SPEC-CXR: Advancing Clinical Safety through Entity-Level Performance Evaluation of Chest X-ray Report Generation

Jung Oh Lee*, Junwoo Cho*, Junha Kim*, Laurent Dillard, Tom van Sonsbeek, Arnaud A. A. Setio, Hyeonsoo Lee, Donggeun Yoo, and Taesoo Kim†

Lunit Inc., Seoul, South Korea
leejung209@gmail.com,{jwcho,jhkim,laurent.dillard,tom.vansonsbeek, arnaud.setio,hslee,dgyoo,taesoo.kim}@lunit.io

Abstract. Automated chest X-ray report generation has great potential to improve healthcare efficiency, but rigorous validation is essential for safe clinical adoption. Existing evaluation metrics focus mainly on report-level scores, failing to provide actionable insights for clinicians.
In this paper, we present SPEC-CXR (Safety-centered Performance Evaluation in Clinical Report for Chest X-Ray), an evaluation framework that integrates entity-level performance assessment with report-level error analysis using a large language model (LLM). In our approach, the LLM extracts and classifies entities—radiological findings and differential diagnoses—from both generated and reference reports based on a carefully curated entity set. Generated reports are then evaluated on entity presence, location, severity, and prior comparison, yielding structured outputs to calculate detailed entity-level scores (F1 for presence and accuracy for location, severity, and comparison).
Our entity-level evaluation shows 91.8% accuracy compared to human evaluation for presence detection and 0.777 Kendall's tau-b correlation for report-level evaluation. Furthermore, our entity-level performance analysis uncovers critical limitations of current state-of-the-art report generation models across diverse entities, highlighting the urgent need for rigorous, safety-oriented evaluation metrics.
Our framework is publicly available and usable: `https://github.com/lunit-io/spec-cxr`.

Keywords: Chest X-ray · Report generation · Report evaluation metric

1 Introduction

In the chest X-ray (CXR) domain, research on vision-language models (VLMs) has surged, aiming to produce accurate radiology reports directly from imaging data [1,3,5,13,17,22,31]. To evaluate these systems, various metrics have been proposed—from traditional natural language generation (NLG) scores [20,28]

* Equal contribution
† Corresponding author

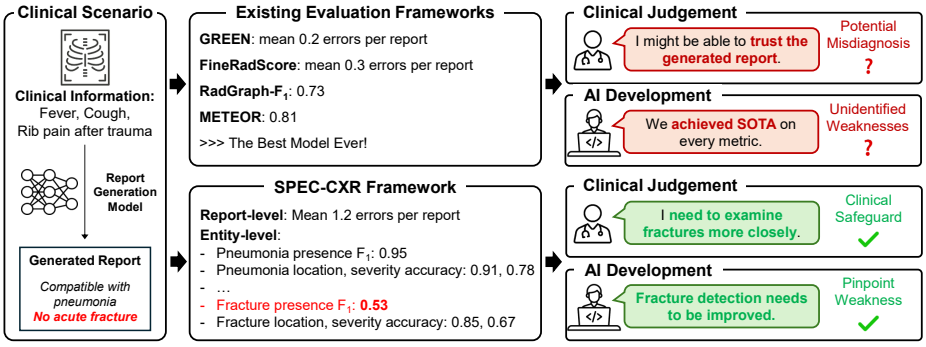

Fig. 1. Comparison of evaluation approaches and their impact on clinical safety and AI development. **Left:** Clinical scenario with AI-generated report identifying pneumonia but no fracture. **Center:** Traditional evaluation frameworks (top) report strong overall performance, potentially masking critical errors in entity-specific findings. In contrast, the SPEC-CXR framework (bottom) highlights a major weakness in fracture detection. **Right:** These different evaluation methods lead to contrasting clinical judgments and AI development priorities.

to entity extraction-based approaches that focus on clinically relevant findings [10,30]. More recently, methods leveraging large language models (LLMs) have been proposed to compare generated and reference reports to score the report pairs using pre-defined criteria [8,15,16,19].

Despite the recent progress of such evaluation methods, accurately estimating the real-world clinical utility of automatic report generation systems remains a critical challenge. Traditional NLG metrics merely quantify sentence similarity without capturing any clinical context. Most LLM-based metrics only quantify report-wise errors, making it hard to assess model-wise detection performance. These limitations make it difficult to assess model reliability across diverse clinical scenarios, preventing real-world clinical adoption.

