CRIMSON: A Clinically-Grounded LLM-Based Metric for Generative Radiology Report Evaluation

Mohammed Baharoon[1,⋆], Thibault Heintz[2,⋆], Siavash Raissi[1,⋆], Mahmoud Alabbad[3], Mona Alhammad[4], Hassan AlOmaish[5], Sung Eun Kim[1], Oishi Banerjee[1], and Pranav Rajpurkar[1,†]

[1] Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[2] Department of Radiation Oncology, Mass General Brigham, Boston, MA, USA
[3] King Fahad Hospital, Al-Ahsa Health Cluster, Al Hofuf, Saudi Arabia
[4] Ras-Tanura General Hospital, Ministry of Health, Eastern Province, Saudi Arabia
[5] Department of Medical Imaging, King Abdulaziz Medical City, Ministry of National Guard, Riyadh, Saudi Arabia

[⋆]These authors contributed equally.
[†]Corresponding author: pranav_rajpurkar@hms.harvard.edu

Abstract. We introduce **CRIMSON**, a clinically grounded evaluation framework for chest X-ray report generation that assesses reports based on diagnostic correctness, contextual relevance, and patient safety. Unlike prior metrics, CRIMSON incorporates full clinical context, including patient age, indication, and guideline-based decision rules, and prevents normal or clinically insignificant findings from exerting disproportionate influence on the overall score. The framework categorizes errors into a comprehensive taxonomy covering false findings, missing findings, and eight attribute-level errors (e.g., location, severity, measurement, and diagnostic overinterpretation). Each finding is assigned a clinical significance level (urgent, actionable non-urgent, non-actionable, or expected/benign), based on a guideline developed in collaboration with attending cardiothoracic radiologists, enabling severity-aware weighting that prioritizes clinically consequential mistakes over benign discrepancies. CRIMSON is validated through strong alignment with clinically significant error counts annotated by six board-certified radiologists in ReXVal (Kendall's $\tau = 0.61$–0.71; Pearson's $r = 0.71$–0.84), and through two additional benchmarks that we introduce. In *RadJudge*, a targeted suite of clinically challenging pass–fail scenarios, CRIMSON shows consistent agreement with expert judgment. In *RadPref*, a larger radiologist preference benchmark of over 100 pairwise cases with structured error categorization, severity modeling, and 1–5 overall quality ratings from three cardiothoracic radiologists, CRIMSON achieves the strongest alignment with radiologist preferences. We release the metric, the evaluation benchmarks, RadJudge and RadPref, and a fine-tuned MedGemma model to enable reproducible evaluation of report generation, all available at https://github.com/rajpurkarlab/CRIMSON.

Keywords: Report Evaluation · Radiology Report Generation

1 Introduction

Automated radiology report generation has advanced rapidly with the emergence of large vision-language models, yet reliable evaluation remains a fundamental challenge [21, 17, 6]. Recent radiology-specific metrics have moved beyond surface-level text similarity and instead assess factual correctness through structured error counting and finding-level comparison [9, 23, 4, 22, 15, 3]. These approaches represent important progress toward clinically meaningful evaluation by explicitly detecting hallucinations and omissions.

Despite these advances, current metrics largely treat detected errors as either uniformly important or binary (significant vs. not significant), and evaluate findings in relative isolation from broader clinical context. In practice, the clinical consequences of errors vary substantially. For instance, failing to report a life-threatening pneumothorax is categorically different from missing age-related aortic calcification. Moreover, the relevance and interpretation of findings depend on a patient's age and indication. Existing evaluation frameworks do not explicitly encode clinical severity as a fine-grained spectrum; instead, they collapse findings into coarse categories, lack sufficient clinical context to determine true significance, or rely on LLM judgments without structured in-context guidelines [9, 4, 15]. Consequently, these frameworks conflate minor, clinically inconsequential discrepancies with omissions that directly impact patient safety.

