Clinically Grounded Agent-based Report Evaluation: An Interpretable Metric for Radiology Report Generation

Radhika Dua[1,2]*, **Young Joon (Fred) Kwon**[4], **Siddhant Dogra**[4‡†], **Daniel Freedman**[4‡], **Diana Ruan**[4‡], **Motaz Nashawaty**[4‡], **Danielle Rigau**[4‡], **Daniel Alexander Alber**[2,5‡], **Kang Zhang**[6,7,8], **Kyunghyun Cho**[1,3], **Eric Karl Oermann**[1,2,4]
[1]Center for Data Science, New York University,
[2]Department of Neurosurgery, NYU Langone Health,
[3]Prescient Design, Genentech,
[4]Department of Radiology, NYU Langone Health,
[5]NYU Grossman School of Medicine, NYU Langone Health
[6]National Clinical Eye Research Center, Eye Hospital, Wenzhou Medical University,
[7]Institute for Clinical Data Science, Wenzhou Medical University,
[8]Institute for AI in Medicine and Faculty of Medicine, Macau University of Science and Technology

Abstract

Radiological imaging is central to diagnosis, treatment planning, and clinical decision-making. Vision-language foundation models have spurred interest in automated radiology report generation (RRG), but safe deployment requires reliable clinical evaluation of generated reports. Existing metrics often rely on surface-level similarity and/or behave as black boxes, lacking interpretability. We introduce **ICARE**(**I**nterpretable and **C**linically-grounded **A**gent-based **R**eport **E**valuation), an interpretable evaluation framework leveraging large language model agents and dynamic multiple-choice question answering (MCQA). Two agents, each with either the ground-truth or generated report, generate clinically meaningful questions and quiz each other. Agreement on answers captures preservation and consistency of findings, serving as interpretable proxies for clinical precision and recall. By linking scores to question–answer pairs, **ICARE** enables transparent, and interpretable assessment. Clinician studies show **ICARE** aligns significantly more with expert judgment than prior metrics, while model comparisons reveal interpretable error patterns.

1 Introduction

Radiology reports are essential for accurate diagnosis, treatment planning, and communication among clinical teams. Traditionally authored by radiologists after interpreting imaging studies such as chest X-rays or CT scans, report writing is time-intensive and demands clinical expertise. As imaging volumes increase and radiologist shortages persist, healthcare systems face pressure that can result in delays and raise the risk of diagnostic errors. In response to these challenges, Automated radiology report generation (RRG) systems have emerged as a promising solution to support clinical workflows.

*Corresponding author. Email: rd3571@nyu.edu
†Siddhant is a part-time employee of a2z Radiology AI and holds stock equity in the company.
‡ These authors contributed equally to this work.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

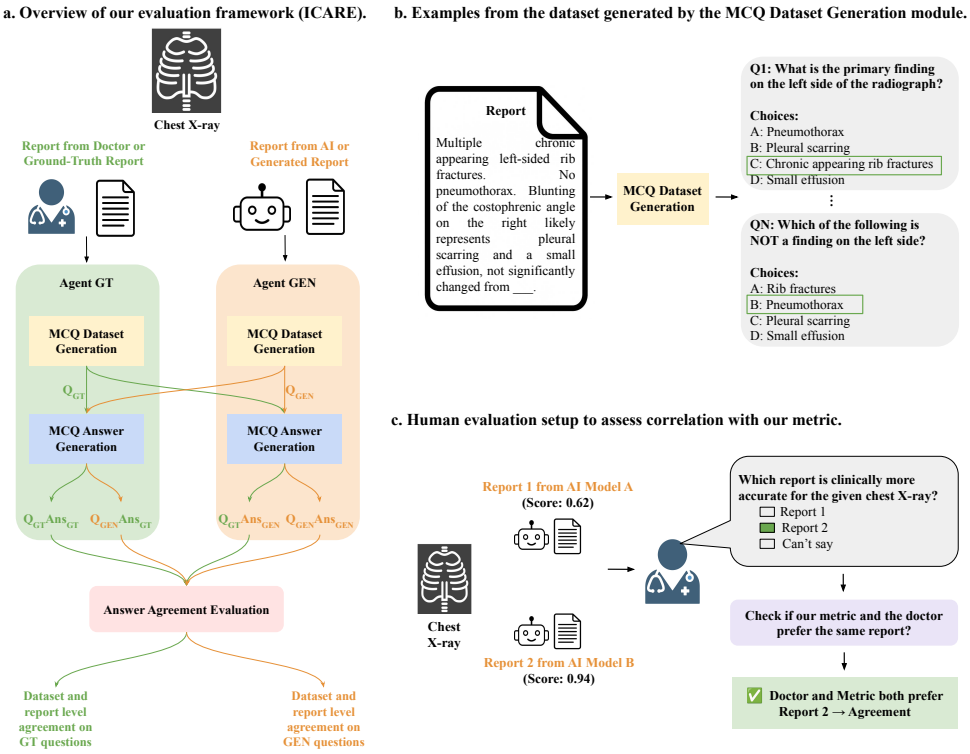


Figure 1: **Overview of our evaluation framework and human validation process. (a) ICARE**: **I**nterpretable and **C**linically-grounded **A**gent-based **R**eport **E**valuation. Two report-aware agents, AgentGT (ground truth) and AgentGEN (generated), independently generate and answer clinically meaningful multiple-choice questions based solely on their respective input reports. The resulting answers are compared through an external agreement module to assess clinical similarity. Agreement on ground-truth questions estimates precision(**ICARE-GT**), while agreement on generated-report questions estimates recall(**ICARE-GEN**). **(b)** Examples of MCQs generated by the framework, capturing diverse findings such as pleural effusion, rib fractures, and cardiomegaly. **(c)** Human evaluation setup to assess alignment between our metric and expert judgment.

