

ReEvalMed: Rethinking Medical Report Evaluation by Aligning Metrics with Real-World Clinical Judgment

Ruochen Li^{1*}, Jun Li^{1,2*}, Bailiang Jian^{1,2}, Kun Yuan^{1,4}, Youxiang Zhu³

¹Technical University of Munich ²Munich Center for Machine Learning

³University of Massachusetts Boston ⁴University of Strasbourg

ruochen.li@tum.de, youxiang.zhu001@umb.edu

Abstract

Automatically generated radiology reports often receive high scores from existing evaluation metrics but fail to earn clinicians’ trust. This gap reveals fundamental flaws in how current metrics assess the quality of generated reports. We rethink the design and evaluation of these metrics and propose a clinically grounded Meta-Evaluation framework. We define clinically grounded criteria spanning clinical alignment and key metric capabilities, including discrimination, robustness, and monotonicity. Using a fine-grained dataset of ground truth and rewritten report pairs annotated with error types, clinical significance labels, and explanations, we systematically evaluate existing metrics and reveal their limitations in interpreting clinical semantics, such as failing to distinguish clinically significant errors, over-penalizing harmless variations, and lacking consistency across error severity levels. Our framework offers guidance for building more clinically reliable evaluation methods. Project link is <https://ruochenli99.github.io/ReEvalMed/>

1 Introduction

Radiology reports constitute a fundamental component of clinical workflows, supporting diagnostic reasoning, treatment planning, and follow-up decisions (Hager et al., 2024; Vrdoljak et al., 2025). The continued advancement of vision-language models has enabled the direct generation of medical reports from imaging data (Li et al., 2023; Hartsock and Rasool, 2024; Chen et al., 2024). Although these generated reports often attain high scores on standard natural language processing (NLP) metrics, such as BLEU (Papineni et al., 2002) and ROUGE-L (Lin, 2004), their clinical adoption has been hindered by the lack of thorough clinical validity and reliability, which undermines clinician trust. This reflects a critical gap where the conventional

NLP metric scores fail to align with real-world clinical utility (Zhang et al., 2025). High-scoring reports can still contain factual inaccuracies, logical inconsistencies, or omissions that could compromise patient safety and care (Hartsock and Rasool, 2024; Wang et al., 2025).

This discrepancy motivates a fundamental re-evaluation of how medical report generation is evaluated (Jing et al., 2025; Wang et al., 2024). Rather than relying exclusively on shallow matching or general language similarity, evaluation metrics should also interpret clinical semantics, such as clinically meaningful differences, and accurately reflect the potential impact of errors on patient care (Gu et al., 2025). Moreover, as metrics are also used for training and benchmarking generative models, a metric that fails to capture what clinicians value may falsely incentivize unsafe outputs and thereby undermine confidence in these models. In Section 2, we analyze current LLM-based metrics and highlight key limitations in both their scoring design and evaluation methodology.

To address these issues, we propose a set of clinically grounded evaluation criteria, detailed in Section 3, that define what constitutes a clinically reliable evaluation metric. These criteria encompass two essential dimensions: (1) *Alignment with clinical needs*, including accurate reporting of description, location, distance, and size; and (2) *Core metric capabilities*, including discriminative ability, robustness to clinically insignificant variations, and monotonic sensitivity to increasing error severity.

Building upon these principles, we introduce a Meta-Evaluation framework in Section 4 with 12 evaluation aspects that serve as probes to assess whether a metric effectively captures clinical semantics. We construct a dataset of ground truth and rewritten reports (GT–ME pairs), annotated with clinical significance labels and explanations across diverse error types and evaluation aspects. Our experiments reveal the strengths and limita-

*Equal contribution.

tions of widely used metrics, offering actionable insights to guide the development of more reliable and clinically aligned evaluation methods. Our contributions are as follows:

- **Rethinking LLM-based metrics.** Our detailed analysis of current LLM-based evaluation metrics reveals essential design flaws and limitations in their evaluation methodology.
- **Clinically grounded evaluation criteria.** We propose a set of evaluation criteria co-developed with clinicians and aligned with established standards organizations such as the Fleischner Society (Farjah et al., 2022), ACR Lung-RADS (Christensen et al., 2024), and SCCT (Leipsic et al., 2014). These criteria provide a solid definition of a clinically meaningful metric, guiding both assessment and future metric design.
- **A unified Meta-Evaluation framework.** We introduce the first comprehensive Meta-Evaluation framework for medical report metrics. This framework enables a more rigorous evaluation by assessing a metric’s alignment with clinical needs and its core capabilities.
- **Empirical comparison of existing metrics.** Using our framework, we empirically benchmark widely used metrics. Our analysis reveals their strengths, limitations, and the underlying factors influencing their performance, offering insights for future development.

2 Rethinking Clinical Report Evaluation: Limitations of Current Metrics

In clinical settings, radiology reports are essential tools for clinicians, underpinning diagnostic reasoning and guiding medical decision-making. Recent advances in vision-language models have led to the development of systems capable of generating radiology reports directly from medical images and contextual inputs, with notable examples including MAIRA-2 (Bannur et al., 2024) and LLM-CXR (Lee et al., 2023). Although generated reports often perform well on standard metrics such as BLEU (Papineni et al., 2002) and ROUGE-L (Lin, 2004), these scores are not indicative of clinical reliability or utility in real-world decision-making.

This discrepancy stems from fundamental limitations inherent in existing evaluation metrics: Conventional approaches, such as BLEU and ROUGE-L, which assess lexical overlap; RadGraph F1 (Jain et al., 2021), which measures entity extraction and

alignment; and CheXbert-based classifiers (Smit et al., 2020), which focus on predefined abnormalities, primarily depend on surface-level matching. These methods lack a deeper understanding of the clinical meaning of the report and struggle to handle the complexity of real-world clinical scenarios. Recently, Large language model (LLM)-based metrics have demonstrated improvements over traditional surface-level matching approaches. Notably, metrics such as GREEN (Ostmeier et al., 2024), GEMA Score (Zhang et al., 2025), MRScore (Liu et al., 2024b), and ReFINE (Liu et al., 2024a), leverage the six error categories defined in the ReXVal dataset to construct structured scoring tables. These scoring schemes not only capture clinically relevant error types, but also integrate subjective aspects, including readability, grammaticality, and coherence, which are commonly employed during both model training and final score computation. However, these metrics still exhibit limitations, revealing a clear gap between their scoring criteria and real-world clinical needs.

Coarse-grained scoring schemes. Many LLM-based evaluation metrics rely on scoring tables derived from the six error categories defined in the ReXVal dataset: (a) false prediction, (b) omission, (c) incorrect location or position, (d) incorrect severity, (e) mention of a comparison not present in referenced impression, and (f) omission of comparison describing a change from previous study.

While these categories are clinically reasonable, they remain coarse and incomplete. Several clinically relevant aspects, such as size, distance, uncertainty expression, internal contradictions, and descriptive accuracy, are not adequately captured. See Section 3 for details.

Questionable evaluation methodology. The clinical validity of many recent evaluation metrics is assessed by computing Pearson Correlations between the metric’s scores and clinician annotations from the ReXVal dataset (Yu et al., 2023b). Specifically, six radiologists annotated 50 ground-truth and generated report sets (each set comprising one ground-truth and four candidate reports), labeling the number of errors per report across six predefined categories. Each error was further classified as clinically significant or insignificant. Metrics are then validated by measuring the Pearson correlation between their predicted scores and the aggregated radiologist-annotated error counts, under the assumption that higher correlations indicate stronger alignment with human preferences.

