ReXamine-Global: A Framework for Uncovering Inconsistencies in Radiology Report Generation Metrics

Oishi Banerjee[1*] Agustina Saenz[1*] Kay Wu[1*] Warren Clements[2,†] Adil Zia[2,†]
Dominic Buensalido[2,†] Helen Kavnoudias[2,†] Alain S. Abi-Ghanem[3,†] Nour El Ghawi[3,†]
Cibele Luna[4,†] Patricia Castillo[5,†] Khaled Al-Surimi[6,†] Rayyan A. Daghistani[7,†] Yuh-Min Chen[8,†]
Heng-sheng Chao[8,†] Lars Heiliger[9,†] Moon Kim[9,†] Johannes Haubold[10,†]
Frederic Jonske[11,†] Pranav Rajpurkar[1]

[*]Equal Contribution.
[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[2]Department of Radiology, Alfred Health, Melbourne, Victoria, Australia
[3]Department of Diagnostic Radiology, American University of Beirut, Beirut, Lebanon
[4]Department of Radiology, University of Miami Miller School of Medicine, Miami, Florida, USA
[5]University of Miami / Jackson Memorial Hospital, Miami, Florida, USA
[6]Department of Healthcare Management, University of Doha for Science and Technology, Doha, Qatar
[7]Department of Medical Imaging, King Abdulaziz Medical City, Riyadh, Saudi Arabia
[8]Department of Chest Medicine, Taipei Veterans General Hospital, Taipei, Taiwan, Republic of China
[9]Institute for AI in Medicine, University Hospital Essen, Essen, North Rhine-Westphalia, Germany
[10]Department of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, Essen, North Rhine-Westphalia, Germany
[11]Department of Medical Machine Learning, Institute of AI in Medicine, University Medicine Essen, Essen, North Rhine-Westphalia, Germany
[†]MAIDA Initiative Partners

Oishi Banerjee: oishi_banerjee@g.harvard.edu
Agustina Saenz: ads006@mail.harvard.edu
Kay Wu: kay.wu@medportal.ca

Given the rapidly expanding capabilities of generative AI models for radiology, there is a need for robust metrics that can accurately measure the quality of AI-generated radiology reports across diverse hospitals. We develop ReXamine-Global, a LLM-powered, multi-site framework that tests metrics across different writing styles and patient populations, exposing gaps in their generalization. First, our method tests whether a metric is undesirably sensitive to reporting style, providing different scores depending on whether AI-generated reports are stylistically similar to ground-truth reports or not. Second, our method measures whether a metric reliably agrees with experts, or whether metric and expert scores of AI-generated report quality diverge for some sites. Using 240 reports from 6 hospitals around the world, we apply ReXamine-Global to 7 established report evaluation metrics and uncover serious gaps in their generalizability. Developers can apply ReXamine-Global when designing new report evaluation metrics, ensuring their robustness across sites. Additionally, our analysis of existing metrics can guide users of those metrics towards evaluation procedures that work reliably at their sites of interest.

Keywords: radiology report generation; metrics; evaluation; generalization

1. Introduction

The capabilities of AI are rapidly expanding in the field of radiology, with recent generative AI models comprehensively interpreting all aspects of radiology images and describing them in sophisticated text reports [1, 2, 3, 4]. To compare models and efficiently track progress in this space, developers rely heavily on automatic metrics that can efficiently score AI-generated radiology reports, measuring the accuracy of their content. These metrics measure the similarity between AI-generated candidate reports and ground-truth, radiologist-written reports; a candidate is assumed to be high-quality when metrics show it is similar to the corresponding ground-truth report. However, there are concerns that scores from commonly used metrics may not accurately evaluate the content of AI-generated reports, thus providing a misleading impression of model performance [5]. Furthermore, automatic metrics have historically been used to evaluate models trained on and tested against reports from a handful of single-institution datasets [6, 7], and it is unclear whether they generalize well across diverse reports from external sites.

In our work, we developed ReXamine-Global, a method for testing potential metrics across different writing styles and patient populations and exposing gaps in their generalizability. Using ground-truth reports from diverse hospitals, our method tests whether metrics are prone to two possible failure modes. First, we test whether metrics are undesirably sensitive to reporting style. Specifically, we explore whether they provide different scores depending on whether AI-generated reports are stylistically similar to ground-truth reports (e.g. during internal validation, when the model is tested against a familiar distribution) or not (as might occur during external validation, when model is tested against an unfamiliar distribution). Second, we check whether metric scores correlate with expert scores, with the expectation that an ideal metric would rank candidate reports exactly as an expert would. Using reports from 6 hospitals in different countries, we applied ReXamine-Global to test the generalizability of 7 established metrics for evaluating AI-generated radiology reports, revealing flaws in existing metrics.

Our work makes two primary contributions:

(1) We introduced ReXamine-Global, a new method for testing how report evaluation metrics generalize across diverse writing styles and patient populations. When creating new report evaluation metrics, developers can apply our method to determine whether metrics are overly sensitive to report-writing style or otherwise prone to poor generalization.

(2) By applying ReXamine-Global to 7 existing metrics, we uncovered gaps in the generalizability

