Automated Chest X-ray Report Generation Remains Unsolved

Xiaoman Zhang[1], Julian Nicolas Acosta[1], Xiaoli Yang[1], Subathra Adithan[1], Luyang Luo[1], Hong-Yu Zhou[1], Joshua Miller[2], Ouwen Huang[2,3,4], Zongwei Zhou[5], Ibrahim Ethem Hamamci[6], Shruthi Bannur[7], Kenza Bouzid[7], Xi Zhang[8], Zaiqiao Meng[8], Aaron Nicolson[9], Bevan Koopman[9], Inhyeok Baek[10], Hanbin Ko[11], Mercy Prasanna Ranjit[12], Shaury Srivastav[12], Sriram Gnana Sambanthan[13] and Pranav Rajpurkar[1]

[1]*Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA*

[2]*Gradient Health, Durham, NC, USA*

[3]*Department of Statistical Science, Duke University, Durham, NC, USA*

[4]*Laplace Institute, Durham, NC, USA*

[5]*Department of Computer Science, Johns Hopkins University, MD, USA*

[6]*University of Zurich, Switzerland*

[7]*Microsoft Research Health Futures, USA*

[8]*University of Glasgow, UK*

[9]*Australian e-Health Research Centre, CSIRO Health and Biosecurity, Brisbane, Australia*

[10]*Seoul National University, South Korea*

[11]*Seoul National University Graduate School, South Korea*

[12]*Microsoft Research India, India*

[13]*Indian Institute of Technology Madras, Chennai, India*

Accurate interpretation of chest radiograph images and generation of narrative reports is essential for patient care but places a heavy burden on radiologists and clinical experts. While AI models for automated report generation show promise, standardized evaluation frameworks remain limited. Here we present the ReXrank Challenge V1.0, a competition in the generation of chest radiograph reports utilizing ReXGradient, the largest test dataset consisting of **10,000** studies across 67 sites. The challenge attracted diverse participants from academic institutions, industry, and independent research teams, resulting in 8 new submissions alongside 16 state-of-the-art models previously benchmarked. Through comprehensive evaluation using multiple metrics, we analyzed model performance across various dimen-

sions: differences between normal and abnormal studies, generalization capabilities across healthcare sites, and error rates in identifying clinical findings. This benchmark reveals that automated chest X-ray report generation remains fundamentally unsolved, with significant performance gaps between normal and abnormal studies, where even top-performing models achieve less than 45% error-free reporting on abnormal cases, and substantial variability across healthcare institutions, indicating that robust, clinically-ready systems require continued development before widespread deployment.

Keywords: Radiology Report Generation; Benchmark.

1. Introduction

Automating radiology report generation has the potential to dramatically improve clinical workflow efficiency, reduce reporting delays, and ensure consistent quality in radiology services in diverse healthcare settings.[1,2] However, creating AI systems that can match the accuracy and nuance of expert radiologists requires overcoming substantial technical and clinical hurdles.

Previous approaches to automated radiology report generation have shown promise, but have not yet achieved the reliability required for widespread clinical applications.[3,4] The field faces several critical obstacles to progress. Existing model comparisons typically employ inconsistent data splits and lack standardized evaluation metrics, preventing reliable comparative analysis across different model architectures. Additionally, commonly used public datasets often serve dual purposes as both training and evaluation data, failing to adequately test the models' ability to generalize to new, unseen distributions.[5–7] To address these challenges, a standardized evaluation framework called ReXrank[8] was recently proposed, establishing a common benchmark for 16 state-of-the-art models across identical evaluation settings. Although this initiative represents an important step forward, further work is needed to push the domain toward clinical applicability.

To further advance the field, we hosted the **ReXrank Challenge V1.0**, a comprehensive competition to evaluate AI-powered chest X-ray report generation models, which ran from December 1, 2024, to March 15, 2025. This challenge builds on the foundation of the ReXrank leaderboard, 8 new submissions were received during the challenge period, each employing distinctive approaches and training sets to the task of generating accurate and clinically relevant radiology reports from chest radiographs. These submissions, along with 16 state-of-the-art models previously benchmarked in the initial ReXrank evaluation, were evaluated as part of the ReXrank Challenge V1.0

In this paper, we present results from the ReXrank Challenge V1.0, providing a comprehensive benchmark for AI-powered chest X-ray report generation. We evaluated all participating models using an expanded set of metrics that included traditional measures (BLEU,[9] BertScore,[10] SembScore,[11] RadGraph,[12] RadCliQ,[12] RaTEScore[13]) as well as newer LLM-based metrics (GREEN,[14] FineRadScore[15]). We perform detailed analyses of model performance across various clinically relevant dimensions, including normal and abnormal studies, performance variations across different medical sites, and error patterns of the model's prediction. These stratified evaluations provide critical insight into the real-world applicability of these models and highlight areas where further development is needed.

Our comprehensive benchmark provides several key insights for medical ML applications:

- Top models achieving over 80-90% no significant error rates on normal studies while struggling with abnormal cases (MedVersa leading with 43.32% no significant error rate but only 11.70% no actionable error rate). This substantial performance gap highlights that automated radiology report generation remains an unsolved problem, particularly for clinically significant abnormal findings.
- Model performance varies dramatically across different healthcare sites, even for top-performing models such as MedVersa, MAIRA-2 and Libra, indicating that achieving robust cross-institutional generalization remains a significant challenge that requires diverse training data and targeted adaptation techniques.
- Current evaluation metrics show low consistency, especially traditional metrics that correlate poorly with newer LLM-based metrics, highlighting the need for evaluation approaches that better align with clinical needs.

2. Related Work

The automated generation of radiology reports has emerged as a significant research area addressing the growing demand for imaging services that exceeds radiologist capacity. The field has evolved from recurrent neural networks to Transformer-based architectures,[5] with recent systems incorporating LLMs such as LLaMA[16] and Vicuna-7B.[17] Non-LLM approaches include BiomedGPT,[18] CheXpertPlus,[6] CvT2DistilGPT2,[19] and RGRG,[20] while LLM-based systems like CheXagent,[21] RaDialog,[22] MAIRA-2,[4] MedVersa,[23] RadFM,[24] and Libra[25] have transformed the field through visual instruction tuning. The evaluation of automated radiology reports presents unique challenges, as traditional natural language generation metrics do not account for clinical significance. Specialized metrics have been developed, including CheXbert,[11] which evaluates reports based on the presence of specific pathologies, and RadGraph-F1,[12] which measures the overlap in clinical entities and relations between generated and reference reports. More recently, ReXrank[8] has been introduced as a public leaderboard and challenge for assessing AI-powered radiology report generation. ReXrank incorporates various datasets, including MIMIC-CXR,[26] IU-Xray,[27] and CheXpert Plus,[6] and employs eight different evaluation metrics to provide a comprehensive assessment of model performance.