Clinical relevance	Significant						Insignificant					
	1	2	3	4	5	6	1	2	3	4	5	6
BERTScore reports	0.540	0.451	0.380	0.398	0.258	0.308	0.163	0.253	0.321	0.313	-0.044	0.270
BLEU reports	0.553	0.414	0.337	0.242	0.387	0.263	0.200	0.209	0.129	0.280	-0.034	-0.032
Radgraph reports	0.454	0.421	0.424	0.278	0.118	0.412	0.238	0.295	-0.026	-0.031	-0.057	0.216
S-Emb reports	0.321	0.443	0.227	0.297	0.434	0.124	0.199	0.210	0.072	-0.028	-0.002	0.128

Table 1: Average pairwise Pearson correlation of significant and insignificant error counts between six radiologists in the ReXVal dataset. Radiologists were presented with a ground-truth report from MIMIC-CXR (Johnson et al., 2019) and a generated report retrieved by a metric (e.g., BERTScore). They only labeled the number of clinically significant and insignificant errors, without providing any explanation.

However, this assumption is problematic. As shown in Table 1, we computed the average pairwise Pearson correlations among the six annotators for each error category across candidate reports. The resulting inter-annotator correlations are notably not high, indicating a lack of consensus among experts regarding the number and significance of errors. Consequently, a high Pearson correlation with these radiologist annotations alone is insufficient to demonstrate the clinical robustness or practical reliability of a metric.

Insufficient alignment with clinical needs. In clinical practice, radiology reports serve as a foundation for diagnosis, treatment planning, and medical decision-making. Clinicians place a premium on factual accuracy across critical aspects and are particularly sensitive to major logical inconsistencies or clinically significant errors. Meanwhile, they are generally tolerant of minor deviations that do not affect patient care, such as anatomically irrelevant details or stylistic variations.

Although the ReXVal dataset distinguishes clinically significant from insignificant errors, it only provides final error counts per category as judged by six annotators. Crucially, it does not document the rationale for these judgments, i.e., why an error was considered significant or insignificant in a given case. This lack of transparency prevents follow-up metrics from learning or modeling the clinical reasoning process behind these annotations. As a result, scoring tables and metrics derived from ReXVal are limited in their ability to truly align with clinician decision-making criteria. They reflect annotation outcomes but not the underlying clinical logic, making them insufficient for capturing the nuanced judgment clinicians apply when evaluating generated report.

3 What Defines a Good Metric for Clinical Report Evaluation?

Based on consultations with clinicians and established clinical guidelines, we identify two essential requirements for effective evaluation metrics. These reflect the practical priorities clinicians consider when interpreting radiology reports and serve as foundational principles for metric design.

3.1 Alignment with Clinical Needs

Clinicians are the primary users of medical reports and rely on them for downstream decisions such as diagnosis and treatment planning. Accordingly, what matters most is the accuracy and clinical reliability of the content. A single clinically significant error, such as misstating a tumor as 4 cm instead of 8 cm, can lead to entirely different clinical actions and potentially cause serious medical harm. By contrast, clinicians are generally tolerant of clinically insignificant deviations, such as the inclusion of benign incidental findings or the use of alternative but semantically equivalent expressions, as long as these do not interfere with diagnostic reasoning or therapeutic decision-making. Motivated by these practical considerations, we define a set of clinically grounded evaluation criteria that reflect the aspects clinicians prioritize when judging report quality. They are listed in the leftmost column of the first ten rows of Criteria Table 2.

Location refers to the precise anatomical site of a lesion or abnormality (e.g., "right upper lobe"). Accurate localization is critical, as different sites often imply different etiologies and treatment strategies. Incorrect location may lead to diagnostic errors or inappropriate interventions.

Severity describes the extent or seriousness of a finding (e.g., "mild," "severe"). Although partially subjective, severity assessments inform clinical urgency and therapeutic prioritization. Incorrect

Aspect	Significant Error	Insignificant Error
Location	GT: Multiple chronic appearing left-sided rib fractures ME: Multiple chronic appearing right rib fractures	GT: New left retrocardiac opacity ME: New opacity behind the heart on the left side
Severity	GT: Heart is mildly enlarged. ME: Severely enlarged heart	GT: Severe cardiomegaly ME: Moderate-to-severe cardiomegaly
Description	GT: An irregular mass with spiculated margins ME: A round, smooth mass	GT: Bibasilar patchy ill-defined opacities ME: Bibasilar faint and poorly marginated opacities
Negation	GT: No evidence of pneumothorax ME: Pneumothorax is present	GT: No pleural effusion is seen ME: There is no definite pleural effusion
Modality	GT: Refer to prior CT torso for full descriptive details of esophageal abnormalities. ME: Refer to prior abdominal ultrasound for details of esophageal abnormalities	GT: Consider chest CT for further evaluation ME: CT can be considered for further assessment
Size Distance	GT: Irregularly marginated 3-cm mass in the lingula ME: Irregularly marginated 8-cm mass in the lingula	GT: ET tube within 1 cm of the carina ME: ET tube within 0.9 cm of the carina
Comparison	GT: Pulmonary edema has improved ME: Pulmonary edema has worsened	GT: No interval change in pleural effusion ME: Pleural effusion is essentially unchanged
Internal Contradiction	GT: The lungs are clear. ME: The lungs are clear. There is consolidation in the right base	GT: No current evidence of larger pleural effusions ME: No current evidence of larger pleural effusions. Minimal pleural effusions may exist
Uncertainty	GT: Whether this is pneumonia is radiographically indeterminate. ME: Pneumonia exists	GT: A possible infiltrate is suggested ME: An infiltrate is likely present
Terminology	GT: A cavitary lesion, suggesting tuberculosis ME: A hole, suggesting infection	GT: A 3-cm mass ME: A 3-cm lesion
Noise	GT: Irregularly marginated 3-cm mass in the lingula has grown ME: 3-cm lingula margins has been growing irregularly	GT: subtle opacity may represent atelectasis ME: subtle opaciti may represent atelectasi
Stylistic Variation	GT: Bilateral left greater than right pleural effusion ME: Fluid accumulation on both sides of the chest, more on the right	GT: lung... pulmonary edema... pleural effusions... ME: pleural effusions... lung...pulmonary edema...

Table 2: Examples of clinically significant and insignificant errors across evaluation aspects, shown as Ground Truth (GT) and Meta-Evaluation Rewrite (ME) pairs.

severity may mislead clinicians regarding the risk level of a condition.

Description captures the morphological characteristics of a lesion, such as its size, shape, and edge properties (e.g., "well-defined," "irregular"). These features are key to differential diagnosis and malignancy assessment, directly affecting downstream clinical decisions.

Negation explicitly indicates the absence of certain abnormalities (e.g., "no pneumothorax"). Accurate negation narrows the differential diagnosis. Errors or omissions in negation may lead to serious misdiagnoses.

Modality awareness assesses whether the report and its recommendations are appropriate given the imaging modality used (e.g., X-ray, CT). Each

modality has different resolution capabilities and clinical applications; failure to account for modality limitations may result in inappropriate conclusions.

Size and distance includes quantitative measurements of lesions (e.g., "4.5 cm") and positional relationships (e.g., "catheter tip 2 cm above the carina"). Such information is crucial for tumor staging, disease assessment, and device placement. Misstatements may directly affect treatment decisions and prognosis.

Comparison and progression captures temporal changes by comparing current findings to prior imaging (e.g., "increased," "unchanged"). It helps evaluate disease progression and treatment response, and guides follow-up planning. Omission or misstatement may disrupt clinical continuity.