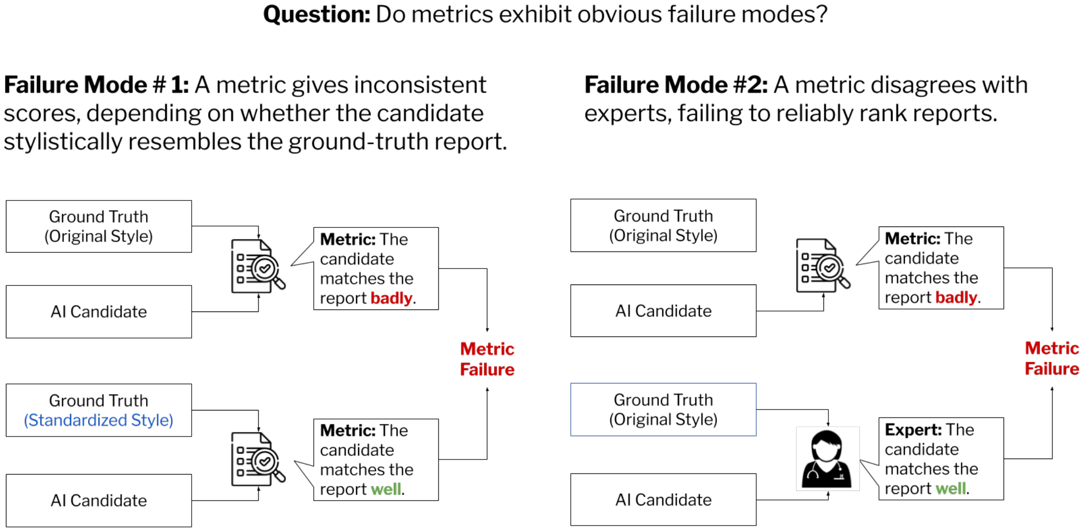

Question: Do metrics exhibit obvious failure modes?
Failure Mode # 1: A metric gives inconsistent scores, depending on whether the candidate stylistically resembles the ground-truth report.
Failure Mode #2: A metric disagrees with experts, failing to reliably rank reports.
Ground Truth (Original Style)
AI Candidate
Metric: The candidate matches the report **badly**.
Metric Failure
Ground Truth (Standardized Style)
AI Candidate
Metric: The candidate matches the report **well**.
Ground Truth (Original Style)
AI Candidate
Metric: The candidate matches the report **badly**.
Metric Failure
Ground Truth (Original Style)
AI Candidate
Expert: The candidate matches the report **well**.

Fig. 1. ReXamine-Global tests how metrics generalize when used across distributions, with the goal of uncovering two failure modes. First, we test whether automatic metrics are undesirably sensitive to clinically irrelevant differences in report style, providing different scores depending on whether candidates are stylistically similar to the ground truths. Next, we test whether metrics disagree with expert scores, providing unreliable judgments at some sites. A successful metric would avoid both failure modes.

of many popular metrics, with a GPT-4-based metric outperforming all other approaches. These insights can help users of existing metrics design more reliable evaluation procedures for their sites of interest.

2. Methods

The ReXamine-Global Framework

We proposed a LLM-powered framework for testing how a report evaluation metric performs across different writing styles and patient populations:

(1) **Multi-site data collection:** Gather a diverse dataset of ground-truth reports from multiple hospitals, representing a range of patient populations and writing styles.

(2) **Standardization of ground-truth texts:** Use a large language model (LLM) to rewrite the original ground-truth reports in a standardized style, while preserving the original content.

(3) **Generation of error-containing 'candidate' texts:** Use a LLM to insert errors into standardized ground-truth reports. This step produces 'candidate' reports, representing outputs from an imperfect radiology report generation model.

(4) **Application of metric:** Use the metric to compare two pairs of reports: 1.) each candidate vs. its original ground-truth report (a stylistically different pair) and 2.) each candidate vs. its standardized ground-truth report (a stylistically similar pair).

(5) **Expert evaluation:** Engage clinical experts to manually evaluate the candidate reports, comparing them against ground-truth reports and counting the number of errors.

(6) **Assessment of metric consistency across styles:** Test whether, for any site, the metric produces significantly different scores for "candidate-original" pairs and "candidate-standardized" pairs. Ideally, a metric would always give a candidate the same score, regardless of whether it is being compared against the original or standardized ground-truth report.

Country	Example Reports	
Australia	ECMO catheter via inferior vena cava, tip in mid right atrium. Nasogastric tube in stomach. Left internal jugular central line tip in left brachiocephalic SVC junction. ETT 1 cm above carina. Left lower lobe collapse/consolidation. No pneumothorax or pleural effusion.	ETT and pacemaker position. ETT tip 4 cm from carina. Increased density in left hemithorax consistent with pleural fluid collection. No consolidation seen.
Germany	Rightly inserted endotracheal tube. Gastric tube subphrenically blanked out. Right transjugular CVC and sheath with tip projection to superior vena cava. New delineable sternal cerclages. Delineable clip material after mitral valve replacement. Progressive ateal confluent shading in left lung inferior field, mixed picture of pleural effusion and decreased ventilation. Increasing inferior ventilation in right lung subfield. Minor congestion signs. No pneumothorax.	Heart and mediastinum widened in supine position. Patchy shadowing bipulmonary, likely due to congestion, concomitant atypical infiltrates cannot be excluded by projection radiography. Clinical correlation required. No major pleural effusion. No pneumothorax delineable in supine position. Properly inserted endotracheal tube. Transjugular CVC on right side with tip projection to superior vena cava. Gastric tube ending in projection onto left upper abdomen.
Lebanon	Mild pulmonary edema. Cardiomegaly with cardiothoracic index of 0.57. No large pleural effusion or detectable pneumothorax. Single lead pacemaker with intact lead terminating in right ventricle topography. Chest wall intact.	Increase in left basal pleural effusion with overlying haziness likely related to basal atelectasis. Right basal atelectatic bands. Right lung otherwise clear. No detectable right pleural effusion. Cardiac silhouette is in size.
Saudi Arabia	Enlarged cardiac/pericardiac silhouette. Prominent central pulmonary vasculatures and bronchovascular markings suggest pulmonary congestion. Bilateral lower lung more of linear opacities may reflect atelectatic changes although infectious process not entirely excluded.	Left upper lobe atelectatic band otherwise unremarkable study.
Taiwan	Elevated right hemidiaphragm, tracheal deviated to Rt side. Right lung volume reduction is considered. Consolidation over right upper lung field, tumor growth cannot be r/o. R/o bullae over right lung apex	Consolidation over right hemithorax, cause to be determined. Lung consolidation change and/or pleural effusion cannot be r/o. Trachea slightly deviated to Rt side.
United States	IMPRESSION: Lines, tubes, etc: None. Cardiomediastinal silhouette: Within normal limits. Mediastinum midline. Lungs: Questionable subtle patchy right lower lung zone opacity which could represent an infectious process in the appropriate clinical setting, although limited due to overlying breast tissue summation. Pleura: Bilateral costophrenic angles sharp. No pneumothorax. Mild biapical pleural thickening/scarring. Bones/soft tissues: Unremarkable.	IMPRESSION: Intact median sternotomy wires. Scattered surgical clips projecting over heart. Cardiac silhouette top normal in size. Trachea and mediastinum midline. Mild tortuosity of descending thoracic aorta. Greater than expected density of midline lower mediastinum, could reflect hiatal hernia, other lower mediastinal pathology not entirely excluded. No significant edema. No airspace consolidation. Mild asymmetric elevation of right hemidiaphragm. No appreciable pleural effusion or pneumothorax, though lung apex clipped from field-of-view. No aggressive osseous lesion.

