THE UNIVERSITY OF CHICAGO

HUMAN-AI DECISION MAKING WITH CASE STUDIES IN RADIOLOGY

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

BY
CHACHA CHEN

CHICAGO, ILLINOIS
AUGUST 2025

Human-AI Decision Making with Case Studies in Radiology

by

Chacha Chen

Abstract

With the explosive progress of machine learning (ML), especially recent foundation models, these advanced systems are increasingly reshaping our daily workflows across domains. This makes human-centered AI research critically important, as it aims to build AI models to better support human tasks and improve decision-making. This thesis focuses on improving human-AI collaboration through both behavioral study and building more effective AI systems. We begin with a theoretical analysis of the interaction between machine learning models and human decisions, which highlights a key insight: human intuition plays a critical role in effective human-AI collaboration. Using prostate cancer diagnosis with MRI as a real-world test bed, we conducted user studies with domain experts to investigate how advanced, human-level ML models are perceived and used in clinical decision-making. Our findings show that experts are often hesitant to adopt AI tools, and even when they do, they struggle to appropriately rely on AI. Importantly, by applying a theoretical framework of human-AI reliance, we identified actionable strategies that help ensure complementary performance (human+AI performance exceeds either alone). In parallel, we explored multimodal large language models for radiology. Starting with an evaluation of out-of-the-box performance of current LLMs (e.g., GPT-4o, Llama) on chest X-ray reporting, we found that, although impressive in general domains, current LLMs perform poorly on specialized medical tasks. Our analysis reliably identified visual understanding as the primary performance bottleneck. Additionally, we proposed a fine-grained tabular based evaluation method with expert curated high-quality data. This benchmark not only enhances the rigor of current evaluation but also holds promise for guiding future model development. My work contributes to broader efforts in adapting foundation models to high-stakes, domain-specific applications. More broadly, my research contributes to the growing understanding of how AI is evolving from simple tools to sophisticated collaborators in knowledge work and specialized fields.

Thesis Advisor: Chenhao Tan

Acknowledgements

First of all, I would like to thank my advisor, Chenhao Tan, for his invaluable mentorship. His works on human-AI decision making inspired me to join the CHAI lab, and I'm deeply proud of what we've accomplished together. I'm especially grateful for the way he leads by example, showing what it means to be a sharp and curious researcher, a kind mentor, and a frequent source of questionable humor. Chenhao brought every member of CHAI together and built a community where we could all grow and thrive. Through this lab, I formed many meaningful relationships that I'll always cherish.

I also want to thank my thesis committee members: Yuxin Chen, Aritrick Chatterjee, and James Evans, whose works I deeply admire. Their guidance, feedback, and support throughout my thesis defense have been incredibly valuable.

This journey would not have been possible (and so much fun) without my labmates in CHAI and many collaborators: Vivian Lai, Jiamin Yang, Han Liu, Yizhou Tian, Shi Feng, Yiming Zhang, Yuyang Jiang, Karen Zhou, Dang Nyugen, Ziyang Guo, Jessica Hullman, Rosa Yangqiaoyu Zhou, Haokun Liu, Karen Zhou, Chao-Chun Hsu, Xiaoyan Bai, and Julia Mendelsohn, Chenghao Yang, Mourad Heddaya. I especially want to thank Vivian, who has been one of my closest friends and was a tremendous help when I first started working on human-centered AI research. Shi has been a mentor I truly admire, someone I could always turn to with questions and who would respond with thoughtful insights. Jiamin, Han, Yizhou, and Yiming helped me see and appreciate the beauty and energy in so many things around us. I appreciate Karen for the countless messages, shared trips, cat tips, and drinks. I'm grateful to Chao-Chun and Iris for treating me like family. Rosa and Haokun always enjoyed my jokes with questionable humor. Ziyang and Jessica helped me tremendously with my

thesis and taught me a great deal about rigor and precision in research. Because of the nature of my research, I've also had the privilege of working with many physician collaborators. I'm especially grateful to Dr. Benjamin M. Mervak for his enthusiasm and constant support. He is always willing to engage with my project and answers my many questions. I also want to thank Aritrick and Aytekin for their support of the prostate MRI project. And though we've never met in person, I want to acknowledge Dr. Emre Cakmakli, who pushed me to do Duolingo every day and also led by example as a role model in consistency.

I also want to thank the many research mentors and friends who have guided and supported me throughout this journey: Ting-Hao Kenneth Huang, Vera Qingzi Liao, Siddharth Suri, Danica Cao Xiao, Frank Ritter, and Jessie Zhenhui Li. I want to thank my friends and colleagues from Penn State University for your support when I first arrived in the States and gave me the courage to pivot my research direction and pursue what I truly care about. Many of you remain my role models to this day. Ting-Hao Kenneth Huang showed me how compassionate, selfless, and fun an advisor can be. I'm especially grateful to my best friend and mentor, Hua Wei, for always being there encouraging and supporting me, and occasionally giving me that gentle kick when I needed it. I also want to thank Guanjie Zheng, whose advice helped me make many important decisions of my life, and Huaxiu Yao, who rented me a room (with AC at a very fair price) of his big Sunnyville house during my (very hot) summer internship in the Bay Area in 2022.

Last but most importantly, I want to thank my parents for their love, compassion, and support. They've always given me the freedom to explore my own path, while also being there whenever I needed a place to pause and feel grounded. I'm also deeply grateful for my encounter with Zhen Lin, my best friend and my source of bravery and safety. I'm very proud of the sanity home we've built together, with Chichi and Deedee.

Contents

Acknowledgements **4**

List of Figures **10**

List of Tables **16**

1 Introduction **22**

2 Prior Work: Theoretical Insights of Human-AI Collaboration **25**

2.1 A Theoretical Framework of Human Understanding and Machine Explanations 25

2.2 Machine Explanations and Human Intuitions 32

2.3 Towards Effective Explanations for Improving Human Understanding . . . 37

2.4 Conclusion 38

3 Human-AI Decision Making for Prostate Cancer MRI Diagnosis **39**

3.1 Introduction 39

3.2 Related work 42

3.3 Methods 44

3.4 Results 49

3.4.1 Performance of Human vs. AI vs. Human+AI Team 49

3.4.2 Analysis of Human-AI Complementarity 51

3.4.3 Measuring Human Reliance on AI 56
3.5 Conclusion 58

4 Evaluation of Frontier Models in Radiology Report Generation 60
4.1 Introduction 60
4.2 Experimental Setup 62
4.3 Results 66
4.4 Conclusion 72

5 CLEAR: A Clinically-Grounded Tabular Framework for Radiology Report Evaluation 74
5.1 Introduction 74
5.2 CLEAR Framework 77
5.3 CLEAR-Bench: Attribute-Level Expert Alignment Dataset 81
5.4 Experiments 83
5.5 Related Work 86
5.6 Conclusion 88

6 Conclusion 89

References 90

A Prior Work Details 111
A.1 A Summary of Recent Empirical Studies 111
A.2 Characterizing Relationship between Core Functions & Human Understandings 115

B AI Prostate Experiments Details 117

B.1 Model Impementation Details . . . 117
B.2 Statistical Test Details . . . 118
B.2.1 Pairwise Comparisons . . . 119
B.3 Demographics . . . 119
B.4 Exit Survey Results . . . 119
B.5 Fine-grained analysis . . . 120
B.6 Ensemble on Common-50 Cases . . . 121
B.7 More Screenshots on User Interface Design . . . 123

C Radiology Report Generation Evaluation Experiments Details 129

C.1 Model Impementation Details . . . 129
C.2 Data . . . 130
C.3 Evaluation Experiment Results . . . 131
C.3.1 Hypothesis Test . . . 131
C.3.2 GPT-4o Evaluation . . . 132
C.3.3 GPT-4-vision-preview Evaluation . . . 133
C.3.4 Llama-3.2-90B-Vision Evaluation . . . 135
C.3.5 Examples of Generated Reports . . . 135
C.4 Human Reader Study . . . 135
C.5 Prompts . . . 142

D CLEAR Evaluation Framework Details 150

D.1 Open-sourced Artifacts . . . 150
D.2 Data Annotation and Curation . . . 150
D.2.1 Label Structure Refinement . . . 151

D.2.2 Expert-in-the-loop Dataset Curation 153
D.2.3 CLEAR-Bench: Expert Ensemble 154
D.2.4 CLEAR-Bench: Attribute Curation 154
D.3 CLEAR: Implementation Details 155
D.4 Template & Terminology List 156

List of Figures

2-1 Illustration of the three fundamental concepts using a binary classification problem. Task decision boundary (dashed line) defines the ground-truth mapping from inputs to labels. Model decision boundary (solid line) defines the model predictions. Model error (highlighted) represents where the model's predictions are incorrect. 26

2-2 Visualizing the relations between core functions, local variables, and human approximations of them. 29

2-3 Causal diagrams visualizing the relationship between a human's local understanding. With the base diagram at the root, we organize its realizations based on different conditions in a two-level decision tree. Undirected dashed lines represent ambiguous causal links. The bidirectional dashed line in subfigure (d) represents the correlation between $\hat{Y}$ and Y potentially induced by the prediction model. Shaded nodes and their edges represent *show* operations. 31

2-4 (a) E is generated from g. (b1) $\hat{Y}$ and $\hat{Y}^H$ are independent given X. (b2) The utility of E: E can improve human understanding of model decision boundary. (c1) E cannot improve human understanding of task decision boundary and model error without human intuitions. (c2) Combined with human intuitions, E can improve task decision boundary and model error. We use orange lines to highlight the links that lead to positive utility of E. We omit links from X to f^H and z^H for simplicity. 34

3-1 Overview of our experiments with radiologists. In study 1, participant radiologists (N=8) reviewed 75 cases in three steps: initial independent diagnosis, review of AI predictions, and final diagnosis. In study 2, we introduce performance feedback to communicate individual radiologist's performance collected from study 1 before the study. Then they reviewed 100 cases with direct AI assistance without independent diagnosis. 42

3-2 Screenshots of the webapp interface for our human study. (a) Fig. 3-2a presents a user interface for patient case evaluation. An AI lesion prediction is highlighted with a red contour in the T2W sequence. On the right, the user's current prediction is shown as "No Cancer," and they are at the stage of evaluating the AI prediction to make a final diagnosis. (b) Fig. 3-2b shows the user interface of the Annotation Panel. The screenshot shows a current annotation of the user. The user can clear the annotation or add new annotations on the canvas. (c) Fig. 3-2c illustrates an example performance feedback page presented to a user before proceeding to Study 2. The page provides a summary of the total number of cases, including counts of correct and incorrect cases, the number of decision changes influenced by AI advice, and whether those changes were correct or incorrect. It also highlights key performance metrics such as accuracy, sensitivity, and specificity, derived from Study 1. To ensure users review the information carefully, they are required to answer attention check questions. 46

3-3 An example of lesion-level annotation comparing human experts (red contour), AI (yellow), and expert annotation from the dataset (green). In this case, the AI successfully detected a lesion which corresponded to a clinically significant prostate cancer in the dataset; our human radiologist did not identify this lesion, and instead annotated a lesion in the transition zone. 51

3-4 Across all scoring rules, a consistent performance pattern is observed: Human-alone < Human+AI (Study 1) < Human+AI (Study 2) < AI-alone. The addition of AI assistance shows improvement over human-alone performance in both studies, yet all human configurations still underperform the AI alone. Bars indicate the mean score under each scoring rule based on the observed decisions, and error bars represent 95% confidence intervals computed using the t-distribution. 52

3-5 Individual radiologists performance compared with the AI model. The model achieves higher performance than all of the radiologists without AI assistance (blue dots). However, with AI assistance, some individual radiologists outperformed the AI model (red and orange dots that are above the curve). 52

3-6 The human-alone score, AI score, human+AI score, rational benchmark, and mis-reliant rational benchmark on the common 50-case subset under scoring rules with different ratios of FPR to FNR. The rational benchmarks (red distributions) achieve no less accuracy than AI decisions (orange distributions) and human-alone decisions (blue distributions), indicating potential for human-AI complementarity. The mis-reliant benchmark (gray distributions) is improved in study 2 than in study 1, indicating that human-AI teams rely more appropriately on AI decisions in study 2. 53

3-7 Human-alone accuracy, AI accuracy, human+AI accuracy, rational benchmark, and mis-reliant rational benchmark on the common 50-case subset for each participant. 55

