.20 ± 0.01	0.48 ± 0.01	0.56 ± 0.02	0.23 ± 0.01	0.60 ± 0.01	0.64 ± 0.02	0.55 ± 0.08
CheXpertPlus_MIMIC	1.13 ± 0.04	0.23 ± 0.01	0.45 ± 0.01	0.59 ± 0.02	0.19 ± 0.01	0.57 ± 0.01	0.68 ± 0.02	**0.61 ± 0.07**
CheXpertPlus_CheX	1.01 ± 0.03	0.20 ± 0.01	0.39 ± 0.01	0.55 ± 0.02	0.21 ± 0.01	0.60 ± 0.01	**0.71 ± 0.02**	0.57 ± 0.07
GPT4V	0.68 ± 0.03	0.08 ± 0.00	0.23 ± 0.01	0.40 ± 0.02	0.16 ± 0.01	0.52 ± 0.01	0.40 ± 0.03	0.53 ± 0.08
Findings								
MedVersa	**1.46 ± 0.03**	0.21 ± 0.01	**0.53 ± 0.01**	0.61 ± 0.02	0.23 ± 0.01	0.65 ± 0.01	0.63 ± 0.02	0.57 ± 0.07
VLCI_IU	1.38 ± 0.04	**0.27 ± 0.01**	0.46 ± 0.01	**0.62 ± 0.02**	**0.29 ± 0.01**	**0.68 ± 0.01**	**0.70 ± 0.02**	0.55 ± 0.07
MAIRA-2	1.30 ± 0.04	0.22 ± 0.01	0.48 ± 0.01	0.60 ± 0.02	0.23 ± 0.01	0.63 ± 0.01	0.19 ± 0.02	0.60 ± 0.07
Cvt2distilgpt2_IU	1.28 ± 0.05	0.24 ± 0.01	0.48 ± 0.01	0.55 ± 0.02	0.27 ± 0.02	0.62 ± 0.01	0.69 ± 0.02	0.56 ± 0.08
RadFM	1.19 ± 0.04	0.20 ± 0.01	0.46 ± 0.01	0.57 ± 0.02	0.23 ± 0.01	0.63 ± 0.01	0.61 ± 0.02	0.57 ± 0.07
CheXpertPlus_CheX_MIMIC	1.18 ± 0.04	0.20 ± 0.01	0.45 ± 0.01	0.59 ± 0.02	0.21 ± 0.01	0.62 ± 0.01	0.65 ± 0.02	0.58 ± 0.07
RGRG	1.17 ± 0.04	0.22 ± 0.01	0.44 ± 0.01	0.60 ± 0.02	0.22 ± 0.01	0.62 ± 0.01	0.67 ± 0.02	0.60 ± 0.07
Cvt2distilgpt2_MIMIC	1.13 ± 0.04	0.20 ± 0.01	0.42 ± 0.01	0.61 ± 0.02	0.21 ± 0.01	0.61 ± 0.01	0.68 ± 0.02	0.61 ± 0.07
RaDialog	1.09 ± 0.03	0.20 ± 0.01	0.44 ± 0.01	0.54 ± 0.02	0.20 ± 0.01	0.59 ± 0.01	0.59 ± 0.02	0.54 ± 0.07
CheXpertPlus_MIMIC	0.99 ± 0.04	0.18 ± 0.01	0.39 ± 0.01	0.59 ± 0.02	0.17 ± 0.01	0.58 ± 0.01	0.66 ± 0.02	**0.62 ± 0.07**
BiomedGPT_IU	0.96 ± 0.05	0.14 ± 0.01	0.38 ± 0.01	0.52 ± 0.02	0.21 ± 0.01	0.54 ± 0.01	0.52 ± 0.02	0.54 ± 0.07
CheXpertPlus_CheX	0.92 ± 0.03	0.16 ± 0.01	0.41 ± 0.01	0.49 ± 0.02	0.15 ± 0.01	0.53 ± 0.01	0.54 ± 0.02	0.55 ± 0.07
VLCI_MIMIC	0.91 ± 0.04	0.14 ± 0.01	0.36 ± 0.01	0.48 ± 0.02	0.22 ± 0.01	0.58 ± 0.01	0.47 ± 0.02	0.49 ± 0.08
CheXagent	0.83 ± 0.05	0.12 ± 0.01	0.35 ± 0.01	0.49 ± 0.02	0.14 ± 0.01	0.50 ± 0.01	0.39 ± 0.03	0.57 ± 0.07
GPT4V	0.71 ± 0.03	0.08 ± 0.00	0.27 ± 0.01	0.41 ± 0.02	0.15 ± 0.01	0.52 ± 0.01	0.65 ± 0.03	0.55 ± 0.08
LLM-CXR	0.49 ± 0.02	0.03 ± 0.00	0.19 ± 0.01	0.06 ± 0.01	0.02 ± 0.00	0.28 ± 0.01	0.03 ± 0.01	0.30 ± 0.06