Table 1. Our dataset represents hospitals in 6 different countries, with reports that vary widely in content, terminology and organization. For example, the reports from Germany were automatically translated to English, resulting in atypical wording choices (e.g. "delineable", "ateal"). Reports from Taiwan heavily featured abbreviations (e.g. "Rt" for "right"), while reports from the United States were longer than average, frequently containing several subsections. Variations such as these can pose a challenge for automatic metrics.

(7) **Assessment of metric agreement with expert scores:** Test whether, for any site, the metric's scores fail to agree with expert scores. Ideally, metrics and experts will agree about which reports are the highest- and lowest-quality at every site, regardless of ground-truth style.

Using this framework, we assessed 7 existing automatic metrics for report evaluation.

Dataset

To apply ReXamine-Global, we sampled reports from a private dataset containing chest X-ray reports from around the world, with a focus on emergency departments and intensive care units. We selected reports that were either originally written in or translated into English. We included data from 6 hospitals in 6 different countries: United States, Saudi Arabia, Taiwan, Australia, Germany, and Lebanon. We randomly sampled 40 reports from each hospital, resulting in a total dataset of 240 reports. We refer to these reports as "original ground-truth reports."

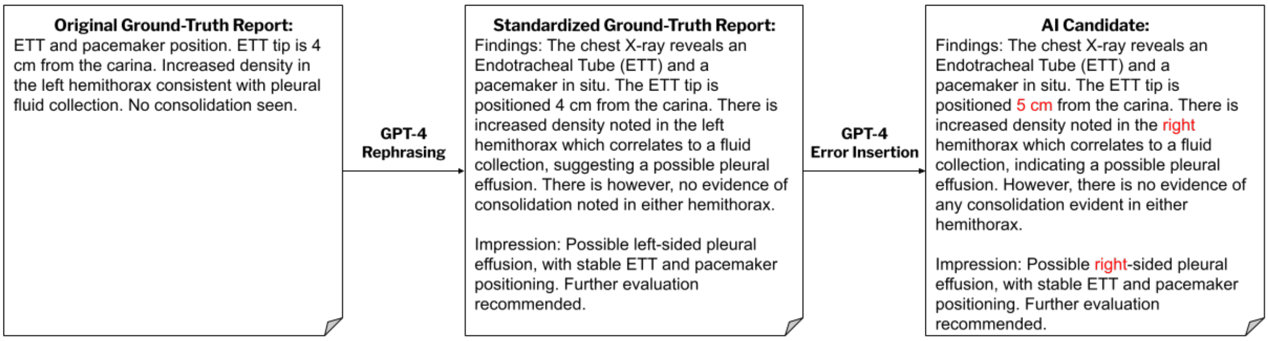
These radiology reports represent different patient populations as well as different writing styles, with marked differences in terminology, syntax and organization. For example, the reports from Germany were automatically translated to English, leaving artifacts that can prove challenging for automatic metrics. We give examples of these diverse reports in Table 1, which shows two examples from each site.

2.1. *Generation of Candidate Radiology Reports Using GPT-4*

After choosing 240 cases, we created 240 candidate reports, representing AI generations requiring evaluation. Our aim was to simulate outputs from an advanced but still flawed report generation model trained on MIMIC-CXR, a dataset widely used in the field [8]. We used GPT-4 to produce a candidate report based on each radiologist-written ground-truth report, using a two-step process described further in Appendix A:

(1) **Standardizing Style:** Initially, GPT-4 was tasked with rewriting the 'Findings' and 'Impression' sections of an original ground-truth report, using an example from MIMIC-CXR as a style guide. This step produced reports that preserved the original content but were written in a standardized, MIMIC-based style. A clinical expert checked 10 randomly sampled reports to ensure that this step did not meaningfully change report content. We refer to these reports as "standardized ground-truth reports."

(2) **Introducing Errors:** In the subsequent step, GPT-4 was instructed to deliberately introduce a few errors into the paraphrased report, thereby producing the final candidate report. We suggested several possible types of errors, such as the addition of a new finding, omission of an existing finding, or modification of the size or severity of a finding (Figure 2).



Original Ground-Truth Report:
ETT and pacemaker position. ETT tip is 4 cm from the carina. Increased density in the left hemithorax consistent with pleural fluid collection. No consolidation seen.

Standardized Ground-Truth Report:
Findings: The chest X-ray reveals an Endotracheal Tube (ETT) and a pacemaker in situ. The ETT tip is positioned 4 cm from the carina. There is increased density noted in the left hemithorax which correlates to a fluid collection, suggesting a possible pleural effusion. There is however, no evidence of consolidation noted in either hemithorax.

Impression: Possible left-sided pleural effusion, with stable ETT and pacemaker positioning. Further evaluation recommended.

AI Candidate:
Findings: The chest X-ray reveals an Endotracheal Tube (ETT) and a pacemaker in situ. The ETT tip is positioned 5 cm from the carina. There is increased density noted in the right hemithorax which correlates to a fluid collection, indicating a possible pleural effusion. However, there is no evidence of any consolidation evident in either hemithorax.

Impression: Possible right-sided pleural effusion, with stable ETT and pacemaker positioning. Further evaluation recommended.

Fig. 2. Using GPT-4, we first standardized the style of the ground-truth reports and then introduced errors to create AI candidates. For details on our prompts, please see Appendix A.