To address these limitations, we introduce CRIMSON, a clinically grounded LLM-based evaluation framework designed to align automated assessment with real-world radiologic reasoning. CRIMSON evaluates reports at the level of individual findings while incorporating full clinical context, including patient age and indication. The framework models a comprehensive taxonomy of errors—false findings, missing findings, attribute-level errors (e.g., location, severity, measurement) and significance labels (urgent, actionable non-urgent, non-actionable, or expected/benign)—defined according to a rubric developed with cardiothoracic radiologists. The severity labels determine weights within a principled scoring formulation that prioritizes clinically consequential errors over benign discrepancies and supports partial credit for partially correct findings under fine-grained attribute rules. We validate CRIMSON through alignment with radiologist-annotated clinically significant error counts in ReXVal [18, 17] (PhysioNet Credentialed Health Data License 1.5.0). We also introduce and validate the metric using RadJudge, a targeted ranking test suite of clinically challenging scenarios, and RadPref, a comprehensive radiologist preference benchmark, demonstrating improved agreement with expert judgment compared to existing metrics.

2 CRIMSON: Clinically-Grounded Report Evaluation

CRIMSON evaluates a candidate radiology report against a reference report by identifying and categorizing discrepancies at the finding level, weighing each error by its clinical significance, and computing a normalized score that reflects clinical preferences. GPT-5.2 [12] is used as the backbone throughout this pipeline. The

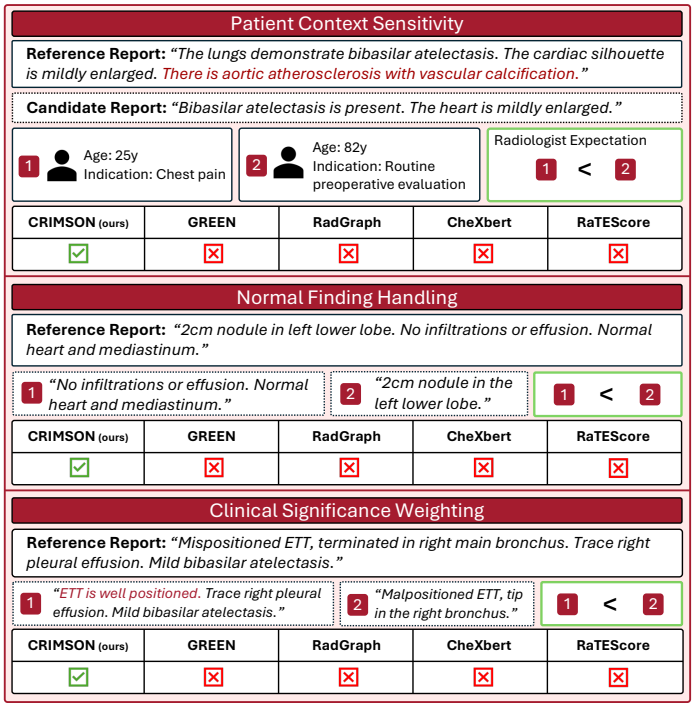

Fig. 1. Representative RadJudge cases illustrating core design principles of CRIMSON. **Top:** Patient context sensitivity. The clinical impact of an omission (e.g., aortic atherosclerosis) varies by age and indication, and CRIMSON adjusts severity accordingly. **Middle:** Normal finding handling. CRIMSON does not reward mentioning normal findings, preventing score inflation. **Bottom:** Clinical significance weighting. Errors are weighted by consequence, prioritizing clinically important findings. In each case, CRIMSON aligns with radiologist expectations, whereas prior metrics fail.

framework operates in three stages: (1) finding extraction and clinical significance assignment, (2) error detection and classification, and (3) severity-aware score computation.

2.1 Finding Extraction and Clinical Significance Assignment

Given a reference report R_{ref} and a candidate report R_{cand}, CRIMSON first extracts all abnormal findings from each report. Normal findings are excluded from evaluation because including them can introduce spurious variability due to stylistic differences across radiologists [2, 1]. Each extracted finding f is assigned a clinical significance weight $w(f)$ based on standard radiological practice and a structured severity framework adapted from [14] with input from attending cardiothoracic radiologists. The clinical significance weight $w(f)$ is defined as $w(f) = 1.0$ if the finding is urgent, 0.5 if actionable but not urgent, 0.25 if