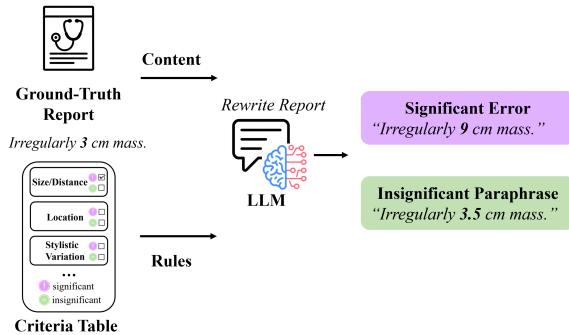


Figure 1: Rewrite report contains a significant error or an insignificant error.

Internal contradiction identifies logical inconsistencies within the same report (e.g., "clear lungs" in one sentence and "left lung infiltrate" in another). It undermines the credibility of the report and can lead to clinical misjudgment or patient harm.

Medical terminology usage assesses whether medical terms are used accurately and clearly. While some stylistic variation is acceptable, reports should avoid too ambiguous or misleading language. Incorrect terminology can impede accurate interpretation and decision-making.

Uncertainty expression involves the use of hedging terms (e.g., "may," "possibly," "suspicious for") to reflect diagnostic uncertainty. Properly expressed uncertainty helps clinicians plan differential diagnoses and additional tests. Omitting or misrepresenting uncertainty may result in overconfident or incorrect decisions.

3.2 Metric Capabilities: Discrimination, Robustness, and Monotonicity

Beyond alignment with clinical needs, a good metric should also possess the following capabilities:

Discriminative ability. Reliable metrics should be capable of distinguishing between clinically acceptable and clinically dangerous reports. Reports containing serious errors that could lead to adverse clinical decisions should receive substantially lower scores.

Robustness ability. Evaluation metrics should exhibit robustness to clinically insignificant variations, meaning they should avoid penalizing reports that differ from the reference only in superficial form, while remaining clinically equivalent in content. Examples are provided in Table 2. Specifically, Grammatical noise refers to grammatical errors, typographical mistakes, or non-standard phrasing that do not affect the underlying clinical

meaning. Stylistic variation refers to differences in expression that do not alter the underlying clinical meaning (e.g., reordering sentences of findings). Such variation may arise from institutional templates, clinician-specific phrasing preferences, or differences in information ordering.

Monotonicity. A well-calibrated metric should exhibit a monotonic response to increasing clinical error severity, with scores consistently decreasing as errors become more serious. This property reflects the metric's ability to not only detect errors, but to differentiate their clinical significance.

4 Meta-Evaluation Framework: Dataset Construction and Metric Assessment

4.1 Dataset and Metric

We evaluate metrics using two primary sources of clinical data: the ReXVal dataset and the MIMIC-CXR dataset. From MIMIC-CXR (Johnson et al., 2019), we randomly sampled 50 radiology reports. Additionally, we selected 20 information-rich reports from ReXVal, as identified by experienced radiologists with extensive clinical expertise.

The evaluation includes the following reference-based metrics from the general NLP community: BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang* et al., 2020), which primarily assess surface-level lexical or semantic similarity. AlignScore (Zha et al., 2023), a factuality-based metric, assesses whether one sentence supports another (entailment).

In addition to these general-purpose metrics, we include several domain-specific metrics tailored for medical report evaluation. RaTEScore (Zhao et al., 2024) is a structured, entity-aware metric specifically designed for medical report evaluation. CheXbert-F1 (Smit et al., 2020) extracts 14 pre-defined thoracic disease labels and classifies their status. The final micro F1 score evaluates only exact label matches. RadGraph-F1 (Jain et al., 2021) converts free-text reports into structured graphs via named entity recognition (NER) and relation extraction (RE). GREEN (Ostmeier et al., 2024) is an LLM-based metric that identifies and explains clinically significant errors in generated reports.

4.2 Discriminative Ability and Robustness

4.2.1 Rewrite Report

We constructed a dataset of paired reports, each consisting of a Ground Truth (GT) report and a

Group	Example
0	"should be repositioned" → "consideration should be given to repositioning"
1	"mild cardiomegaly" → "mild enlargement of the cardiac silhouette"
2	"right lower and left upper lobes" → "left lower and right middle lobes"
3	"No current evidence of pleural effusions, pulmonary edema, or pneumonia" → "consolidation in right middle lobe; moderate right-sided pleural effusion; small left apical pneumothorax"
4	Report first states "The previous right internal jugular vein catheter was removed", then later fabricates "malpositioned right internal jugular catheter"

Table 3: Examples of rewritten errors grouped by severity, where a higher group ID indicates greater severity.

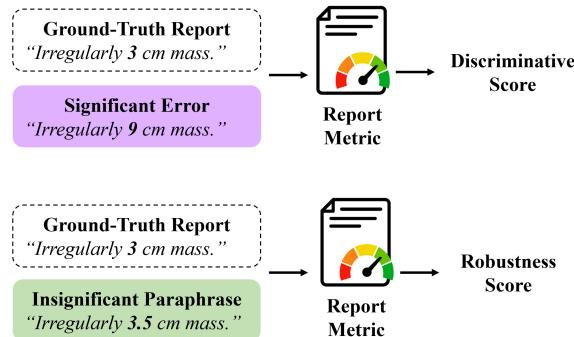


Figure 2: Discriminative Score and Robustness Score.

corresponding Meta-Evaluation Rewrite (ME), to test whether metrics can distinguish clinically significant from insignificant errors and remain robust to clinically irrelevant variations.

We categorized the differences between GT and ME into three primary error types. **Omission:** A clinically relevant fact present in the GT is missing in the ME. **Fabrication:** The ME introduces a clinically relevant fact that is not present in the GT. **Inaccuracy:** The same clinical fact is described inconsistently between GT and ME.

For each evaluation aspect listed in the Criteria Table 2, we prompted DeepSeek-R1 (Guo et al., 2025) to rewrite the GT report by modifying only one targeted aspect. The rewrites yield paired samples that contain either clinically significant or clinically insignificant errors. For example, as shown in Figure 1, under the Size/Distance aspect, an example clinically significant inaccuracy could be:

GT: "Irregularly marginated 3 cm mass."

ME: "Irregularly marginated 9 cm mass."

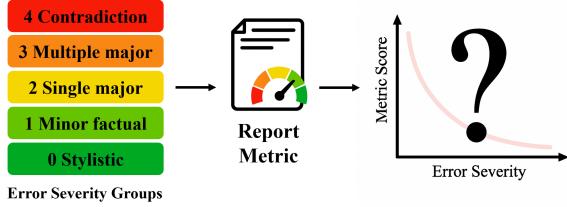
For those clinically high-impact aspects: Location, Severity, Description, and Comparison / Progression, we generated 10 significant and 10 insignificant error pairs for each of the three error types (omission, fabrication, inaccuracy). For the remaining aspects, we randomly sampled across er-

ror types while ensuring that each aspect contained 10 significant and 10 insignificant pairs. In total, the dataset comprises 400 expert-validated GT-ME report pairs.

To preserve semantic fidelity and ensure isolation of the targeted error, we retained substantial contextual content in both GT and ME reports for significant error pairs, ensuring that the introduced change is the primary deviation. In contrast, for insignificant error pairs, we intentionally reduced the amount of surrounding context and kept only the modified part when appropriate, to prevent metrics from giving high scores just because the reports look similar overall. All generated report pairs were reviewed and validated by experienced clinicians to ensure consistency with real-world clinical understanding and relevance, with explanations provided alongside each pair.