3-8 Comparison of Human-AI Decision Alignment and Accuracy. Blue shading indicates frequency of cases for each scenarios; percentages showing diagnostic accuracy for scenario. Accuracy is the highest in the follow-AI group for both studies. 57

4-1 Evaluation overview. In Experiment 1, we evaluate the out-of-box capability of GPT-4V. We further decompose the task into medical image reasoning (Experiment 2) and report synthesis (Experiment 3). 61

4-2 95% Bootstrap confidence interval of example 3 conditions for MIMIC-CXR. When zero falls into the interval, at 95% confidence level, we cannot reject the null hypothesis that GPT-4V labels j-th condition independent of which condition group this study belongs to. 69

5-1 A comparison of existing metrics with CLEAR. Yellow highlights indicate the main evaluation mechanism for each type of metric. Red underlining marks an erroneous term in the candidate report, in contrast to the black underlined term in the ground-truth report, which the designed metric fails to evaluate. 75

5-2 **CLEAR Framework.** Given a pair of ground-truth and candidate reports, we first assesses whether the candidate report can accurately identify a set of medical observations in the **label extraction module**. For each correctly identified positive condition, the **description extraction module** further evaluates the report's ability to describe the condition across five attributes: `first occurrence`, `change`, `severity`, `descriptive location`, and `recommendation`. Finally, the **scoring module** compiles and outputs the evaluation metrics. 75

B-1 Login page. 123

B-2 Consent page. 125

B-3 Toy demonstration example page. 126

B-4 Exit survey for study 1. 127

B-5 Exit survey for study 2. 128

C-1 95% Bootstrap confidence interval of top 6 conditions in our sample for GPT-4-vision-preview. . . . 138

C-2 95% Bootstrap confidence interval of top 5 conditions in our sample for GPT-4o. . . . 138

C-3 Correlations between distributions of Pr(Pos) for different condition groups (GPT-4-vision-preview). . . . 139

C-4 Correlations between distributions of Pr(Pos) for different condition groups (GPT-4o). . . . 139

C-5 Overlap count heatmap within groudtruth condition groups. The overlap between any two groups is relatively small compared to the original group sizes, which relate to the counts on the diagonal, ensuring that this overlap does not compromise the robustness of our results. . . . 139

C-6 Human reader study instructions. . . . 141

D-1 4-type labeling criteria in MIMIC. . . . 151

D-2 Our refined 5-type labeling criteria during expert annotation. . . . 152

D-3 Interface for Label Annotation. . . . 153

D-4 Interface for Attribute Curation. . . . 154

D-5 Instruction Template for Label Annotation Task . . . 157

D-6 Instruction Template for Attribute Curation Task . . . 157

D-7 Prompt 1 . . . 158

D-8 Prompt 2 . . . 159

D-9 Prompt 3 . . . 159

D-10 Prompt 4 . . . 160

D-11 Prompt 5: Additional Notes: location/descriptor/status/note are a list of example key words or phrases for each condition collected from radiologists, such as (e.g., compressive, segmental, focal, terminal, peripheral, etc.). . . . 161

D-12 Prompt 6 163

D-13 Prompt 7 163

D-14 o1-mini prompt 164

List of Tables

2.1 A summary of notations. 29

3.1 Performance comparison between AI, Human, and Human+AI for identifying csPCa from MRI scans. For each metric, the means, 95% confidence intervals, and number of instances are reported. The reported values and instance counts represent averages across eight radiologists. All confidence intervals are derived using bootstrap methods. p-values are calculated using the bootstrap z-test with a significance threshold of $\alpha = 0.05$. 49

3.2 Participants' reliance on AI and appropriate reliance for them in the common 50-case subset under the evaluation of accuracy ($\mu = 0.5$). 55

4.1 An index to prompts used in all of our experiments. 63

4.2 Direct report generation performance comparison. GPT-4V shows a significant performance gap compared to SOTA, and the results are consistent across the five prompting strategies. Open sourced Llama3.2 performs similarly compared with GPT-4V. Examples of generated reports across different prompts can be found in Appendix C.3.5. 65

4.3 Image reasoning performance of GPT-4V on MIMIC-CXR. The model performs poorly in identifying medical conditions from chest X-ray images. 67

4.4 χ^2-test for homogeneity of label distribution across different condition groups. When p-value is smaller than 0.0001, at 0.01% significance level, we can reject the null hypothesis that different groups follow the same label distribution. 69

4.5 Performance in report generation with groundtruth conditions. Although GPT-4V's performance improves significantly, it still underperforms finetuned Llama-2, especially in matching the writing style of groundtruth reports. 69

4.6 Human evaluation of radiology report quality. From the perspective of radiologists, GPT-4V underperforms relative to the finetuned Llama-2, particularly in fine-grained likert scale metrics of diagnostic accuracy, completeness, and clartiy/readability. 71

4.7 Comparison of example reports from human annotation. 71

5.1 An overview of our expert-curated fine-grained attributes in CLEAR. 77

5.2 Evaluation of the label extraction module. CLEAR outperforms existing labelers across all metrics in identifying both positive and negative conditions. Specifically, larger models perform better at capturing positive conditions, while techniques such as 5-shot prompting and supervised fine-tuning significantly improve the detection of negative conditions. 82

5.3 Evaluation of the description extraction module. Expert ratings are averaged across all samples (0 = incorrect, 0.5 = partially correct, 1 = correct). According to radiologists' clinical judgment, CLEAR can accurately extract attribute-level information from free-text reports. Additionally, GPT-4o is consistently preferred over Llama-3.1-8B-Instruct, though Llama performs reasonably well, especially on `descriptive location`, and remains a low-cost, open-source option. 84

5.4 Pearson correlation between CLEAR and expert scores. All of automated metrics generated by CLEAR show strong alignment with expert evaluations. 86

A.1 A summary of recent empirical studies measuring human understanding with machine explanations. The papers are sorted by time, starting form the newest. Note: columns g, z, and f mean model decision boundary, model error, and task decision boundary respectively. ✓(or ✗) means the study measures (or does not measure) the corresponding type of human understanding. 113

B.1 AI model performance. 117

B.2 Performance across different scoring rules and conditions. Values are reported as the mean with 95% confidence intervals (CI). 118

B.3 Statistical comparisons across scoring rules (μ values) for different contrasts. t test and Benjamini–Hochberg adjusted p-values. 118

B.4 Degrees of freedom for corresponding t-tests in Table B.3. 118

B.5 Study 1 fine-grained subgroup performance. 121

B.6 Study 2 fine-grained subgroup performance. 121

B.7 Study 1: Cases where human agreed with AI and decision was kept. 121

B.8 Study 1: Cases where human agreed but AI initially but still changed decision against AI. 122

B.9 Study 1: cases where human disagreed with AI but kept original decision. . 122

B.10 Study 1: cases where human disagreed with AI but followed AI advice. . . 122

B.11 Finegrained analysis for Study 2: (1) When Human disagrees with AI, human are prone to errors (accuracy is lower than 50%); (2) Human is better at identifying AI false positives than identifying false negatives, i.e., humans are better at catching AI's false alarms than its missed cases. 122

B.12 Performance comparison between AI, Human, Human-ensemble, Human+AI, and human+AI ensemble (study 1 and 2) for the common 50-case subset. . 123

B.13 Rational agent's decision rule on the common 50-case subset under the evaluation of different scoring rules. 124

C.1 Summary of actual sample size across different experiments. 132

C.2 Label distribution of top 5 conditions (MIMIC-CXR). 132

C.3 Direct report generation performance comparison for IU X-ray findings and impressions (GPT-4o). 133

C.4 Performance comparison for MIMIC-CXR findings and impressions (GPT-4o). 133

C.5 Performance comparison for ChexPert Plus findings and impressions (GPT-4o). 134

C.6 Direct report generation performance comparison for MIMIC-CXR findings and impressions (GPT-4-vision-preview). 134

C.7 Direct report generation performance comparison for IU X-RAY findings and impressions (GPT-4-vision-preview). 135

C.8 Image reasoning performance of GPT-4-vision-preview on chest X-ray images. 135

C.9 Image reasoning performance of GPT-4o on chest X-ray images. 136

C.10 χ^2-test for homogeneity of label distribution across different condition groups (GPT-4o). 136

C.11 χ^2-test for homogeneity of label distribution across different condition groups (GPT-4-vision-preview). 136

C.12 Performance in report generation with groundtruth conditions (GPT-4-vision-preiew). 136

C.13 Direct report generation performance comparison for MIMIC-CXR findings and impressions (Llama-3.2-90B-Vision-Instruct). 137

C.14 Direct report generation performance comparison for IU X-RAY findings and impressions (Llama-3.2-90B-Vision-Instruct). 137

C.15 Performance comparison for Chexpert Plus findings and impressions (Llama-3.2-90B-Vision-Instruct). . . . 138

C.16 Comparison of generated reports across different prompting strategies for one study. . . . 140

C.17 Prompt 1.1 Basic generation: direct report generation based on chest X-ray images. . . . 142

C.18 Prompt 1.2 Indication enhancement: providing the indication section. . . . 142

C.19 Prompt 1.3 Instruction enhancement: providing information on medical condition labels. . . . 143

C.20 Prompt 1.4 Chain-of-Thought: step 1 - medical condition labeling; step 2 - report synthesis. . . . 144

C.21 Prompt 1.5 Few-shot: few-shot in-context learning given a few examples (MIMIC). . . . 145

C.22 Prompt 1.5 Few-shot: few-shot in-context learning given a few examples (IU X-RAY). . . . 146

C.23 Prompt 2.1 Image reasoning: medical condition labeling from chest X-ray images (2-class). . . . 147

C.24 Prompt 2.2 Image reasoning: medical condition labeling from chest X-ray images (4-class). . . . 148

C.25 Prompt 3.1 Report synthesis: report generation using provided positive and negative conditions. . . . 149

C.26 Prompt of finetuned LLaMA-2 report synthesis given groundtruth labels . . 149

D.1 Hyperparameter search results. GAS denotes the number of gradient-accumulation steps, LR the learning rate, and Epochs the total training epochs. . . . 155

D.2 Standard API pricing per 1M tokens for GPT-4o and o1-mini models, based on Azure OpenAI pricing: https://azure.microsoft.com/en-us/pricing/details/cognitive-services/openai-service/#pricing. 156

D.3 Key Words List for Location Prompt (extracted using GPT-4o, then discussed and confirmed by two radiologists) 162

Chapter 1

Introduction

The rapid progress of machine learning, particularly the emergence of large-scale foundation models, is transforming both everyday life and professional workflows. Frontier foundation models like GPT-4 and Deepseek-R1 have demonstrated remarkable capabilities in natural language understanding, image generation, and complex reasoning tasks. As AI systems grow increasingly capable, they underscore the importance of human-centered AI research, which aims to study the interaction between humans and AI as well as build models that can improve human decision-making. The challenge lies in understanding how humans interact with these advanced models, how AI-generated information is interpreted and acted upon, and how human and AI strengths can be combined to achieve complementary outcomes. Developing a deeper understanding of human-AI collaboration is critical for unlocking the full potential of these advanced models and safely integrating them into our day-to-day workflows.

This thesis builds on prior work that introduced a theoretical framework of human understanding and machine explanations, which provided useful insights into human-AI collaboration. It formally demonstrates that human task-specific intuition plays a crucial role in helping humans make effective use of machine-generated explanations, thereby enabling complementary performance, i.e., human-AI teams outperform both human experts and AI models working alone. The framework offers a structured approach to understanding how human interact with AI models and explanations, providing a foundation to future human-AI study designs.

Secondly, we present a case study of human-AI collaboration in prostate cancer diagnosis, where the AI model has already achieved superhuman performance. Through direct collaboration with radiologists, we found that human-AI teams outperformed human clinicians alone but still underperformed compared to AI due to a lack of trust in the model recommendations. Interestingly, while performance feedback did not significantly improve the overall diagnostic accuracy of human-AI teams, it did increase human reliance on AI, highlighting the complex dynamics between trust and performance in human-AI collaboration. Using a recent theoretical framework for reliance, we identify a simple strategy that could improve human-AI complementarity by adjusting reliance patterns. This study provides empirical evidence of the complexities and negative results that can arise in human-AI collaboration, while also pointing to promising future directions for improving complementary performance.