Figure 1 shows that awareness of entity-level performance helps clinicians better identify the model's reliability for each entity compared to a single report-level score. Existing entity extraction-based metrics [10,11,23,30] take a structured approach by identifying and comparing clinically meaningful entities in the generated and reference reports. While this improves interpretability and alignment with clinical relevance, most existing methods extract entities in a free-form manner without relying on a pre-defined set. This makes it difficult to perform consistent comparisons across examples or models. Additionally, these metrics often aggregate scores over all extracted entities, obscuring model weakness in specific clinical categories. Our method addresses these gaps by integrating report-level error analysis with entity-level scoring based on a carefully curated set of clinical entities, enabling a more precise and safety-oriented evaluation of report generation models.

In this paper, we present an evaluation framework for automated report generation systems that employs LLMs to assess overall report accuracy and measures performance across an extensive set of categorized clinical entities. This offers clinicians granular, clinically meaningful insights for real world adoption. Specifically, our framework enables fine-grained examination of entity-level performance, which reveals critical shortcomings in recent report generation models, underscoring the need for rigorous, safety-focused evaluations to ensure real-world use of automatic radiology reporting systems. In addition, our method quantifies report-level error counts, showing high correlation with clinician-annotated errors on the ReXVal dataset [27].

1. A comprehensive and customizable set of clinical entities for ensuring safe automatic CXR radiology reporting, with category of radiological findings and differential diagnoses, curated by expert radiologists.
2. An evaluation framework for LLM generated reports grounded on our entity definition, providing both report-level scores and fine-grained quantitative metrics on presence, location, severity and prior comparison for each entity.
3. Identification of limitations in recent report generation methods, providing guidance to the community to build more clinically safe methods.

2 Method

2.1 Overview of the Proposed Evaluation Framework

Figure 2 presents our proposed framework integrating both entity-level and report-level evaluation in CXR report generation. In this framework, we specifically define entities as radiological findings and differential diagnoses, where anatomical structures and descriptive terms are excluded. We provide reference and candidate reports to an LLM, which evaluates errors for all entities present in both reports. The generated text from the LLM is structured in a pre-defined JSON format to facilitate systematic error analysis. The following sections describe our approach in detail.

2.2 SPEC-CXR: Safety-Centered Performance Evaluation in Clinical Reports for Chest X-Ray

Entity Category Three board-certified radiologists developed comprehensive entity sets based on anatomical and clinical expertise to categorize all possible chest X-ray findings and differential diagnoses. The entity set includes a total of 89 findings and 33 differential diagnoses (Figure 3). Unlike previous approaches such as CheXpert with 14 findings [9] or the UMLS terminology used in PadChest dataset [2,4], we designed our entity set to be mutually exclusive while covering all findings and differential diagnoses.

Although we propose this entity set for benchmark compatibility, the set is easily customizable by the user. In Figure 2 (b), we treat each entity as a distinct, named field in a Pydantic [6] model. This structure decouples the evaluation logic from the specific list of entities, allowing seamless extensions to incorporate new or institution-specific entities as clinical needs evolve.