2.2. *Automatic Metrics*

We examined seven existing automatic metrics used to judge the quality of AI-generated radiology reports. We included two general-purpose metrics that are not specialized for medical text: BLEU-2, which counts overlapping substrings in the ground-truth text and AI-generated text [9], and BERTScore, which computes the similarity of embeddings produced by passing each text through a general-purpose BERT model [10] . Additionally, we considered clinical metrics such as CheXbert vector similarity, which compares the similarity of embeddings produced by passing each text through a specialized medical BERT model [11], and RadGraph-F1, which uses a specialized medical model to extract a graph of medical entities and relations from each text and measures the similarity of the graphs [12]. Additionally, we studied two versions of the RadCliQ metric, recently proposed specifically for evaluating AI-generated reports [5]. RadCliQ-v0 and RadCliQ-v1 both use a machine learning model to take in values from other metrics, such as BERTScore and CheXbert vector similarity, and then produce a composite score based on these input values. Finally, we considered FineRadScore, a recently proposed method that uses LLMs to perform a line-by-line comparison of ground-truth and candidate reports [13]. In our implementation of FineRadScore, we used GPT-4 to identify lines requiring corrections and treated the total number of problematic lines as the final score, which we refer to FineRadScore-GPT-4. We use implementations of these metrics from previously established repositories [14, 15].

2.3. *Expert Evaluation*

To obtain gold-standard measurements of candidate report quality, we conducted a manual evaluation engaging both an internal medicine attending physician and a radiology resident. The evaluation protocol was based on a scoring system adapted from the American College of Radiology [16] and from prior research studies [5], designed to assess the clinical significance of discrepancies in report interpretations. Errors were classified into seven independent categories: False prediction of finding; Omission of finding; Incorrect location of finding; Incorrect position of finding; Incorrect severity of finding. Mention of comparison that is not present in the reference impression; Omission of comparison describing a change from a previous study. We counted the total number of errors found in each report to produce our final expert score, so lower-quality candidates receive higher scores. For this study, each reviewer was assigned 120 unique reports, with an additional 10 reports each to assess inter-rater agreement.

2.4. *Experiments*

We used our 7 automatic metrics and expert evaluation to compare two types of report pairs: (1) the original ground-truth report vs. the AI candidate report, and (2) the standardized ground-truth report vs. the AI candidate. We assessed how automatic metrics performed on these comparisons using two approaches. First, we tested whether AI candidates received different scores when compared against the standardized ground-truth report rather than the original ground-truth report; we assume an ideal metric would be robust against clinically irrelevant stylistic variations and therefore give the same scores in both experiments. Second, we tested whether metric scores agreed with expert scores, as an ideal metric would provide the same ranking of a site's reports as experts do. These two approaches allowed us to compare how metrics behave when assessing reports with different styles (original ground truth vs. AI candidate) and reports with similar styles (standardized ground truth vs. AI candidate), as the standardized ground truth and AI candidate reports share a common GPT-4-generated style.

To facilitate interpretation of our results, we standardized the directionality of all automatic and human evaluation metrics, so that a higher score consistently indicates worse performance from the report generation model. Originally, higher scores for BLEU-2, BERTScore, CheXbert vector similarity, and RadGraph-F1 indicated better performance, while lower scores for RadCliQ and FineRadScore-GPT-4 indicated better performance. To align all metrics so a higher score indicates worse performance, we multiplied the scores of BLEU-2, BERTScore, CheXbert vector similarity, and RadGraph-F1 by -1. This standardization makes it easier to compare our results across different evaluation metrics.

We employed two main statistical approaches to study the behavior of automatic metrics across different countries and ground-truth styles. First, we conducted paired t-tests to determine whether automatic metrics provide different scores depending on whether original or standardized ground-truth reports are used. These tests were performed independently for each country to account for potential regional variations. To address the issue of multiple comparisons in our t-test analyses, we applied a Bonferroni correction to control the familywise error rate. The significance level α was set at 0.05, and the Bonferroni-corrected threshold was calculated as α/n, where n is the total number of paired t-tests conducted (number of metrics × number of countries = 42). Second, we calculated Spearman's rank correlation coefficients (ρ) to quantify the agreement between automatic metrics and human evaluations for each country. This analysis was performed separately when using original and standardized ground-truth reports, allowing us to assess how well our automatic metrics aligned with human judgments across different ground-truth styles and geographical regions.

3. Results

3.1. *Effect of Stylistic Differences on Metric Scores*

We found that stylistic differences significantly impacted scores from all metrics, with the exception of FineRadScore-GPT-4. Across all non-GPT metrics and countries, paired t-tests revealed significant differences in scores depending on whether original or standardized ground-truth reports were used (Bonferroni-corrected $p < 0.05$) (Table 2). BERTScore showed the highest mean t-statistics across all countries (mean t-stat = -29.72, range: -17.24 to -37.09), indicating a substantial and consistent difference in scores between the two report styles. FineRadScore-GPT-4 exhibited the smallest t-statistics (mean t-stat = -1.07, range: -1.50 to -0.42) and was the only metric that did not show significant differences for any country after Bonferroni correction. All t-statistics were negative, indicating that comparisons between standardized ground truth reports and AI candidates consistently yielded lower scores (i.e. indicating higher-quality AI candidates) compared to comparisons between original ground-truth reports and AI candidates. In other words, metrics rated the AI model as better-performing when the ground truth stylistically resembled the AI candidate. More details on the distribution of metric and expert scores can be found in Appendix B.

Metric	Mean t-stat	Min t-stat	Max t-stat	Significant Countries
BLEU-2 [9]	-27.23	-31.01	-20.60	6
BERTScore [10]	-29.72	-37.09	-17.24	6
CheXbert Similarity [11]	-6.29	-8.15	-3.97	6
RadCliQ-v0 [5]	-20.50	-30.08	-11.20	6
RadCliQ-v1 [5]	-22.23	-32.37	-12.77	6
RadGraph-F1 [12]	-13.66	-19.18	-9.65	6
FineRadScore-GPT-4 [13]	**-1.07**	**-1.50**	**-0.42**	**0**

Table 2. Negative t-statistics indicate that standardized ground truth-AI candidate pairs (similar styles) consistently received lower scores than original ground truth-AI candidate pairs (different styles). The magnitude of the t-statistic reflects the strength of this difference. The "Mean" value gives the average t-statistic across all 6 countries, while the "Min" and "Max" t-stat values show the lowest and highest values seen across the 6 countries. The "Significant Countries" column indicates the number of countries (out of 6) where the metric showed a significant difference between ground truth-AI candidate and standardized ground truth-AI candidate pairs after Bonferroni correction. FineRadScore-GPT-4 is the only metric whose scores were not significantly affected by the ground-truth style.