Table 4: Comprehensive evaluation of medical report generation models on the CheXpert Plus datasets. Models are ranked by 1/RadCliQ-v1. Model evaluation results with 95% confidence intervals (mean ± CI) under normality assumption. The best results for each metric are shown in **bold**.

Model	1/RadCliQ-v1 ↑	BLEU-2 ↑	BertScore ↑	SembScore ↑	RadGraph ↑	RaTEScore ↑	GREEN ↑	1/FineRadScore ↑
Findings + Impression								
CheXpertPlus_CheX	**0.51 ± 0.07**	**0.14 ± 0.01**	**0.02 ± 0.02**	0.38 ± 0.03	0.07 ± 0.02	0.49 ± 0.02	0.36 ± 0.05	0.35 ± 0.11
CheXpertPlus_CheX_MIMIC	0.51 ± 0.07	0.14 ± 0.01	0.01 ± 0.02	**0.39 ± 0.03**	**0.07 ± 0.02**	**0.50 ± 0.02**	**0.38 ± 0.04**	**0.36 ± 0.12**
MedVersa	0.49 ± 0.06	0.09 ± 0.01	0.01 ± 0.02	0.34 ± 0.03	0.05 ± 0.01	0.45 ± 0.02	0.33 ± 0.05	0.35 ± 0.11
CheXpertPlus_MIMIC	0.48 ± 0.06	0.10 ± 0.01	0.00 ± 0.02	0.32 ± 0.03	0.05 ± 0.01	0.43 ± 0.02	0.29 ± 0.04	0.35 ± 0.11
RadFM	0.44 ± 0.05	0.07 ± 0.01	-0.04 ± 0.02	0.23 ± 0.03	0.03 ± 0.01	0.39 ± 0.02	0.14 ± 0.03	0.34 ± 0.09
GPT4V	0.43 ± 0.05	0.06 ± 0.01	-0.07 ± 0.02	0.21 ± 0.02	0.03 ± 0.01	0.39 ± 0.01	0.18 ± 0.04	0.33 ± 0.10
Findings								
CheXpertPlus_CheX_MIMIC	**0.81 ± 0.12**	0.15 ± 0.03	0.34 ± 0.04	**0.40 ± 0.05**	**0.21 ± 0.03**	**0.50 ± 0.03**	**0.27 ± 0.05**	0.35 ± 0.18
MAIRA-2	0.79 ± 0.10	**0.16 ± 0.03**	**0.36 ± 0.03**	0.35 ± 0.04	0.19 ± 0.03	0.48 ± 0.02	0.27 ± 0.05	**0.35 ± 0.18**
CheXpertPlus_CheX	0.79 ± 0.10	0.15 ± 0.03	0.34 ± 0.03	0.38 ± 0.04	0.19 ± 0.03	0.49 ± 0.03	0.24 ± 0.05	0.34 ± 0.20
MedVersa	0.72 ± 0.10	0.13 ± 0.02	0.32 ± 0.03	0.34 ± 0.05	0.15 ± 0.02	0.47 ± 0.03	0.24 ± 0.05	0.34 ± 0.18
RaDialog	0.71 ± 0.09	0.13 ± 0.02	0.31 ± 0.03	0.35 ± 0.05	0.14 ± 0.02	0.45 ± 0.02	0.21 ± 0.04	0.33 ± 0.17
RGRG	0.67 ± 0.11	0.15 ± 0.02	0.32 ± 0.04	0.27 ± 0.05	0.14 ± 0.02	0.45 ± 0.03	0.22 ± 0.04	0.34 ± 0.17
CheXpertPlus_MIMIC	0.66 ± 0.10	0.14 ± 0.02	0.29 ± 0.03	0.29 ± 0.05	0.13 ± 0.03	0.43 ± 0.03	0.24 ± 0.05	0.34 ± 0.19
CheXagent	0.64 ± 0.11	0.12 ± 0.02	0.28 ± 0.04	0.27 ± 0.04	0.12 ± 0.02	0.43 ± 0.02	0.18 ± 0.05	0.34 ± 0.19
Cvt2distilgpt2_MIMIC	0.63 ± 0.08	0.12 ± 0.02	0.27 ± 0.03	0.27 ± 0.04	0.12 ± 0.02	0.42 ± 0.03	0.21 ± 0.05	0.35 ± 0.19
VLCI_MIMIC	0.59 ± 0.09	0.12 ± 0.02	0.23 ± 0.03	0.25 ± 0.04	0.10 ± 0.02	0.38 ± 0.03	0.17 ± 0.05	0.33 ± 0.20
Cvt2distilgpt2_IU	0.58 ± 0.11	0.08 ± 0.02	0.27 ± 0.03	0.15 ± 0.05	0.10 ± 0.02	0.38 ± 0.03	0.15 ± 0.05	0.33 ± 0.20
RadFM	0.57 ± 0.08	0.08 ± 0.02	0.23 ± 0.03	0.22 ± 0.04	0.08 ± 0.01	0.40 ± 0.03	0.10 ± 0.03	0.33 ± 0.16
GPT4V	0.57 ± 0.08	0.08 ± 0.01	0.21 ± 0.03	0.23 ± 0.04	0.08 ± 0.02	0.41 ± 0.02	0.15 ± 0.05	0.34 ± 0.20
VLCI_IU	0.56 ± 0.09	0.11 ± 0.02	0.22 ± 0.03	0.17 ± 0.05	0.09 ± 0.02	0.42 ± 0.03	0.19 ± 0.05	0.34 ± 0.20
BiomedGPT_IU	0.55 ± 0.08	0.02 ± 0.01	0.20 ± 0.03	0.24 ± 0.04	0.06 ± 0.01	0.35 ± 0.03	0.12 ± 0.04	0.32 ± 0.18
LLM-CXR	0.52 ± 0.06	0.04 ± 0.01	0.16 ± 0.02	0.21 ± 0.04	0.04 ± 0.01	0.32 ± 0.02	0.02 ± 0.01	0.29 ± 0.13