4.2.2 Discriminative and Robustness Score

We applied multiple existing evaluation metrics to all constructed report pairs. For each metric, we separately computed the average scores for clinically significant and clinically insignificant error pairs, as shown in Figure 2. The Discriminative Score is defined as the average score assigned to clinically significant error pairs. Lower scores indicate that the metric effectively penalizes critical errors, reflecting strong discriminative ability. The Robustness Score is defined as the average score assigned to clinically insignificant error pairs. Higher scores suggest that the metric tolerates minor, clinically irrelevant variations, indicating desirable robustness. Together, these two scores offer a comprehensive assessment of each metric’s capacity to distinguish between clinically meaningful and negligible errors. We further report confidence intervals to present a more transparent view of each metric’s consistency and variability.



4.3 Test Monotonicity

To assess whether evaluation metrics exhibit monotonic sensitivity to increasing error severity, we conducted a controlled test using a subset of the ReXVal dataset. We selected four GT reports and, for each, constructed corresponding ME reports with varying levels of error severity. These were organized into five groups (Group 0–4), each containing the same four GT–ME pairs, with severity increasing incrementally from minor stylistic variations to severe logical contradictions. Examples for each group are provided in Table 3, and the overall grouping design is illustrated in Figure 3.

Group 0: Stylistic Variation (Clinically Insignificant). Contains purely stylistic or linguistic changes that do not alter clinical meaning. Examples include hedging expressions, synonymous reformulations, or reordering descriptive sentences.

Group 1: Minor Factual Errors (Clinically Insignificant). Involves minor factual inaccuracies that are clinically negligible and do not affect diagnostic interpretation or treatment decisions.

Group 2: Single Error (Clinically Significant). Introduces a single clinically significant error that may plausibly affect downstream clinical decisions.

Group 3: Multiple Errors (Clinically Significant). Introduces multiple (typically three) clinically significant errors, collectively increasing the risk of diagnostic misguidance.

Group 4: Logical Contradiction (Severe Errors). Introduces internal inconsistencies or logical contradictions, such as describing the presence of a structure previously stated to be absent. Such contradictions severely undermine clinician trust in the report and are considered critical failures.

For each group, we computed the average metric score across its four GT–ME pairs for each evaluation metric under consideration. We then analyzed the score trajectory across severity levels to assess whether the metric exhibits a monotonic decreasing trend. A consistent decline in score as

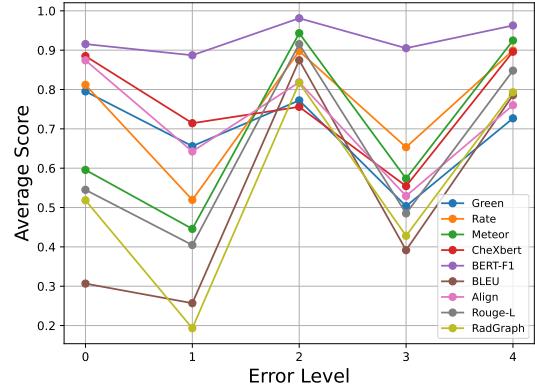


Figure 4: Metric scores vs. clinical error severity. Ideally, metric scores should decrease monotonically with increasing error severity.

error severity increases indicates that the metric is well-calibrated. It appropriately reflects the clinical impact of errors through its scoring behavior.

5 Results and Analysis

5.1 Discriminative Ability and Robustness

We assess each metric’s ability to differentiate clinically significant and insignificant errors using the Discriminative Score and Robustness Score, as summarized in Tables 4 and 5.

General NLP metrics lack clinical knowledge and aspect sensitivity. BLEU, METEOR, and ROUGE-L all exhibit poor robustness, frequently penalizing stylistic or structural variations that are clinically harmless. This stems from their reference-based design, which focuses on surface-level lexical or token overlap. BLEU shows relatively stronger discriminative ability, particularly for omission-related errors, but its extreme sensitivity to surface overlap leads to unjustified penalties. METEOR and ROUGE-L perform poorly on both scores, indicating limited clinical applicability. Overall, these metrics lack domain-specific understanding and fail to distinguish clinically significant from insignificant errors.

BERT-F1 achieves a high Robustness Score, indicating strong tolerance to clinically insignificant variations. However, it also yields a relatively high Discriminative Score, suggesting that it fails to adequately penalize clinically significant errors. This implies that while BERT-F1 is resistant to superficial changes, it lacks the sensitivity needed to distinguish harmful clinical deviations.

AlignScore, a factuality-based metric, has the lowest overall Discriminative Score as well as a low

Metric ↓	Comparison/Progression			Cont.	Description			Location			Mod.	Neg.	Noise	Severity			S/D	S/V	Unc.	Term.	Overall	CI
	E	F	O		S	E	F	O	E	F	O			E	F	O						
Align	61.03	71.54	88.81	65.02	59.92	73.91	93.43	50.94	52.66	95.01	55.40	79.22	74.96	71.17	80.88	96.58	77.60	49.94	72.52	59.38	71.50	[65.06, 77.93]
BERT-F1	98.40	97.54	96.55	96.03	97.56	98.29	97.35	99.46	98.41	97.27	95.06	97.72	94.89	99.19	98.39	97.97	99.12	93.47	97.73	95.90	97.31	[96.63, 98.00]
BLEU	86.23	81.10	67.78	76.01	83.44	87.99	79.28	87.95	81.19	71.34	68.11	82.99	60.81	90.02	85.53	80.26	91.60	58.80	75.53	73.63	78.48	[74.37, 82.59]
CheXbert	96.67	94.10	96.23	85.62	80.71	93.87	89.96	99.89	93.94	93.44	80.89	69.22	85.82	95.44	95.27	99.27	94.38	79.11	60.27	69.42	87.68	[82.79, 92.56]
GREEN	67.89	77.90	78.44	68.08	69.45	83.14	77.88	66.55	74.57	74.90	69.17	80.55	73.98	73.98	82.73	74.67	83.87	51.15	83.81	73.21	74.30	[70.90, 77.69]
METEOR	94.17	93.47	78.95	86.24	90.16	97.92	86.36	96.32	97.88	84.78	79.86	90.59	81.28	95.33	98.15	91.03	97.22	73.11	89.34	85.82	89.40	[86.23, 92.57]
RadGraph	85.66	86.06	85.63	85.57	77.03	86.07	79.79	90.47	84.51	79.61	72.60	89.24	68.72	87.00	88.40	87.96	92.01	58.22	86.33	63.08	81.35	[77.28, 85.42]
RaTE	89.49	89.42	84.74	87.66	89.96	95.98	90.48	93.11	95.68	84.69	88.15	91.69	79.76	90.38	95.29	91.02	96.78	78.43	92.67	78.64	89.20	[86.77, 91.63]
Rouge-L	91.32	90.14	86.10	83.31	89.75	93.94	91.10	94.23	91.46	89.07	77.17	88.67	75.44	94.51	93.01	93.33	95.84	69.02	88.18	82.39	87.90	[84.77, 91.02]

Table 4: Discriminative Score (clinically significant errors). Lower values indicate better discrimination of errors. **Cont.:** Contradiction. **Mod.:** Modality. **Neg.:** Negation. **S/D:** Size/distance. **S/V:** Stylistic Variation. **Unc.:** Uncertainty. **Term.:** Terminology. **E:** Inaccuracy error. **F:** Fabrication error. **O:** Omission error. **S:** Inaccuracy, fabrication, and omission errors are randomly and evenly distributed.