3. Overview

3.1. *Challenge Design*

The ReXrank Challenge, which ran from December 1, 2024, to March 15, 2025, was designed as a comprehensive competition for evaluating AI-powered chest X-ray report generation models. This challenge aimed to benchmark model performance across multiple critical dimensions, including linguistic quality, clinical accuracy, and generalization capability across diverse clinical settings.

3.2. *Datasets*

The ReXrank Challenge leveraged both public and private datasets for robust report generation assessment. The private ReXGradient dataset, comprising 10,000 studies collected from 67 U.S. medical institutions, represents one of the largest and most geographically diverse evaluation sets available. For public evaluation, we utilized the official test splits of MIMIC-CXR (2,347 studies)[26] and IU-Xray (590 studies),[27] along with CheXpert Plus's validation set (200 studies),[6] as no official test split was available. Each study may contain multiple images, all of which were utilized in our evaluation. We specifically focused on the generation of "Findings" section of reports and filtered out any reports that lacked this section, reducing the CheXpert Plus evaluation set to 62 studies with available findings sections.

3.3. *Models*

During the challenge period, we received 8 model submissions from 5 institutions, including Libra,[25] CXRMate,[28] CXRMate-ED,[29] CXRMate-RRG24,[30] RadPhi3.5Vision,[31] DD-LLaVA-X, MoERad-IU, and MoERad-MIMIC. We compared these submissions with **16** previously benchmarked report generation models from 10 different institutions:[8] BiomedGPT-IU,[32] CheXagent,[21] CheXpertPlus-CheX,[6] CheXpertPlus-CheX-MIMIC,[6] CheXpertPlus-MIMIC,[6] CvT2DistilGPT2-IU,[5] CvT2DistilGPT2-MIMIC,[5] GPT4V,[33] LLM-CXR,[34] MAIRA-2,[4] MedVersa,[23] RadFM,[24] RaDialog,[22] RGRG,[20] VLCI-IU,[7] and VLCI-MIMIC.[7] The detailed evaluation setup of each models can be found in the appendix.

3.4. *Metrics*

The ReXrank Challenge employed **8** distinct metrics to comprehensively assess the quality of generated radiology reports. These included traditional text generation metrics such as BLEU-2[9] and BERTScore,[10] alongside domain-specific metrics designed for radiology report evaluation, including SembScore,[11] RadGraph-F1,[35] RadCliQ-v1,[35] and RaTEScore.[13] The framework also incorporated recently developed LLM-based metrics, including GREEN,[14] FineRadScore,[15] which focus on identifying clinically significant errors. Each metric evaluated different aspects of the generated reports, from textual similarity to clinical accuracy, enabling a thorough and multifaceted assessment of model performance. The detailed information can be found in the appendix.

4. Results

4.1. *Model Results Summary*

The comprehensive evaluation across multiple datasets revealed several key insights about current chest X-ray report generation capabilities. No single model dominated across all metrics, suggesting that different evaluation metrics capture complementary aspects of report quality from linguistic similarity (BLEU, BertScore) to clinical accuracy (SembScore, RadCliQ-V1, RadGraph, RATEScore) and error detection (FineRadScore, GREEN). MedVersa emerged as the top-performing model on both ReXGradient and MIMIC-CXR datasets, achieving the

highest 1/FineRadScore (0.475 and 0.365, respectively) and excelling in multiple metrics including BertScore and SembScore. On the IU-Xray dataset, CheXpertPlus-MIMIC achieved the highest 1/FineRadScore (0.622), while CXRMate-ED led on the CheXpert Plus dataset (0.367). Among the newly submitted models, Libra and CXRMate-ED demonstrated strong performance, with Libra achieving the highest BLEU score (0.246) on MIMIC-CXR and CXRMate-ED showing remarkable consistency across diverse institutional datasets. MoERad-IU excelled on the IU-Xray dataset, achieving the highest scores in BLEU (0.277), SembScore (0.641), RadGraph (0.341), and 1/RadCliQ-V1 (1.922). Notably, specialist models trained on specific datasets generally performed best on their corresponding test sets, highlighting the challenge of developing models that generalize across institutions.

Table 1. Comprehensive evaluation of medical report generation models on ReXGradient. Models are ranked by 1/FineRadScore. The best results for each metric are shown in **bold** (n=10,000).

Model-Name	BLEU	BertScore	SembScore	RadGraph	1/RadCliQ-V1	RATEScore	GREEN	1/FineRadScore
MedVersa	0.210	0.431	**0.498**	0.202	1.008	0.527	0.532	**0.475**
MAIRA-2	0.205	**0.436**	0.462	0.187	0.963	0.559	0.531	**0.475**
Libra	0.176	0.408	0.474	0.169	0.913	0.544	0.549	0.473
CXRMate-ED	0.202	0.398	0.415	0.187	0.872	0.564	0.518	0.472
CheXpertPlus-MIMIC	0.154	0.341	0.440	0.131	0.778	0.501	0.517	0.471
MoERad-IU	**0.227**	0.434	0.446	**0.247**	**1.018**	**0.575**	0.494	0.468
CvT2DistilGPT2-MIMIC	0.186	0.374	0.458	0.175	0.866	0.522	0.510	0.468
CXRMate	0.169	0.363	0.479	0.174	0.863	0.545	**0.550**	0.466
CheXpertPlus-CheX-MIMIC	0.169	0.372	0.440	0.153	0.828	0.516	0.486	0.463
DD-LLaVA-X	0.166	0.387	0.469	0.174	0.886	0.542	0.504	0.459
CXRMate-RRG24	0.150	0.327	0.462	0.152	0.792	0.518	0.408	0.458
RadPhi3.5Vision	0.209	0.383	0.488	0.169	0.891	0.544	0.453	0.458
RGRG	0.190	0.391	0.470	0.169	0.888	0.540	0.487	0.458
CvT2DistilGPT2-IU	0.176	0.394	0.405	0.166	0.839	0.518	0.467	0.456
CheXagent	0.093	0.305	0.366	0.080	0.674	0.428	0.241	0.455
RaDialog	0.188	0.402	0.450	0.158	0.876	0.522	0.435	0.454
VLCI-MIMIC	0.158	0.310	0.400	0.122	0.720	0.487	0.473	0.453
VLCI-IU	0.214	0.365	0.466	0.213	0.894	0.571	0.532	0.451
BioMedGPT-IU	0.099	0.317	0.437	0.157	0.771	0.472	0.388	0.450
MoERad-MIMIC	0.145	0.351	0.406	0.116	0.756	0.508	0.431	0.446
RadFM	0.157	0.365	0.392	0.135	0.775	0.504	0.406	0.437
GPT4V	0.075	0.214	0.337	0.138	0.629	0.470	0.497	0.428
CheXpertPlus-CheX	0.144	0.361	0.428	0.124	0.785	0.475	0.407	0.414
LLM-CXR	0.043	0.182	0.142	0.029	0.507	0.317	0.044	0.326