Furthermore, this thesis also explores the diagnosis and evaluation of multimodal large language models (LLMs) for radiology. We began by assessing the out-of-the-box performance of state-of-the-art models such as GPT-4o and LLaMA 3 on chest X-ray radiology report generation. While these models demonstrate impressive capabilities in general domains, our results reveal that they still fall *short* on specialized medical tasks. Through a detailed error analysis, we identified that visual understanding is the key bottleneck limiting model performance. Notably, when provided with essential findings, these models can synthesize coherent and clinically usable radiology reports, reflecting their strong textual capabilities but limited visual comprehension ability. To address limitations in existing evaluation practices, we proposed a fine-grained, tabular-based evaluation framework using expert-curated, high-quality annotations. This benchmark enables a more comprehensive and clinically validated assessment of model capabilities, and serves as a practical tool for guiding future improvements in multimodal model development.

More broadly, this research contributes to the growing understanding of how AI is transitioning from simple tools to sophisticated collaborators in knowledge work and specialized domains. The insights gained from this work are not only applicable to radiology but also extend to other expert-level tasks.

The remainder of this thesis is structured as follows: Chapter 2 presents the theoretical framework for human understanding and machine explanations. Chapter 3 discusses the case study on AI-assisted prostate cancer diagnosis and human-AI reliance. Chapter 4 introduces the diagnosis and evalaution of MLLMs for radiology report generation.

Chapter 2

Prior Work: Theoretical Insights of Human-AI Collaboration

In human-AI collaboration with machine explanations, ***task-specific human intuitions are essential for improving human understanding and decision-making***. While explanations derived from machine learning models can enhance human understanding of the model's decision boundary (i.e., how the model makes predictions), they cannot improve understanding of the task's decision boundary (i.e., the true relationship between inputs and outputs) or model errors unless the human possesses task-specific intuitions. This chapter establishes a theoretical framework showing that without human intuitions about the task, explanations alone cannot lead to better decision-making or complementary performance (i.e., human-AI performance exceeding either human or AI alone). Empirical studies confirm that people rely more on AI predictions when they lack task-specific intuitions, but when explanations align with human intuitions, agreement with model predictions improves. These findings highlight that *effective human-AI collaboration requires incorporating human task-specific knowledge to interpret and benefit from machine explanations*.

2.1 A Theoretical Framework of Human Understanding and Machine Explanations

Three Core Concepts for Measuring Human Understanding

In this section, we identify three key concepts of interest in human-AI decision making: task decision boundary, model decision boundary, and model error. We present high-level definitions of these concepts and formalize them later.[1]

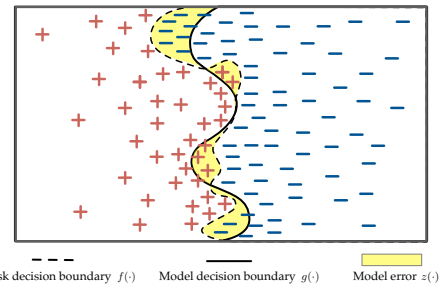


Figure 2-1: Illustration of the three fundamental concepts using a binary classification problem. Task decision boundary (dashed line) defines the ground-truth mapping from inputs to labels. Model decision boundary (solid line) defines the model predictions. Model error (highlighted) represents where the model's predictions are incorrect.

We use a two-dimensional binary classification problem to illustrate the three concepts of interest (Fig. 2-1). *Task decision boundary*, as represented by the dashed line, defines the mapping from inputs to ground-truth labels: inputs on the left are positive and the ones on the right are negative. *Model decision boundary*, as represented by the solid line, determines model predictions. Consequently, the area between the two boundaries is where the model makes mistakes. This yellow highlighted background captures *model error*, i.e., where the model prediction is incorrect. With a perfect model, the model decision boundary would be an exact match of the task decision boundary, and model error never happens. [2]

To the best of our knowledge, we are not aware of any existing quantitative behavioral measure of human understanding that does not belong to one of these three concepts of interest. Building on a recent survey [59], we identify 30 papers that: 1) use machine learning models and explanations with the goal of improving human understanding; and 2) conduct empirical human studies to evaluate human understanding with quantitative metrics. Although human-subject experiments vary in subtle details, the three concepts allow us to organize existing work into congruent categories. We provide a reinterpretation of existing behavioral measures using the three concepts below; a detailed summary is in appendix A.1.

Based on the three concepts above, we introduce a theoretical framework of human

[1] In this work, we omit subjective measures.

[2] We present a deterministic example for ease of understanding and one can interpret this work with deterministic functions in mind. In general, one can also think of model decision boundary, task decision boundary, and model error probabilistically.

understanding in the context of human-AI decision making. We do not discuss machine explanations yet; instead, we formalize the relationship between task decision boundary, model decision boundary, and model error, as well as human understanding of them. This framework enables a rigorous discussion on human understanding as well as the underlying assumptions/interventions that shape the relationship between those understanding.

Defining Core Functions and Human Understanding of Them

Formally, the three concepts of interest are functions defined w.r.t. a prediction problem and a machine learning model:

- **Task decision boundary** is a function $f : \mathbb{X} \rightarrow \mathbb{Y}$ that represents the groundtruth mapping from an input X to the output Y.
- **Model decision boundary** is another function $g : \mathbb{X} \rightarrow \mathbb{Y}$ that represents our ML model which outputs a prediction $\hat{Y}$ given an input. g is usually trained to be an approximation of f. We assume that we are given a model g; the training process of g (and the connection between f and g) is not crucial for this work.
- **Model error** represents the model's error; it is an indicator of whether the model prediction differs from the groundtruth for an input: $z(X, f, g) = \mathbb{I}[f(X) \neq g(X)], \forall X \in \mathbb{X}$. We use $z(X)$ for short when the omitted arguments f and g are clear from context, which maps an input X to whether the model makes an error Z.

We call them *core* functions as they underpin human understanding. We refer to the outputs of core functions for an instance X, Y, $\hat{Y}$, and Z as the three core variables. Note that the core functions do not involve any people; they exist even in absence of human understanding.

We use f^H, g^H, and z^H to denote the human's *subjective* approximations of the core functions, each of them being a function with the same domain and codomain as its objective counterpart. These human approximations can be interpreted as mental models, influenced by the human's knowledge (both on the prediction problem and the ML model), and can change over time as the human-AI interaction progresses.

We can rephrase common cooperative tasks in human-AI decision making in terms of the core functions and human understanding grouped by stakeholders:

- For **decision makers** such as doctors, judges, and loan officers, the main goal is to improve their understanding of task decision boundary (f^H).
- For **decision subjects** such as patients, defendants, and loan applicants, the object of interest can differ even for these three examples. Patients care about the task decision boundary more, while defendants and loan applicants may care about the model decision boundary and especially model error, and would like to figure out how they can appeal model decisions.
- **Model developers** might be most interested in model error, and the eventual goal is to change the model decision boundary.
- For **algorithm auditors**, the main goal is to figure out whether the model decision boundary and model error conform to laws/regulations.

The distance between core functions and their human approximations can be used as a measure for human understanding. Since human approximations are theoretical constructs that only exist in the human brain, we need to perform user studies to measure them. For example, we can ask a human to guess what the model would have predicted for a given input X; the human's answer $\hat{Y}^H$ characterizes their *local* understanding of the model decision boundary. In the rest of the paper, one can interpret "human understanding" with this particular measurement of human approximations. Perfect human understanding thus refers to 100% accuracy in such measurement.

In the interest of space, we assume that the approximation functions remain static and examine a human's local understanding with our framework in the main paper; in other words, we assume that the human employs a consistent mental model for all instances and only reason about Y^H, Z^H, $\hat{Y}^H$. We note that improving human global understanding is often the actual goal in many applications. Table 2.1 summarizes the notations for core functions and human understandings.

Variable	Function	Description
X	—	Input instance
Y	$f : \mathbb{X} \to \mathbb{Y}$	Task decision boundary
$\hat{Y}$	$g : \mathbb{X} \to \mathbb{Y}$	Model decision boundary
Z	$z : \mathbb{X} \to \mathbb{Z}$	Model error
Y^H	$f^H : \mathbb{X} \to \mathbb{Y}$	Human understanding of the task decision boundary
$\hat{Y}^H$	$g^H : \mathbb{X} \to \mathbb{Y}$	Human understanding of the model decision boundary
Z^H	$z^H : \mathbb{X} \to \mathbb{Z}$	Human understanding of the model error
H	—	Task-specific human intuitions
E	—	Machine explanations

Table 2.1: A summary of notations.

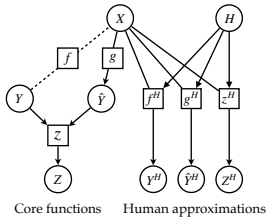


Figure 2-2: Visualizing the relations between core functions, local variables, and human approximations of them.

Causal Graph Framework for Core Functions

To reason about human understanding, we need to understand how core functions relate to each other, and how interventions may affect human understanding. To do so, we adapt causal directed acyclic graphs (causal DAGs) to formalize a causal model for the core functions of human understanding. We start with a simple diagram (Fig. 2-2) without assumptions about human intuitions.[3]

Let us first look at core functions on the left in Fig. 2-2. We use a functional view to represent $\hat{Y} = g(X)$: we add a functional node (g in a square) on the edge from X to $\hat{Y}$ to indicate that g controls the causal link from X to $\hat{Y}$. g is treated as a parent of $\hat{Y}$. As X is the input of g and does not affect g, there is no arrowhead from X to g. Alternatively, one can use a parametric view and use a node θ to capture all variables in g and add θ as a parent of $\hat{Y}$, in addition to X. We use the functional view because it simplifies the visualization, but it deviates slightly from the standard causal diagrams. X and Y are connected with a dashed line through f since we do not assume the causal direction between them. Z is the binary indicator of whether Y and $\hat{Y}$ are different. According to d-separation [94], $\hat{Y}$ is independent of Y given X and g. Z is a collider for Y and $\hat{Y}$, so knowing Z and $\hat{Y}$ entails Y in binary classification.

Next, in Fig. 2-2 on the right, we introduce *task-specific* intuitions, H, that defines human mental models of the core functions. We emphasize *task-specific* to capture intuitions about the current problem, as opposed to generic intuitions such as that humans can interpret saliency maps or humans can update their understanding over time. Fig. 2-2(b) shows a base

[3]Throughout this paper, X in the diagrams refer to a test instance that the model has *not* been trained on.

version of how human intuitions relate to human understanding of core variables. For now, we do not make any assumptions about human intuitions, we simply connect human intuition with their understanding through f^H, g^H, z^H. As H directly influence f^H, g^H, z^H, there is an arrowhead in the links from H to f^H, g^H, z^H. Later, we will discuss more realistic instantiations, e.g., Z^H when $\hat{Y}$ is given.

Looking together at Fig. 2-2, d-separation suggests that human approximation of core variables are independent from core variables given X, without extra assumptions about human intuitions. Therefore, a key goal of our work is to articulate what assumptions we make and how they affect the causal diagrams.

A New Operator Next, we analyze human local understanding ($Y^H, \hat{Y}^H, Z^H$ on the right of Fig. 2-2). Without extra assumptions, the causal direction between $Y^H, \hat{Y}^H, Z^H$ is unclear, because their generation process is controlled by the human brain, a black box. We visualize this ambiguity by connecting nodes with undirected dashed links in Fig. 2-3(a) as the base diagram.

The base diagram is not useful in its current state; in order to use the diagram to reason about human understanding, we need *realizations* of the base diagram where dashed links are replaced by solid, directional links. No realization is universally correct, and each realization requires certain assumptions or interventions, which we refer to as *conditions*.