Recent methods cover a broad range of modeling approaches. These include vision–language models like Flamingo-CXR [19], CNN-LSTM architectures with attention [17], and knowledge-enhanced models that incorporate structured medical information [11]. More recent systems such as MAIRA-1 [8], MAIRA-2 [2], LLAVA-Rad [3], and Radialog [15] integrate domain-specific vision encoders with large language models. Other approaches like MedPaLM-M [16] and MedVersa [23] highlight the importance of scaling, instruction tuning, and factuality evaluation. These models aim to reduce the clinical burden, improve clarity, and increase the scalability of radiology services.

Before such systems can be deployed safely in practice to assist radiologists, it is necessary to rigorously evaluate whether the generated reports are comparable to expert-written reports. This raises a fundamental question: ***"what are the essential criteria that an evaluation metric must satisfy to be clinically useful?"*** We argue that three properties are necessary. The first is **semantic understanding**, where the metric should determine if two reports express the same clinical information regardless of phrasing. The second is **interpretability**, where users should be able to trace the score back to specific clinical content. The third is **scalability**, where the method must be efficient and usable across large datasets.

Manual review by medical experts is reliable but not scalable. This has led to a growing interest in automated metrics. While many existing metrics are scalable, they often fall short in semantic understanding or interpretability. BLEU [14], ROUGE [12], and BERTScore [21] focus on surface-level similarity or embeddings without clinical context. Metrics like F1-CheXpert [9], SembScore [18], and F1-RadGraph [20] rely on structured labels but offer limited transparency. RadCliQ [20] aggregates multiple scores without clarity. More recent methods such as GREEN [13], FineRadScore [7], RaTEScore [22], G-Rad [4], and RadFact [2] use large language models but still lack interpretability in how scores are derived.

To this end, we introduce **ICARE** (**I**nterpretable and **C**linically-grounded **A**gent-based **R**eport **E**valuation), a clinically grounded evaluation framework for radiology report generation. Our method uses two agents, one reading the ground-truth report and the other reading the generated report. Each agent uses a language model to generate multiple-choice questions based on its input report. After filtering generic questions, each agent answer both sets of questions. Agreement on questions written from the ground-truth report reflects **precision**, which measures whether the generated report preserves clinically important information. Agreement on questions written from the generated report reflects **recall**, which measures whether additional information is clinically consistent with the ground-truth. The final scores are derived from these answer agreements and are linked to specific clinical content. This framework satisfies the three necessary criteria for clinical evaluation. It captures semantic understanding by checking whether the two reports lead to the same clinical conclusions. It is interpretable, since the similarity score is based on clear question–answer pairs. It is also scalable, because the entire pipeline is automated and can be applied to large datasets.

We validate **ICARE** on multiple report generation models and conduct human evaluation studies with board-certified clinicians. These studies show that **ICARE** correlates more strongly with expert preferences compared to prior metrics, demonstrating its utility as a clinically meaningful evaluation method. We also cluster the questions semantically to analyze common failure modes such as missing or hallucinated findings. This reveals patterns in model behavior and provides clinical insight.

To sum up, our contributions include:

- We introduce **ICARE**, a clinically grounded evaluation framework that uses dual report-aware agents to generate and answer multiple-choice clinical questions. The resulting agreement-based precision and recall scores are scalable, semantically grounded, and interpretable, as they are tied directly to explicit clinical questions and answers.
- Through human evaluation studies with board-certified clinicians, we show that **ICARE** aligns more closely with clinician preferences than prior metrics. This supports its utility for evaluating the clinical fidelity of report generation systems.
- We perform semantic clustering of generated questions to analyze model behavior and reveal common failure modes such as omissions and hallucinations. This highlights how **ICARE** provides interpretable insights into the limitations of current report generation models.

2 A clinically grounded, interpretable framework for evaluating radiology report generation

We propose **ICARE**, an evaluation framework composed of two report-aware agents designed to assess the clinical similarity between a generated radiology report and its ground-truth counterpart. The evaluation consists of three stages: MCQ Dataset Generation within each agent, MCQ Answer Generation within each agent, and Answer Agreement Evaluation across agents (Figure 1).

2.1 Step 1: MCQ Dataset Generation (within each agent)

We generate multiple-choice questions (MCQs) independently for the ground-truth and generated reports, with each processed by a dedicated agent:

- Agent_{GT} receives the ground-truth report R_{GT}
- $\text{Agent}_{\text{GEN}}$ receives the generated report R_{GEN}

Using a large language model (LLAMA 3.1 70B model [6]), each agent generates n multiple-choice questions (MCQs) targeting clinical content. The questions are designed to probe various clinical aspects, such as the location, severity, or presence of findings.

Each MCQ consists of:

- A question prompt Q related to the report content,
- Four answer choices $\{a^1, a^2, a^3, a^4\}$,
- A correct answer $a^* \in \{a^1, a^2, a^3, a^4\}$.