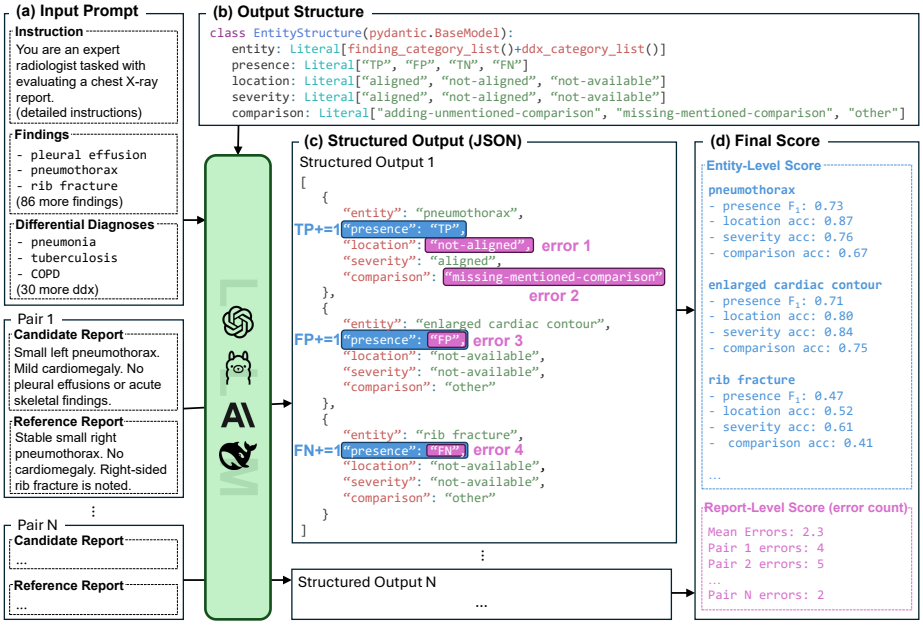

Fig. 2. Our approach to evaluate chest X-ray reports. (a) The input prompt consists of descriptive instructions, an entity set containing findings and differential diagnoses. (b) The output structure categorizes entities based on predefined attributes. This is given to an LLM to force the generation structure. (c) The structured output in JSON format encodes errors related to different entities. (d) The final scoring mechanism constructs a confusion matrix and counts errors to assign entity-level and report-level scores.

Structured Output from LLM Recent LLMs show strong capabilities in structured output generation [24,25], allowing for more organized and interpretable results by constraining responses to pre-defined formats. Our pre-defined output structure is illustrated as a Pydantic [6] model in Figure 2 (b). Specifically, the LLM evaluates four attributes of each entity: *presence*, *location*, *severity*, and *comparison*. For presence, the LLM categorizes into one of `TP` (true positive), `FP` (false positive), `TN` (true negative), or `FN` (false negative). Location and severity are evaluated as `aligned`, `not-aligned`, or `not-available`, while comparison is assessed as `adding-unmentioned-comparison`, `missing-mentioned-comparison`, or `other`. The generated text is returned in a JSON format (Figure 2 (c)).

Entity-Level and Report-Level Metrics As shown in Figure 2 (d), the final score can be computed at two levels: entity and report. For entity-level performance, we calculated the F_1 score for entity presence using the structured outputs from all reference-candidate report pairs, and we used accuracy as the metric for the evaluations of location, severity, and comparison. The report-

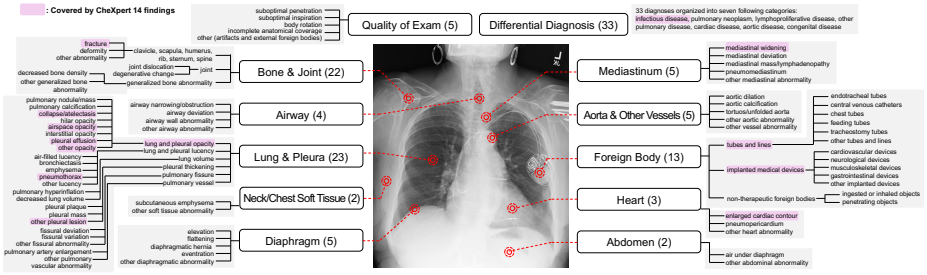

Fig. 3. The comprehensive entity classification system developed for SPEC-CXR. Entities are organized into key anatomical categories, exam quality, and differential diagnoses, with detailed abnormalities extending far beyond the CheXpert 14 findings (purple). The framework enables systematic evaluation across 89 findings and 33 differential diagnoses.

level score is defined as the total error count in the generated report. This error count is derived from the number of occurrences of `FP`, `FN`, `not-aligned`, `adding-unmentioned-comparison`, and `missing-mentioned-comparison` in the LLM's structured output.

3 Experiments and Results

3.1 Datasets

MIMIC-CXR We utilized MIMIC-CXR [12] ReXrank test split [29], which contains 2,347 paired images and free-text reports. This dataset is used to evaluate the coverage of our entity set against report contents and to assess entity-level performance using reports generated by state-of-the-art models.

ReXVal The ReXVal dataset [27] consists of 200 reference-candidate report pairs (50 reference reports with 4 candidate reports each), evaluated by six radiologists who counted the number of errors in each pair. This dataset enables us to assess how well report evaluation metrics align with human evaluation. We used this dataset to validate our report-level metrics and compare them with other existing metrics.