Condition 1—emulation vs. discovery. To delineate the feasibility of various base diagram realizations, we introduce two conditions. The first condition is an assumption about human knowledge, i.e., that the human has perfect knowledge about task decision boundary; in other words, f^H perfectly matches f and $Y^H = Y$ for all inputs. Problems where human labels are used as ground truth generally satisfy this condition, e.g., *topic classification*, reading comprehension, and object recognition. We follow Lai et al. [61] and call them *emulation* task, in the sense that the model is designed to emulate humans; by contrast, discovery problems are the ones humans do not have perfect knowledge of task decision boundary (e.g., *deceptive review detection* recidivism prediction).[4] It follows

[4]Emulation and discovery can be seen as two ends of a continuous spectrum. The emulation vs. discovery

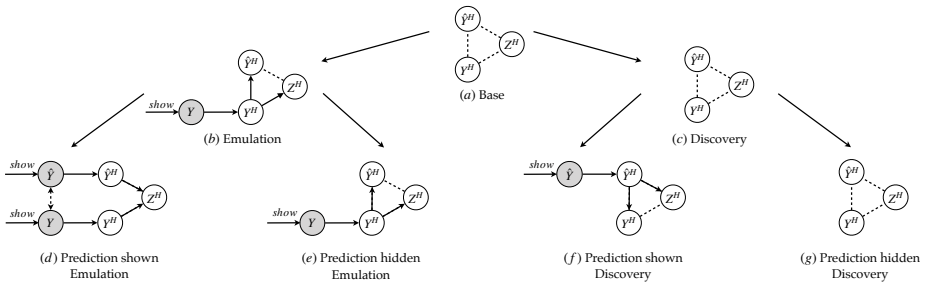


Figure 2-3: Causal diagrams visualizing the relationship between a human's local understanding. With the base diagram at the root, we organize its realizations based on different conditions in a two-level decision tree. Undirected dashed lines represent ambiguous causal links. The bidirectional dashed line in subfigure (d) represents the correlation between $\hat{Y}$ and Y potentially induced by the prediction model. Shaded nodes and their edges represent *show* operations.

that human understanding of task decision boundary is irrelevant in emulation tasks, but achieving complementary performance is a key goal in discovery tasks.

Condition 2—prediction shown vs. hidden. An alternative condition is an intervention that presents the model prediction $\hat{Y}$ to the human. Given this information, a rational human would gain a perfect understanding of the *local* model decision boundary and always predict $\hat{Y}^H = \hat{Y}$.

The *show* operator. To describe the effect of applying these conditions, we introduce a new syntax for manipulating causal diagrams: the *show* operator. When *show* is applied to a core variable, that information becomes available to the human. For example, $show(\hat{Y})$ means that the human can see the model prediction for X. This operation draws an equivalence between the core variable and the human approximated counterpart, assuming that the human is rational.

We introduce the new *show* operator as opposed to the standard *do* operator for two reasons. First, *show* operator introduces new variables to the causal diagram instead of setting the value of an existing variable (e.g., Y and $\hat{Y}$ in Fig. 2-3 and E in §2.2). Second, the *show* operator can change the causal diagram as we reason about human understanding, including changing edges and variables. Notation-wise, *show* allows us to specify the

categorization determines the set of causal diagrams that applies to the problem; this decision is at the discretion of practitioners that design experiments using our framework.

condition for human approximations; for example, $Y^H_{show(\hat{Y})}$ denotes the local understanding of task decision boundary given predicted label $\hat{Y}$.

In Fig. 2-3, applying *show* operator leads to two changes: it adds a link from a core variable to the corresponding human approximation, and it removes influences from other human approximations. For example, under the emulation condition visualized in Fig. 2-3b, $show(Y)$ adds a link from Y to Y^H and removes all other edges going into Y^H, effectively disambiguating the relation between Y^H and the two other variables.

2.2 Machine Explanations and Human Intuitions

Explanations of machine predictions can provide richer information about the model than predicted labels and are hypothesized to improve human understanding of core variables. In this section, we use our framework to discuss the utility and limitations of machine explanations. We first show that without assumptions about human intuitions, explanations can improve human understanding of model decision boundary, but not task decision boundary or model error. As a result, *complementary performance in discovery tasks is impossible*. We then discuss possible ways that human intuitions can allow for effective use of explanations and lay out several directions for improving the effectiveness of explanations. Our analyses highlight the importance of articulating and measuring human intuitions in leveraging machine explanations to improve human understanding.

Limitations of Explanations without Human Intuitions

Existing explanations are generated from g (Fig. 2-4(a)). We first introduce explanation (E) to our causal diagram. Since the common goal of explanation in the existing literature is to explain the underlying mechanism of the model, E is derived from g and thus we argue that explanation should have only one parent, g, among the core functions. For example, gradient-based methods use gradients from g to generate explanations [111, 7]. Both LIME [101] and SHAP [74] use local surrogate models to compute importance scores, and the local surrogate model is based on g. Counterfactual explanations [79, 117] typically identify

examples that lead to a different predicted outcome from g. In all of these explanation algorithms, there is no connection between E and f or u.

In addition, there should be no connection between E and task-specific intuitions, H. Conceptually, only task-agnostic human intuitions are incorporated by existing algorithms of generating explanations. It is well recognized that humans cannot understand all parameters in a complex model, so promoting sparsity can be seen as incorporating some human intuition. Similarly, the underlying assumption for transparent models is that humans can fully comprehend a certain class of models, e.g., decision sets [63] or generalized linear models [80, 22]. In counterfactual explanations, it is assumed that by contrasting similar examples, people can recognize the differentiating feature and thus derive feature importance [55]. However, none of these assumptions about the human intuitions are about task decision boundary, model error, or human understanding of them.

Now we discuss the effect of explanations on human understanding without assuming any task-specific human intuitions (i.e., without adding new edges around H).

Explanations can improve human understanding of model decision boundary, but cannot improve human understanding of task decision boundary or model error. We start with the cases *where predicted labels are not shown*. Fig. 2-4(b1) shows the subgraph related to $\hat{Y}$ and $\hat{Y}^H$ from Fig. 2-2. Without explanations, $\hat{Y}$ and $\hat{Y}^H$ are independent given X. Fig. 2-4(b2) demonstrates the utility of machine explanations. Because of the shared parent (g) with $\hat{Y}$, the introduction of E *can* improve human understanding of model decision boundary, $\hat{Y}^H$. *Note that our discussion on improvement is concerned with the upper bound of understanding assuming that humans can rationally process information if the information is available.* This improvement holds regardless of the assumption about Y^H (i.e., both in emulation and discovery tasks).

When predicted labels are shown, improving human local understanding of model decision boundary is irrelevant, so we focus on task decision boundary and model error. In emulation tasks ($show(Y)$), and once provided with predicted labels ($show(\hat{Y})$), humans would achieve perfect accuracy at approximating the three core variables. Because this

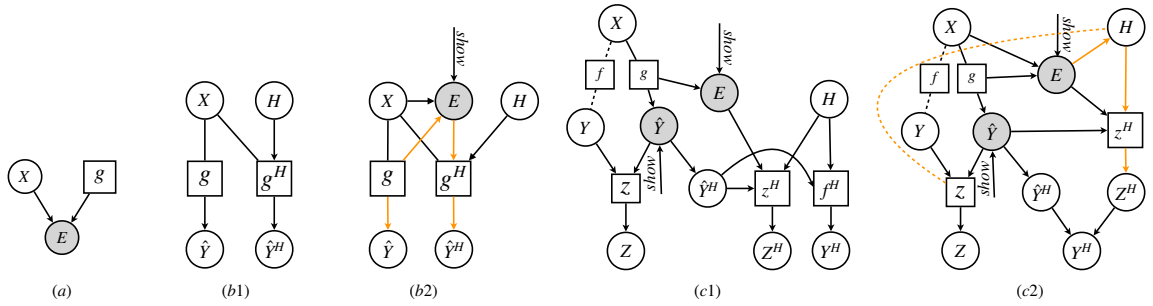


Figure 2-4: (a) E is generated from g. (b1) $\hat{Y}$ and $\hat{Y}^H$ are independent given X. (b2) The utility of E: E can improve human understanding of model decision boundary. (c1) E cannot improve human understanding of task decision boundary and model error without human intuitions. (c2) Combined with human intuitions, E can improve task decision boundary and model error. We use orange lines to highlight the links that lead to positive utility of E. We omit links from X to f^H and z^H for simplicity.

perfect local understanding also holds in emulation tasks without machine explanations, explanations have no practical utility in this setting. That is, machine explanations cannot help humans achieve better approximation than showing predicted labels in local understanding. Note that this is only true for local human understanding, explanations can still potentially improve global understanding, which explains the success of model debugging in an emulation task, topic classification in Ribeiro et al. [101].

In comparison, Fig. 2-4(c1) shows the diagram for the more interesting case, discovery tasks. Explanations are often hypothesized to improve human decision making, i.e., bringing Y^Hcloser to Y. However, if we do not make assumptions about human intuitions, although E can affect f^H, it cannot bring any additional utility over showing $\hat{Y}$. The reason is that d-separation indicates that given the prediction g and X, the explanation E is independent of Y and Z. That is, E cannot provide any extra information about Y (task decision boundary) and Z (model error) beyond the model. Moreover, the model cannot provide any better approximation of Y than $\hat{Y}$. Alternatively, we can also think of the functional form. If we cannot make any assumption about H, $g^H_{show(E,\hat{Y})}$ is no different from $g^H_{show(\hat{Y})}$. *It is plausible for a person to follow machine predictions when they have no intuitions about a task at all.* Therefore, complementary performance is impossible without extra assumptions about human intuitions.

As a concrete example, consider the case of deceptive review detection with an alien

who does not understand English (i.e., guaranteeing that there is no intuition about this task). Machine explanation such as feature importance cannot provide meaningful information to the decision maker, i.e., the alien.

Explanation + Human Intuitions

Next, we discuss how explanations can be integrated with human intuition to achieve an improved understanding in discovery tasks (recall that $Z^H$and $Y^H$are entailed in emulation tasks when $\hat{Y}$is shown). We have seen that E itself does not reveal more information about Y or Z beyond g. Therefore, an important role of E is in shaping human intuitions. We present two possible ways.

Activating prior knowledge about model error. E can activate prior human knowledge that can reveal information about model error (Fig. 2-4(c2)). We examine two sources of such prior knowledge that is concerned with *what* information should be used and *how*. First, human intuitions can evaluate *relevance*, i.e., whether the model leverages relevant information from the input based on the explanations. For example, human intuitions recognize that "chicago" should not be used for detecting deceptive reviews or that race should not be used for recidivism prediction, so a model prediction relying on these signals may be more likely wrong. The manifestation of relevance depends on the explanation's form: feature importance directly operates on pre-defined features (e.g., highlighting race for tabular data or a word in a review), example-based explanations or counterfactual explanations narrow the focus of attention to a smaller (relevant) area of the input. Note that the intuition of relevance only applies to the input and does not consider the relation between the input and output.

Second, human intuitions can evaluate *mechanism*, i.e., whether the relationship between the input and the output is valid. Linear relationship is a simple type of such relation: human intuitions can decide that education is negatively correlated with recidivism, and thus that a model making positive predictions based on education is wrong. In general, mechanisms can refer to much more complicated (non-linear) relations between (intermediate) inputs

and labels.

Fig. 2-4(c2) illustrates such activations in causal diagrams. The link from E to H highlights the fact that human intuitions when E is shown are different from H without E because these intuitions about model error would not have been useful without machine explanation. We refer to H in Fig. 2-4(c2) as $H_{show(E)}$. *If* $H_{show(E)}$ is correlated with z (indicated by the dash link), then Z^H is no longer independent from Z (e.g., education should be negatively correlated with recidivism) and can thus improve Y^H because Z is a collider for Y and $\hat{Y}$, leading to complementary performance. It is important to emphasize that this potential improvement depends on the quality of $H_{show(E)}$ (e.g., whether education is actually negatively correlated with recidivism). The lack of useful task-specific human intuitions can explain the limited human-AI performance in deceptive review detection [62].

Expanding human intuitions. Another way that explanations can improve human understanding is by expanding human intuitions. Consider the example of "Chicago" as an important indicator for deceptive reviews in Lai et al. [61]. "Chicago" is reliably associated with deceptive reviews in this dataset are for two reasons: 1) people are less likely to provide specific details when they write fictional texts (theory I); 2) deceptive reviews in this dataset are written by crowdworkers on mechanical Turk for hotels in Chicago (fact II). Highlighting the word for "Chicago" (relevance) and its connection with deceptive reviews (mechanism) is counterintuitive to most humans because this is not part of common human intuitions. But if machine explanations can expand human intuitions and help humans derive theory I, this can lead to improvement of the human understanding of task decision boundary (i.e., humans develop new knowledge from machine explanations). Formally, the key change in the diagram for this scenario is that E influences human intuitions in the next time step H_{t+1}.

2.3 Towards Effective Explanations for Improving Human Understanding

Machine explanations are only effective if we take into account human intuitions. We encourage the research community to advance our understanding of task-specific human intuitions, which are necessary for effective human-AI decision making. We propose the following recommendations.

Articulating and measuring human intuitions. It is important to think about how machine explanations can be tailored to either leveraging prior human knowledge or expanding human intuitions, or other ways that human intuitions can work together with explanations.