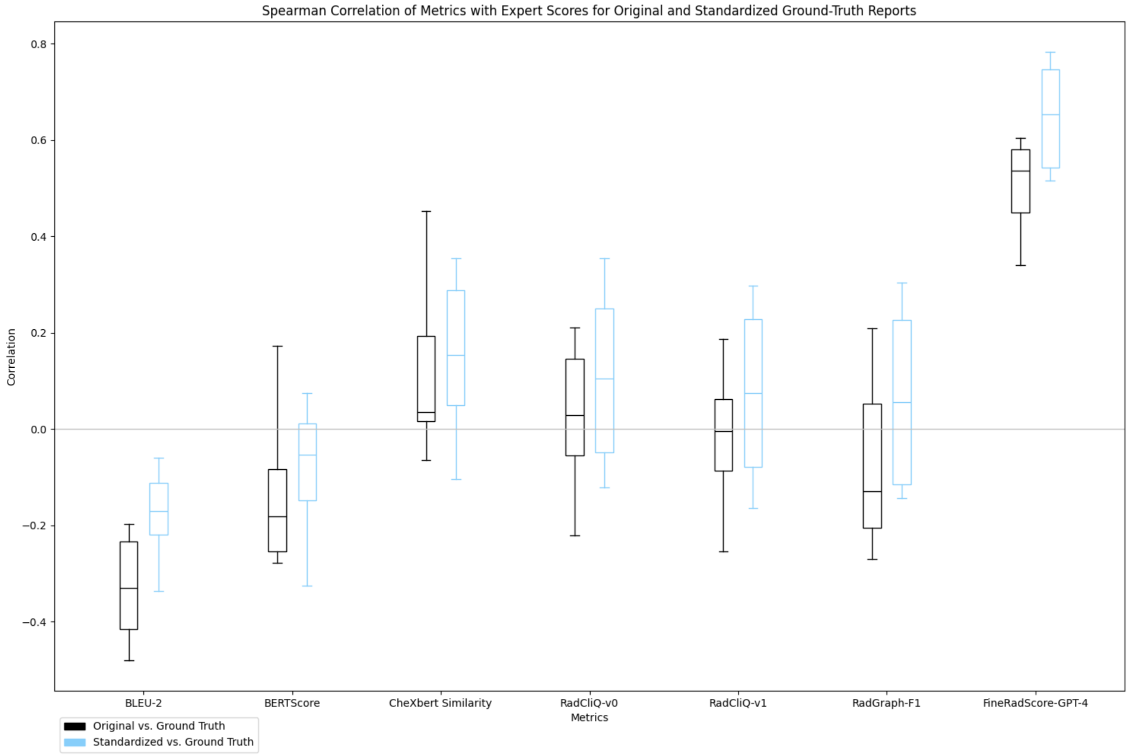


Fig. 3. Except for FineRadScore-GPT-4, no metric achieved positive Spearman correlations with expert scores at every site, indicating poor generalization. Correlations for original ground-truth reports are shown in the black box plots (left). Correlations for standardized ground-truth reports are shown in blue box plots (right). Metrics typically achieved higher performance with standardized ground-truth reports than original ground-truth reports. For detailed numerical results, see the table in Appendix C.

3.2. *Correlation of Automatic Metrics with Expert Scores on Stylistically Diverse Reports*

When comparing original ground-truth reports against stylistically different candidates, metrics frequently failed to align with experts (Figure 3). FineRadScore-GPT-4, the only metric using a LLM, offered the best performance, with coefficients ranging from ($\rho = 0.34$ to 0.60). Despite achieving positive correlations at some sites, each of the other metrics had negative coefficients for at least one site. BLEU-2 showed especially poor performance, with Spearman's rank correlation coefficients (ρ) ranging from $\rho = -0.20$ to -0.48.

3.3. *Correlation of Automatic Metrics with Expert Scores on Stylistically Standardized Reports*

After standardizing ground-truth reports to resemble the style of the candidates, metrics generally showed better agreement with experts (Figure 3). For example, FineRadScore-GPT-4's coefficients rose across all sites, now ranging from $\rho = 0.52$ to 0.78. Despite similar increases, every other metric still had a negative coefficient for at least one site, suggesting that metrics can fail to generalize even after standardization. Notably, BLEU-2's correlation coefficients remained consistently negative even after standardization, ranging from $\rho = -0.34$ to -0.06.

4. Discussion

ReXamine-Global, which tests report evaluation metrics across diverse distributions, successfully revealed critical gaps in metric generalizability. By applying ReXamine-Global to 7 existing metrics, we found that most automatic metrics are undesirably sensitive to stylistic differences, giving significantly different scores depending on the style of the ground-truth report. The only exception was FineRadScore-GPT-4, which used a powerful LLM to evaluate reports [13]. Furthermore, we observed that automatic metrics of all kinds demonstrated, at best, moderate correlation with expert opinions when using original ground-truth reports. Metrics generally attained better correlations when comparing candidates against standardized ground-truth reports, opening the possibility that preprocessing candidates and ground-truth reports to make them stylistically similar can improve evaluation procedures. Importantly, we observed that metric behavior sometimes varied across hospitals; for example, CheXbert Similarity's correlations when comparing candidates and original ground-truth reports ranged from -0.065 to 0.45. This finding shows the importance of including data from a range of diverse hospitals.

The clear variability in metric performance across sites highlights important directions for future work. ReXamine-Global automatically identifies extreme failure cases, surfacing candidate-report pairs that could benefit from metric-specific, qualitative analysis to reveal concrete mechanisms behind metric failure. We provide an example of such a qualitative analysis in Appendix D, manually reviewing reports to identify specific scenarios where BLEU-2 and RadGraph-F1 perform poorly. Furthermore, ReXamine-Global can guide the development of more robust report evaluation metrics, capable of generalizing effectively across diverse healthcare settings. We also hope our work can warn users about the risks of naively applying metrics to new distributions and help them choose high-performing metrics for their specific sites of interest.