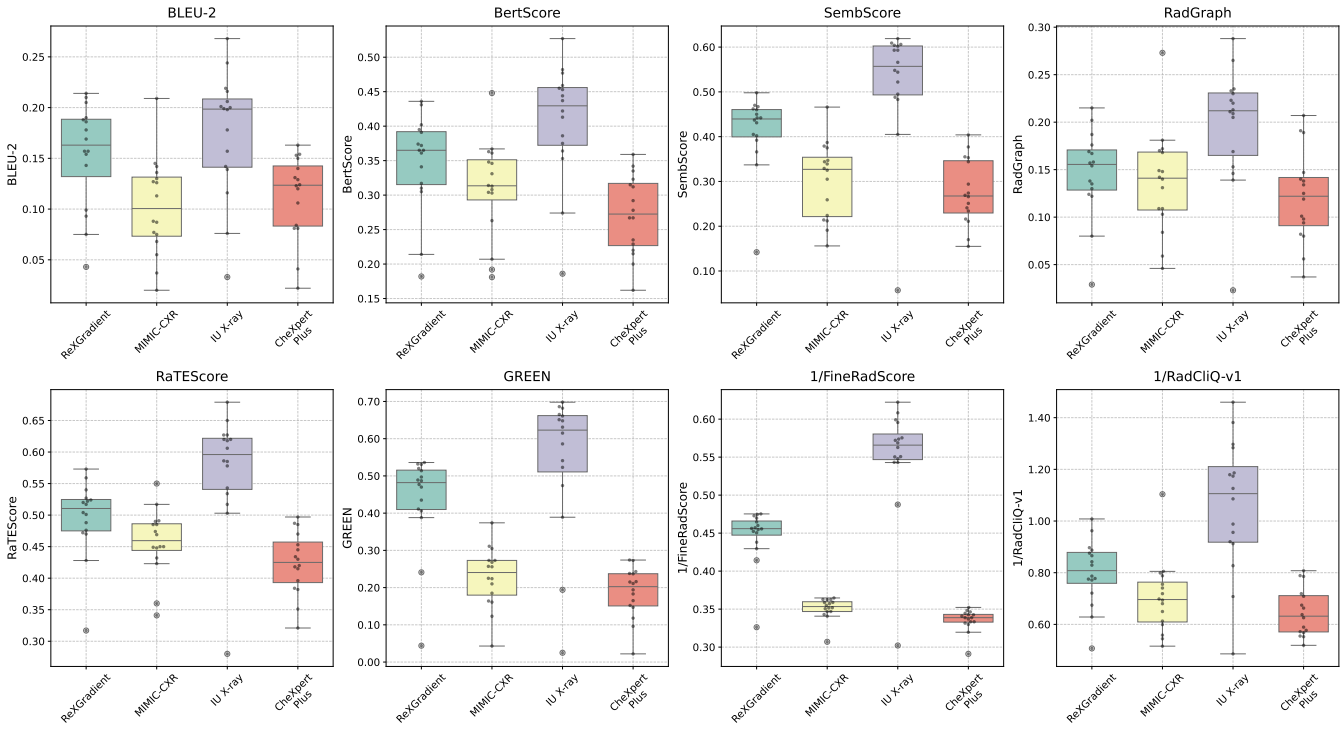

Figure 3: Distribution of evaluation metrics across different datasets: ReXGradient, and MIMIC-CXR, IU X-ray, CheXpert Plus. Box plots show the variation in model performance for each metric. For consistency in visualization, we plot the reciprocals (1/x) of FineRadScore and RadCliQ-v1, so higher values indicate better performance across all metrics.

to VLCI_MIMIC (0.91 ± 0.04), while VLCI_MIMIC performs better on MIMIC-CXR (0.68 ± 0.02 vs 0.60 ± 0.02 for VLCI_IU).

Analysis of Benchmark Datasets. Moreover, the results displayed in the tables show that ReXGradient serves as a dataset where models consistently exhibit minimal confidence intervals of 0.01 for most models on the RadCliQ-v1 metric, thereby supporting its utility as a reliable benchmark for medical report generation models. Figure 3 illustrates the distribution of evaluation metrics. The IU X-ray dataset, while always showing high-performance scores (the best model achieving a 1/RadCliQ-v1 score of 1.46 ± 0.03 on the findings sections), suggests that it may be too simplistic or lacking in complexity necessary for rigorous model differentiation. In contrast, CheXpert Plus shows lower overall performance (the best model obtaining a 1/RadCliQ-v1 score of 0.81 ± 0.12 on the findings sections) with higher variance, potentially indicating dataset distribution shifts or noise.

Analysis of Benchmark Datasets. We further analyze different datasets for benchmarking report generation tasks. ReXGradient serves as a dataset where models consistently exhibit minimal confidence intervals of 0.01 for most models on the RadCliQ-v1 metric, thereby supporting its utility as a reliable benchmark for medical report generation models. Figure 3 illustrates