First, we need to make these assumptions about human intuitions explicit so that the research community can collectively study them rather than repeating trial-and-error with the effect of explanations on an end outcome such as task accuracy. We recommend the research community be precise about the type of tasks, the desired understanding, and the required human intuitions to achieve success with machine explanations.

Second, to make progress in experimental studies with machine explanations, we need to develop ways to either control or measure human intuitions. This can be very challenging in practice. To illustrate a simple case study, we will present an experiment where we control and measure human intuitions in human-AI decision making.

Incorporate f and z into explanations. An important premise for explanations working together with human intuitions is that machine explanations capture the mechanism or the relevance underlying the model. Indeed, faithfulness receives significant interest from the ML community for the sake of explaining the mechanisms of a model. However, faithfulness to g alone is insufficient to improve human understanding of task decision boundary and the model error.

In order to effectively improve human understanding of f and z, it would be useful to explicitly incorporate f and z into the generation process of E. For example, a basic way

to incorporate z is to report the error rate in a development set. In the case of deceptive review detection, it could be when "Chicago" is used as an important feature, the model is 90% accurate. This allows humans to have access to part of model error and have a more accurate Z^H.

To summarize, we emphasize the following three takeaways:

- Current machine explanations are mainly about the model and its utility for human understanding of the task decision boundary and the model error is thus limited.
- Human intuitions are a critical component to realize the promise of machine explanations in improving human understanding and achieving complementary performance.
- We need to articulate our assumptions about human intuitions and measure human intuitions, and incorporate human intuitions, f, and z into generating machine explanations.

2.4 Conclusion

In this part of the work, we propose the first theoretical work to formally characterize the interplay between machine explanations and human understanding. We identify core concepts of human understanding and reveal the utility and limitations of machine explanations. By focusing on explaining the model, current machine explanations cannot improve human understanding of task decision boundary and model error in discovery tasks. Our work highlights the important role of human intuition. First, we recommend the research community explicitly articulate human intuitions involved in research hypotheses. Hypotheses such as "explanations improve human decisions" cannot contribute generalizable insights, because they can hold or fail depending on human intuitions. Second, we identify future directions for algorithmic development and experimental design. We need to take into account task-specific human intuitions in algorithms that generate machine explanations and develop methods to measure human intuitions and characterize the changes resulting from machine explanations in experimental design.

Chapter 3

Human-AI Decision Making for Prostate Cancer MRI Diagnosis

3.1 Introduction

AI holds promise for improving human decision making in a wide range of domains [60, 54, 100, 105, 3]. Radiology is a representative example as AI outperforms or shows comparable performance with experts [41, 122, 103, 99, 104, 56, 93, 78]. Rather than complete automation, there is growing consensus that AI's optimal role in the near future will serve as an assistance tool for human radiologists in clinical decision making [65, 2, 83, 39]. On the one hand, legal and regulatory challenges stand in the way of full automation. On the other hand, human AI collaboration has the potential to achieve *complementary performance*, where human experts can leverage their contextual knowledge and expertise to correct AI mistakes in ways that could surpass either human or AI performance alone.

However, the actual utility of integrating AI assistance tools in clinical settings remain poorly understood. In particular, very few studies examine the effectiveness of AI assistance in real clinical decision-making with domain experts [88, 6]. In this work, we conduct an in-depth collaboration with radiologists and focus on the case of prostate cancer diagnosis. Prostate cancer diagnosis with magnetic resonance imaging (MRI) remains one of the most difficult tasks for radiologists—even experienced ones—and inter-reader variability is high [28, 25]. Such complexity makes prostate MRI an ideal testbed for studying how AI assistance may complement human expertise. If AI can help reduce radiologists' mistakes

here, it is plausible that similar technology could be effective in other radiology tasks as well.

We run human studies with domain experts to understand AI tool integration in radiology workflow, particularly for challenging diagnoses like prostate cancer. Our investigation focuses on three key questions:

Q1: *Can AI-assistance help humans achieve higher diagnostic performance than either human experts or AI systems alone?*

Q2: *How does different workflow design influence human reliance on AI?*

Q3: *What actionable strategies can we recommend to human decision-makers to ensure complementary performance?*

To answer these questions, we conducted human subject experiments with domain experts, specifically board-certified radiologists (N=8), focusing on prostate cancer diagnosis with AI assistance. We first trained a state-of-the-art AI model [45] for prostate cancer detection from MRI scans. The AI model is able to provide both diagnostic predictions and lesion annotation maps for positive cases as assistance for radiologists. To simulate real-world clinical practice, we designed and implemented two distinct workflows, see Fig. 3-1 for an overview of the design of our human studies. Building on existing tools for teaching prostate cancer diagnosis, we also developed a web-based diagnostic platform that enables radiologists to review MRI scans and annotate suspicious cancer lesions seamlessly.

In Study 1, radiologists each evaluated 75 cases in a three-step process. For each case, they first made independent diagnoses, which helped us to establish baseline human performance. Then, they were shown the AI's predictions. In the final step, they are asked to finalize their decisions after reviewing AI predictions. In Study 2, we introduced a novel element: before starting their evaluations, radiologists first received detailed individual performance feedback from Study 1, as shown in the screenshot in Fig. 3-2c. This feedback included various metrics of their own performance, AI's performance, and their AI-assisted performance. To ensure engagement with this feedback, participants completed attention checks about their performance metrics before proceeding with new cases. This design

allowed us to systematically examine how performance awareness influences radiologists' interaction with AI assistance. Moreover, for each case diagnosis, AI assistance was provided directly to radiologists without them making independent diagnosis.

These two distinct workflows represent common scenarios in the deployment of AI assistance tools in clinical practice and their evolution over time. Study 1 simulates an approach often regarded as responsible, as it allows radiologists to form independent opinions before consulting AI predictions. This approach may be particularly relevant during early deployments, since radiologists may prefer minimal intervention to exercise caution. Over time, the performance information will become available in a local scenario that retains the same distribution of doctors and patients as in the earlier integration of AI tools. Through the design of Study 2, we can investigate how both the timing of AI assistance and awareness of comparative performance metrics influence diagnostic accuracy and radiologists' integration of AI recommendations.

Our findings are consistent with prior studies on human-AI decision making. Human+AI outperforms human alone, showcasing the positive utility of AI assistance. However, Human+AI underperforms AI alone, largely driven by under-reliance. Although performance feedback and upfront AI assistance nudged radiologists to incorporate AI predictions more frequently, we did not observe statistically significant improvements in metrics. To better understand these dynamics and underlying human behaviors, we further analyzed our data using the theoretical reliance framework proposed by Guo et al. [38]. This analysis revealed that changes in workflow—specifically, performance feedback and upfront AI assistance, led to increased reliance on AI. However, this improved reliance does not translate to improved performance, which means radiologists did not appropriately rely on AI when it was correct. More importantly, the framework also helped us identify tangible strategies for achieving human-AI complementarity, where combined decisions outperform either human or AI alone. One such strategy involves only predicting a positive case when both the human and AI agree on its presence. These findings are especially promising and distinguish our work from prior studies, as they highlight actionable directions for designing more effective human-AI collaboration workflows in high-stakes decision-making settings. In contrast,

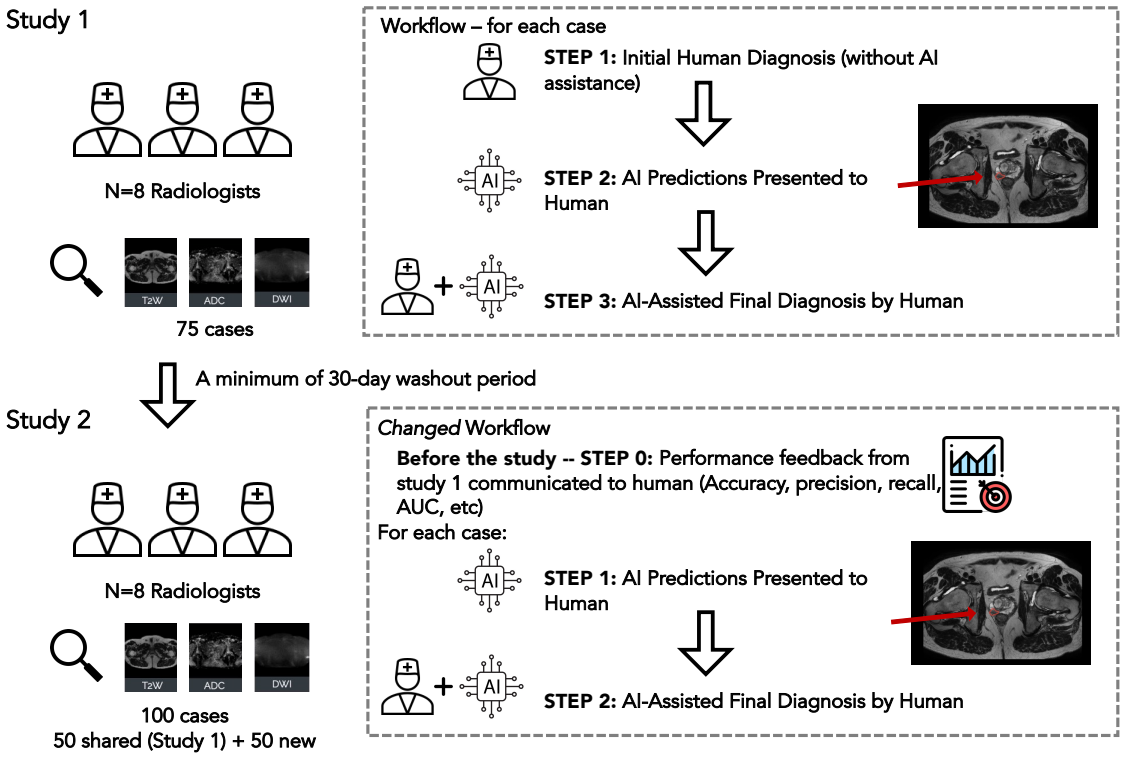


Figure 3-1: Overview of our experiments with radiologists. In study 1, participant radiologists (N=8) reviewed 75 cases in three steps: initial independent diagnosis, review of AI predictions, and final diagnosis. In study 2, we introduce performance feedback to communicate individual radiologist's performance collected from study 1 before the study. Then they reviewed 100 cases with direct AI assistance without independent diagnosis.

previous work has often focused on characterizing human behavior patterns without offering concrete recommendations or actionable guidance for system design.

3.2 Related work

Human-AI decision making. There is a growing interest in the research community to augment human decision making with AI assistance [60]. Typically, the tasks of interest are situated in high-stakes domains such as medicine, law, and finance, where AI-assisted decisions can have significant consequences. However, due to constraints related to resources and the simplicity of participant recruitment, the majority of empirical studies in this area are conducted with crowdworkers or laypeople without expertise. For instance, instead of involving real judges, researchers have explored recidivism prediction as a testbed for Human-AI decision making using crowdworkers [13, 62, 36]. Similarly, in the medical

domain, experiments on disease diagnosis have been conducted with laypeople, such as students [63]. In finance, studies have utilized crowdworkers for tasks like income prediction [129], loan approval [36], and sales forecasting [30]. In some cases, researchers have substituted real-world tasks with entirely artificial ones to facilitate experimentation with crowdworkers, such as alien medicine recommendation [58].

While crowdworkers offer a convenient participant pool, it remains unclear if findings based on these populations generalize to domain experts in real cases. In our work, we work directly with domain experts.

Human-AI decision making with experts in the clinical context. There have been several studies with healthcare professionals in the clinical context, but experiments focused on human-AI complementary performance remain limited. While several studies have shown that AI assistance can improve diagnostic accuracy [110, 107, 46, 106, 76], the experts behavior in human-AI collaboration are underexamined. Existing research also reveals complex performance trade-offs: some studies reveal important trade-offs, such as improved sensitivity at the cost of reduced specificity [51, 90]. Some studies explicitly demonstrated that the performance of human-AI performance falls short of AI alone [96, 52]. To the best of our knowledge, the only work that achieves complementary performance is Steiner et al. [110], which demonstrated that algorithm-assisted pathologists outperformed both the algorithm and pathologists in detecting breast cancer metastasis. However, human specificity is 100% on that task, suggesting a relatively easy task for domain experts.

In summary, human-AI decision making with domain experts, especially for complementary performance, remains underexplored. In light of this gap, our study aims to provide an in-depth analysis of both human+AI team performance and domain expert behavior in a difficult, real-world clinical setting.