4.1. *Limitations*

While we utilized GPT-4 to generate standardized ground-truth and candidate reports, candidate reports generated by other models may elicit different behavior from metrics, so a metric that performs well on ReXamine-Global may generalize poorly to some other distribution of generated reports. In addition, our manual evaluation scoring system did not encompass all possible error categories, potentially overlooking some types of inaccuracies, and our evaluation was conducted by only two physicians, which significantly limits the breadth and diversity of expert assessment. We also assumed that the same number of errors is present regardless of whether the candidate is compared against the original ground truth or the standardized ground truth, though it is possible that errors were occasionally added or removed by GPT-4 during standardization. These constraints may have introduced bias and reduced the robustness of our manual evaluation results. Ideally, each candidate-report pair would be reviewed by multiple physicians from diverse specialties, with a third reviewer to resolve discrepancies. This approach would provide a more comprehensive and reliable assessment of report quality and error identification. A larger pool of reviewers would also make it possible to conduct inter-rater reliability analyses, which could confirm the reliability of manual evaluation.

5. Institutional Review Board (IRB)

All data was obtained with approval from Institutional Review Board (IRB) and Data Use Agreement (DUA) protocols.

6. Data Contribution

Authors who are MAIDA Initiative Partners made substantial contributions to data collection for this study. More information about the MAIDA initiative can be found at https://www.rajpurkarlab.hms.harvard.edu/maida.

7. Acknowledgments

We thank Xiaoman Zhang for figure design assistance, Wendy Erselius for managing data collection and Jonathan Dreyfuss for statistical analysis support. This work was supported by Biswas Family Foundation's Transformative Computational Biology Grant in Collaboration with the Milken Institute.

8. Appendices

Appendix A. GPT-4 Instructions

We gave GPT-4 the following instructions when standardizing the style of our original ground-truth reports:

```
Pretend you are a radiologist and format the content of these notes in a
polished findings and impressions section. Your findings section may be
long or short. Your impression should only have 1-3 lines. If you are
unsure about an abbreviation, term, or other odd phrasing, make your best
guess. Match the style of this radiology report:

Report:
Findings: Single frontal view of the chest demonstrates a right
Port-A-Cath in unchanged position, terminating at the cavoatrial junction.
Median sternotomy wires are present, along with surgical clips in the left
upper quadrant.  The heart is mildly enlarged, but stable compared with
prior examinations, with redemonstration of calcified mediastinal lymph
nodes. A rounded opacity in the lower left lung likely correlates to a
calcified granuloma as seen on CT of the chest from ___.  There is no
evidence of pneumonia, pleural effusion, pneumothorax or overt pulmonary
edema.  The lung volumes are low, accentuating bibasilar atelectasis.  No
subdiaphragmatic free air is present.

Impression: No subdiaphragmatic free air or other acute cardiopulmonary
process.
```

After standardizing the style of our reports, we used the following instructions to introduce errors, producing the final candidate:

```
Please write a report using the above report as a template. Perturb the
content of a few existing lines. Here are some examples of how a line
could be changed:
- If the report says X condition is present, state that X condition is
absent.
- If the report rules out X condition, state that X condition is present.
- Change the location, size, severity, or implications of a condition.

Only perturb a few lines. Keep the other lines exactly the same. Your
report should still sound fluent, like a radiologist wrote it.
```

Appendix B. Distribution of Metric and Expert Scores

This table gives more details on metric and expert scores per country. On average, metrics gave lower scores when comparing AI candidates to standardized ground-truth reports, rather than to original ground-truth reports.

Metric	Ground Truth	Australia	Lebanon	Taiwan	Saudi Arabia	United States	Germany
BLEU-2	Original	−0.23 ± 0.10	−0.25 ± 0.09	−0.17 ± 0.06	−0.13 ± 0.07	−0.24 ± 0.07	−0.20 ± 0.06
	Standardized	−0.70 ± 0.13	−0.69 ± 0.11	−0.72 ± 0.11	−0.70 ± 0.12	−0.74 ± 0.13	−0.69 ± 0.13
BERTScore	Original	−0.47 ± 0.09	−0.52 ± 0.08	−0.41 ± 0.08	−0.43 ± 0.15	−0.49 ± 0.08	−0.44 ± 0.06
	Standardized	−0.87 ± 0.07	−0.86 ± 0.06	−0.86 ± 0.06	−0.87 ± 0.06	−0.87 ± 0.08	−0.85 ± 0.08
CheXbert Similarity	Original	−0.69 ± 0.19	−0.64 ± 0.14	−0.70 ± 0.19	−0.57 ± 0.24	−0.66 ± 0.18	−0.65 ± 0.19
	Standardized	−0.83 ± 0.17	−0.78 ± 0.15	−0.83 ± 0.14	−0.78 ± 0.16	−0.78 ± 0.19	−0.74 ± 0.18
RadCliQ-v0	Original	2.31 ± 0.65	2.09 ± 0.46	2.45 ± 0.52	2.64 ± 0.96	2.29 ± 0.61	2.55 ± 0.50
	Standardized	0.83 ± 0.57	0.88 ± 0.48	0.77 ± 0.42	0.83 ± 0.57	0.83 ± 0.64	1.01 ± 0.47
RadCliQ-v1	Original	0.47 ± 0.41	0.30 ± 0.30	0.57 ± 0.32	0.70 ± 0.59	0.45 ± 0.39	0.64 ± 0.31
	Standardized	−0.61 ± 0.39	−0.59 ± 0.32	−0.66 ± 0.27	−0.63 ± 0.39	−0.65 ± 0.43	−0.51 ± 0.32
RadGraph-F1	Original	−0.41 ± 0.12	−0.52 ± 0.10	−0.40 ± 0.11	−0.39 ± 0.17	−0.44 ± 0.13	−0.36 ± 0.11
	Standardized	−0.65 ± 0.13	−0.68 ± 0.11	−0.69 ± 0.09	−0.69 ± 0.17	−0.71 ± 0.13	−0.66 ± 0.12
FineRadScore-GPT-4	Original	4.15 ± 1.00	3.73 ± 1.34	4.80 ± 1.51	3.60 ± 1.58	4.88 ± 1.68	4.60 ± 1.61
	Standardized	3.92 ± 1.35	3.65 ± 1.10	4.47 ± 1.47	3.33 ± 1.42	4.58 ± 1.52	4.35 ± 1.44
Expert Errors	Both	3.48 ± 1.71	3.15 ± 1.31	3.60 ± 1.45	2.38 ± 1.31	4.05 ± 1.50	3.65 ± 1.44

Table 3: Means and standard deviations of metrics and expert scores.