4.2. *Performance Comparison Across Normal and Abnormal Studies*

Our analysis of model performance on the ReXGradient dataset reveals significant differences when comparing normal versus abnormal chest X-ray studies. The results are summarized in the Appendix tables. For abnormal studies, MedVersa emerged as the top-performing model, achieving the highest scores in SembScore (0.425), RadGraph (0.169), 1/RadCliQ-v1 (0.856), and 1/FineRadScore (0.396). This suggests MedVersa has superior capability in detecting and describing pathological findings, which is critical for clinical utility. In normal studies, MAIRA-2 achieved the highest 1/FineRadScore (0.734), indicating strong clinical accuracy with minimal required corrections, and it ranked 4th in abnormal studies with 1/FineRadScore (0.392). Models like MoERad-IU demonstrated superior performance across multiple metrics for normal studies, achieving the highest scores in BLEU (0.353), BertScore (0.532), RadGraph (0.367), 1/RadCliQ-v1 (2.061), and RATEScore (0.710). While for abnormal studies,

Table 2. Comprehensive evaluation of medical report generation models on MIMIC-CXR. Models are ranked by 1/FineRadScore. The best results for each metric are shown in **bold**($n = 2,347$).

Model-Name	BLEU	BertScore	SembScore	RadGraph	1/RadCliQ-V1	RATEScore	GREEN	1/FineRadScore
MedVersa	0.209	**0.448**	**0.466**	**0.273**	**1.103**	**0.550**	**0.374**	**0.365**
CheXpertPlus-MIMIC	0.145	0.361	0.375	0.170	0.788	0.485	0.311	0.363
CheXpertPlus-CheX-MIMIC	0.142	0.367	0.379	0.181	0.805	0.490	0.305	0.363
CvT2DistilGPT2-MIMIC	0.126	0.331	0.329	0.149	0.719	0.432	0.268	0.362
Libra	**0.246**	0.431	0.405	0.222	0.944	0.529	0.351	0.362
CXRMate	0.198	0.367	0.423	0.220	0.870	0.521	0.338	0.362
DD-LLaVA-X	0.154	0.348	0.402	0.182	0.801	0.505	0.301	0.361
RaDialog	0.127	0.363	0.387	0.172	0.799	0.485	0.273	0.359
MAIRA-2	0.088	0.308	0.339	0.131	0.694	0.517	0.224	0.359
CXRMate-RRG24	0.198	0.367	0.423	0.220	0.870	0.521	0.338	0.359
CXRMate-ED	0.208	0.383	0.396	0.223	0.872	0.531	0.327	0.358
VLCI-MIMIC	0.136	0.304	0.305	0.140	0.680	0.450	0.256	0.357
RadPhi3.5Vision	0.223	0.386	0.431	0.207	0.888	0.534	0.294	0.356
CheXagent	0.113	0.346	0.347	0.148	0.741	0.474	0.257	0.355
MoERad-MIMIC	0.163	0.341	0.334	0.143	0.726	0.465	0.240	0.354
RGRG	0.130	0.348	0.344	0.168	0.755	0.491	0.273	0.352
CheXpertPlus-CheX	0.077	0.314	0.325	0.142	0.698	0.469	0.225	0.351
RadFM	0.087	0.313	0.259	0.109	0.650	0.450	0.185	0.351
VLCI-IU	0.075	0.263	0.212	0.109	0.599	0.449	0.210	0.347
CvT2DistilGPT2-IU	0.055	0.303	0.191	0.103	0.613	0.448	0.164	0.347
MoERad-IU	0.064	0.321	0.213	0.122	0.643	0.455	0.174	0.347
GPT4V	0.068	0.207	0.214	0.084	0.558	0.423	0.161	0.343
BioMedGPT-IU	0.020	0.192	0.224	0.059	0.544	0.360	0.123	0.341
LLM-CXR	0.037	0.181	0.156	0.046	0.516	0.341	0.043	0.307

Table 3. Comprehensive evaluation of medical report generation models on IU-Xray. Models are ranked by 1/FineRadScore. The best results for each metric are shown in **bold** ($n = 590$).

Model-Name	BLEU	BertScore	SembScore	RadGraph	1/RadCliQ-V1	RATEScore	GREEN	1/FineRadScore
CheXpertPlus-MIMIC	0.178	0.386	0.593	0.169	0.988	0.585	0.661	**0.622**
CvT2DistilGPT2-MIMIC	0.199	0.422	0.609	0.209	1.126	0.606	0.682	0.608
MAIRA-2	0.219	0.477	0.604	0.233	1.298	0.627	0.194	0.599
CXRMate-RRG24	0.245	0.456	0.638	0.302	1.458	0.666	0.680	0.598
CXRMate-ED	0.225	0.464	0.557	0.249	1.220	0.655	0.685	0.597
RGRG	0.216	0.437	0.602	0.223	1.174	0.620	0.665	0.596
Libra	0.192	0.461	0.616	0.210	1.226	0.623	0.674	0.593
MoERad-IU	**0.277**	0.525	**0.641**	**0.341**	**1.922**	**0.684**	0.665	0.587
MoERad-MIMIC	0.171	0.420	0.559	0.178	1.020	0.603	0.584	0.579
CheXpertPlus-CheX-MIMIC	0.198	0.453	0.593	0.211	1.179	0.618	0.648	0.576
DD-LLaVA-X	0.189	0.443	0.600	0.233	1.204	0.636	0.671	0.574
CheXagent	0.116	0.353	0.488	0.139	0.827	0.503	0.389	0.574
RadFM	0.200	0.459	0.566	0.230	1.187	0.627	0.615	0.572
MedVersa	0.206	**0.527**	0.606	0.235	1.460	0.650	0.631	0.569
CXRMate	0.181	0.418	0.625	0.213	1.146	0.637	**0.730**	0.565
CvT2DistilGPT2-IU	0.244	0.482	0.548	0.265	1.283	0.620	0.686	0.563
RadPhi3.5Vision	0.248	0.433	0.607	0.220	1.166	0.634	0.597	0.552
VLCI-IU	0.268	0.455	0.619	0.288	1.381	0.679	0.698	0.551
GPT4V	0.076	0.274	0.405	0.146	0.708	0.517	0.651	0.550
CheXpertPlus-CheX	0.157	0.413	0.495	0.153	0.920	0.534	0.541	0.548
BioMedGPT-IU	0.142	0.375	0.522	0.213	0.956	0.543	0.523	0.543
RaDialog	0.201	0.444	0.544	0.205	1.086	0.586	0.586	0.543
VLCI-MIMIC	0.139	0.364	0.483	0.220	0.913	0.578	0.474	0.488
LLM-CXR	0.033	0.186	0.057	0.023	0.486	0.280	0.025	0.302

it ranked 13th with a 1/FineRadScore of 0.379 and a 1/RadCliQ-v1 score of 0.755, showing a substantial performance drop when handling pathological cases. Most models performed substantially better on normal cases compared to abnormal ones, with average 1/RadCliQ-v1 scores approximately 40-50% higher for normal studies. This pattern is consistent with pre-