3.3 Methods

Dataset

We used public data from the PI-CAI challenge[1] for training and testing. The dataset originally contained 1500 cases, which we filtered down to 1411 cases by excluding cases from the same patients to avoid data leakage. We ensure that all testing cases are biopsy-confirmed. Our AI model was trained on 1211 cases, including 365 (30.1%) clinically significant prostate cancer (csPCa) cases. For study 1, the testing set includes 75 cases, of which 23 (30.6%) are csPCa. Study 2 consists of 100 cases, with 32 (32%) being csPCa. For each patient case, we used T2-weighted (T2W), diffusion-weighted imaging (DWI), and apparent diffusion coefficient (ADC) sequences as inputs for both AI and human studies. 50 cases were shared between study 1 and study 2, which allows us to directly compare performance metrics across both studies on this shared subset.

Labels/annotations. Case labels were obtained from three sources: biopsy-confirmed results (from systematic, magnetic resonance-guided biopsy, or prostatectomy), human-expert annotations, and AI-derived annotations [15]. Out of the original 1500 cases, 1001 has biopsy confirmed case-level labels. Out of the 425 positive cases, 220 have human expert annotations, with the remaining annotated by AI. We prioritized human expert annotations when available, defaulting to AI annotations otherwise. Ground truth case-level labels are approximately accurate, with 66.7% (1001/1500) cases having biopsy results. Lesion-level annotations are less accurate due to the practical challenges of annotating all lesions in the large dataset. For all of our testing patient cases, case-level labels are derived from biopsy results. Lesion-level annotations are derived by experts (trained investigators and resident, supervised by expert radiologists), using all available clinical data. This includes MRI scans, diagnostic reports (radiology and pathology), and whole-mount prostatectomy specimens or other biopsy results when available.

[1] https://pi-cai.grand-challenge.org/DATA/

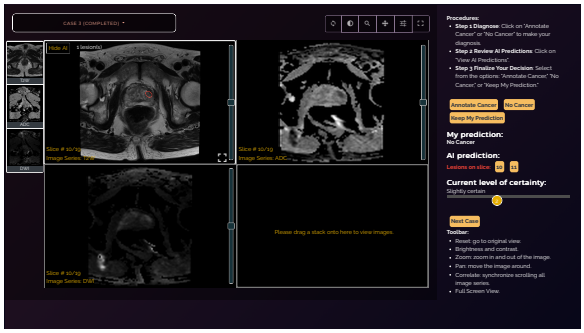
AI model & performance

We use the established nnU-Net model [45, 15] as our AI model, trained from scratch with our own splits. We ensure that all testing examples have pathology groundtruth. Training examples have a mixture of different types of labels: pathology groundtruth, human expert labeled csPCa and delineation of the lesion area, and AI-labeled csPCa and lesion area [104]. The AI model achieves an AUROC of 0.910 in the training set, 0.730 and 0.790 respectively for the study 1 and study 2 testing set. Note that all testing examples have pathology groundtruth while as training sample have a mixture of pseudo labels. For comprehensive details on the AI model's training configurations and performance metrics, please refer to appendix B.1.

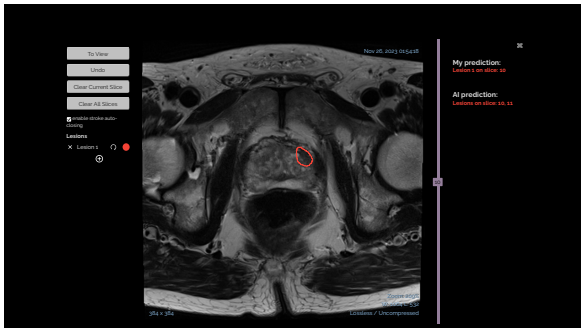
Human-AI Decision Making Interface

We developed a webapp to conduct the human-study. Participants can log in with their name and email. They will see a consent page when they log in for the first time. Once they give the consent, they will enter the study and see our study interface. A screenshot of the consent page can be found in appendix Fig. B-2. Our human study is approved by the Institutional Review Board (IRB).

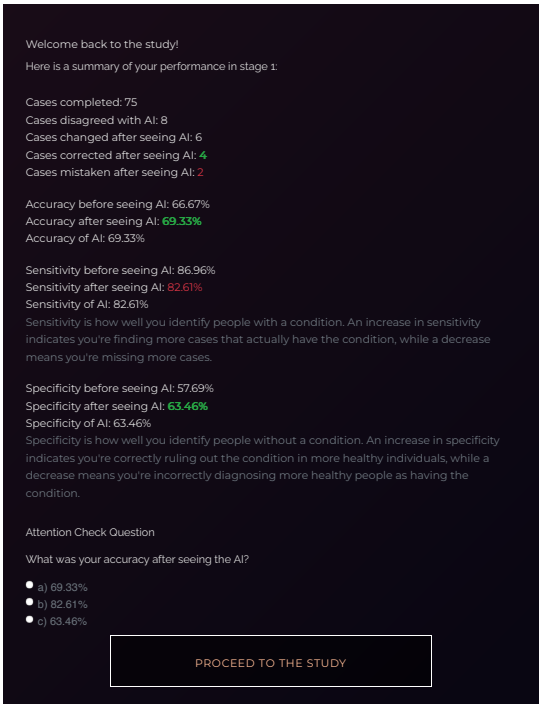
Study interface. Our study interface has three major components: the View Panel on the left, the Control Panel on the right, and the Annotation Panel as a pop-up in the center of the screen. The interface is shown in Fig. 3-2a. In the View Panel, we display three image sequences (T2W, ADC, BWI) from the MRI scans of the current case. In the Control Panel, participants are informed about the current study (study 1 or 2) and provided with control buttons to make decisions or proceed to the next steps. Binary case-level AI predictions are also presented in this panel. Participants make their own predictions by clicking the buttons (‘Annotate Cancer" for positive cases and “No Cancer” for negative cases) and indicate their confidence level using a sliding bar. If a participant believes the case is positive, they click the "Annotate Cancer" button, which triggers a pop-up window (Annotation Panel)



(a) Patient case review interface.



(b) Lesion annotation panel.



(c) Performance feedback page.

Figure 3-2: Screenshots of the webapp interface for our human study. (a) Fig. 3-2a presents a user interface for patient case evaluation. An AI lesion prediction is highlighted with a red contour in the T2W sequence. On the right, the user's current prediction is shown as "No Cancer," and they are at the stage of evaluating the AI prediction to make a final diagnosis. (b) Fig. 3-2b shows the user interface of the Annotation Panel. The screenshot shows a current annotation of the user. The user can clear the annotation or add new annotations on the canvas. (c) Fig. 3-2c illustrates an example performance feedback page presented to a user before proceeding to Study 2. The page provides a summary of the total number of cases, including counts of correct and incorrect cases, the number of decision changes influenced by AI advice, and whether those changes were correct or incorrect. It also highlights key performance metrics such as accuracy, sensitivity, and specificity, derived from Study 1. To ensure users review the information carefully, they are required to answer attention check questions.

displaying enlarged images from the T2W sequence of the current case, allowing participants to annotate the suspicious lesion areas. Participants can annotate any suspicious lesions by freely drawing on any image slice, using the sidebar to navigate between slices. The annotation interface is illustrated in Fig. 3-2b.

Performance feedback. In Study 2, the first page after the login page will be the performance feedback page, as shown in Fig. 3-2c. This page provides detailed individual

feedback on their performance from Study 1. The feedback includes both case counts and performance metrics. Specifically, we present the total number of cases completed by the participant, the number of cases where their prediction disagreed with the AI's prediction, and the number of times they changed their decision after viewing the AI's advice. Among these decision changes, we further highlight how many were correct and how many were mistaken after incorporating the AI's input. For performance metrics, we provide accuracy, sensitivity, and specificity. These metrics are shown for the participant's diagnoses before and after reviewing AI predictions, as well as for the AI's performance alone. This breakdown allows participants to see the impact of the AI on their decision-making and compare their independent performance with AI. At the bottom of the feedback page, we ask an attention check question to ensure participants review the information carefully. The attention question is a single-answer multiple-choice question that asks for the value of one of the performance metrics displayed on the page.

Exit survey. As the final step in both studies, participants are required to complete an exit survey. The survey for Study 1 collects demographic information and participants' opinions on AI. The survey for Study 2 gathers their thoughts on the performance feedback provided and revisits their opinions on AI. Screenshots of these surveys are included in the appendix Fig. B-4 and Fig. B-5.

Experimental Design

To evaluate the effectiveness of AI assistance, we conduct two studies with practicing radiologists ($N = 8$). An overview of our experimental workflow is shown in Fig. 3-1.

Participant demographics, including experience levels, are detailed in Appendix B.3. Participants are recruited through interest forms distributed at the annual conference of RSNA (Radiological Society of North America), one of the largest radiology conferences in the world. We also use snowball recruiting, where participants refer colleagues and peers in their network. All participants are practicing radiologists and come from different regions (US and Europe), and all US-based participants are board-certified.

Study conditions. Our experiments include three main conditions to evaluate radiologist performance:

- Human-only (Study 1): Independent diagnosis without AI assistance.
- Human+AI (Study 1): Diagnosis made after independent diagnosis and reviewing AI predictions.
- Human+AI (Study 2): Diagnosis made with AI predictions shown upfront, with prior feedback on individual performance metrics at the beginning of the study.

In Study 1, participants complete 75 test cases. After logging in and signing the consent form, we provide a toy case to familiarize participants with the interface and workflow. For each of the test cases, participants first make an independent diagnosis (human-only condition). Then they review the AI prediction and annotations. Participants have a chance to update and finalize their diagnosis before moving on to the next case (Human+AI condition for Study 1).

Between Study 1 and Study 2, we set a minimum memory wash-out period of 30 days to eliminate any recall effects. The actual period varies because participants complete the study at their own pace.

In Study 2, participants begin by reviewing a summary of their performance metrics from the Human+AI condition in Study 1. This feedback includes key metrics and interaction statistics to encourage reflection on their interaction with AI. To ensure engagement, participants answer an attention check question about the feedback before proceeding. Study 2 consists of 100 cases, 50 randomly sampled from Study 1 and 50 new cases from a separate test pool. Different from Study 1, AI predictions and annotations are shown upfront, and participants either accept the AI diagnosis or make modifications (Human+AI condition for Study 2).

Both studies conclude with an exit survey.

Table 3.1: Performance comparison between AI, Human, and Human+AI for identifying csPCa from MRI scans. For each metric, the means, 95% confidence intervals, and number of instances are reported. The reported values and instance counts represent averages across eight radiologists. All confidence intervals are derived using bootstrap methods. p-values are calculated using the bootstrap z-test with a significance threshold of $\alpha = 0.05$.

	Per-patient Analysis						
	Study 1				Study 2		
	AI	Human	Human+AI	P (AI>Human)[1] P (Human+AI>Human) P (AI>Human+AI)	AI	Human+AI	P (Human+AI>Human) P (AI>Human+AI)
AUROC	0.730 [0.686, 0.772]	0.674 [0.627, 0.719]	0.701 [0.656, 0.746]	0.023*/0.033*/0.131	0.790 [0.751, 0.829]	0.732 [0.689, 0.776]	0.036*/0.005*
Accuracy	69.3% [0.647, 0.738] 52/75	63.2% [0.585, 0.677] 47/75	66.2% [0.615, 0.708] 50/75	0.013*/0.009*/0.103	76.0% [0.718, 0.800] 76/100	69.6% [0.650, 0.743] 70/100	0.026*/0.003*
Sensitivity (Recall)	82.6% [0.757, 0.891] 19/23	78.3% [0.708, 0.853] 18/23	80.4% [0.732, 0.874] 18/23	0.171/0.207/0.299	87.5% [0.815, 0.930] 28/32	83.2% [0.765, 0.896] 27/32	0.163/0.111
Specificity	63.5% [0.577, 0.690] 33/52	56.5% [0.507, 0.622] 29/52	59.9% [0.542, 0.655] 31/52	0.021*/0.009*/0.125	70.6% [0.651, 0.759] 48/68	63.2% [0.575, 0.691] 43/68	0.052/0.006*
NPV	89.2% [0.847, 0.933] 33/37	85.9% [0.803, 0.904] 29/34	88.0% [0.826, 0.919] 31/36	0.081/0.108/0.220	92.3% [0.886, 0.958] 48/52	89.3% [0.842, 0.932] 43/48	0.159/0.052
PPV (Precision)	50.0% [0.431, 0.569] 19/38	44.7% [0.378, 0.509] 18/41	47.1% [0.403, 0.537] 18/39	0.014*/0.012*/0.105	58.3% [0.514, 0.654] 28/48	51.9% [0.447, 0.585] 27/52	0.066/0.003*
	Per-lesion Analysis[2]						
	Study 1[3]				Study 2		
	AI	Human	Human+AI	P (AI>Human) P (Human+AI>Human) P (AI>Human+AI)	AI	Human+AI	P (Human+AI>Human) P (AI>Human+AI)
Accuracy	35.4% [0.307, 0.403] 17/48	25.7% [0.212, 0.297] 13/53	28.5% [0.240, 0.330] 15/51	0.001*/0.168/0.019*	36.9% [0.323, 0.417] 24/65	33.8% [0.292, 0.385] 22/66	0.005*/0.170
Sensitivity (Recall)	73.9% [0.675, 0.800] 17/23	58.4% [0.509, 0.658] 13/23	63.4% [0.561, 0.706] 15/23	0.001*/0.176/0.015*	72.7% [0.665, 0.787] 24/33	67.4% [0.608, 0.737] 22/33	0.036*/0.121
PPV (Precision)	40.5% [0.353, 0.456] 17/42	31.5% [0.261, 0.361] 13/43	34.4% [0.290, 0.394] 15/43	0.005*/0.202/0.045*	42.9% [0.377, 0.482] 24/56	40.6% [0.350, 0.456] 22/55	0.006*/0.247