3.2 Implementation Details

Structured output is a crucial component of our framework, as it forms the basis for both report-level and entity-level metrics. We successfully enforced structured output by combining GPT's structured output capabilities and the instructor package [14] with Claude's API. For fine-tuned open-source LLMs, outlines package [24] supports `JSONLogitsProcessor` for structured generation. To validate

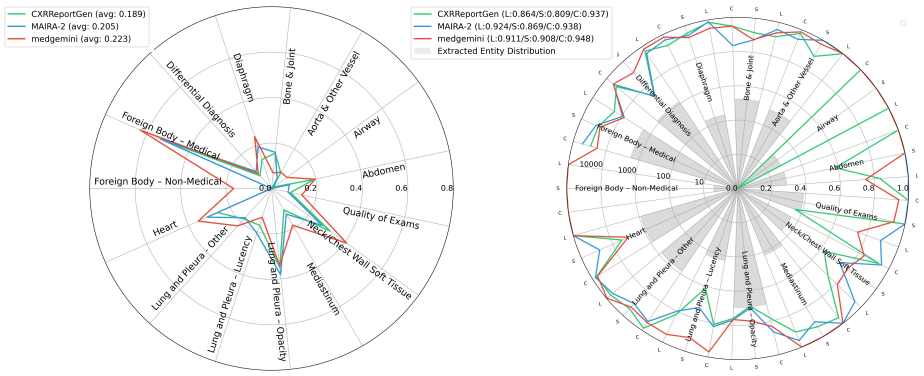

Fig. 4. Entity-level evaluation results comparing state-of-the-art models visualized through spider charts. The **left** chart displays F1 scores for entity presence detection, while the **right** chart shows accuracy scores for attribute prediction: Location (L), Severity (S), and Comparison (C). The grey shaded areas in the background represent the logarithmic distribution of extracted entity frequencies by LLM, providing context for the relative prevalence of each entity type. In the legend, we report macro-averaged scores for each metric, calculated by taking the mean across all 122 entities. Absence of a plot point indicates insufficient data to evaluate the metric.

structured output, we resolved entity duplicates using majority voting or, in the case of ties, by retaining the first occurrence.

To further enable accessibility and transparency of our framework, we fine-tuned Llama-3.1-8B-Instruct and DeepSeek-R1-Distill-Qwen-32B using a dataset generated by OpenAI's o3-mini. The training data consists of 3,019 examples from the training split of MIMIC-CXR, with generated reports by MAIRA-2 [1].

3.3 Entity-Level Evaluation on Report Generation Model

Entity-Level Performance Analysis Figure 4 presents entity-level evaluation results of 3 SOTA report generation models: CXRReportGen [18], MAIRA-2 [1], and Med-Gemini [22]. For visual comparison, we used spider charts to capture four key aspects: presence (F_1 score), location (accuracy), severity (accuracy), and comparison (accuracy). To enhance visual interpretability, we plot performance for 15 entities according to the hierarchical entity structure shown in Figure 3, rather than displaying individual results for all 122 entities. In evaluating attribute prediction (location, severity, and comparison), we count only the true positive (`TP`) cases from the presence detection step, thereby avoiding double penalization.

Our framework's entity-level evaluation enables detailed comparative analysis of model performance, revealing both strengths and limitations of each report generation system. For instance, while MedGemini demonstrates superior overall performance across most categories, it exhibits notably lower performance on Bone & Joint entities compared to other models. Such granular per-

Table 1. Comparison of evaluation metrics on ReXVal dataset showing tau-b scores (*: fine-tuned models).

Report Evaluation Metric	LLM	Tau-b value
RadCliQ [26]	-	0.615
RaTEScore [30]	BERT* [7], BioLORD* [21]	0.527
GREEN [19]	GPT-4	0.640
FineRadScore	o3-mini	0.692 ± 0.028
FineRadScore [8]	Claude-3 Opus	0.738
SPEC-CXR	GPT-4o-mini	0.597 ± 0.035
	GPT-4o	0.656 ± 0.014
	o1	0.758 ± 0.006
	o3-mini	0.734 ± 0.004
	Claude-3.5 Haiku	0.619 ± 0.014
	Claude-3.5 Sonnet	$\mathbf{0.777} \pm 0.004$
	Llama-3.1-8B-Instruct*	0.721 ± 0.008
	DeepSeek-R1-Distill-Qwen-32B*	0.715 ± 0.009

formance analysis facilitates targeted improvements in model development and helps identify domain-specific limitations that might be overlooked by aggregate metrics alone. Additionally, our entity-level evaluation revealed that state-of-the-art models still struggle with many clinically important entities. Even the best-performing model achieved an F1 score below 0.2 for presence detection in 9 out of 15 grouped entities, despite demonstrating relatively high performance predicting location and severity across all entities. These results highlight the substantial challenges that remain in developing a clinically adaptable report generation model.