the distribution of evaluation metrics. The IU X-ray dataset, while always showing high-performance scores (the best model achieving a 1/RadCliQ-v1 score of 1.46 ± 0.03 on the findings sections), suggests that it may be too simplistic or lacking in complexity necessary for rigorous model differentiation. In contrast, CheXpert Plus shows lower overall performance (the best model obtaining a 1/RadCliQ-v1 score of 0.81 ± 0.12 on the findings sections) with higher variance, potentially indicating dataset distribution shifts or noise. The MIMIC-CXR dataset, being widely used in model training, generally reflects the same distribution as most models' training data, making it less effective for evaluating model generalization capabilities. Based on these observations, ReXGradient proves to be the most reliable benchmark dataset, offering meaningful differentiation of model performance while testing true generalization abilities across different medical institutions.

Conclusion

We present ReXrank, the first comprehensive leaderboard for evaluating AI-powered chest X-ray report generation systems. Through extensive evaluation of **16** state-of-the-art models across **4** diverse datasets and **8** different metrics, our benchmark provides a clear summarization of current report generation models. We show that ReXGradient serves as a reliable benchmark with consistent evaluation metrics and meaningful model differentiation. ReXrank continuously accepts model submissions and updates the leaderboard, and we plan to extend this framework to additional medical imaging modalities in the future.

Acknowledgments

This work was supported by Biswas Family Foundation's Transformative Computational Biology Grant in Collaboration with the Milken Institute.