[1] p-values compare the performance of different conditions using bootstrap z-test. In Study 1, a paired test is conducted on 75 cases, where each case is evaluated by both Human Alone and Human+AI. In Study 2, an unpaired test is performed, comparing the performance on 75 Human Alone cases and 100 Human+AI cases.
[2] Note that the lesion-level analysis should be interpreted with caution compared to the per-patient analysis. Since lesion-level analysis excludes true negatives (TNs), we only calculate metrics that do not rely on TNs, i.e. accuracy, sensitivity and PPV.
[3] For study 1 lesion-level human results, one radiologist's results were excluded because they used our annotation tool incorrectly.

3.4 Results

3.4.1 Performance of Human vs. AI vs. Human+AI Team

We recruited eight radiologists to participate in both studies. In Study 1, each radiologist independently reviewed 75 prostate MRI cases, resulting in 600 case reviews. Following a

washout period of at least one month, the same radiologists participated in Study 2, where each reviewed 100 cases (including 50 cases from Study 1) for an additional 800 reviews. We introduced a modified human-AI collaboration workflow in Study 2. Specifically, we provided radiologists with individual performance feedback from Study 1 and upfront AI predictions. Across both studies, we collected 1,400 case reviews under four conditions: human-only, two variants of human+AI collaboration (Human+AI (Study 1) and Human+AI (Study 2)), and AI-alone.

Because clinical costs of false positives and negatives differ, we used a V-shaped scoring rule derived from Kleinberg et al. [53] to evaluate performance in our binary decision space. This scoring rule flexibly weights false-positive and false-negative costs with a tunable parameter μ, allowing us to reflect varying clinical priorities. The parameter μ (where $0 \leq \mu \leq 1$) tunes the severity of false-positive vs false-negative penalties: $\mu = 0.5$ collapses to plain *accuracy*, $\mu = 0$ reproduces *specificity*, and $\mu = 1$ reproduces *sensitivity*. This approach enables comprehensive evaluation across clinical scenarios where different error types carry distinct consequences. Traditional performance measures including accuracy, AUROC, sensitivity, and specificity are also presented in Appendix Table 3.1.

According to Fig. 3-4, our findings reveal a consistent performance trend across all scoring rules: human-alone < human+AI collaboration < AI-alone. Under the $\mu = 0.5$ accuracy scoring rule, performance in Human+AI (Study 1) (0.662, [0.624, 0.700]) and Human+AI (Study 2) (0.696, [0.664, 0.728]) increased over human-alone (0.632, 95% CI: [0.593, 0.670]), but remained below AI-alone (0.731, [0.708, 0.755]). These 95% confidence intervals were computed using the t-distribution based on group-level standard errors. The observed ordering was consistent across all scoring rules. Full numerical results are provided in Appendix Table B.2. These differences were statistically significant: human-alone < AI ($p < 0.001$), Human+AI (Study 1) < AI ($p < 0.01$), Human+AI (Study 2) < AI ($p < 0.1$), and human-alone < Human+AI (Study 2) ($p < 0.05$), for the majority of the scoring rules. All reported p-values are derived from one-sided t-tests with unequal variances; degrees of freedom were calculated using Welch's approximation and ranged from approximately 1050 to 1650 across all scoring rules and pairwise comparisons, reflecting varying group sizes

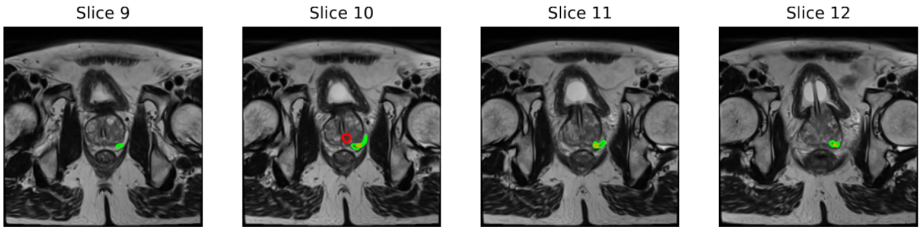


Figure 3-3: An example of lesion-level annotation comparing human experts (red contour), AI (yellow), and expert annotation from the dataset (green). In this case, the AI successfully detected a lesion which corresponded to a clinically significant prostate cancer in the dataset; our human radiologist did not identify this lesion, and instead annotated a lesion in the transition zone.

and variance differences. Full hypothesis test details, including p-values, and degrees of freedom, are reported in Appendix B.2, Table B.3, and Table B.4, respectively.

Fig. 3-3 provides an example from Study 2 in which the AI correctly identified a biopsy-confirmed lesion that was missed by radiologists, even with AI assistance.

Despite some improvement with AI assistance, no human configurations, whether assisted or not, matched the standalone AI performance. The observed positive trend from Study 1 to Study 2 suggests that changes in workflow can influence human performance; however, these gains were limited in magnitude and not statistically significant. These results indicate a key challenge of human-AI collaboration: simply providing AI outputs is not sufficient to realize better clinical decision-making. The consistently lower performance of human-involved conditions across both workflow designs points to deeper issues, such as how radiologists interpret, trust, and incorporate AI guidance, that may limit the effectiveness of assistance. We explore these behavioral and workflow dynamics further in section 3.4.2 and section 3.4.3, where we analyze radiologists' reliance patterns in greater detail.

3.4.2 Analysis of Human-AI Complementarity

In this section, we analyze the extent to which we observed human-AI complementarity in our studies. We interpret empirically-observed complementarity among human-AI teams by comparing it to the theoretical best-attainable benchmark, defined as the performance of an idealized, Bayesian rational agent with access to the human and AI predictions.

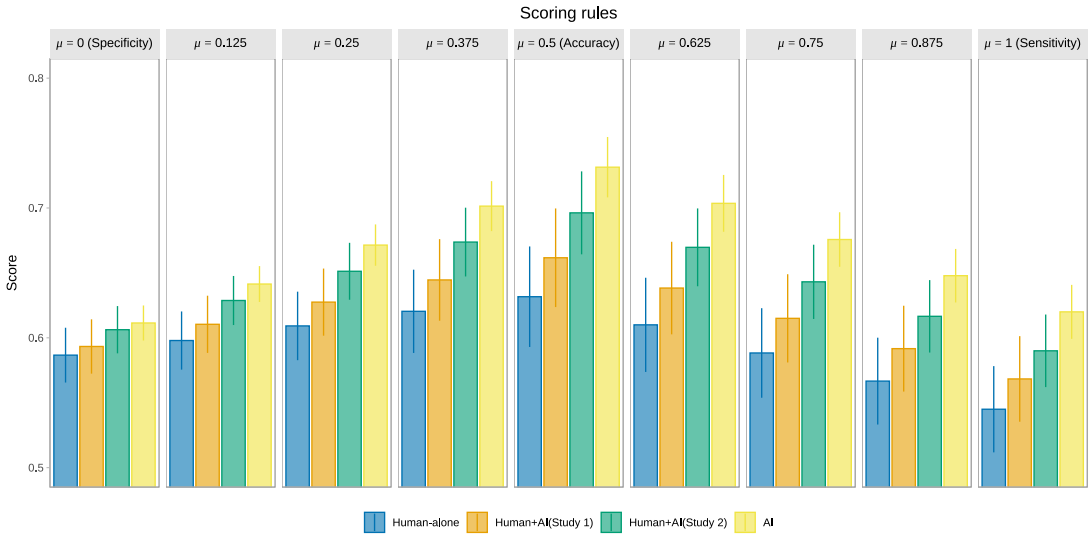
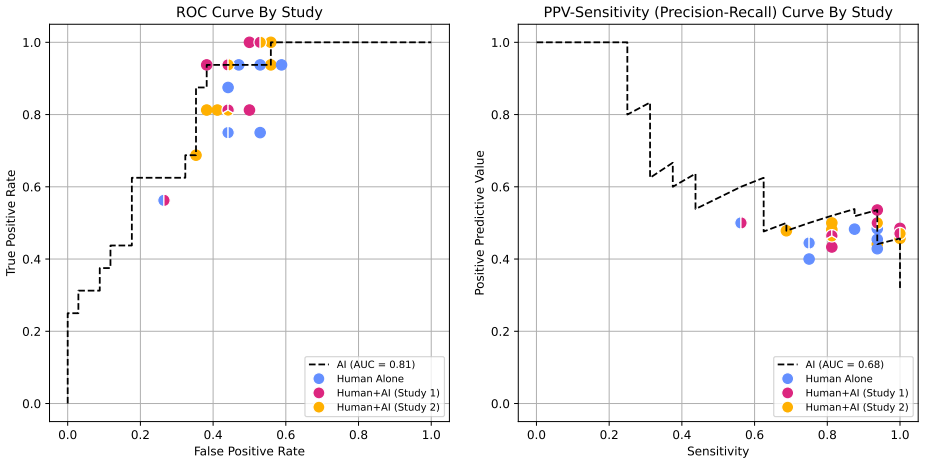


Figure 3-4: Across all scoring rules, a consistent performance pattern is observed: Human-alone < Human+AI (Study 1) < Human+AI (Study 2) < AI-alone. The addition of AI assistance shows improvement over human-alone performance in both studies, yet all human configurations still underperform the AI alone. Bars indicate the mean score under each scoring rule based on the observed decisions, and error bars represent 95% confidence intervals computed using the t-distribution.



(a) ROC Curve (b) PR Curve

Figure 3-5: Individual radiologists performance compared with the AI model. The model achieves higher performance than all of the radiologists without AI assistance (blue dots). However, with AI assistance, some individual radiologists outperformed the AI model (red and orange dots that are above the curve).

Observations of Empirical Complementarity

Individual human radiologists can occasionally achieve complementary performance.