Table 4. Comprehensive evaluation of medical report generation models on CheXpert Plus. Models are ranked by 1/FineRadScore. The best results for each metric are shown in **bold** ($n = 62$).

Model-Name	BLEU	BertScore	SembScore	RadGraph	1/RadCliQ-V1	RATEScore	GREEN	1/FineRadScore
CXRMate-ED	0.157	0.324	0.316	0.175	0.723	0.498	0.265	**0.367**
RadPhi3.5Vision	**0.198**	0.353	**0.437**	0.217	**0.860**	0.510	0.243	0.356
MAIRA-2	0.163	**0.359**	0.355	0.189	0.788	0.485	0.273	0.352
CXRMate-RRG24	0.157	0.315	0.411	**0.218**	0.801	**0.521**	**0.276**	0.350
CheXpertPlus-CheX-MIMIC	0.153	0.335	0.404	0.207	0.808	0.497	0.274	0.348
CXRMate	0.135	0.289	0.327	0.138	0.678	0.425	0.235	0.348
CvT2DistilGPT2-MIMIC	0.124	0.267	0.266	0.119	0.626	0.420	0.215	0.346
CheXpertPlus-MIMIC	0.140	0.292	0.294	0.134	0.663	0.430	0.238	0.344
Libra	0.165	0.343	0.318	0.171	0.738	0.477	0.265	0.344
DD-LLaVA-X	0.085	0.318	0.385	0.172	0.753	0.476	0.206	0.343
MoERad-MIMIC	0.122	0.267	0.300	0.120	0.641	0.434	0.166	0.343
CheXpertPlus-CheX	0.150	0.342	0.377	0.191	0.786	0.487	0.237	0.343
MedVersa	0.129	0.323	0.344	0.147	0.719	0.470	0.243	0.343
CheXagent	0.123	0.278	0.269	0.125	0.638	0.434	0.183	0.341
MoERad-IU	0.075	0.284	0.175	0.102	0.595	0.390	0.127	0.341
VLCI-IU	0.106	0.220	0.170	0.094	0.555	0.418	0.194	0.339
GPT4V	0.081	0.215	0.234	0.082	0.568	0.415	0.152	0.339
RGRG	0.154	0.315	0.274	0.140	0.674	0.453	0.216	0.337
RadFM	0.081	0.235	0.216	0.080	0.572	0.396	0.096	0.333
RaDialog	0.131	0.312	0.353	0.138	0.709	0.445	0.211	0.333
CvT2DistilGPT2-IU	0.084	0.267	0.155	0.098	0.577	0.382	0.147	0.332
VLCI-MIMIC	0.120	0.229	0.251	0.101	0.589	0.384	0.165	0.330
BioMedGPT-IU	0.022	0.200	0.241	0.056	0.552	0.351	0.118	0.320
LLM-CXR	0.041	0.162	0.211	0.037	0.519	0.321	0.022	0.291

vious research showing that AI systems typically find it easier to describe normal anatomical structures than to detect and characterize pathological findings.

4.3. *Error Rates Analysis Across Normal and Abnormal Studies*

Figure 1 presents a comprehensive comparison of error rates across multiple AI models when generating radiology reports for both abnormal and normal chest X-ray studies from the ReXGradient dataset. The visualization reveals several important patterns. All models demonstrate significantly higher no actionable error rates for normal studies compared to abnormal ones, with most models achieving 60-75% no actionable error rates for normal studies but only 5-15% for abnormal studies. For abnormal studies, MedVersa performs best with a 43.32% no significant error rate, closely followed by CXRMate-ED (43.05%) and MAIRA-2 (41.98%). However, when examining the no actionable error rate (dark blue bars), these values drop dramatically to only 11.70%, 13.00%, and 10.78% respectively, highlighting how frequently these models make at least some errors in abnormal cases. Performance on normal studies is consistently stronger across all models. MAIRA-2 achieves the highest no significant error rate at 91.95%, followed by Libra (89.10%) and MedVersa (89.01%). The no-error rates for normal studies are also substantially higher, with many models exceeding 60-70%. There's considerable variability in model performance rankings between abnormal and normal categories. While some models like MedVersa and MAIRA-2 maintain strong performance in both categories, others show notable differences. For instance, MoERad-IU performs relatively well on normal studies (90.21% no significant error rate) but drops significantly for abnormal studies (38.06%). This analysis underscores the persistent challenge in developing AI systems that can accurately report abnormal findings in chest X-rays.