Formally, the questions are:

$$Q_{\text{GT}} = \{Q_{\text{GT},i}\}_{i=1}^{n},$$
$$Q_{\text{GEN}} = \{Q_{\text{GEN},j}\}_{j=1}^{n}$$

Generating questions from both reports ensures the evaluation captures unique and overlapping information, identifying both omissions and hallucinations.

Filtering clinically meaningful questions. To retain only report-dependent questions, we use LLAMA 3.1 to answer each question both with and without the report. Let $P_{\text{with}}(Q_k, R)$ and $P_{\text{without}}(Q_k)$ denote the accuracy of answering question Q_k with and without access to report R, respectively. We retain questions where $P_{\text{with}}(Q_k, R) = 1$ and $P_{\text{without}}(Q_k) = 0$, ensuring that only clinically grounded questions requiring the report are included.

Formally, the filtered sets of questions are defined as:

$$Q_{\text{filtered,GT}} = \{Q_{\text{GT},i} \mid P_{\text{with}} = 1 \text{ and } P_{\text{without}} = 0\}$$
$$Q_{\text{filtered,GEN}} = \{Q_{\text{GEN},j} \mid P_{\text{with}} = 1 \text{ and } P_{\text{without}} = 0\}$$

These sets form the final MCQ datasets. They contain diverse, clinically specific questions such as location and severity of findings (Figure 1b).

Bias mitigation. To prevent positional bias in answer options, we randomly shuffle the choices $\{a^1, a^2, a^3, a^4\}$ both before and after filtering. This ensures a uniform distribution across correct answer positions.

2.2 Step 2: MCQ Answer Generation (within each agent)

Following MCQ dataset generation, each agent independently answers both filtered question sets using its respective report. This results in four answer sets:

- $A_{\text{GT}}(Q_{\text{filtered,GT}})$, $A_{\text{GT}}(Q_{\text{filtered,GEN}})$: answers by Agent_{GT} using R_{GT}
- $A_{\text{GEN}}(Q_{\text{filtered,GT}})$, $A_{\text{GEN}}(Q_{\text{filtered,GEN}})$: answers by $\text{Agent}_{\text{GEN}}$ using R_{GEN}

Each agent answers questions using only its assigned report, ensuring independent operation and that all answers reflect only the information present in that report.

2.3 Step 3: Answer Agreement Evaluation (across agents)

We evaluate answer agreement to quantify clinical similarity between reports.

Report-level agreement. We compute agreement scores separately for each report, providing a fine-grained, interpretable view of performance. For a given report r, the report-level agreement scores are defined as:

$$S_{\text{GT},r} = \frac{1}{|Q_{\text{filtered,GT},r}|} \sum_{Q_k} \mathbb{I}(A_{\text{GT}}(Q_k) = A_{\text{GEN}}(Q_k)),$$
$$S_{\text{GEN},r} = \frac{1}{|Q_{\text{filtered,GEN},r}|} \sum_{Q_k} \mathbb{I}(A_{\text{GT}}(Q_k) = A_{\text{GEN}}(Q_k)),$$

where $Q_{\text{filtered,GT},r}$ and $Q_{\text{filtered,GEN},r}$ denote the subsets of questions associated with report r. $\mathbb{I}(\cdot)$ is the indicator function that returns 1 if both agents provide the same answer and 0 otherwise. These report-level scores provide a similarity score based on the number of questions that are answered in agreement for every report.

Dataset-level agreement. To compute overall similarity across all reports, we aggregate agreement over the entire dataset. For questions derived from the ground-truth and generated reports respectively, the agreement scores are defined as:

$$S_{\text{GT}} = \frac{1}{|Q_{\text{filtered,GT}}|} \sum_{Q_k \in Q_{\text{filtered,GT}}} \mathbb{I}(A_{\text{GT}}(Q_k) = A_{\text{GEN}}(Q_k)),$$

$$S_{\text{GEN}} = \frac{1}{|Q_{\text{filtered,GEN}}|} \sum_{Q_k \in Q_{\text{filtered,GEN}}} \mathbb{I}(A_{\text{GT}}(Q_k) = A_{\text{GEN}}(Q_k)).$$

These scores provide a detailed quantitative measure of the similarity between R_{GT} and R_{GEN}. The per-report **ICARE** scores, $S_{\text{GT},r}$ and $S_{\text{GEN},r}$, allow us to analyze similarity on a report-by-report basis, while the dataset-level **ICARE** scores, S_{GT} and S_{GEN}, capture aggregate agreement across all questions.

Interpretation of agreement scores.

- **ICARE**-GT reflects the degree to which clinically important information is preserved in the generated report (analogous to precision).
- **ICARE**-GEN reflects the degree to which additional information introduced in the generated report is consistent with the ground-truth (analogous to recall).

2.4 Summary

ICARE is a clinically grounded and interpretable evaluation framework that captures semantic similarity through structured question answering and answer agreement. It links evaluation scores to specific question–answer pairs, enabling interpretable, report-level analysis. By comparing answers to questions derived from both ground-truth and generated reports, **ICARE** enables precise and scalable evaluation of clinical fidelity.

