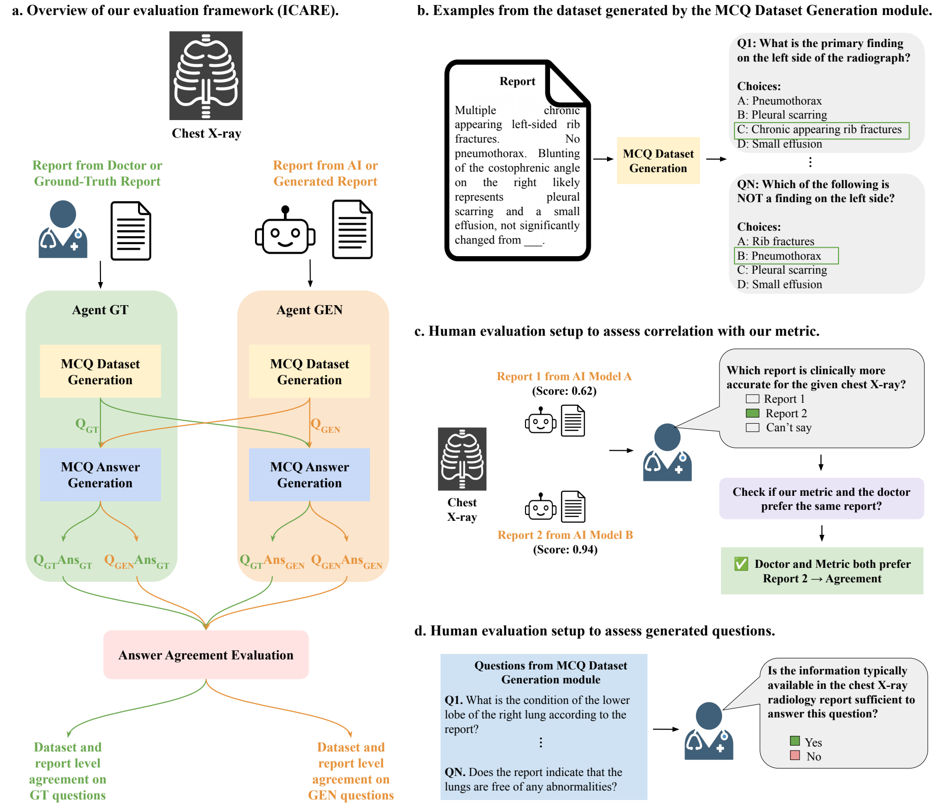
Clinically Grounded Agent-based Report Evaluation: An Interpretable Metric for Radiology Report Generation

Radiological imaging plays a central role in diagnosis, treatment planning, and clinical decision-making. The emergence of vision-language foundation models has generated enormous interest in the use of artificial intelligence to read imaging and generate reports describing imaging findings. However, automated radiology report generation (RRG) suffers from a difficulty in evaluating generated reports for clinical accuracy, which presents a major challenge to safe deployment. Existing metrics often rely on surface-level similarity and/or behave as black boxes and lack clinical interpretability.

To address this challenge, we developed **ICARE**(**I**nterpretable and **C**linically-grounded **A**gent-based **R**eport **E**valuation), an interpretable evaluation framework that uses large language model agents and dynamic multiple-choice question answering (MCQA). Two agents, assigned either the ground-truth or generated report, generate clinically meaningful MCQA and then quiz each other. The answer agreement on these questions reflects whether critical findings are preserved or consistent, acting as interpretable proxies for clinical precision and recall. By linking evaluation scores to specific clinical question–answer pairs, **ICARE** offers transparent and interpretable assessment of report quality while preserving semantic understanding and scalability.

ICARE demonstrates strong alignment with expert radiologists' preferences across report comparisons, outperforming existing metrics such as BLEU, BERTScore, and RadGraph in capturing clinically relevant differences. Extensive clinician validation studies confirm the quality and relevance of generated questions, while controlled perturbation experiments show **ICARE** is sensitive to clinical content and robust across evaluation seeds. Cluster-level analysis further reveals interpretable model error patterns (e.g., omission vs. hallucination of findings), offering insight beyond a single aggregate score.

ICARE is fully automated, scalable to large datasets, adaptable to varied imaging modalities and other medical text generation tasks (e.g., pathology reports or discharge summaries), and facilitates post-deployment monitoring even without ground-truth data. By making evaluation transparent and clinically meaningful, **ICARE** supports the development of safer systems in healthcare and strengthens trust in automated radiology report generation.



a. Overview of our evaluation framework (ICARE).

Chest X-ray
Report from Doctor or Ground-Truth Report
Report from AI or Generated Report
Agent GT
Agent GEN
MCQ Dataset Generation
MCQ Dataset Generation
Q_{GT}
Q_{GEN}
MCQ Answer Generation
MCQ Answer Generation
$Q_{GT}Ans_{GT}$
$Q_{GEN}Ans_{GT}$
$Q_{GT}Ans_{GEN}$
$Q_{GEN}Ans_{GEN}$
Answer Agreement Evaluation
Dataset and report level agreement on GT questions
Dataset and report level agreement on GEN questions

b. Examples from the dataset generated by the MCQ Dataset Generation module.

Multiple chronic appearing left-sided rib fractures. No pneumothorax. Blunting of the costophrenic angle on the right likely represents pleural scarring and a small effusion, not significantly changed from ___.

Q1: What is the primary finding on the left side of the radiograph?

Choices:
A: Pneumothorax
B: Pleural scarring
C: Chronic appearing rib fractures
D: Small effusion

⋮

QN: Which of the following is NOT a finding on the left side?

Choices:
A: Rib fractures
B: Pneumothorax
C: Pleural scarring
D: Small effusion

c. Human evaluation setup to assess correlation with our metric.

Which report is clinically more accurate for the given chest X-ray?
□ Report 1
🟩 Report 2
□ Can't say

Check if our metric and the doctor prefer the same report?

✅ **Doctor and Metric both prefer Report 2 → Agreement**

d. Human evaluation setup to assess generated questions.

Questions from MCQ Dataset Generation module

Q1. What is the condition of the lower lobe of the right lung according to the report?

⋮

QN. Does the report indicate that the lungs are free of any abnormalities?

Is the information typically available in the chest X-ray radiology report sufficient to answer this question?

🟩 Yes
🟥 No

Figure 1: Overview of our evaluation framework and human validation process.