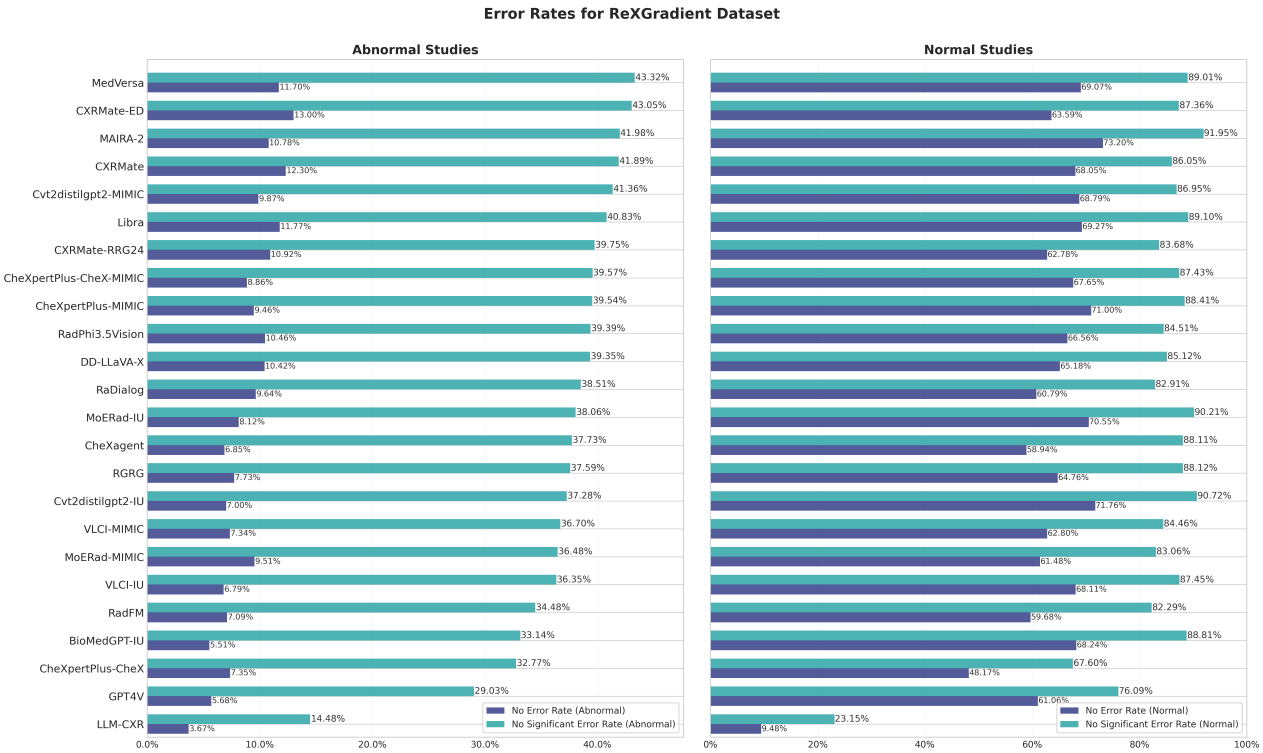

Fig. 1. Comparison of error rates for AI-generated radiology reports on the ReXGradient Dataset. The chart shows no-actionable-error rates (dark blue) and no-significant-error rates (teal) for both abnormal studies (left) and normal studies (right).

4.4. *Analysis of Results on Different Sites*

Figure 2 shows model performance variability across the 30 different healthcare sites represented in the ReXGradient dataset. As shown, MAIRA-2 and Libra demonstrate the most robust performance across sites, ranking in the top positions for the majority of sites. MAIRA-2 particularly excels, achieving first place in 8 sites and rarely falling below 6th place except for a few outliers. CheXpertPlus-MIMIC and MedVersa show generally strong performance but with noticeable variability. MedVersa, for instance, ranks first in several sites (particularly sites 1, 14, 19) but unexpectedly drops to much lower rankings (17th, 19th, 20th) in others (sites 10, 22, 28). The considerable performance fluctuation of models across different sites highlights the challenge of developing AI systems that generalize well across diverse clinical settings. For example, CXRMate performs excellently in some sites (ranked 1st in sites 4, 14, 16) but poorly in others (21st in institution 26). The substantial rank variability for most models (except the very top and bottom performers) suggests that current AI approaches for radiology report generation face significant challenges in generalizing across different institutional data distributions, which may reflect variations in patient demographics, imaging protocols, or radiologist reporting styles. Some models show similar performance patterns across certain groups of sites, potentially indicating similarities in those sites' data characteristics. For example, MAIRA-2 performs particularly well in sites 0-6 but shows more variable performance in sites 10-15. This analysis underscores the importance of evaluating AI models

across diverse institutional datasets before deployment in clinical settings, as performance can vary dramatically depending on the specific healthcare environment.

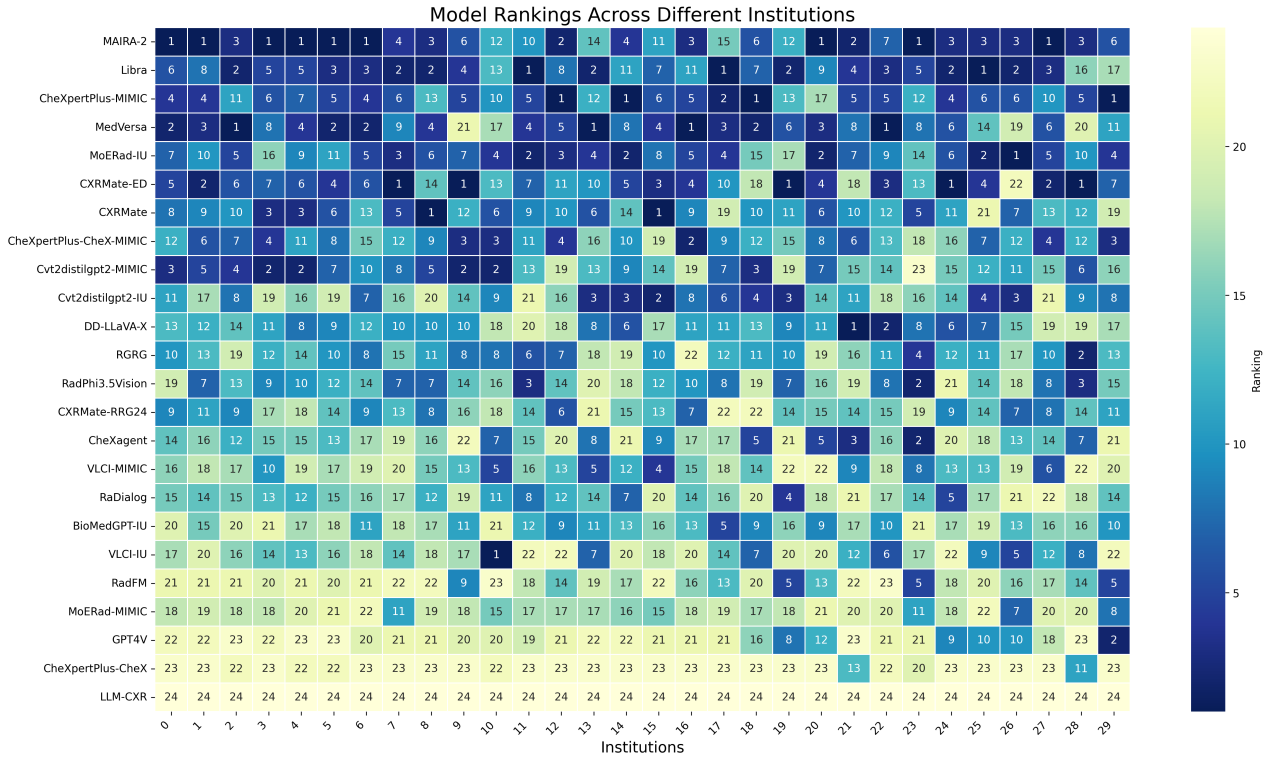

Fig. 2. Model Rankings Across Different Sites. The heatmap displays the performance ranking of 24 AI models (rows) across top 30 different healthcare sites (columns) in the ReXGradient dataset, with darker blue representing better rankings (1 being the best).