Validation of Analysis Results To better validate our entity-level evaluation, we conducted a manual assessment of 50 randomly sampled reference-candidate report pairs and their corresponding evaluation results. We measured the accuracy of our framework by comparing its automated assessments against manual annotations, achieving high accuracy rates across all metrics: presence (91.8%), location (87.2%), severity (84.5%), and comparison (75.3%). These results validate the reliability of our entity-level evaluation, though relatively lower accuracy for comparison attributes remains an area for improvement. Manual review showed that comparison statements are harder to evaluate due to language ambiguity. The LLM makes two main errors: First, when a report starts with "compared with 2024-06-23 CXR," it incorrectly treats all following statements as comparisons, even those standing independently. Second, it misses subtle comparison words like "improved" or "worsened." Despite these challenges, we included comparison results because they are important in radiology evaluation. We believe sophisticated prompt tuning can mitigate the issue.

Table 2. Comparison of different entity sets and their impact on human alignment.

Entity Set	Full Coverage (%)	Tau-b Value	# Entities
CheXpert [9]	56.8	0.676 ± 0.016	14
UMLS (PadChest) [4]	73.7	0.694 ± 0.032	210
Our Entity Set	81.9	0.734 ± 0.004	122

3.4 Alignment with Human Evaluation

We evaluated our framework's report-level score alignment with human judgment using the ReXVal dataset. We calculated the Kendall's tau-b correlation coefficient (tau-b) between the average total error count from six radiologists and our report-level score. For LLM-based metrics, we reported means and standard deviations across five inference runs to account for output variability.

Table 1 compares different report evaluation metrics on the ReXVal dataset using tau-b values to measure alignment with human judgments. Our proposed metric, SPEC-CXR, demonstrated strong performance across different LLMs, with the highest score achieved by SPEC-CXR (Claude-3.5 Sonnet) at 0.777. Metrics like GREEN and FineRadScore also rely on LLMs to identify errors in report pairs, similar to our approach. However, they differ in how they assign final scores, where the clinical significance of errors are assigned solely on LLM judgments. In contrast, SPEC-CXR's scoring method appears to better capture how humans compare reports, leading to a noticeable performance gap. Additionally, fine-tuned open-source models demonstrated strong results, offering a more practical and accessible alternative to proprietary models like GPT-4o.

3.5 Entity Set Analysis and Validation

Assessment of Report Coverage To evaluate the adequacy of our entity set in capturing report contents, we analyzed its coverage on the MIMIC-CXR test set. We prompted LLM to extract entities (Figure 3) from the report, and the LLM categorizes whether each sentence is entailed by the extracted entities or not. We then measured the percentage of reports in which all sentences were fully entailed by different entity sets. As shown in the second column of table 2, our entity set achieved full coverage of 81.9%. This indicates that our entity set can fully extract the mentioned abnormal findings and differential diagnoses in 81.9% of reports. The results show superior ability of our entity set to capture findings and differential diagnoses more precisely than existing alternatives.

Impact on Alignment with Human Evaluation We investigated how different entity sets affect human alignment of report-level scores on ReXVal by comparing tau-b values across various entity sets. Third column of the Table 2 shows the comparison between CheXpert's 14 findings, UMLS, and our entity set. The results are obtained from o3-mini. Despite UMLS containing more entities, it showed lower human alignment, demonstrating that deliberate entity set definition is more important than simply including more findings.

4 Conclusion

In this work, we introduced SPEC-CXR, a comprehensive evaluation framework for automated chest X-ray report generation, addressing the critical need for safety-centered performance assessment. By integrating report-level error analysis with entity-level evaluation using an LLM, our approach provides granular insights into model performance across diverse clinical entities. Experimental results demonstrate strong alignment between SPEC-CXR and human expert assessments, validating our framework's reliability. Moreover, our analysis highlights substantial limitations in current state-of-the-art CXR report generation models, emphasizing the necessity for more rigorous, clinically meaningful evaluation metrics. Despite these contributions, our work has limitations that present opportunities for future research. Validation should be extended with larger, more diverse datasets, and incorporating a weighting mechanism could better reflect the clinical importance of findings so that not all errors are considered equal. We believe SPEC-CXR guides the development of safer AI-driven radiology reporting systems by identifying areas for improvement and providing a structured framework for benchmarking model performance.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., Ilse, M., Pérez-García, F., Salvatelli, V., Sharma, H., et al.: Maira-2: Grounded radiology report generation. arXiv preprint arXiv:2406.04449 (2024)
2. Bodenreider, O.: The unified medical language system (umls): Integrating biomedical terminology. Nucleic Acids Research **32**(Database issue), D267–D270 (January 2004). https://doi.org/10.1093/nar/gkh061
3. Bu, S., Li, T., Yang, Y., Dai, Z.: Instance-level expert knowledge and aggregate discriminative attention for radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14194–14204 (2024)
4. Bustos, A., Pertusa, A., Salinas, J.M., de la Iglesia-Vayá, M.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis **66**, 101797 (2020)
5. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056 (2020)
6. Colvin, S., Jolibois, E., Ramezani, H., Garcia Badaracco, A., Dorsey, T., Montague, D., Matveenko, S., Trylesinski, M., Runkle, S., Hewitt, D., Hall, A., Plot, V.: Pydantic (Jan 2025), https://github.com/pydantic/pydantic
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)