3 Experiment and Results

3.1 Experimental Setup

We evaluate our framework on the IU X-ray dataset[5] using three pretrained report generation models: CheXpertPlus trained on MIMIC[1][‡] , CheXpertPlus trained on both CheXpertPlus and MIMIC[1],[§] , and MAIRA2[2].[¶] To ensure fair comparison, we recompute all baseline metrics. BLEU-2, BERTScore, SembScore, RadGraph, and 1/RadCliqQ-v1 are evaluated using a consolidated codebase,[‖] , and GREEN using its official repository.[**]Although our method is broadly applicable, we focus on chest X-rays due to their clinical importance in cardiopulmonary diagnosis and widespread availability.

3.2 Alignment of Agreement Scores with Human Judgment

Clinician study design. To assess the clinical validity of our MCQA-based agreement scores, we conducted a human study with six board-certified clinicians. We evaluated 154 samples, with each sample independently reviewed by three clinicians. As shown in Figure. 1 (c), each clinician was shown a chest X-ray and two corresponding AI-generated reports. They were instructed to select the report that more accurately described the X-ray based on clinical content only, ignoring style or formatting. If both reports were equally accurate or the difference was negligible, they could select "Can't say."

Does our metric align with clinician judgments?. Figure. 2 (a) shows clinician preferences across report pairs, grouped by **ICARE** score gap (the difference in **ICARE**-AVG between the two reports), using three plots: Indecision Rate (left), Alignment Rate (center), and Misalignment Rate (right). Each point represents a score gap bin, with point size reflecting the number of samples. When

[‡]https://huggingface.co/IAMJB/mimic-cxr-findings-baseline

[§]https://huggingface.co/IAMJB/chexpert-mimic-cxr-findings-baseline

[¶]https://huggingface.co/microsoft/maira-2

[‖]https://github.com/rajpurkarlab/CXR-Report-Metric

[**]https://github.com/ostmeier/green

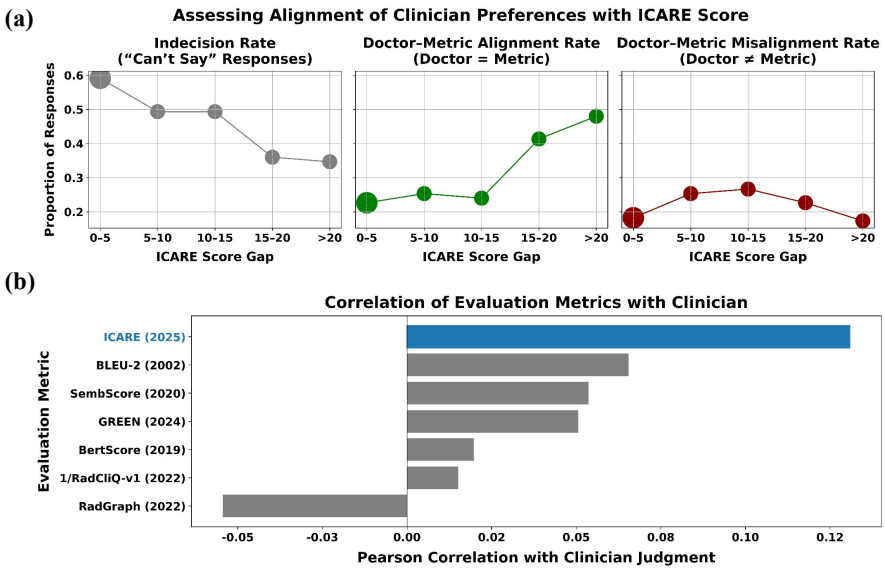


Figure 2: **Alignment of evaluation scores with expert preference. (a)** Clinician preferences across report pairs, shown in three plots: Indecision Rate (left), Alignment Rate (center), and Misalignment Rate (right), grouped by **ICARE** score gap. When the score gap is small, clinicians often selected "Can't say," indicating uncertainty between similarly scored reports. As the gap increases, "Can't say" responses decrease and alignment rises, showing that both clinicians and the metric are more confident in distinguishing report quality. Misalignment remains low throughout. These trends highlight that the metric aligns with expert judgment and reflects meaningful clinical differences. Dot size reflects the number of samples in each **ICARE** score gap bin. **(b)** Correlation between preferences based on clinician judgement and different automatic evaluation metrics. Our metric **ICARE** shows the strongest correlation, indicating that it most closely captures clinician judgments of report quality across samples.

the score gap is small, the Indecision Rate is high. Clinicians often select "Can't say," indicating difficulty in distinguishing between similarly scored reports. As the score gap increases, "Can't say" responses decrease, and the Alignment Rate rises. This shows that clinicians increasingly prefer the report with the higher **ICARE** score. The Misalignment Rate remains low across all bins, suggesting that strong disagreements between the metric and clinician judgment are rare. Overall, these trends show that **ICARE** captures differences in report quality that matter to clinicians and aligns well with expert judgment.

Does our metric align better than prior metrics?. To more directly compare our metric with prior evaluation methods, we analyzed how well each metric's preferences align with individual clinician responses. For each doctor response on a report pair, we assigned a label of +1 if the doctor preferred Report 1, -1 if they preferred Report 2, and 0 if they selected "Can't say." Similarly, for each metric, we assigned a label of +1 if Report 1 received a higher score, -1 if Report 2 did, and 0 if the scores were equal. These labels are ordinal, capturing both agreement and the severity of disagreement. A perfect match occurs when the human and metric labels are identical. A full mismatch (e.g., +1 vs. -1) reflects strong disagreement, while a partial mismatch (e.g., +1 vs. 0) reflects a weaker conflict where one party is undecided. We computed Pearson correlation between these label vectors across 459 individual responses. As shown in Figure. 2(b), **ICARE** achieves the highest correlation with expert preferences. Despite the proliferation of recent evaluation metrics, most exhibit weak or negative alignment with clinicians, underscoring the need for clinically grounded assessment of evaluation metrics themselves. These findings reinforce that our metric not only provides interpretable clinical signals but also most closely reflects how radiologists judge report quality.