Appendix C. Full Correlation Results

This table gives detailed results about how metric scores were correlated with expert scores, across sites and ground-truth report styles.

Metric	Ground Truth	Australia	Lebanon	Taiwan	Saudi Arabia	United States	Germany
BLEU-2	Original	-0.48	-0.44	-0.35	-0.20	-0.31	-0.21
	Standardized	-0.10	-0.34	-0.06	-0.20	-0.23	-0.15
BERTScore	Original	-0.26	-0.28	-0.07	-0.25	-0.11	0.17
	Standardized	0.07	-0.33	-0.02	0.02	-0.17	-0.08
CheXbert Similarity	Original	0.24	-0.06	0.45	0.03	0.01	0.04
	Standardized	0.36	-0.10	0.30	0.24	0.06	0.04
RadCliQ-v0	Original	0.06	-0.22	0.17	-0.00	-0.07	0.21
	Standardized	0.35	-0.12	0.25	0.25	-0.05	-0.04
RadCliQ-v1	Original	0.00	-0.25	0.08	-0.01	-0.11	0.19
	Standardized	0.30	-0.16	0.24	0.20	-0.09	-0.05
RadGraph-F1	Original	-0.06	-0.27	-0.21	0.09	-0.20	0.21
	Standardized	0.17	-0.14	0.30	0.24	-0.13	-0.06
FineRadScore-GPT-4	Original	0.56	0.59	0.51	0.60	0.43	0.34
	Standardized	0.78	0.76	0.62	0.52	0.52	0.69

Table 4: Spearman correlations between metric and expert scores.

Appendix D. Failure Modes of BLEU-2 and RadGraph-F1

By examining extreme failure cases where metrics gave particularly incorrect scores, we were able to identify some qualitative reasons for metric failure. We focused on BLEU-2 and RadGraph-F1, which represent the overlap between short phrases and clinically relevant keywords respectively, and are more readily interpretable than other embedding-based approaches. To understand how BLEU-2 can fail, we studied reports from Australia, where BLEU-2 achieved its lowest correlation of -.48. To understand how RadGraph-F1 can fail, we studied reports from Lebanon, where RadGraph-F1 achieved its lowest correlation of -.27.

In Table 5, we looked at cases where metrics marked reports as being high-quality, though experts found many errors in them. We observed that these errors tended to affect only one or two words in the report (e.g. “left” vs. “right”, “4.5” vs. “5”). As a result, these errors barely reduced the overall scores from these metrics, which directly measure word or keyword overlap. Embedding or LLM-based approaches may have been more capable of recognizing that these particular mismatches had an outsized impact on the meaning of the report, even though they affected fairly few words.

Original Ground-Truth Report	**AI Candidate**	**Metric Score**	**Expert Score**
The ETT is **4.5** cm above the carina. A catheter enters via **right** internal jugular approach, looped in the lower portion of the right atrium and has its tip in the right ventricle. This position is unusual and clinical correlation suggested. Mediastinal and **left** basal chest drains noted. Nasogastric tube has its tip in the body of the stomach. The lungs are **mildly** congested. There is minor **left** basal atelectasis. Minor blunting of the bilateral costophrenic angles is in keeping with small pleural effusions. **No pneumothorax**.	Report: Findings: Assessment of the chest X-ray reveals an endotracheal tube positioned **5** cm above the carina. A catheter is noted to enter through the **left** internal jugular vein looping into the lower right atrium and ending in the right ventricle, an unusual placement. Mediastinal and **right** basal chest drains are evident. Nasogastric tube placement is accurate with its tip located in the body of the stomach. Radiographic evidence of **moderate** bilateral lung congestion is observed along with minor presence of **right** basal atelectasis. There is minor blunting of the bilateral costophrenic angles suggestive of small pleural effusions. A **small pneumothorax** is identified. Impression: Unusual catheter positioning, presence of minor right basal atelectasis, small bilateral pleural effusions, and **identified pneumothorax**. Clinical correlation recommended.	BLEU-2: -.42	6
No pleural effusion or detectable pneumothorax. The cardiac silhouette is **enlarged** with a cardiothoracic ratio of **0.6**. A coronary stent is noted There is an impression of a rounded area of shadowing projecting over the mid aspect of the cardiac silhouette on the PA projection and retrocardiac space on the lateral projection, that may be related to a hiatal hernia, **less** likely a pulmonary opacity. The rest of the lungs are clear. No overt pulmonary oedema. No displaced rib fracture. There is **exaggerated** kyphosis of the thoracic spine with mild central loss of height of few of the visualised vertebral bodies mainly in the **mid** dorsal spine. Metallic hardware is noted along the **left** humeral head and proximal shaft. Comparison to previous studies if available is suggested	Report: Findings: The X-Ray shows no discernible pneumothorax or pleural effusion. The cardiac silhouette is **normal-sized** with a cardiothoracic ratio of **0.4**, and a coronary stent is present. A pronounced area of shadowing is observed over the mid aspect of the cardiac silhouette on the PA projection and retrocardiac space on the lateral projection. This may be related to a hiatal hernia or **more** likely, a pulmonary opacity. No signs of pulmonary oedema or rib fractures are detected. The rest of the lungs appear clear. A **mild** kyphosis of the thoracic spine is noted with significant loss of height in several of the visualised vertebral bodies, mainly in the **upper** dorsal spine. Metallic evidence is seen along the **right** humeral head and proximal shaft. Impression: Normal cardiac size and probable pulmonary opacity noted. Suggest comparison with previous studies for further evaluation.	RadGraph-F1:-.62	5

Table 5. These candidates received low metric scores despite containing many errors. Errors are highlighted in bold. Scores for these metrics have been multiplied by -1, so higher scores indicate worse performance.