8. 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)
9. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)
10. 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)
11. Jiang, Y., Chen, C., Wang, S., Li, F., Tang, Z., Mervak, B.M., Chelala, L., Straus, C.M., Chahine, R., Armato III, S.G., et al.: Clear: A clinically-grounded tabular framework for radiology report evaluation. arXiv preprint arXiv:2505.16325 (2025)
12. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
13. Li, Y., Wang, Z., Liu, Y., Wang, L., Liu, L., Zhou, L.: Kargen: Knowledge-enhanced automated radiology report generation using large language models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 382–392. Springer (2024)
14. Liu, J.: Instructor: Structured outputs for llms (2024), `https://github.com/instructor-ai/instructor`
15. Liu, Y., Li, Y., Wang, Z., Liang, X., Liu, L., Wang, L., Zhou, L.: Er2score: Llm-based explainable and customizable metric for assessing radiology reports with reward-control loss. arXiv preprint arXiv:2411.17301 (2024)
16. Liu, Y., Wang, Z., Li, Y., Liang, X., Liu, L., Wang, L., Zhou, L.: Mrscore: Evaluating radiology report generation with llm-based reward system. arXiv preprint arXiv:2404.17778 (2024)
17. Liu, Z., Zhu, Z., Zheng, S., Zhao, Y., He, K., Zhao, Y.: From observation to concept: A flexible multi-view paradigm for medical report generation. IEEE Transactions on Multimedia (2023)
18. Microsoft: Cxrreportgen: Grounded report generation model for chest x-rays. `https://ai.azure.com/catalog/models/CxrReportGen` (2025)
19. Ostmeier, S., Xu, J., Chen, Z., Varma, M., Blankemeier, L., Bluethgen, C., Michalson, A.E., Moseley, M., Langlotz, C., Chaudhari, A.S., et al.: Green: Generative radiology report evaluation and error notation. arXiv preprint arXiv:2405.03595 (2024)
20. 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)
21. Piran, Z., Cohen, N., Hoshen, Y., Nitzan, M.: Disentanglement of single-cell data with biolord. Nature Biotechnology **42**(11), 1678–1683 (2024)
22. Saab, K., Tu, T., Weng, W.H., Tanno, R., Stutz, D., Wulczyn, E., Zhang, F., Strother, T., Park, C., Vedadi, E., et al.: Capabilities of gemini models in medicine. arXiv preprint arXiv:2404.18416 (2024)
23. Tagawa, Y., Momoki, Y., Nakano, N., Ozaki, R., Taniguchi, M., Hori, M., Tomiyama, N.: Finding-centric structuring of japanese radiology reports and analysis of performance gaps for multiple facilities. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational

Linguistics: Human Language Technologies (Volume 3: Industry Track). pp. 70–85 (2025)
24. Willard, B.T., Louf, R.: Efficient guided generation for llms. arXiv preprint arXiv:2307.09702 (2023)
25. Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., Wang, Y., Chen, E.: Large language models for generative information extraction: A survey. Frontiers of Computer Science **18**(6), 186357 (2024)
26. 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)
27. 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)
28. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)
29. Zhang, X., Zhou, H.Y., Yang, X., Banerjee, O., Acosta, J.N., Miller, J., Huang, O., Rajpurkar, P.: Rexrank: A public leaderboard for ai-powered radiology report generation. arXiv preprint arXiv:2411.15122 (2024)
30. 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)
31. Zhou, H.Y., Adithan, S., Acosta, J.N., Topol, E.J., Rajpurkar, P.: A generalist learner for multifaceted medical image interpretation. arXiv preprint arXiv:2405.07988 (2024)