3.3 Quantitative evaluation of our metric on radiology report generation models

Dataset-level results. We evaluated three radiology report generation models using a combination of established automatic metrics and our clinically grounded evaluation framework based on answer agreement. As shown in Figure. 3(a), we compute three **ICARE** scores: **ICARE**-GT, using questions produced from the ground-truth report; **ICARE**-GEN, using questions from the generated report;

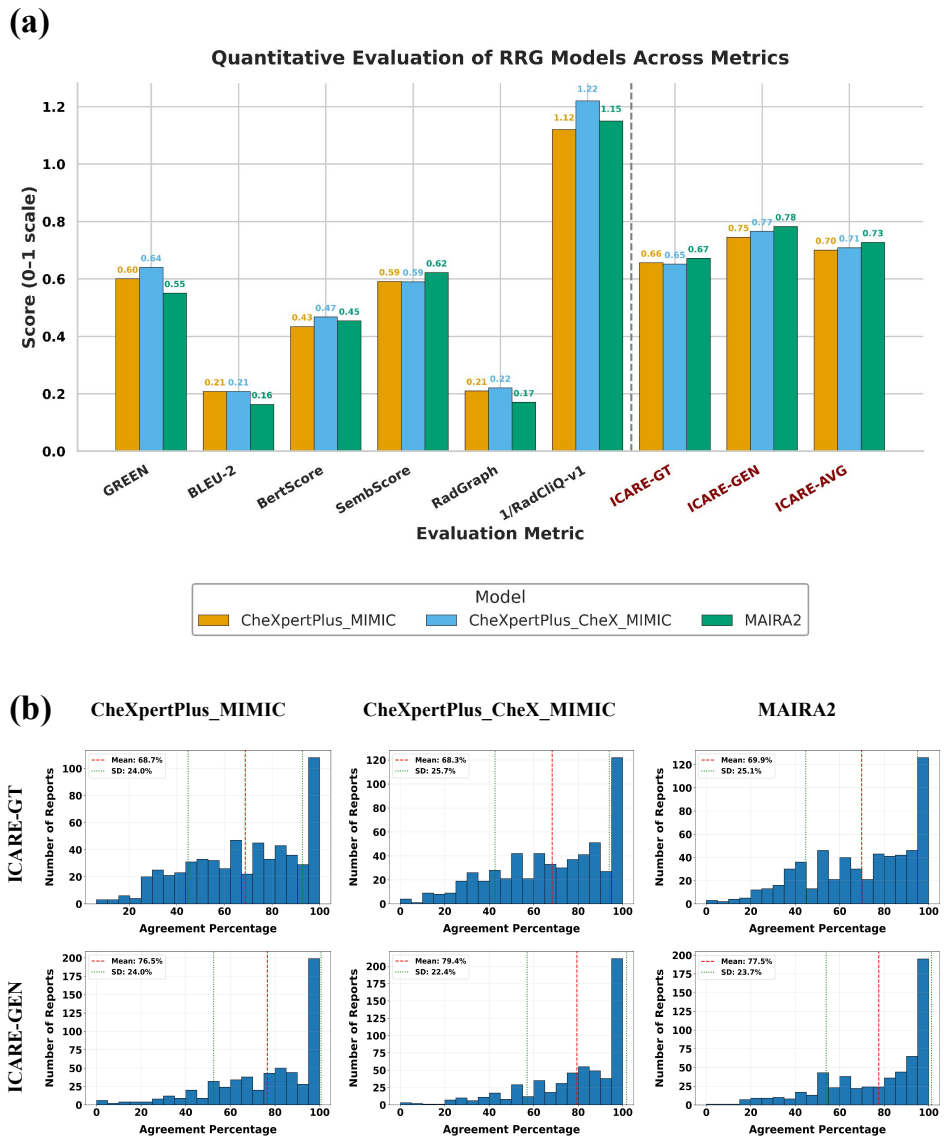
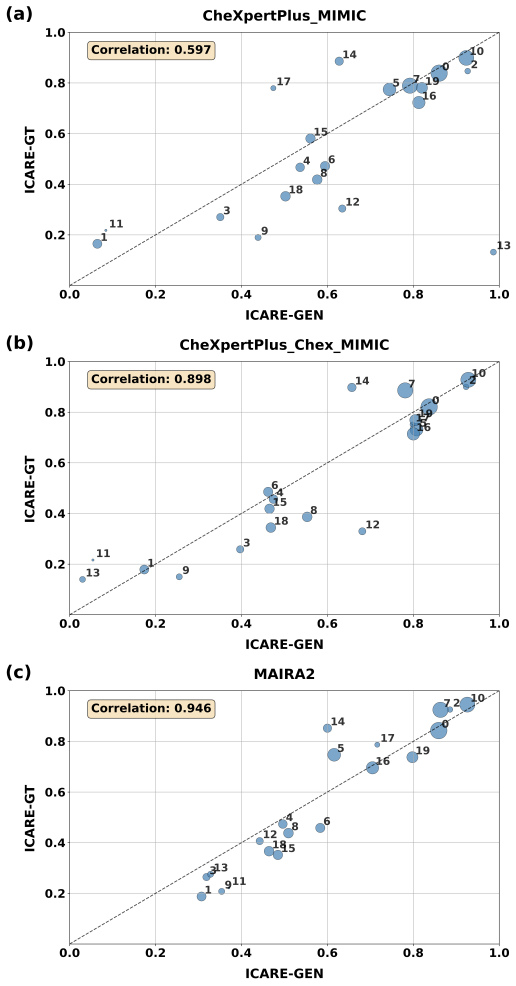