Disclosures

O.H and J.M are founders and hold equity in Gradient Health, a private company focused on health data accessibility and availability for commercial research. Gradient Health provided the Private Dataset used in this work and did not provide funding for this research and had no role in its design, execution, or publication. The Private Dataset's 4x downsampled version is available under the https://gradienthealth.io/terms-of-use/.

References

Alfarghaly, O.; Khaled, R.; Elkorany, A.; Helal, M.; and Fahmy, A. 2021. Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, 24: 100557.

Bannur, S.; Bouzid, K.; Castro, D. C.; Schwaighofer, A.; Bond-Taylor, S.; Ilse, M.; Pérez-García, F.; Salvatelli, V.; Sharma, H.; Meissen, F.; et al. 2024. MAIRA-2: Grounded Radiology Report Generation. *arXiv preprint arXiv:2406.04449*.

Bruls, R.; and Kwee, R. 2020. Workload for radiologists during on-call hours: dramatic increase in the past 15 years. *Insights into imaging*, 11: 1–7.

Chambon, P.; Delbrouck, J.-B.; Sounack, T.; Huang, S.-C.; Chen, Z.; Varma, M.; Truong, S. Q.; Chuong, C. T.; and Langlotz, C. P. 2024. CheXpert Plus: Hundreds of Thousands of Aligned Radiology Texts, Images and Patients. *arXiv preprint arXiv:2405.19538*.

Chen, W.; Liu, Y.; Wang, C.; Zhu, J.; Zhao, S.; Li, G.; Liu, C.-L.; and Lin, L. 2023. Cross-Modal Causal Intervention for Medical Report Generation. *arXiv preprint arXiv:2303.09117*.

Chen, Z.; Song, Y.; Chang, T.-H.; and Wan, X. 2020. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*.

Chen, Z.; Varma, M.; Delbrouck, J.-B.; Paschali, M.; Blankemeier, L.; Van Veen, D.; Valanarasu, J. M. J.; Youssef, A.; Cohen, J. P.; Reis, E. P.; et al. 2024. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*.

Demner-Fushman, D.; Kohli, M. D.; Rosenman, M. B.; Shooshan, S. E.; Rodriguez, L.; Antani, S.; Thoma, G. R.; and McDonald, C. J. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2): 304–310.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

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

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

Johnson, A. E.; Pollard, T. J.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Mark, R. G.; and Horng, S. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1): 317.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, 2.

Lee, S.; Kim, W. J.; Chang, J.; and Ye, J. C. 2023. LLM-CXR: Instruction-Finetuned LLM for CXR Image Understanding and Generation. *arXiv preprint arXiv:2305.11490*.

Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. 2022. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12009–12019.

Nicolson, A.; Dowling, J.; and Koopman, B. 2023. Improving chest X-ray report generation by leveraging warm starting. *Artificial intelligence in medicine*, 144: 102633.

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

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

Pellegrini, C.; Özsoy, E.; Busam, B.; Navab, N.; and Keicher, M. 2023. RaDialog: A large vision-language model for radiology report generation and conversational assistance. *arXiv preprint arXiv:2311.18681*.

Smit, A.; Jain, S.; Rajpurkar, P.; Pareek, A.; Ng, A. Y.; and Lungren, M. P. 2020. CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. *arXiv preprint arXiv:2004.09167*.

Tanida, T.; Müller, P.; Kaissis, G.; and Rueckert, D. 2023. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7433–7442.

Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*.

Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; and Zhang, L. 2021a. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 22–31.

Wu, J. T.; Agu, N. N.; Lourentzou, I.; Sharma, A.; Paguio, J. A.; Yao, J. S.; Dee, E. C.; Mitchell, W.; Kashyap, S.; Giovannini, A.; et al. 2021b. Chest imagenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*.

Yang, Z.; Li, L.; Lin, K.; Wang, J.; Lin, C.-C.; Liu, Z.; and Wang, L. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1): 1.

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

Zhang, K.; Zhou, R.; Adhikarla, E.; Yan, Z.; Liu, Y.; Yu, J.; Liu, Z.; Chen, X.; Davison, B. D.; Ren, H.; et al. 2024. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 1–13.

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

Zhao, W.; Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2024. RaTEScore: A Metric for Radiology Report Generation. *medRxiv*, 2024–06.

Zhou, H.-Y.; Adithan, S.; Acosta, J. N.; Topol, E. J.; and Rajpurkar, P. 2024. A Generalist Learner for Multifaceted Medical Image Interpretation. *arXiv preprint arXiv:2405.07988*.