4.5. *Analysis of Current Evaluation Metrics*

Our analysis of the correlation between different evaluation metrics across multiple chest X-ray datasets reveals important insights about their relationships and potential redundancies. Figure 3 presents the correlation matrices for ReXGradient, MIMIC-CXR, IU X-ray, and CheXpert Plus datasets. We observe strong positive correlations (>0.75) between traditional text matching and NER-based metrics, particularly between BertScore and 1/RadCliQ-v1 (0.90), as well as between RadGraph and 1/RadCliQ-v1 (0.84) on IU X-ray. This suggests these conventional metrics may capture similar aspects of report quality. However, the LLM-based metrics, GREEN and 1/FineRadScore, show weaker correlations with traditional metrics (mostly < 0.60) and moderate correlation with each other (0.59), suggesting they capture unique dimensions of report quality that text matching and NER-based approaches miss. Notably, correlation patterns remain remarkably consistent across all four datasets, indicating the robustness of these relationships regardless of the data source. These findings highlight the complementary nature of LLM-based and traditional evaluation approaches, suggesting that comprehensive assessment of radiology reports requires metrics from both categories.

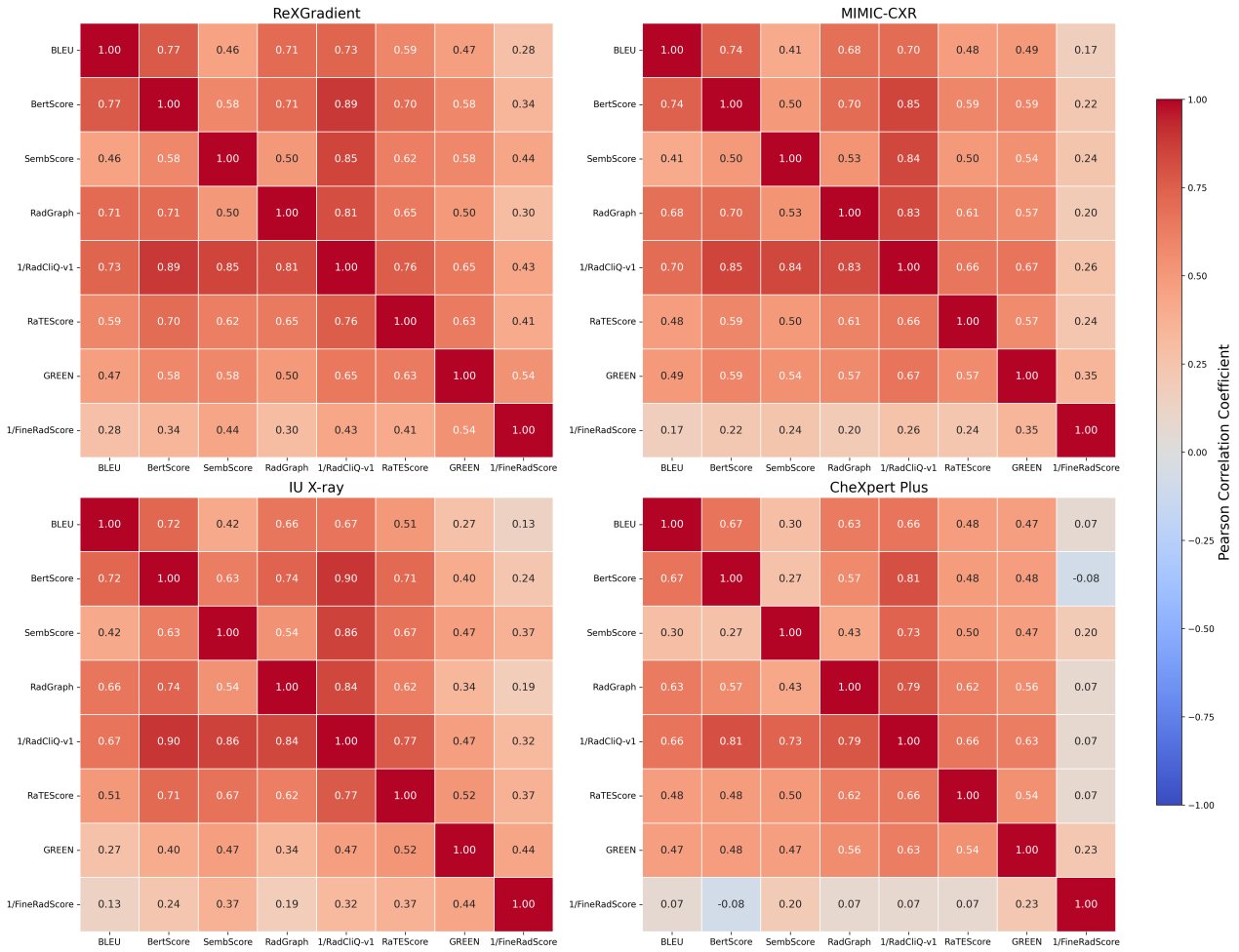

Fig. 3. Pearson correlation coefficients between eight evaluation metrics for the ReXGradient, MIMIC-CXR, IU X-ray, and CheXpert Plus datasets.

5. Discussion

Our comprehensive evaluation of AI-powered chest X-ray report generation models offers important insights into the current state of automated radiology reporting and reveals several key implications for both technical development and clinical implementation. The ReXrank Challenge demonstrates substantial progress in the field, with top-performing models like MedVersa, MAIRA-2, and Libra achieving impressive metrics across diverse institutional datasets. However, the significant performance gap between normal and abnormal studies—with even the best models showing no significant error rates below 45% for abnormal cases compared to over 90% for normal studies, highlights a fundamental challenge: current AI systems excel at identifying normal studies but struggle significantly with accurate pathology detection and characterization.

Technical Implications. Our analysis of model performance across multiple healthcare sites reveals substantial variability, with most models showing inconsistent rankings across different sites. This finding underscores the critical importance of developing robust cross-institutional generalization capabilities. Models trained on diverse datasets (e.g., MedVersa, MAIRA-2) demonstrated more consistent performance compared to those trained on homogeneous datasets (e.g., Libra), suggesting that exposure to varied reporting styles and patient populations may improve generalization. Our evaluation metric analysis reveals important insights for model assessment. The relatively weak correlation between LLM-based metrics (GREEN, FineRadScore) and general lexical metrics (BLEU, BERTScore) suggests they capture complementary aspects of report quality. This reveals a significant limitation of general lexical metrics that prioritize stylistic consistency over clinical accuracy. There remains substantial room for improvement in developing metrics that better align with clinical needs, such as HeadCT-ONE,[36] which utilizes ontology normalization to make evaluation more robust to variations in radiological language while allowing customizable weighting of different clinical entities.