Figure 3: Quantitative results. **(a)** Comparison of model performance across standard metrics and our agreement-based evaluation (**ICARE**). Our metric captures clinically meaningful differences in model behavior by quantifying both the preservation of reference findings (**ICARE**-GT) and the consistency of additional content (**ICARE**-GEN). MAIRA2 achieves the highest agreement across all variants. **(b)** Report-level distribution of **ICARE**-GT and **ICARE**-GEN scores across models and question sources. **ICARE**-GEN scores (agreement on generated-report questions) are generally higher, while **ICARE**-GT scores (agreement on ground-truth questions) show greater variability, reflecting omissions in clinical content in the generated reports.

and their average, **ICARE**-AVG. MAIRA2 achieves the highest values across all three, indicating stronger alignment with clinically meaningful content.

Beyond separating models, our metric reveals clinically interpretable patterns in model behavior. Across all models, **ICARE**-GT scores (agreement on questions derived from the ground-truth report) are consistently lower than **ICARE**-GEN scores (agreement on questions derived from the generated report). This indicates that models are more likely to omit relevant clinical findings than to introduce unsupported content. Traditional metrics generally fail to capture this distinction, whereas our approach enables targeted assessment of both types of errors.

While recent report generation models demonstrate promising performance, our findings reveal that they still fall short of reliably capturing the full clinical content of radiologists-written reports. These limitations are often obscured by existing evaluation metrics, which tend to emphasize surface-level



(d)

Cluster ID	Cluster Name
0	Lung Condition Assessment
1	Thoracic Spine Degeneration
2	Pleural Fluid Presence
3	Diaphragm Position Abnormality
4	Pulmonary Vascular Condition
5	Heart Size Normalcy
6	Lung Opacity and Infiltrates
7	Pleural Effusion Presence
8	Additional Radiologic Findings
9	Aortic Condition
10	Pneumothorax Diagnosis
11	Surgical Clip Location
12	Lung Lobe Abnormality
13	Calcified Granuloma Location
14	Cardiac and Mediastinal Contours
15	Osseous Structure Condition
16	Focal Airspace Opacity
17	Lung Clearness and Normalcy
18	Lung Volume Condition
19	Heart Abnormality Presence

Figure 4: **Cluster-level ICARE score analysis across RRG models.** **(a)**, **(b)**, and **(c)** show scatterplots of **ICARE** scores for semantically grouped clinical question clusters across three models. The x-axis shows **ICARE**-GEN (reflecting the consistency of added content), and the y-axis shows **ICARE**-GT (reflecting preservation of key findings). Point size reflects the number of questions in each cluster. **(d)** provides descriptive cluster names. Clusters below the diagonal indicate omission-dominated errors, while those above indicate hallucinations. Most larger clusters representing common clinical concepts show high agreement and lie near the top right, indicating strong performance on frequently seen findings. A few smaller clusters with low agreement reflect rarer or subtle findings with limited clinical impact. CheXpertPlus variants display several omission-heavy clusters (e.g., clusters 12 and 13), while MAIRA2 shows a compact, balanced distribution near the diagonal, suggesting stronger clinical fidelity and more reliable report generation.

similarity, predefined entity structures, or learned preferences rather than clinical completeness. In contrast, our evaluation provides a structured and interpretable signal that directly reflects clinical fidelity. By linking model performance to specific preserved and omitted findings, our framework offers a complementary and necessary perspective for assessing radiology report generation systems and supports the development of more clinically robust models.

Report-level results. We analyze the distribution of report-level **ICARE** scores (**ICARE**-GT and **ICARE**-GEN) across different model variants, as shown in Figure. 3(b). For each model, we separately evaluate questions generated from ground-truth reports and from generated reports. Our findings reveal that for all models, a large majority of reports achieve high **ICARE** scores, indicating substantial clinical similarity between ground-truth and generated reports.

When using ground-truth reports as reference, the distribution of **ICARE**-GT scores is slightly wider, with a small proportion of reports exhibiting lower agreement. In contrast, when using generated reports as reference, **ICARE**-GEN scores are more concentrated toward higher values, suggesting that generated reports may not fully capture the richer clinical details present in ground-truth reports. This pattern is consistent across all evaluated models.

Among the model variants, MAIRA-2 achieves the highest mean report-level **ICARE**-AVG, followed by CheXpertPlus_CheX_MIMIC and CheXpertPlus_MIMIC. These trends are aligned with the dataset-level results and support the observation that ground-truth reports contain richer clinical content compared to automatically generated reports. Overall, the distribution plots confirm that our MCQA-based evaluation captures clinically meaningful differences at the individual report level.