In Table 6, we examined cases where metrics marked reports as being low-quality, though experts found few errors. Here, we found that stylistic differences in how normal findings were described substantially impacted BLEU-2’s performance. In the first example, the predicted report was much longer than the ground-truth report, in part because it was more explicit in ruling out abnormalities ("no additional abnormalities", "no other acute cardiopulmonary conditions"). This discrepancy reflects realistic variation among radiologists. Some radiologists may prefer concise reports which imply that unmentioned abnormalities are absent, while other radiologists write longer reports that clearly rule out many abnormalities. Unfortunately, BLEU-2 performed poorly when facing this stylistic diversity and found little overlap due to the substantial difference in wording, even though the ground truth and candidate conveyed nearly identical content.

RadGraph-F1 failed due to even subtler differences in writing style. The RadGraph pipeline extracts keywords and labels them based on their function in the report, and the RadGraph-F1 score quantifies how often the labeled keywords from the ground-truth and candidate reports match. While the names of abnormalities and anatomical locations stayed largely consistent between the ground truth and AI candidate, other differences in wording led to markedly different sets of keywords (e.g. "within", "acceptable", and "range" vs. "normal"), causing RadGraph-F1 to indicate that these reports do not match well. Additionally, we observed discrepancies in how the RadGraph model labeled keywords across stylistically different reports. For example, "infiltrates" was labeled as being "located at 'pneumonic'" in the ground-truth report, yet it was labeled as "modifying 'pneumonic'" in the AI candidate, an inconsistency that further harmed RadGraph-F1's performance.

Original Ground-Truth Report	**AI Candidate**	**Metric Score**	**Expert Score**
Bilateral **lower** hilar peribronchial thickening in this setting is likely to represent bronchitis.	Report: Findings: The scrutinized chest X-ray exhibits bilateral **mid** hilar peribronchial thickening. The pattern of these findings is frequently seen in instances of bronchitis. No additional abnormalities or significant changes in comparison to preceding examinations are identified. Impression: The noted markers are suggestive of bronchitis. No other acute cardiopulmonary conditions have been detected.	BLEU-2: -.09	1
Impression **No consolidation or definite pneumonic infiltrates**. No pneumothorax or pleural effusion. The cardiomediastinal silhouette is normal. The visualised bones are unremarkable.	Report: Findings: Single frontal view of the chest shows **evidence of consolidation and immediate pneumonic infiltrates**. There are no signs of pneumothorax or pleural effusion. The cardiomediastinal silhouette falls within the acceptable range. Evaluation of the visible bones does not reveal any conspicuous anomalies. Impression: **Chest X-ray exhibits acute cardiopulmonary changes** and still does not present any bone abnormalities.	RadGraph-F1:-.35	1

Table 6. These candidates received high metric scores despite containing almost no errors. Errors are highlighted in bold. Scores for these metrics have been multiplied by -1, so higher scores indicate worse performance.

References

[1] Hong-Yu Zhou et al. *A Generalist Learner for Multifaceted Medical Image Interpretation.* 2024. arXiv: 2405.07988 [cs.CV]. URL: https://arxiv.org/abs/2405.07988.

[2] Stephanie L Hyland et al. "MAIRA-1: A Specialised Large Multimodal Model for Radiology Report Generation". In: *arXiv preprint arXiv:2311.13668* (2023). URL: http://arxiv.org/abs/2311.13668.

[3] Tim Tanida et al. "Interactive and Explainable Region-Guided Radiology Report Generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 2023. URL: http://openaccess.thecvf.com/content/

CVPR2023/html/Tanida_Interactive_and_Explainable_Region-Guided_Radiology_Report_Generation_CVPR_2023_paper.html.

[4] Tao Tu et al. "Towards Generalist Biomedical AI". In: *arXiv preprint arXiv:2307.14334* (2023). URL: http://arxiv.org/abs/2307.14334.

[5] Feiyang Yu et al. "Evaluating Progress in Automatic Chest X-Ray Radiology Report Generation". In: *Patterns* (2023). URL: https://doi.org/10.1016/j.patter.2023.100802.

[6] Alistair E W Johnson et al. "MIMIC-CXR, a de-Identified Publicly Available Database of Chest Radiographs with Free-Text Reports". In: *Scientific Data* 6.1 (2019), p. 317.

[7] Dina Demner-Fushman et al. "Preparing a Collection of Radiology Examinations for Distribution and Retrieval". In: *Journal of the American Medical Informatics Association: JAMIA* 23.2 (2016), pp. 304–310.

[8] Alistair E. W. Johnson et al. *MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs*. 2019. arXiv: 1901.07042 [cs.CV]. URL: https://arxiv.org/abs/1901.07042.

[9] Kishore Papineni et al. "Bleu: A Method for Automatic Evaluation of Machine Translation". In: *Annual Meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

[10] Tianyi Zhang et al. "BERTScore: Evaluating Text Generation with BERT". In: *arXiv preprint arXiv:1904.09675* (2019). URL: http://arxiv.org/abs/1904.09675.

[11] Akshay Smit et al. *CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT*. 2020. arXiv: 2004.09167 [cs.CL]. URL: https://arxiv.org/abs/2004.09167.

[12] Saahil Jain et al. *RadGraph: Extracting Clinical Entities and Relations from Radiology Reports*. 2021. arXiv: 2106.14463 [cs.CL]. URL: https://arxiv.org/abs/2106.14463.

[13] Alyssa Huang et al. *FineRadScore: A Radiology Report Line-by-Line Evaluation Technique Generating Corrections with Severity Scores*. 2024. arXiv: 2405.20613 [cs.CL]. URL: https://arxiv.org/abs/2405.20613.

[14] Feiyang Yu. *Evaluating Progress in Automatic Chest X-Ray Radiology Report Generation*. https://github.com/rajpurkarlab/CXR-Report-Metric. 2023.

[15] Alyssa Huang. *FineRadScore*. https://github.com/rajpurkarlab/FineRadScore. 2024.

[16] Shlomit Goldberg-Stein et al. "ACR RADPEER Committee White Paper with 2016 Updates: Revised Scoring System, New Classifications, Self-Review, and Subspecialized Reports". In: *Journal of the American College of Radiology: JACR* 14.8 (2017), pp. 1080–1086.