not actionable and not urgent, and 0.0 if expected or benign. Findings classified as *urgent* correspond to abnormalities requiring immediate intervention or indicating life-threatening conditions, such as a tension pneumothorax. *Actionable non-urgent* findings are those that alter patient management but are not immediately critical, including nodules, moderate pleural effusions, or consolidations. *Not actionable* are findings that represent minimal clinical impact but remain worth documenting, such as a cervical rib or appropriately positioned support devices. *Expected benign* findings include expected or age-appropriate changes with no impact on care, such as degenerative spine changes or a tortuous aorta. We separate non-actionable from expected benign findings because reporting of expected benign findings varies substantially by radiologist style; penalizing based on them introduces unnecessary randomness. Clinical significance assignment incorporates patient context when available, including age and indication. For example, aortic calcification in a 75-year-old patient is classified as expected benign, whereas the same finding in a 25-year-old patient may be considered actionable non-urgent due to atypical early onset.

2.2 Error Taxonomy and Classification

CRIMSON characterizes discrepancies through three primary error categories: false findings, missing findings, and attribute errors. False findings are abnormal findings present in R_{cand} but absent from R_{ref}, representing a hallucination. Missing findings are abnormal findings present in R_{ref} but absent from R_{cand}, representing diagnostically meaningful omissions.

Findings that appear in both reports are considered "matched." For such findings, CRIMSON evaluates attribute-level correctness across eight dimensions: (1) anatomical location or laterality, (2) severity or extent, (3) morphological descriptors, (4) quantitative measurements, (5) certainty level, (6) diagnostic underinterpretation, (7) overinterpretation, and (8) temporal or comparison descriptors.

Each attribute error e is assigned a severity-based weight $w_{\text{attr}}(e)$, where $w_{\text{attr}}(e) = 0.5$ if the error is labeled as significant, and $w_{\text{attr}}(e) = 0.0$ if it is labeled as negligible. Significant attribute errors are those that could alter treatment decisions or patient management, whereas negligible errors correspond to clinically inconsequential differences. For example, incorrect lung laterality is considered significant, whereas positional differences within the same lobe, such as 'apical' vs 'lateral' are considered negligible. For pulmonary nodules smaller than 6 mm, measurement discrepancies exceeding 2 mm are considered significant; for nodules 6 mm or larger, discrepancies exceeding 4 mm are considered significant, reflecting established practice [8]. Changes in severity descriptors that affect urgency, such as "small" versus "large," are significant, whereas as "small" versus "tiny" are negligible. A single matched finding may contain multiple attribute errors, each evaluated independently.

2.3 Severity-Aware Scoring

The framework produces a score in the range of $(-1, 1]$ that can be easily interpreted in clinical workflows. The scale is grounded at 0, which corresponds to a normal candidate report. This reflects a practical assumption that a radiologist begins from a normal template and modifies the report by adding abnormal findings. A score greater than zero indicates that the candidate report contains more correct findings than errors after severity weighting. A score equal to zero indicates that the report is no more informative than submitting a normal template, except when the reference report is also normal, in which case a correct normal report receives a score of 1. A score less than zero indicates that the report contains more errors than correct findings, implying that a radiologist would likely spend more effort correcting it than editing a normal template. The upper bound of 1 represents a perfect report with no missed findings, no false positives, and no significant attribute errors. Negative scores approach -1 asymptotically because errors are theoretically unbounded: a candidate can always become worse by introducing additional false findings.

For each matched finding m_i, let its clinical significance weight be $w_i = w(m_i)$. Attribute-level penalties are aggregated as $E_{\text{attr},i} = \sum_j w_{\text{attr}}(e_{j,i})$, where $e_{j,i}$ referes to attribute error j for finding i, and total credit, C, across matched findings is $C = \sum_{i \in \text{matched}} w_i \cdot \frac{w_i}{w_i + E_{\text{attr},i}}$. Let $W_{\text{ref}} = \sum_{f \in R_{\text{ref}}} w(f)$ denote the total weighted clinical significance of the reference report, and let $E_{\text{false}} = \sum_{f \in \text{false}} w(f)$ denote the weighted sum of false positive findings. The raw score is defined as:

$$S = \begin{cases} \dfrac{C - E_{\text{false}}}{W_{\text{ref}}} & \text{if } W_{\text{ref}} > 0, \\ -E_{\text{false}} & \text{if } W_{\text{ref}} = 0 \text{ and } E_{\text{false}} > 0. \\ 1 & \text{if } W_{\text{ref}} = 0 \text{ and } E_{\text{false}} = 0, \end{cases} \tag{1}$$