3.4 Question Categorization and Cluster-level Analysis

To better understand how clinical content influences evaluation outcomes, we performed a cluster-level analysis of **ICARE**-GT and **ICARE**-GEN scores across different categories of clinical questions. We first collected all unique questions generated across all models, and seeds. Using MedCPT[10], a medical-domain language model, we computed embeddings for each question and applied K-means clustering to group them into 20 semantically coherent clusters. From each cluster, five representative questions were sampled, and LLAMA 3.1 70B language model was prompted to generate a descriptive name for each cluster. These cluster names are shown in Figure 4d and reflect a diverse range of clinical concepts.

Each cluster is visualized as a point in a scatterplot (Figure 4), with **ICARE**-GEN scores on the x-axis and **ICARE**-GT scores on the y-axis. Point size indicates the number of questions in the cluster. We observe substantial variation across clusters. Larger clusters tied to common findings such as pleural effusion and cardiomegaly show high agreement on both axes, indicating strong performance on frequently observed content. Smaller clusters representing rarer or subtle findings exhibit lower scores and have less clinical impact due to their limited size.

Systematic error patterns emerge from cluster positions relative to the diagonal. Clusters below the diagonal, where **ICARE**-GT exceeds **ICARE**-GEN, indicate omission dominated errors, suggesting that important information from the ground-truth report is missing in the generated report. Clusters above the diagonal signal hallucinations, where unsupported content appears in the generated report. These omission patterns are prominent in the CheXpertPlus_MIMIC and CheXpertPlus_CheX_MIMIC models, especially in clusters 12 and 13. In contrast, MAIRA2 shows a more compact, balanced distribution near the diagonal, suggesting higher clinical fidelity and more consistent report quality.

Overall, these findings highlight that model performance varies across clinical categories. Cluster-level analysis offers a clinically meaningful and interpretable lens to assess model behavior, highlighting which content is preserved, omitted, or hallucinated, revealing both strengths and common failure modes.

4 Conclusions

We presented **ICARE**, a clinically grounded, interpretable framework for evaluating radiology report generation. Unlike existing metrics that either rely on surface-level similarity or operate as opaque black boxes, **ICARE** leverages dual report-aware agents and multiple-choice question answering to directly capture clinical fidelity. By separately measuring agreement on ground-truth derived questions (**ICARE**-GT, proxy for precision) and generated-report questions (**ICARE**-GEN, proxy for recall), the framework disentangles omissions from hallucinations, offering interpretable insights into model behavior. Our human evaluations show that **ICARE** aligns more closely with expert radiologists' judgments than prior metrics. Beyond providing a single score, **ICARE** highlights category-level trends, revealing that models tend to preserve common findings while omitting rarer or subtler abnormalities. This demonstrates that even the strongest models miss clinically important details that are often obscured by traditional metrics. Therefore, **ICARE** not only evaluates how well models perform but also pinpoints where they fail, providing actionable feedback for improvement. While our experiments focused on chest X-ray reports, the framework generalizes to other imaging modalities and clinical text tasks, enabling post-deployment monitoring even without ground-truth references.

References

[1] Chexpert plus: Hundreds of thousands of aligned radiology texts, images and patients. URL `https://api.semanticscholar.org/CorpusID:270123479`.

[2] Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando P'erez-Garc'ia, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Prasanna Ranjit, Shaury Srivastav, Julia Gong, Fabian Falck, Ozan Oktay, Anja Thieme, Matthew P. Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. Maira-2: Grounded radiology report generation. *ArXiv*, abs/2406.04449, 2024. URL https://api.semanticscholar.org/CorpusID:270357817.

[3] Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Hassan Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu-Hsin Wei, Tristan Naumann, Muhao Chen, Matthew P. Lungren, Serena Yeung-Levy, Curtis P. Langlotz, Sheng Wang, and Hoifung Poon. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature communications*, 16 1:3108, 2024. URL https://api.semanticscholar.org/CorpusID:268379244.

[4] Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Hassan Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chun yue Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu-Hsin Wei, Tristan Naumann, Muhao Chen, Matthew P. Lungren, Serena Yeung-Levy, Curtis P. Langlotz, Sheng Wang, and Hoifung Poon. Training small multimodal models to bridge biomedical competency gap: A case study in radiology imaging. *ArXiv*, abs/2403.08002, 2024. URL https://api.semanticscholar.org/CorpusID:270591813.

[5] Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza M. Rodriguez, Sameer Kiran Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association : JAMIA*, 23 2:304–10, 2015. URL https://api.semanticscholar.org/CorpusID:16941525.

[6] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aur'elien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cris tian Cantón Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab A. AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriele Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guanglong Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Ju-Qing Jia, Kalyan Vasuden Alwala, K. Upasani, Kate Plawiak, Keqian Li, Ken-591 neth Heafield, Kevin R. Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuen ley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melissa Hall Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Niko lay Bashlykov, Nikolay Bogoychev, Niladri S. Chatterji, Olivier Duchenne, Onur cCelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasić, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Ro main Sauvestre, Ron nie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sa hana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Chandra Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Vir ginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whit ney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yiqian Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zhengxu Yan, Zhengxing Chen, Zoe Papakipos, Aaditya K. Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adi Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau

James, Ben Maurer, Benjamin Leonhardi, Po-Yao (Bernie) Huang, Beth Loyd, Beto de Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Da mon Civin, Dana Beaty, Daniel Kreymer, Shang-Wen Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzm'an, Frank J. Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory G. Sizov, Guangyi Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Han Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kaixing(Kai) Wu, U KamHou, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, A Lavender, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navy ata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pe dro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollár, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sung-Bae Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Andrei Poenaru, Vlad T. Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xia Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models. *ArXiv*, abs/2407.21783, 2024. URL `https://api.semanticscholar.org/CorpusID:271571434`.