Metric ↑	Comparison/Progression			Cont.	Description			Location			Mod.	Neg.	Noise	Severity			S/D	S/V	Unc.	Term.	Overall	CI
	E	F	O		S	E	F	O	E	F	O			E	F	O						
Align	52.22	38.03	76.92	51.59	36.60	27.00	94.29	88.71	51.73	83.13	63.45	61.08	95.70	51.77	57.55	83.52	25.82	74.77	80.52	66.73	63.06	[53.67, 72.45]
BERT-F1	91.25	93.83	93.57	95.18	89.08	95.18	95.61	94.53	95.14	97.35	88.66	88.91	92.99	90.02	93.36	97.55	93.81	89.32	94.19	90.21	92.99	[91.75, 94.22]
BLEU	29.76	52.95	37.36	56.30	23.91	57.79	58.05	57.20	58.46	61.89	21.99	24.90	40.55	0.00	45.28	56.27	49.11	29.81	45.45	30.74	41.89	[34.61, 49.16]
CheXbert	94.93	92.36	97.48	84.57	83.06	91.40	90.02	95.35	88.83	99.25	86.55	83.83	93.61	94.61	88.64	98.93	94.95	80.19	89.68	80.03	90.41	[87.82, 93.00]
GREEN	81.67	80.00	76.67	67.17	75.00	63.33	75.00	78.24	66.33	73.33	91.00	82.50	84.17	60.00	56.67	70.00	90.00	65.05	77.50	81.67	74.76	[70.60, 78.93]
METEOR	60.02	88.01	71.33	84.39	53.87	86.68	73.48	79.94	92.30	82.22	44.03	49.20	70.98	41.61	83.43	81.62	81.99	50.28	77.57	51.13	70.20	[63.02, 77.39]
RadGraph	57.11	72.47	67.83	73.18	25.94	63.94	66.82	61.34	67.27	78.55	44.34	39.08	52.50	44.07	56.08	74.91	64.33	43.80	83.46	34.16	58.56	[51.66, 65.46]
RaTE	64.30	84.60	71.32	83.57	63.11	85.55	89.81	78.03	82.62	84.02	63.25	64.57	75.49	75.15	79.71	80.50	78.13	72.26	88.58	63.80	76.42	[72.52, 80.31]
Rouge-L	54.47	73.03	78.89	74.44	46.94	77.59	82.83	71.92	76.38	86.66	42.55	46.60	64.72	61.61	69.22	87.32	71.47	35.54	67.98	48.91	65.95	[59.26, 72.65]

Table 5: Robustness Score (clinically insignificant errors). Higher values indicate greater robustness to clinically irrelevant variations. Metrics with bold indicate top performance in that column.

Robustness Score, indicating that it cannot reliably distinguish clinically significant from insignificant errors. This stems from its lack of medical knowledge, which limits its capacity to capture clinically relevant relationships, particularly in the presence of subtle but significant semantic shifts.

Non-LLM medical-specific metrics often consider too few aspects or rely on rigid matching. RaTE, although structured and medical-entity-based, lacks sufficient medical grounding and fails to capture clinical priorities or severity distinctions in a nuanced way. While it demonstrates reasonably high robustness, its Discriminative Score is unexpectedly higher than its Robustness Score, reflecting an overreliance on surface-level entity matching and an inability to capture deeper clinical semantics or error impact. CheXbert-F1 gave consistently high scores to both types of errors, failing to reflect severity. This is because its final F1 score evaluates only exact matches against 14 predefined thoracic disease labels, ignoring semantic variations, contextual cues, and clinically equivalent paraphrases. RadGraph-F1, by contrast, shows a high Discriminative Score but a low Robustness Score. While theoretically powerful, it is overly sensitive to entity boundaries, relation formats, and exact phrasing. Even semantically equivalent rewrites, such as reordering or lexical variation, may reduce the score due to graph-matching failures, which highlights its poor robustness to

stylistic variation and clinical equivalence.

LLM-based medical-specific metrics often rely heavily on ReXVal error categories, with a scoring framework that covers only a limited set of error types and lacks comprehensive coverage of clinically relevant semantic dimensions. This limitation extends to several newer LLM-based metrics that inherit similarly simplified scoring tables. GREEN achieves a better Discriminative Score than most metrics, performing particularly well on omissions of description and severity, indicating its ability to penalize clinically significant errors. However, it still suffers from a notably low Robustness Score, suggesting a tendency to over-penalize minor, clinically irrelevant differences and thus limiting its practical reliability.

Overall, the results show that while some metrics perform well in either discrimination or robustness, none excel at both. Moreover, most existing metrics do not capture clinical semantics, which is essential to evaluate reports from the perspective of a clinician. These findings highlight the need for clinically grounded, error-aware evaluation.

5.2 Monotonicity

As illustrated in Figure 4, although different metrics exhibit varying absolute scores, their overall trends across severity groups are remarkably consistent. From Group 0 to Group 1, all metrics show a decreasing trend, indicating that each metric can

reliably distinguish stylistic variations (clinically negligible) from minor factual errors (still clinically insignificant but of slightly higher concern). Similarly, from Group 2 to Group 3, we observe a consistent decline across all metrics, suggesting that they are sensitive to the increased number of clinically significant errors, even if this sensitivity may stem more from the extent of textual changes than from a true understanding of clinical severity. These patterns indicate that existing metrics possess some limited ability to capture differences in error severity, particularly when differences are accompanied by large textual modifications.

However, two key transitions reveal important weaknesses. From Group 1 to Group 2, all metrics unexpectedly show an increase in scores, implying a failure to distinguish clinically significant single errors from clinically insignificant factual deviations. This reversal can be attributed to our data set design: insignificant errors in Group 1 often involve larger surface-level changes (which do not affect clinical interpretation), while significant errors in Group 2 are more localized, with most of the surrounding context preserved, potentially misleading surface-based metrics. From Group 3 to Group 4, metrics again show an increase in scores, reflecting their difficulty in detecting logical contradictions. This is likely because logical contradictions in Group 4 are introduced via small, localized insertions (e.g., contradicting earlier statements with a single sentence), while Group 3 reports contain multiple significant edits across the text. Metrics that rely heavily on overall textual similarity may struggle to penalize these subtle but clinically critical inconsistencies.

In summary, although existing metrics are sensitive to gross differences in error severity, they struggle with fine-grained distinctions, particularly in separating clinically significant from insignificant errors and in detecting logical contradictions.

6 Conclusion

In this paper, we rethink the design and evaluation of the existing metrics for medical report generation, arguing that effective metrics evaluation should not rely solely on coarse radiologists' counting-based annotations.

To address this, we introduce a clinically grounded Meta-Evaluation framework and show that many existing metrics fail to capture clinical semantics, a critical requirement for evaluations

that align with clinical judgement. To ground our Meta-Evaluation framework, we define clinical semantics as the fine-grained, decision-informing criteria within medical reports. We create an expert-annotated dataset of GT-ME report pairs to simulate a broad spectrum of clinically relevant scenarios and real-world diagnostic needs. Our multi-dimensional framework enables a rigorous evaluation of clinical alignment and core metric capabilities, including a metric's discriminative ability, its robustness to clinically insignificant variations, and its monotonic sensitivity to increasing error severity. Our findings uncover a critical misalignment between existing metrics and clinical needs.

General NLP metrics lack both clinical knowledge and aspect sensitivity. Likewise, existing medical-specific metrics often suffer from insufficient aspects and rigid matching. Although these metrics reflect thoughtful and valuable design efforts, our in-depth Meta-Evaluation concludes that the limitations of these metrics stem from the omission of clinically critical aspects during their formulation. We view these metrics as a strong foundation for further improvement and encourage future research to incorporate clinical aspects more explicitly. Our Meta-Evaluation framework can function as a diagnostic lens to pinpoint where current metrics fall short and serve as an important stepping stone toward developing more clinically aligned evaluation tools.

Limitations

Scalability of dataset construction. The current pipeline for building the Meta-Evaluation dataset requires manual verification to ensure clinical accuracy, limiting scalability compared to other medical domains (Sun et al., 2016, 2024). A more automated and scalable framework is needed.