To bound the negative range while preserving relative ordering, let $A = E_{\text{false}} - C$, which represents the excess weighted errors relative to correct findings. The final score is:

$$\text{CRIMSON} = \begin{cases} S & \text{if } S \geq 0 \\ -\dfrac{A}{1 + A} & \text{if } S < 0 \end{cases} \tag{2}$$

3 Results

We perform three complementary forms of validation: correlation with radiologist-annotated clinically significant error counts (Section 3.1), a radiologist-guided pass–fail clinical judgment test on RadJudge (Section 3.2), and large-scale radiologist preference alignment on RadPref (Section 3.3).

RadJudge	CRIMSON	GREEN	RadGraph	CheXbert	RaTEScore	RadCliQ-v1	ROUGE-L	BLEU	BERTScore
1 False Finding Penalization	3	0	1	0	1	3	1	1	0
2 Patient Context Sensitivity	3	0	0	0	0	0	0	0	0
3 Normal Finding Handling	3	0	0	2	0	0	0	0	0
4 Paraphrase Robustness	3	1	1	2	0	0	0	1	0
5 Clinical Practicality	3	2	0	2	1	0	0	1	0
6 Location Error Handling	3	3	0	0	0	0	0	0	0
7 Measurement Error Sensitivity	3	1	1	0	1	2	1	1	2
8 Diagnostic Precision	3	1	0	0	3	1	0	0	3
9 Clinical Significance Weighting	3	0	0	2	2	2	0	0	1
10 Partial Credit Assignment	3	2	2	0	0	0	0	0	0
Total (/30)	**30**	**10**	**5**	**8**	**8**	**8**	**2**	**4**	**6**

Fig. 2. RadJudge results. For each case, metrics are evaluated based on whether their relative ranking of multiple candidate reports agrees with the expected ordering determined with agreement across three attending cardiothoracic radiologists. Each category contains three cases; entries are cases passed (out of 3), with totals out of 30.

3.1 Correlation with Radiologist-Annotated Significant Errors

We evaluated CRIMSON on 50 cases from ReXVal [18] annotated by six board-certified radiologists. We computed Kendall's τ and Pearson r correlations between automatic metric scores and radiologist-derived clinically significant error counts. As shown in Table 1, CRIMSON demonstrates strong alignment with these expert annotations. Furthermore, error counts (E) exhibited even stronger alignment, and severity-weighted errors (Weighted E) achieved the highest correlations overall, demonstrating that explicitly modeling clinical consequence further improves agreement with expert judgment.

3.2 Radiologist-Guided Clinical Judgment Test

We also developed RadJudge, a targeted pass–fail test suite reflecting real-world radiologist intuition. RadJudge comprises 30 curated cases across 10 clinically nuanced categories in which multiple candidate reports are compared and independently reviewed by three cardiothoracic radiologists, with agreement required to establish the reference preference. A metric passes if it ranks reports in accordance with expert judgment or assigns equivalent scores when radiologists deem them clinically indistinguishable, defined as differences within a threshold of 0.01. Three representative cases are shown in Figure 1. The suite probes challenging scenarios, including urgent omissions versus benign hallucinations, context-dependent findings, diagnostic over- and under-interpretation, and situations reflecting the clinical reality of imperfect reference reports that omit localization or age-expected benign findings.

Table 1. Kendall τ and Pearson r correlations (95% CI) between automatic metrics and radiologist-derived clinically significant error counts ($n = 50$). Columns refer to different candidate reports on ReXVal, each of which was chosen to optimize a specific metric [18]. GREEN E and CRIMSON E denote the total (unweighted) error count, while CRIMSON Weighted E applies clinical severity–based weighting to errors. All other metrics were calculated using RadEval [16]. *CRIMSON results are averaged across 5 runs due to non-deterministic API outputs.