[7] Alyssa Huang, Oishi Banerjee, Kay Wu, Eduardo Pontes Reis, and Pranav Rajpurkar. Fineradscore: A radiology report line-by-line evaluation technique generating corrections with severity scores. *ArXiv*, abs/2405.20613, 2024. URL `https://api.semanticscholar.org/CorpusID:270199552`.

[8] Stephanie L. Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Mercy Prasanna Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, Noel Codella, Matthew P. Lungren, Maria Teodora Wetscherek, Ozan Oktay, and Javier Alvarez-Valle. Maira-1: A specialised large multimodal model for radiology report generation. *ArXiv*, abs/2311.13668, 2023. URL `https://api.semanticscholar.org/CorpusID:265445382`.

[9] Jeremy A. Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David Andrew Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, C. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and A. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI Conference on Artificial Intelligence*, 2019. URL `https://api.semanticscholar.org/CorpusID:58981871`.

[10] Qiao Jin, Won Kim, Qingyu Chen, Donald C. Comeau, Lana Yeganova, John Wilbur, and Zhiyong Lu. Biocpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39 11, 2023. URL `https://api.semanticscholar.org/CorpusID:259316759`.

[11] Kaveri Kale, Pushpak Bhattacharyya, and Kshitij Sharad Jadhav. Replace and report: Nlp assisted radiology report generation. In *Annual Meeting of the Association for Computational Linguistics*, 2023. URL https://api.semanticscholar.org/CorpusID:259309063.

[12] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*, 2004. URL https://api.semanticscholar.org/CorpusID:964287.

[13] Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael E. Moseley, Curtis P. Langlotz, Akshay S. Chaudhari, and Jean-Benoit Delbrouck. Green: Generative radiology report evaluation and error notation. *ArXiv*, abs/2405.03595, 2024. URL https://api.semanticscholar.org/CorpusID:269605082.

[14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*, 2002. URL https://api.semanticscholar.org/CorpusID:11080756.

[15] Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Nassir Navab, and Matthias Keicher. Radialog: A large vision-language model for radiology report generation and conversational assistance. *ArXiv*, abs/2311.18681, 2023. URL https://api.semanticscholar.org/CorpusID:265506090.

[16] K. Singhal, Shekoofeh Azizi, Tao Tu, Said Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather J. Cole-Lewis, Stephen J. Pfohl, P A Payne, Martin G. Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, P. A. Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Greg S. Corrado, Yossi Matias, Katherine Hui-Ling Chou, Juraj Gottweis, Nenad Tomašev, Yun Liu, Alvin Rajkomar, Joëlle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620:172 – 180, 2022. URL https://api.semanticscholar.org/CorpusID:255124952.

[17] Mehreen Sirshar, Muhammad Faheem Khalil Paracha, Muhammad Usman Akram, Norah Saleh Alghamdi, S. Zainab Yousuf Zaidi, and Tatheer Fatima. Attention based automated radiology report generation using cnn and lstm. *PLoS ONE*, 17, 2022. URL https://api.semanticscholar.org/CorpusID:245801513.

[18] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, A. Ng, and Matthew P. Lungren. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Conference on Empirical Methods in Natural Language Processing*, 2020. URL https://api.semanticscholar.org/CorpusID:215827807.

[19] Ryutaro Tanno, David G. T. Barrett, Andrew Sellergren, Sumedh Ghaisas, Sumanth Dathathri, Abigail See, Johannes Welbl, Karan Singhal, Shekoofeh Azizi, Tao Tu, Mike Schaekermann, Rhys May, Roy Lee, SiWai Man, Zahra Ahmed, S. Sara Mahdavi, Danielle Belgrave, Vivek Natarajan, Shravya Shetty, Pushmeet Kohli, Po-Sen Huang, Alan Karthikesalingam, and Ira Ktena. Consensus, dissensus and synergy between clinicians and specialist foundation models in radiology report generation. *ArXiv*, abs/2311.18260, 2023. URL https://api.semanticscholar.org/CorpusID:265506498.

[20] Feng Yu, Masahiro Endo, Rayan Krishnan, Ian Pan, Andrew Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, H. H. Lee, Zohreh Hossein Abad, Andrew Y. Ng, C. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4, 2022. URL https://api.semanticscholar.org/CorpusID:251950682.

[21] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675, 2019. URL https://api.semanticscholar.org/CorpusID:127986044.

[22] W. Zhao, C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie. Ratescore: A metric for radiology report generation. In *Conference on Empirical Methods in Natural Language Processing*, 2024. URL https://api.semanticscholar.org/CorpusID:270699665.

[23] Hong-Yu Zhou, Julián Nicolás Acosta, Subathra Adithan, Suvrankar Datta, Eric J. Topol, and Pranav Rajpurkar. Medversa: A generalist foundation model for medical image interpretation. 2024. URL https://api.semanticscholar.org/CorpusID:269756808.