Limited evaluation of metric interpretability. Interpretability and error localization are essential for clinical users to understand and trust evaluation metrics. Although our criteria table supports fine-grained error attribution, we do not directly assess the metric's interpretability in this study. Future work will explore this dimension more explicitly.

Limited Evidence for Modality Transferability. Our criteria and Meta-Evaluation framework are designed to be modality-agnostic and theoretically generalizable to other domains (e.g., CT, MRI), but we have not validated them beyond CXR. Future work will explore broader imaging settings.

References

Satanjeev Banerjee and Alon Lavie. 2005. **METEOR: An automatic metric for MT evaluation with improved correlation with human judgments**. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Srinidhi Bannur, Kamel Bouzid, Daniel C. Castro, Rayan Krishnan, Nisha Pillai, Roger Zou, Adam P. Harrison, Le Lu, and Pranav Rajpurkar. 2024. **Maira-2: Grounded radiology report generation**. *arXiv preprint arXiv:2406.04449*. Version 1.

Zhe Chen, Midhun Varma, Jing Xu, and 1 others. 2024. A vision-language foundation model to enhance efficiency of chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*.

James Christensen, Aaron E. Prosper, Charles C. Wu, and 1 others. 2024. ACR lung-RADS v2022: Assessment categories and management recommendations. *Journal of the American College of Radiology*, 21(3):473–488.

Farhood Farjah, Sarah E. Monsell, Rebecca Smith-Bindman, and 1 others. 2022. Fleischner society guideline recommendations for incidentally detected pulmonary nodules and the probability of lung cancer. *Journal of the American College of Radiology*, 19(11):1226–1235.

Di Gu, Yuying Gao, Yuxuan Zhou, Yuxuan Song, Haoran Mo, Shikang Zhang, Yuntao Zhuang, Wei Xiong, Yawen Zeng, Xiaodan Liang, and Wei Shen. 2025. **Radalign: Advancing radiology report generation with vision-language concept alignment**. *Computing Research Repository*, arXiv:2501.07525. Version 1.

Dong Guo, Daming Yang, Haotian Zhang, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Patrick Hager, Felix Jungmann, Ryan Holland, Ansh Kalra, Benjamin J. Lengerich, Simon Kohl, Michael Roberts, Olivia L. Cardenas, Liwei Jiang, Owkin Inc., Trevor Back, Jure Leskovec, Mate Leng, and Pranav Rajpurkar. 2024. **Evaluation and mitigation of the limitations of large language models in clinical decision-making**. *Nature Medicine*, 30(9):2613–2622.

Isaac Hartsock and Ghalib Rasool. 2024. **Vision-language models for medical report generation and visual question answering: A review**. *Frontiers in Artificial Intelligence*, 7:1430984.

A. Heiman, X. Zhang, E. Chen, and 1 others. 2025. Factchexcker: Mitigating measurement hallucinations in chest x-ray report generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 30787–30796.

Allen Huang, Oindrila Banerjee, Kai Wu, and 1 others. 2024. Fineradscore: A radiology report line-by-line evaluation technique generating corrections with severity scores. *arXiv preprint arXiv:2405.20613*.

Interventional Medicine Center Association, CHA. 2021. Expert consensus for diagnosis and treatment of esophageal cancer based on artificial intelligence platform. *Chinese Journal of Interventional Radiology (Electronic Edition)*, 9(3):235–246.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, and 1 others. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.

Peiye Jing, Kwonjoon Lee, Zihan Zhang, Kexin Pei, Eric Zelikman, Yixin Liu, Jure Leskovec, and Pranav Rajpurkar. 2025. **Reason like a radiologist: Chain-of-thought and reinforcement learning for verifiable report generation**. *Computing Research Repository*, arXiv:2504.18453. Version 1.

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

Seungjun Lee, Won J. Kim, Jaemin Chang, Jaehoon Oh, Sungho Shin, Sungjun Kwon, Hyoungseok Kim, Hwanjun Song, Kyungwoo Song, Sung Ju Hwang, Jung-Woo Ha, and Hyun Oh Song. 2023. **Llm-cxr: Instruction-finetuned llm for cxr image understanding and generation**. *Computing Research Repository*, arXiv:2305.11490. Version 1.

Jonathon Leipsic, Suhny Abbara, Stephan Achenbach, and 1 others. 2014. SCCT guidelines for the interpretation and reporting of coronary CT angiography: A report of the society of cardiovascular computed tomography guidelines committee. *Journal of Cardiovascular Computed Tomography*, 8(5):342–358.

Chunyuan Li, Ching-Yao Wong, Shiyang Zhang, Yufei Wang, Talia Ringer, Jiaming Song, Jindong Gu, Hsiu-Kai Tsou, Jiali Guo, Xiang Gao, Yujia Xie, Ekin Dogus Cubuk, Mohammad Shoeybi, Anima Anandkumar, Yizhou Sun, Anshul Kundaje, Kai-Wei Chang, Yuxin Chen, Yizhou Zhang, and Yue Zhao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

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

Yunyi Liu, Yingshu Li, Zhanyu Wang, Xinyu Liang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2024a. Er2score: Llm-based explainable and customizable metric for assessing radiology reports with reward-control loss. *arXiv preprint arXiv:2411.17301*.

Yunyi Liu, Zhanyu Wang, Yingshu Li, Xinyu Liang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2024b. Mrscore: Evaluating radiology report generation with llm-based reward system. *arXiv preprint arXiv:2404.17778*.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and 1 others. 2024. Green: Generative radiology report evaluation and error notation. *arXiv preprint arXiv:2405.03595*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Nida Qadir, Shyoko Sahetya, Laveena Munshi, and 1 others. 2024. An update on management of adult patients with acute respiratory distress syndrome: An official american thoracic society clinical practice guideline. *American Journal of Respiratory and Critical Care Medicine*, 209(1):24–36.

Rafael Rami-Porta, Kohei K. Nishimura, Daniel J. Giroux, and 1 others. 2024. The international association for the study of lung cancer lung cancer staging project: Proposals for revision of the tnm stage groups in the forthcoming (ninth) edition of the tnm classification for lung cancer. *Journal of Thoracic Oncology*, 19(7):1007–1027.

M. E. Roberts, N. M. Rahman, N. A. Maskell, and 1 others. 2023. British thoracic society guideline for pleural disease. *Thorax*, 78(Suppl 3):s1–s42.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Parrekk, Andrew Y Ng, and Matthew P Lungren. 2020. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*.

Xiaohua Sun, Jing Yang, Ming Sun, and 1 others. 2016. A benchmark for automatic visual classification of clinical skin disease images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 206–222, Cham. Springer International Publishing.

Xiaohua Sun, Yanan Yao, Shuo Wang, and 1 others. 2024. ALICE benchmarks: Connecting real world re-identification with the synthetic. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

A. Vogel, S. L. Chan, L. A. Dawson, and 1 others. 2025. Hepatocellular carcinoma: ESMO clinical practice guideline for diagnosis, treatment and follow-up. *Annals of Oncology*, 36(5):491–506.

Josip Vrdoljak, Zvonimir Boban, Marina Vilović, Josip Bašić, Anamarija Tokić, Stipe Ćurko, Ana Marija Slišković, Maja Baretić, Marijo Parćina, Tea Vukušić Rukavina, and Luka Vučemilo. 2025. A review of large language models in medical education, clinical decision support, and healthcare administration. *Healthcare*, 13(6):603.

Chenxi Wang, Weizhe Zhou, Soumya Ghosh, Shreyas Sudhakar, Karan Aggarwal, Mengzhou Li, Jay Urbain, Matthew B. A. McDermott, and Ping Zhang. 2024. Semantic consistency-based uncertainty quantification for factuality in radiology report generation. *Computing Research Repository*, arXiv:2412.04606. Version 1.