Metric	**Correlation with Significant Error Counts**			
	CheXbert		**BERTScore**	
	Kendall τ	Pearson r	Kendall τ	Pearson r
RadGraph [5]	0.41 [0.19,0.61]	0.59 [0.41,0.75]	0.54 [0.36,0.68]	0.65 [0.51,0.78]
BLEU [10]	0.49 [0.30,0.65]	0.60 [0.47,0.72]	0.36 [0.16,0.54]	0.48 [0.32,0.63]
BERTScore [19]	0.52 [0.35,0.67]	0.65 [0.52,0.78]	0.49 [0.30,0.66]	0.60 [0.44,0.74]
GREEN [9]	0.62 [0.46,0.75]	0.75 [0.64,0.86]	0.67 [0.54,0.78]	0.70 [0.59,0.80]
ROUGE-L [7]	0.58 [0.44,0.71]	0.71 [0.60,0.81]	0.54 [0.37,0.70]	0.62 [0.46,0.75]
CheXbert [13]	0.46 [0.26,0.63]	0.45 [0.18,0.70]	0.30 [0.08,0.51]	0.34 [0.09,0.60]
RaTEScore [23]	0.39 [0.17,0.57]	0.52 [0.31,0.69]	0.49 [0.32,0.65]	0.56 [0.37,0.73]
RadCliQ-v1 [17]	0.34 [0.21,0.46]	0.34 [0.19,0.52]	0.35 [0.21,0.48]	0.35 [0.22,0.53]
CRIMSON*	0.68 [0.54,0.79]	0.84 [0.76,0.90]	0.71 [0.60,0.80]	0.82 [0.74,0.89]
GREEN E	0.71 [0.59,0.81]	0.75 [0.65,0.86]	0.75 [0.63,0.85]	0.85 [0.75,0.92]
CRIMSON E	0.73 [0.61,0.83]	0.88 [0.79,0.94]	0.72 [0.62,0.80]	0.86 [0.77,0.93]
CRIMSON Weighted E	**0.78** [0.67,0.86]	**0.90** [0.85,0.95]	**0.80** [0.71,0.87]	**0.91** [0.88,0.95]
	RadGraph		**BLEU**	
	Kendall τ	Pearson r	Kendall τ	Pearson r
RadGraph	0.59 [0.46,0.71]	0.60 [0.50,0.72]	0.64 [0.50,0.75]	0.75 [0.65,0.84]
BLEU	0.13 [0.09,0.34]	0.23 [0.02,0.39]	0.52 [0.34,0.68]	0.67 [0.53,0.79]
BERTScore	0.46 [0.29,0.61]	0.54 [0.39,0.68]	0.58 [0.41,0.72]	0.72 [0.60,0.83]
GREEN	0.62 [0.48,0.74]	0.65 [0.53,0.77]	0.70 [0.55,0.83]	0.79 [0.68,0.89]
ROUGE-L	0.54 [0.38,0.67]	0.60 [0.49,0.70]	0.67 [0.52,0.79]	0.80 [0.70,0.87]
CheXbert	0.29 [0.08,0.48]	0.33 [0.08,0.55]	0.18 [0.05,0.40]	0.23 [0.04,0.47]
RaTEScore	0.57 [0.42,0.70]	0.62 [0.45,0.78]	0.54 [0.39,0.68]	0.67 [0.54,0.79]
RadCliQ-v1	0.12 [0.05,0.28]	0.06 [0.11,0.28]	0.28 [0.11,0.43]	0.16 [0.01,0.53]
CRIMSON*	0.61 [0.45,0.75]	0.71 [0.53,0.85]	0.67 [0.54,0.79]	0.81 [0.71,0.89]
GREEN E	0.71 [0.59,0.81]	0.75 [0.65,0.86]	**0.80** [0.71,0.88]	**0.88** [0.82,0.93]
CRIMSON E	0.73 [0.61,0.83]	0.86 [0.78,0.92]	0.74 [0.61,0.84]	0.87 [0.78,0.93]
CRIMSON Weighted E	**0.77** [0.67,0.85]	**0.86** [0.80,0.93]	0.78 [0.69,0.86]	**0.88** [0.82,0.93]