Clinical Implications. The observed error rates, particularly in abnormal studies where fewer than 45% of cases are error-free, indicate that the technology is not yet ready for fully autonomous reporting. Nevertheless, these models show considerable promise as assistive tools that can generate preliminary report drafts for radiologist verification and refinement. This collaborative approach has the potential to simultaneously reduce radiologist workload and enhance clinical accuracy, as multiple recent studies have substantiated .[3,37,38]

Future Directions Our findings suggest several promising directions for advancing automated radiology report generation:

- **Improving abnormality detection.** Future research should focus specifically on enhancing model performance for abnormal studies, perhaps through preference fine-tuning,[39] or two-stage approaches combining classification/segmentation with report generation.[40]
- **Cross-institutional robustness:** Developing techniques to improve generalization across diverse clinical settings represents a critical research priority. This might include domain adaptation methods, adversarial training to reduce institutional bias, or federated learning approaches that preserve privacy while leveraging multi-institutional data.[41] Initiatives such as MAIDA,[42] which coordinates data sharing from 69 hospitals across 28 countries, offer promising frameworks for collaborative data sharing that could help address these challenges while ensuring diverse representation across global healthcare environments.
- **Enhanced evaluation frameworks:** The complementary nature of different evaluation metrics suggests value in developing more comprehensive assessment frameworks that combine strengths of traditional and LLM-based approaches while addressing their respective limitations.

- **Human-AI collaboration:** Given the performance gap between normal and abnormal cases, research into effective human-AI collaboration workflows could help optimize the clinical utility of these systems while ensuring patient safety.

Limitations. Our study has several important limitations to consider. First, while ReXGradient represents one of the largest and most diverse evaluation datasets to date, it still cannot fully capture the heterogeneity of chest X-ray imaging and reporting practices globally. Cultural and regional variations in radiology practice may impact the generalizability of our findings. Second, our evaluation framework did not systematically account for the presence of prior studies. Some models are designed to incorporate prior imaging and reports as additional context, but our standardized evaluation did not consistently leverage this capability across all models. Additionally, ground truth reports often contain references to prior studies and temporal changes that may be difficult for models to generate without access to the relevant prior information. This limitation may disadvantage models designed to utilize longitudinal data and could explain some performance variations across different clinical scenarios. Third, our evaluation relied on comparing AI-generated reports to human-written references, which inherently assumes that the reference reports are optimal. However, inter-radiologist variability in reporting style and content means that valid alternative phrasings or observations might be unfairly penalized by reference-based metrics. Fourth, while we employed multiple evaluation metrics to assess different aspects of report quality, these metrics cannot fully capture the clinical utility of generated reports. Important factors like actionability, clinical relevance, and communication effectiveness are challenging to quantify with automated metrics. Finally, our analysis focused on English-language reports, limiting generalizability to non-English healthcare settings. Language-specific nuances and reporting conventions may impact both model performance and metric reliability in other languages. Despite these limitations, the ReXrank Challenge provides valuable insights into the current capabilities and limitations of AI-powered chest X-ray report generation, establishing important benchmarks and highlighting critical areas for future research and development.

Acknowledgement

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

References

1. V. M. Rao, M. Hla, M. Moor, S. Adithan, S. Kwak, E. J. Topol and P. Rajpurkar, Multimodal generative ai for medical image interpretation, *Nature* **639**, 888 (2025).
2. M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol and P. Rajpurkar, Foundation models for generalist medical artificial intelligence, *Nature* **616**, 259 (2023).
3. R. Tanno, D. G. Barrett, A. Sellergren, S. Ghaisas, S. Dathathri, A. See, J. Welbl, C. Lau, T. Tu, S. Azizi *et al.*, Collaboration between clinicians and vision–language models in radiology report generation, *Nature Medicine* **31**, 599 (2025).
4. S. Bannur, K. Bouzid, D. C. Castro, A. Schwaighofer, S. Bond-Taylor, M. Ilse, F. Pérez-García, V. Salvatelli, H. Sharma, F. Meissen *et al.*, Maira-2: Grounded radiology report generation, *arXiv preprint arXiv:2406.04449* (2024).

5. A. Nicolson, J. Dowling and B. Koopman, Improving chest x-ray report generation by leveraging warm starting, *Artificial intelligence in medicine* **144**, p. 102633 (2023).
6. P. Chambon, J.-B. Delbrouck, T. Sounack, S.-C. Huang, Z. Chen, M. Varma, S. Q. Truong, C. T. Chuong and C. P. Langlotz, Chexpert plus: Hundreds of thousands of aligned radiology texts, images and patients, *arXiv preprint arXiv:2405.19538* (2024).
7. W. Chen, Y. Liu, C. Wang, J. Zhu, S. Zhao, G. Li, C.-L. Liu and L. Lin, Cross-modal causal intervention for medical report generation, *arXiv preprint arXiv:2303.09117* (2023).
8. X. Zhang, H.-Y. Zhou, X. Yang, O. Banerjee, J. N. Acosta, J. Miller, O. Huang and P. Rajpurkar, Rexrank: A public leaderboard for ai-powered radiology report generation, *arXiv preprint arXiv:2411.15122* (2024).
9. K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, (Association for Computational Linguistics, 2002).
10. T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger and Y. Artzi, Bertscore: Evaluating text generation with bert, *arXiv preprint arXiv:1904.09675* (2019).
11. A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng and M. P. Lungren, Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert, *arXiv preprint arXiv:2004.09167* (2020).
12. S. Jain, A. Agrawal, A. Saporta, S. Q. Truong, D. N. Duong, T. Bui, P. Chambon, Y. Zhang, M. P. Lungren, A. Y. Ng *et al.*, Radgraph: Extracting clinical entities and relations from radiology reports, *arXiv preprint arXiv:2106.14463* (2021).
13. W. Zhao, C. Wu, X. Zhang, Y. Zhang, Y. Wang and W. Xie, Ratescore: A metric for radiology report generation, *medRxiv* , 2024 (2024).
14. S. Ostmeier, J. Xu, Z. Chen, M. Varma, L. Blankemeier, C. Bluethgen, A. E. Michalson, M. Moseley, C. Langlotz, A. S. Chaudhari *et al.*, Green: Generative radiology report evaluation and error notation, *arXiv preprint arXiv:2405.03595* (2024).
15. A. Huang, O. Banerjee, K. Wu, E. P. Reis and P. Rajpurkar, Fineradscore: A radiology report line-by-line evaluation technique generating corrections with severity scores, *arXiv preprint arXiv:2405.20613* (2024).
16. H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, Llama: Open and efficient foundation language models, *arXiv preprint arXiv:2302.13971* (2023).
17. L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing *et al.*, Judging llm-as-a-judge with mt-bench and chatbot arena, *Advances in Neural Information Processing Systems* **36**, 46595 (2023).
18. Y. Luo, J. Zhang, S. Fan, K. Yang, Y. Wu, M. Qiao and Z. Nie, Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine, *arXiv preprint arXiv:2308.09442* (2023).
19. O. Alfarghaly, R. Khaled, A. Elkorany, M. Helal and A. Fahmy, Automated radiology report generation using conditioned transformers, *Informatics in Medicine Unlocked* **24**, p. 100557 (2021).
20. T. Tanida, P. Müller, G. Kaissis and D. Rueckert, Interactive and explainable region-guided radiology report generation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (Association for Computational Linguistics, 2023).
21. Z. Chen, M. Varma, J.-B. Delbrouck, M. Paschali, L. Blankemeier, D. Van Veen, J. M. J. Valanarasu, A. Youssef, J. P. Cohen, E. P. Reis *et al.*, Chexagent: Towards a foundation model for chest x-ray interpretation, *arXiv preprint arXiv:2401.12208* (2024).
22. C. Pellegrini, E. Özsoy, B. Busam, N. Navab and M. Keicher, Radialog: A large vision-language model for radiology report generation and conversational assistance, *arXiv preprint arXiv:2311.18681* (2023).
23. H.-Y. Zhou, S. Adithan, J. N. Acosta, E. J. Topol and P. Rajpurkar, A generalist learner for