Xueyan Wang, Guilherme Figueiredo, Ruoxi Li, Peng Zhang, Lei Zhang, Heye Zhang, Georges B. Koenig, Xin Yi, and Ziyue Xu. 2025. A survey of deep-learning-based radiology report generation using multimodal inputs. *Medical Image Analysis*, page 103627.

David E. Wood. 2015. National comprehensive cancer network (nccn) clinical practice guidelines for lung cancer screening. *Thoracic Surgery Clinics*, 25(2):185–197.

Fei Yu, Makoto Endo, Raghav Krishnan, and 1 others. 2023a. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9).

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, EKU Fonseca, Henrique Lee, Zahra Shakeri, Andrew Ng, and 1 others. 2023b. Radiology report expert evaluation (rexval) dataset.

Juan M. Zambrano Chaves, Sheng-Chieh Huang, Yuncheng Xu, and 1 others. 2025. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(1):3108.

Yiqing Zha, Yilun Yang, Ruixue Li, and 1 others. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zhenxuan Zhang, Kinhei Lee, Weihang Deng, Huichi Zhou, Zihao Jin, Jiahao Huang, Zhifan Gao, Dominic C Marshall, Yingying Fang, and Guang Yang. 2025. Gema-score: Granular explainable multi-agent score for radiology report evaluation. *arXiv preprint arXiv:2503.05347*.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Ratescore: A metric for radiology report generation. *arXiv preprint arXiv:2406.16845*.

A Appendix: Aspect Explanation

These aspect-level scores enable clinicians and researchers to assess whether existing metrics truly capture clinical semantics. Our dataset contains the explanation for each GT-ME pair.

Location errors can directly affect treatment decisions, especially in surgery or local therapy. Misreporting lesion sites may lead to incorrect staging, inappropriate treatment, or unnecessary procedures. Significant(S): Confusing "medial right apex" with "medial left and right apex" in a lung cancer case changes the interpretation from a unilateral lesion to bilateral involvement. This could result in a staging upshift (e.g., from M0 to M1a (Rami-Porta et al., 2024)), causing clinicians to abandon curative options (e.g., surgery or SBRT) and turn to systemic therapies or additional diagnostics. Insignificant(I): Omitting "right" in "blunting of the costophrenic angle on the right" is often tolerable. Clinicians can easily identify the affected side through imaging comparison, and this does not typically alter clinical decisions.

Disease severity often guides treatment choices. Misreporting severity can lead to suboptimal decisions. For instance, choosing surgery when medication suffices, or delaying surgery when urgently needed. S: The original report described focal consolidation with stable cardiomegaly. The modified version added "diffuse severe opacities bilaterally," falsely suggesting acute deterioration (e.g., ARDS or severe pneumonia). This may prompt unnecessary escalation of treatment (e.g., broad-spectrum antibiotics, aggressive ventilator settings); trigger additional diagnostics (e.g., bronchoscopy, BAL) (Qadir et al., 2024) I: Adding "mild elevation of the right hemidiaphragm" to a report on small bilateral effusions usually has no clinical consequence. This finding is common and non-specific, often due to benign factors like body position, mild diaphragm laxity, or small effusions. It rarely requires intervention and does not alter management.

Lesion morphology (e.g., margin, internal texture, and structural characteristics) is essential for tumor diagnosis and staging. Errors in description can lead to misclassification of malignancy and inappropriate treatment. S: Changing "irregularly marginated" and "has grown" to "smoothly marginated" and "no change" understates malignant risk (Wood, 2015). This may mislead clinicians into assuming the lesion is benign, resulting

in: delayed diagnostic workup (e.g., CT, functional evaluation); missed early treatment opportunities; potential progression of undiagnosed cancer. I: Replacing "irregularly marginated" with "poorly defined mass with uneven edges" conveys a similar clinical implication—both suggest malignancy and warrant further evaluation. Thus, such phrasing differences do not affect clinical interpretation or decision-making.

Negative findings are essential for ruling out differential diagnoses and avoiding unnecessary interventions. Misreporting such information can lead to overtreatment and increased medical burden. S: Changing "no pneumothoraces" to "right pneumothorax" introduces a critical false positive. In patients with heart failure, this may trigger emergency responses such as chest tube placement, delay proper heart failure management, and risk unnecessary invasive procedures (Roberts et al., 2023). I: Omitting mention of food content in the esophagus has little clinical impact. Such findings are common and do not typically influence diagnostic or therapeutic decisions.

Different diseases require appropriate **imaging modalities**. Incorrect statements about what can be seen on a given modality, or misleading follow-up recommendations, may result in wasted resources and misinformed clinical decisions. S: Claiming that "esophageal mural thickening is clearly delineated on X-ray" is incorrect—such findings require CT (Interventional Medicine Center Association, CHA, 2021). This may cause confusion, unnecessary concern, and inappropriate reliance on suboptimal diagnostic imaging. I: Referring to "CT" without specifying "chest CT" is generally acceptable. Physicians can interpret the intent correctly based on prior reports and standard diagnostic pathways in clinical contexts.

Quantitative descriptors (Heiman et al., 2025), such as lesion size or device position, are critical for diagnosis, staging, and treatment planning. Misstatements may lead to overtreatment, under-treatment, or delays in care. S: Describing a 3-cm mass as a "very large mass" may falsely suggest a tumor ≥ 5 cm, resulting in higher T-stage classification under lung cancer TNM criteria (Rami-Porta et al., 2024). This can lead clinicians to: abandon curative surgery due to perceived inoperability; overestimate disease aggressiveness; order unnecessary invasive procedures or specialist referrals. I: Adding a normal cardiothoracic ratio (e.g., 0.49) to a report about rib fractures has minimal clinical

relevance. It does not impact fracture management or related clinical decisions.

Temporal comparisons are key to evaluating disease progression and treatment response. In conditions like hepatocellular carcinoma (HCC), small changes in lesion size over time can redefine response categories (e.g., partial response vs. progression) (Vogel et al., 2025), directly influencing clinical decisions. S: Changing the interpretation from "pulmonary edema improved" to "worsened" misleads clinicians into believing the patient is deteriorating. This may lead to: escalation of medication (e.g., higher diuretic doses); unnecessary ICU monitoring or imaging; increased healthcare costs and patient anxiety. I: Replacing "less severe" with "slightly improved" conveys the same clinical direction (i.e., improvement). It does not alter risk assessment or treatment planning.

Major internal contradictions can undermine the credibility of a radiology report and may render the findings clinically unreliable. Such errors are often flagged in quality control and may directly endanger patient safety. S: Stating "lungs are clear" immediately after describing large pleural effusion, atelectasis, and possible consolidation introduces a severe inconsistency. This may cause clinicians to question the report's validity and hesitate to act on its findings. I: Saying "no clear current evidence of chronic pulmonary changes" after suggesting chronic changes introduces a mild inconsistency, but not a true contradiction. Clinicians can still interpret the statement within clinical contexts and proceed appropriately.

Expressions of uncertainty in radiology reports guide clinicians toward cautious decision-making, including further testing or observation. Replacing uncertain language with unjustified certainty can mislead treatment and compromise patient safety. S: Changing "may represent atelectasis or pneumonia" to a definitive "represents atelectasis," and describing possible free air as confirmed, may lead clinicians to: dismiss infection unnecessarily (e.g., withholding antibiotics); initiate premature surgical interventions based on presumed pneumoperitoneum. This compromises diagnostic objectivity and risks inappropriate treatment. I: Phrases like "may be" vs. "appears to be" both reflect clinical uncertainty and do not meaningfully alter diagnostic interpretation or next steps.