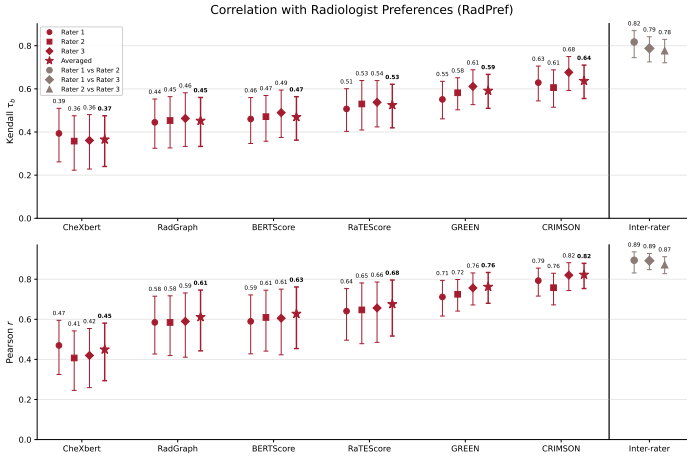

Fig. 3. Radiologist Preference Alignment (RadPref). Correlation between metric score and radiologist rating differences across 100 pairwise cases. Each point corresponds to a case comparing two candidate reports for the same reference report.

As shown in Figure 2, CRIMSON is the only metric that correctly solves all 30 out of 30 cases, consistently ranking candidate reports in accordance with expert radiologist judgment. In contrast, all prior metrics perform substantially worse, correctly resolving fewer than 35% of cases, highlighting their limited ability to capture nuanced clinical judgment.

3.3 Radiologist Preference Alignment

Correlation with error counts may not fully capture radiologist preference, as experts don't weigh different types of errors equally in overall judgment. To directly assess preference alignment, we introduce RadPref, a benchmark of 100 cases, each containing a reference report from ReXGradient-160K [20] (Non-Commercial Data Access and Use Agreement (v1, April 7, 2025)) and two candidate reports randomly generated using diverse regimes: report generation with MedGemma [11], randomly sampled reports, BERT similarity–matched reports, and LLM-based editing, addition, or removal of findings. Each candidate was rated on a 1–5 scale by three cardiothoracic radiologists based on overall clinical quality and correctness relative to the referene report. Scores were computed separately for both candidates.

Figure 3 shows that CRIMSON demonstrates the strongest alignment with radiologist pairwise preferences among all evaluated metrics. Across Kendall's τ_b and Pearson r, CRIMSON consistently outperforms prior approaches and approaches inter-rater radiologist agreement. These findings indicate that CRIMSON more faithfully reflects expert clinical preference in relative report quality.

4 Conclusion

We introduce CRIMSON, a clinically grounded and severity-aware framework for fine-grained radiology report evaluation that explicitly models patient context, diagnostic consequence, and structured attribute-level errors. By incorporating clinical significance weights and score normalization, CRIMSON aligns automated evaluation with real-world radiologist reasoning. Across correlation with radiologist-annotated significant error counts, RadJudge, and RadPref, CRIMSON consistently demonstrates stronger agreement with expert judgment than prior metrics, establishing a robust and interpretable foundation for clinically meaningful evaluation of report generation systems.

Acknowledgments. This research was supported in part by Nebius AI and the Harvard Medical School Dean's Innovation Award for Accelerating Foundation Model Research.

Disclosure of Interests. The authors declare no competing interests.