multifaceted medical image interpretation, *arXiv preprint arXiv:2405.07988* (2024).
24. C. Wu, X. Zhang, Y. Zhang, Y. Wang and W. Xie, Towards generalist foundation model for radiology, *arXiv preprint arXiv:2308.02463* (2023).
25. X. Zhang, Z. Meng, J. Lever and E. S. Ho, Libra: Leveraging temporal images for biomedical radiology analysis, *arXiv preprint arXiv:2411.19378* (2024).
26. A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark and S. Horng, Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports, *Scientific data* **6**, p. 317 (2019).
27. D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma and C. J. McDonald, Preparing a collection of radiology examinations for distribution and retrieval, *Journal of the American Medical Informatics Association* **23**, 304 (2016).
28. A. Nicolson, J. Dowling, D. Anderson and B. Koopman, Longitudinal data and a semantic similarity reward for chest x-ray report generation, *Informatics in Medicine Unlocked* **50**, p. 101585 (2024).
29. A. Nicolson, S. Zhuang, J. Dowling and B. Koopman, The impact of auxiliary patient data on automated chest x-ray report generation and how to incorporate it, *arXiv preprint arXiv:2406.13181* (2024).
30. A. Nicolson, J. Liu, J. Dowling, A. Nguyen and B. Koopman, e-health csiro at rrg24: entropy-augmented self-critical sequence training for radiology report generation, *arXiv preprint arXiv:2408.03500* (2024).
31. M. Ranjit, S. Srivastav and T. Ganu, Radphi-3: Small language models for radiology, *arXiv preprint arXiv:2411.13604* (2024).
32. K. Zhang, R. Zhou, E. Adhikarla, Z. Yan, Y. Liu, J. Yu, Z. Liu, X. Chen, B. D. Davison, H. Ren *et al.*, A generalist vision–language foundation model for diverse biomedical tasks, *Nature Medicine* , 1 (2024).
33. Z. Yang, L. Li, K. Lin, J. Wang, C.-C. Lin, Z. Liu and L. Wang, The dawn of lmms: Preliminary explorations with gpt-4v (ision), *arXiv preprint arXiv:2309.17421* **9**, p. 1 (2023).
34. S. Lee, W. J. Kim, J. Chang and J. C. Ye, Llm-cxr: Instruction-finetuned llm for cxr image understanding and generation, *arXiv preprint arXiv:2305.11490* (2023).
35. F. Yu, M. Endo, R. Krishnan, I. Pan, A. Tsai, E. P. Reis, E. K. U. N. Fonseca, H. M. H. Lee, Z. S. H. Abad, A. Y. Ng *et al.*, Evaluating progress in automatic chest x-ray radiology report generation, *Patterns* **4** (2023).
36. J. N. Acosta, X. Zhang, S. Dogra, H.-Y. Zhou, S. Payabvash, G. J. Falcone, E. K. Oermann and P. Rajpurkar, Headct-one: Enabling granular and controllable automated evaluation of head ct radiology report generation, *arXiv preprint arXiv:2409.13038* (2024).
37. J. N. Acosta, S. Dogra, S. Adithan, K. Wu, M. Moritz, S. Kwak and P. Rajpurkar, The impact of ai assistance on radiology reporting: A pilot study using simulated ai draft reports, *arXiv preprint arXiv:2412.12042* (2024).
38. E. K. Hong, B. Roh, B. Park, J.-B. Jo, W. Bae, J. Soung Park and D.-W. Sung, Value of using a generative ai model in chest radiography reporting: a reader study, *Radiology* **314**, p. e241646 (2025).
39. D. Hein, Z. Chen, S. Ostmeier, J. Xu, M. Varma, E. P. Reis, A. E. Michalson, C. Bluethgen, H. J. Shin, C. Langlotz *et al.*, Preference fine-tuning for factuality in chest x-ray interpretation models without human feedback, *arXiv preprint arXiv:2410.07025* (2024).
40. P. R. Bassi, M. C. Yavuz, K. Wang, X. Chen, W. Li, S. Decherchi, A. Cavalli, Y. Yang, A. Yuille and Z. Zhou, Radgpt: Constructing 3d image-text tumor datasets, *arXiv preprint arXiv:2501.04678* (2025).
41. P. R. Bassi, W. Li, Y. Tang, F. Isensee, Z. Wang, J. Chen, Y.-C. Chou, Y. Kirchhoff, M. R. Rokuss, Z. Huang *et al.*, Touchstone benchmark: Are we on the right way for evaluating ai

algorithms for medical segmentation?, *Advances in Neural Information Processing Systems* **37**, 15184 (2024).
42. A. Saenz, E. Chen, H. Marklund and P. Rajpurkar, The maida initiative: establishing a framework for global medical-imaging data sharing, *The Lancet Digital Health* **6**, e6 (2024).