Accurate use of **medical terminology** is essential for precise communication and diagnostic clarity. Substituting standard terms with vague or non-

professional expressions can obscure clinical meaning and delay appropriate care. S: Replacing "pleural effusion" with "pleural empyema" constitutes a critical error. This falsely suggests a localized infection requiring urgent drainage (e.g., chest tube placement), potentially leading to unnecessary invasive procedures, prolonged hospital stays, and inappropriate antibiotic use, while the actual cause (e.g., heart failure, malignancy) is overlooked. Similarly, substituting "atelectasis" with a phrase like "possible mass" incorrectly raises suspicion for malignancy, potentially triggering unnecessary biopsies, CT scans, and significant patient anxiety. I: Using lay terms like "breathing tube" for "endotracheal tube" or "feeding tube" for "enteric tube" does not affect clinical interpretation when tube position and anatomy are described clearly. These variations preserve the report's medical accuracy.

Minor linguistic errors, such as typos or grammatical mistakes, are common in clinical reports and typically do not affect interpretation. However, when noise alters the meaning of a sentence, it can mislead clinical decisions. S: Changing "most likely due to low lung volumes and positioning" to "unlikely the cause" reverses the interpretation of a key finding. This may prompt unnecessary concern over a mediastinal abnormality, leading to further testing or referrals. I: Typos like "lingulas", "growed", "studys", "atelectasi", "adenocarcinomia" are linguistically incorrect but do not affect the core diagnostic message. Clinicians can readily infer the intended meaning without clinical misunderstanding.

Stylistic differences, when semantically equivalent, usually do not impact clinical decisions. However, poor phrasing, ambiguous emphasis, or incorrect wording can reduce report clarity or lead to clinical misjudgment. S: Changing a report that indicates clinical improvement (e.g., improving edema, resolving effusions, low lung volumes) to one that suggests deterioration (e.g., no edema, persistent effusions, normal lung inflation) reverses the overall interpretation. These conflicting signals may mislead treatment evaluation and disrupt appropriate care. I: Changes in sentence order or phrasing using a different template do not alter diagnostic content. Both versions communicate the same findings and support the same clinical interpretation.

B Appendix: Additional Clarification

CheXbert-F1 classifies each of 14 thoracic conditions into four categories: Positive, Negative, Uncertain, and Blank. The final score is computed as the micro-averaged F1 across all condition–label pairs. While this offers more granularity than binary classification, the metric still relies on discrete label matching and is limited in several ways: (1) it cannot detect clinically important nuances such as changes in severity (e.g., "small effusion" vs. "large effusion"), (2) it ignores how multiple findings interact or contradict each other (e.g., stating "no pleural effusion" in one sentence and "moderate right pleural effusion" in another), and (3) it is insensitive to paraphrasing, hedging, or indirect language that may shift the clinical implication without altering the label category. Thus, while CheXbert is valuable for structured disease extraction, it lacks the semantic depth required to evaluate subtle but clinically significant variations in report generation. This concern has also been echoed in recent work—the GEMA score (Zhang et al., 2025).

We did not include certain metrics in our Meta-Evaluation framework for the following reasons. CheXprompt (Zambrano Chaves et al., 2025) and RadCliQ (Yu et al., 2023a) output structured error counts, such as per ReXVal category, rather than a scalar score, which makes them incompatible with our pairwise evaluation. As previously discussed, simple error counting is also inherently limited. Similarly, FineRadScore (Huang et al., 2024) employs an LLM to classify and explain errors line by line, but it does not provide a single numerical output and is highly dependent on sentence-level formulations.

Furthermore, since our goal is not to rank metrics in an absolute sense, we chose not to conduct statistical significance tests such as t-tests. Instead, we aim to demonstrate that many existing metrics lack the ability to capture clinical semantics and to provide guidance for future metric design.

Data Availability. Our dataset is derived from MIMIC-CXR and ReXVal, both of which are distributed under the PhysioNet Credentialed Health Data License 1.5.0. Our work uses only these publicly available, credentialed datasets, and our derived dataset does not involve new data collection from patients. As a result, we cannot openly redistribute the report texts. Instead, we provide annotation guidelines, error taxonomy, and processing

scripts so that credentialed users can reproduce our dataset from their own copies of MIMIC-CXR and ReXVal.

We have also initiated the process of submitting our derived dataset to PhysioNet for controlled release under the same license, ensuring compliance with patient privacy and reproducibility standards. Updated information can be found at <https://github.com/ruochenli99/ReEvalMed>.

Intended Use. Our use of MIMIC-CXR and ReXVal strictly followed their intended purpose of research-only use. The derived dataset inherits the same restrictions and is provided solely for research, not for clinical decision-making or commercial applications.

Privacy and Safety. MIMIC-CXR and ReXVal have been de-identified by the dataset providers in compliance with HIPAA. No personally identifying information or offensive content is present in these datasets. Our derived dataset contains only de-identified report texts and annotation labels, and does not introduce any additional personally identifying or offensive content.

Documentation. All evaluation metrics used in this work were implemented via publicly available repositories, each with its own documentation and license (e.g., MIT, Apache 2.0). We used the official or widely adopted implementations without modification to ensure reproducibility, and we provide references to the original papers.

MIMIC-CXR and ReXVal, which consist of English radiology reports (findings sections) from chest X-rays collected at a large U.S. academic medical center. Our derived dataset focuses on the findings of the report. No demographic attributes of patients are included.

Experiment Details. For generating rewritten reports, we used the DeepSeek-R1 7B model with fixed prompts and standard decoding settings, run locally on an NVIDIA H100 80GB GPU. Nonetheless, regardless of the model used, all generated modifications were carefully reviewed and validated by clinical experts to ensure correctness and clinical plausibility. The evaluation metrics were used via their publicly released implementations without modification, relying on their default model sizes and parameters. All experiments are lightweight and inference-only.

Descriptive Statistics. We report confidence intervals for each evaluation metric. For the Discriminative Score and Robustness Score, results are computed as the mean over 10 samples per

evaluation aspect.

Use of AI Assistants. ChatGPT and Grammarly were used to support writing and editing tasks, including drafting LaTeX tables, formatting references, and suggesting wording for writing paper.

C Appendix: Future Work

Prior evaluation assessments, often based on a single correlation coefficient, may obscure important limitations. Our proposed aspect-based Meta-Evaluation framework aims to explicitly test whether metrics can distinguish between clinically significant and insignificant errors across a wide range of real-world report variations, offering a path toward more clinically aligned metric design.

Moving forward, we hope future metric development can incorporate these clinical aspects more explicitly. For example:

Knowledge infusion (e.g., integrating domain-specific ontologies or structured clinical guidelines) may help metrics reason about subtle but clinically meaningful variations.

Chain-of-thought prompting or step-by-step reasoning could guide LLM-based metrics to better assess the semantic consistency and clinical implications of generated content.

Agent-based debate or multi-agent deliberation may offer a way to simulate clinical decision-making dynamics when evaluating borderline cases or conflicting evidence.

An automatic and scalable workflow is critical, as high-quality dataset construction in clinical domains is inherently labor-intensive due to the need for expert validation. However, we designed our annotation protocol with future scalability in mind. Specifically, we found that LLM-based rewriting (guided by structured prompts) can generate clinically realistic error types across dimensions (e.g., severity, location, description). Current generations are generally acceptable to clinicians upon review. They significantly reduce clinical workload, especially when paired with targeted expert review rather than full manual rewriting. In future iterations, we plan to explore more robust semi-automated pipelines combining LLM generation and selective expert validation. This can enable scalable benchmark extension without sacrificing quality.