References

1. Baharoon, M., Ma, J., Fang, C., Toma, A., Wang, B.: Exploring the design space of 3d mllms for ct report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 237–246. Springer (2025)
2. Baharoon, M., Raissi, S., Jun, J.S., Heintz, T., Alabbad, M., Alburkani, A., Kim, S.E., Kleinschmidt, K., Alhumaydhi, A.O., Alghamdi, M.M.G., et al.: Radgame: An ai-powered platform for radiology education. arXiv preprint arXiv:2509.13270 (2025)
3. Chaves, J.M.Z., Huang, S.C., Xu, Y., Xu, H., Usuyama, N., Zhang, S., Wang, F., Xie, Y., Khademi, M., Yang, Z., et al.: Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation. arXiv preprint arXiv:2403.08002 (2024)
4. Huang, A., Banerjee, O., Wu, K., Reis, E.P., Rajpurkar, P.: Fineradscore: A radiology report line-by-line evaluation technique generating corrections with severity scores. arXiv preprint arXiv:2405.20613 (2024)
5. Jain, S., Agrawal, A., Saporta, A., Truong, S.Q., Duong, D.N., Bui, T., Chambon, P., Zhang, Y., Lungren, M.P., Ng, A.Y., et al.: Radgraph: Extracting clinical entities and relations from radiology reports. arXiv preprint arXiv:2106.14463 (2021)
6. Li, R., Li, J., Jian, B., Yuan, K., Zhu, Y.: Reevalmed: Rethinking medical report evaluation by aligning metrics with real-world clinical judgment. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 11823–11837 (2025)
7. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
8. MacMahon, H., Naidich, D.P., Goo, J.M., Lee, K.S., Leung, A.N., Mayo, J.R., Mehta, A.C., Ohno, Y., Powell, C.A., Prokop, M., et al.: Guidelines for management of incidental pulmonary nodules detected on ct images: from the fleischner society 2017. Radiology **284**(1), 228–243 (2017)

9. Ostmeier, S., Xu, J., Chen, Z., Varma, M., Blankemeier, L., Bluethgen, C., Md, A.E.M., Moseley, M., Langlotz, C., Chaudhari, A.S., et al.: Green: Generative radiology report evaluation and error notation. In: Findings of the association for computational linguistics: EMNLP 2024. pp. 374–390 (2024)
10. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
11. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
12. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
13. Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A.Y., Lungren, M.P.: Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. arXiv preprint arXiv:2004.09167 (2020)
14. Tian, K., Hartung, S.J., Li, A.A., Jeong, J., Behzadi, F., Calle-Toro, J., Adithan, S., Pohlen, M., Osayande, D., Rajpurkar, P.: Refisco: Report fix and score dataset for radiology report generation. PhysioNet (2023)
15. Wang, Z., Luo, X., Jiang, X., Li, D., Qiu, L.: Llm-radjudge: Achieving radiologist-level evaluation for x-ray report generation. arXiv preprint arXiv:2404.00998 (2024)
16. Xu, J., Zhang, X., Abderezaei, J., Bauml, J., Boodoo, R., Haghighi, F., Ganjizadeh, A., Brattain, E., Van Veen, D., Meng, Z., et al.: Radeval: A framework for radiology text evaluation. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 546–557 (2025)
17. Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E.P., Fonseca, E.K.U.N., Lee, H.M.H., Abad, Z.S.H., Ng, A.Y., et al.: Evaluating progress in automatic chest x-ray radiology report generation. Patterns **4**(9) (2023)
18. Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E.P., Fonseca, E., Lee, H., Shakeri, Z., Ng, A., et al.: Radiology report expert evaluation (rexval) dataset (2023)
19. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)
20. Zhang, X., Acosta, J.N., Miller, J., Huang, O., Rajpurkar, P.: Rexgradient-160k: A large-scale publicly available dataset of chest radiographs with free-text reports. In: Machine Learning for Health 2025
21. Zhang, X., Acosta, J.N., Yang, X., Adithan, S., Luo, L., Zhou, H.Y., Miller, J., Huang, O., Zhou, Z., Hamamci, I.E., et al.: Automated chest x-ray report generation remains unsolved. In: Biocomputing 2026: Proceedings of the Pacific Symposium. pp. 236–250. World Scientific (2025)
22. Zhang, Z., Lee, K., Deng, W., Zhou, H., Jin, Z., Huang, J., Gao, Z., Marshall, D.C., Fang, Y., Yang, G.: Gema-score: Granular explainable multi-agent score for radiology report evaluation. arXiv preprint arXiv:2503.05347 (2025)
23. Zhao, W., Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Ratescore: A metric for radiology report generation. arXiv preprint arXiv:2406.16845 (2024)