We evaluate individual radiologists and AI-assisted radiologists against AI model using both

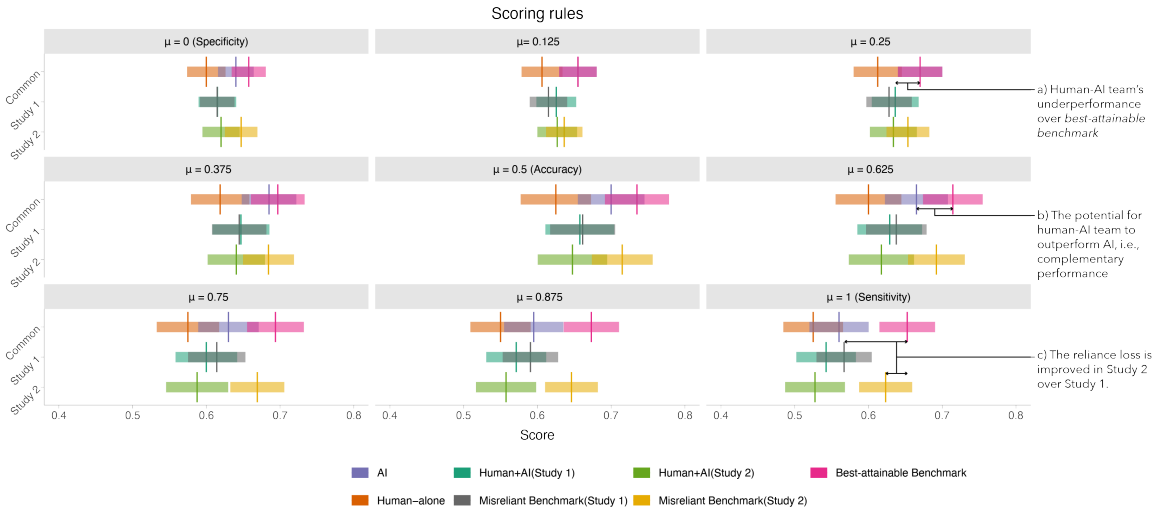


Figure 3-6: The human-alone score, AI score, human+AI score, rational benchmark, and mis-reliant rational benchmark on the common 50-case subset under scoring rules with different ratios of FPR to FNR. The rational benchmarks (red distributions) achieve no less accuracy than AI decisions (orange distributions) and human-alone decisions (blue distributions), indicating potential for human-AI complementarity. The mis-reliant benchmark (gray distributions) is improved in study 2 than in study 1, indicating that human-AI teams rely more appropriately on AI decisions in study 2.

receiver operating characteristic (ROC) and precision-recall (PR) curves in the common cases between Study 1 and Study 2. As shown in Fig. 3-5, and consistent with prior discussions, the AI curve generally outperforms individual radiologists (represented by blue dots). Additionally, AI-assisted radiologists in both studies (red and orange dots) are generally positioned above individual radiologists (blue dots) in both figures, indicating that AI assistance helps improve radiologists' performance. We highlight that there are cases where AI-assisted radiologists outperform the AI curve, as shown by the red and orange dots above the AI curve. This is a promising finding as it suggests that AI assistance could potentially augment human performance to achieve complementary performance (Human+AI > human and Human+AI > AI).

Theoretical Complementarity

Formalization of theoretical best-attainable benchmark. We define the best-attainable benchmark given the observed human-alone and AI decisions, following Guo et al. [38], as

the performance of an idealized, Bayesian rational agent who decides between the human-alone and AI decisions on each decision trial. This agent starts with knowledge of the prior probability that the human versus the AI decision is correct. Upon observing the human-alone decision d^H and AI decision d^{AI} on each trial, the rational agent chooses the one with a higher expected payoff under the evaluation of a given scoring rule S that maps from a decision and state to a numeric score ($S : D \times \Theta \rightarrow \mathbb{R}$):

$$d^r(d^H, d^{AI}) = \underset{d \in \{d^H, d^{AI}\}}{\arg\max} \; \mathbf{E}_{\theta \sim \hat{\pi}(\theta | d^H, d^{AI})} \left[S(d, \theta) \right] \tag{3.1}$$

The rational agent always achieves the best-attainable performance given the information reflected in the human-alone decisions and AI decisions. We can therefore use the expected performance of the rational agent over the decision trials given to humans to identify whether complementarity performance is possible in a task and how much of an improvement it can yield over the better performing of the agents in isolation.

We apply this method to the common 50-case subset of the experimental data in studies 1 and 2. Our results include two levels of analysis: aggregate analysis, with different scoring rules (as shown in Fig. 3-6), and individual analysis, with the scoring rule set to accuracy, $\mu = 0$ (as shown in Fig. 3-7).

Human-AI teams have the potential to achieve complementary performance. The best attainable performance achieved by the rational agent benchmark (pink distribution) never under-performs the AI alone (orange distribution) nor humans alone (blue distribution), as shown in Fig. 3-6b. This confirms that the human-AI teams in our studies have the potential to achieve complementarity. More specifically, Table B.13 shows the strategies adopted by the rational agent to achieve complementarity under the scoring rules we used in our studies. Heuristically, we identify the following strategies to be effective:

- When $\mu \in \{0.4, 0.5, 0.6, 0.75, 1\}$, the rational agent only report positive decisions when both the human-alone and AI decisions are positive, otherwise it reports negative decisions.

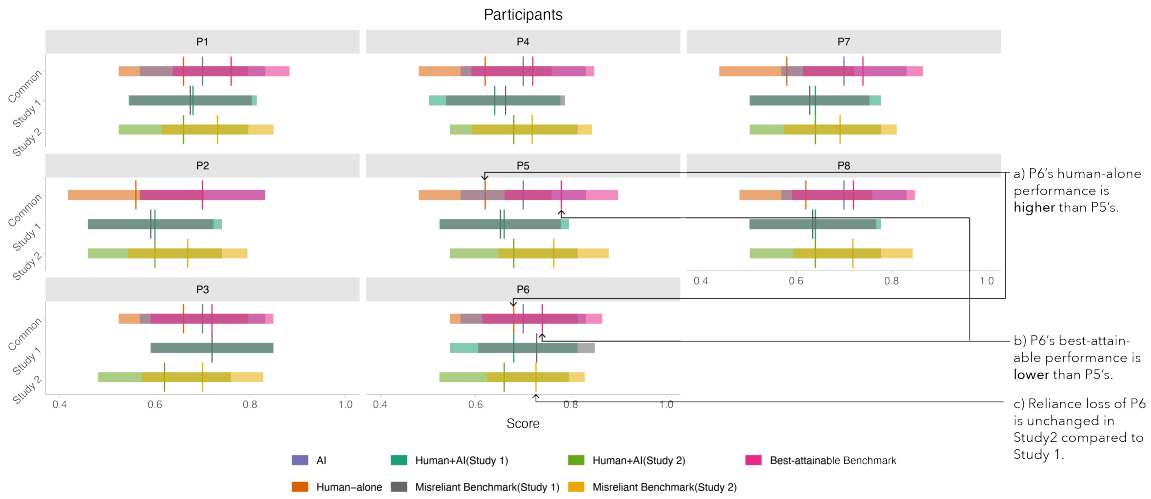


Figure 3-7: Human-alone accuracy, AI accuracy, human+AI accuracy, rational benchmark, and mis-reliant rational benchmark on the common 50-case subset for each participant.

Table 3.2: Participants' reliance on AI and appropriate reliance for them in the common 50-case subset under the evaluation of accuracy ($\mu = 0.5$).

	Participants							
	P1	P2	P3	P4	P5	P6	P7	P8
Reliance in Study 1	0.07	0.12	0.75	0.19	0.13	0.19	0.21	0.06
Reliance in Study 2	0.36	0.41	0.50	0.50	0.56	0.57	0.50	0.50
Appropriate reliance	0.50	0.53	0.75	0.44	0.63	0.24	0.71	0.44

- When $\mu = 0.25$, the rational agent always follows the AI decisions.
- When $\mu = 0$, corresponding to *specificity*, the rational agent only report negative decisions when both the human-alone and AI decisions are negative, otherwise it reports positive decisions.

Different radiologists have varying potential for complementarity, which is not predictable from their human-alone performance. As shown in Fig. 3-7, the rational agent's strategies on every participant achieve no less accuracy than the AI and humans alone. This suggests that the rational agent simulation can be used to identify strategies for individual radiologists that help them achieve complementarity. However, the extent to which superior performance to human or AI alone is possible varies across radiologists. Moreover, we observe that the radiologist having higher accuracy alone does not necessarily

lead to more potential for complementarity. For example, as shown by Fig. 3-7a and b, P6 has better human-alone accuracy than P5, but P5 has a higher rational benchmark than P3 and P6. This helps illustrate how the benchmarks for combined human-AI performance truly capture the notion of *complementation* of information; i.e., they are not simply an additive function of the human-alone and AI-alone performance.

3.4.3 Measuring Human Reliance on AI

We evaluate human reliance on AI within our studies using the rational agent framework. Following Guo et al. [38], we define the reliance level γ of a decision-maker as the overall probability that they choose the AI decision, conditional on the decision-maker facing different recommendations from the human-alone decision and the AI decision, i.e.,

$$\gamma = \Pr[d = d^{AI} | d^{AI} \neq d^{H}] \tag{3.2}$$

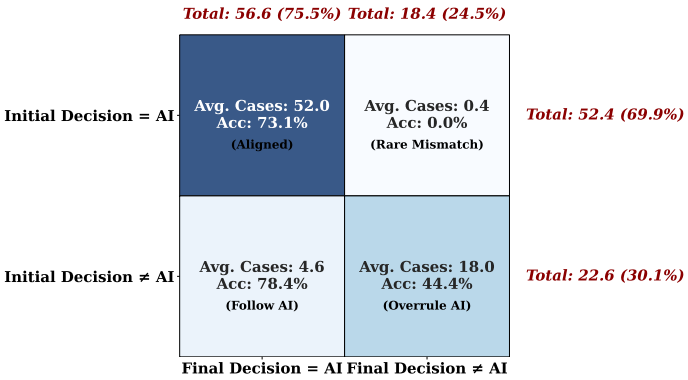
This definition targets a conditional probability, because the reliance level cannot be defined when the human alone makes the same decisions as the AI. As the rational agent defines the best-attainable performance, their reliance on AI also defines the "approximate reliance", i.e.,

$$\gamma_{opt} = \Pr[d^{r}(d^{H}, d^{AI}) = d^{AI} | d^{AI} \neq d^{H}] \tag{3.3}$$

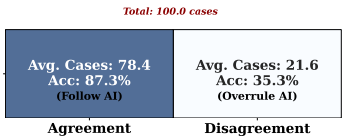
To decouple the impact of reliance from other confounders, we define the mis-reliant rational agent, representing a rational agent who maximizes expected performance but is constrained to have the same reliance as γ^{b}, with the decision rule as

$$\begin{aligned} d^{m}(d^{H}, d^{AI}) = & \underset{d \in \{d^{H}, d^{AI}\}}{\arg\max} \ \mathbf{E}_{\theta \sim \hat{\pi}(\theta | d^{H}, d^{AI})} \left[S(d, \theta) \right] \\ & \text{s.t. } \Pr\left[d^{m}(d^{H}, d^{AI}) = d^{AI} | d^{AI} \neq d^{H} \right] = \gamma^{b} \end{aligned} \tag{3.4}$$

where we use $d^{m}(\cdot, \cdot)$ to denote the decision rule by the mis-reliant rational agent and the probability Pr is measured under the estimated distribution $\hat{\pi}(d^{H}, d^{AI})$.



(a) Study 1 (total cases n=75; radiologists N=8). Top-2 frequent groups are aligned and overrule AI. When there is a disagreement in the initial decision, radiologists are more likely to overrule AI predictions. However, Accuracy in the follow-AI group is higher than the 'Aligned' and 'Overrule AI' groups ($p = 0.04^*$).



(b) Study 2 (total cases n=100; radiologists n=8). Compared with study 1, radiologists are more likely to follow AI when the AI is shown directly without them making their own initial decision. Accuracy is also higher in the 'follow AI' group compared with 'overrule AI' ($p = 0.00^*$).

Figure 3-8: Comparison of Human-AI Decision Alignment and Accuracy. Blue shading indicates frequency of cases for each scenarios; percentages showing diagnostic accuracy for scenario. Accuracy is the highest in the follow-AI group for both studies.

The performance of the mis-reliant rational agent quantifies the best-attainable performance for a given observed reliance level from radiologists. As defined in Equation (3.4), the mis-reliant rational agent always chooses to go with AI decisions on the γ^b ratio of instances that have the biggest advantage of going with AI over human. The difference between the mis-reliant rational agent and the rational agent arises from the suboptimal reliance level γ^b, which we define as *reliance loss*.

Similar to our analysis in Section 3.4.2, we use the common 50-case subset and present the expected empirical performance of the mis-reliant rational agent. We also analyze the performance of the mis-reliant rational agent in aggregation level and at the individual level (Fig. 3-6 and Fig. 3-7 respectively). The choices of scoring rules are the same with Section 3.4.2.

Upfront AI and prior feedback *increase* reliance on AI. As shown in Fig. 3-6, the performance of the mis-reliant rational agent (gray distribution) in Study 2 is higher than their performance in Study 1. This indicates that the rational agent loses less performance from being constrained to the reliance level as the participants in study 2 than in study 1.

This greater willingness to defer to the AI is also shown in Fig. 3-8. The results indicate that performance feedback and upfront AI assistance leads to higher rate of human-AI agreement (78.4% "follow AI" vs. 75.5% final human-AI agreement from study 1). Moreover, "follow AI" group shows higher accuracy (87.3%) compared with "overrule AI" group (35.3%), as well as sensitivity (92.1% vs. 36.6%), and specificity (72.4% vs. 34.8%). This slightly higher adoption rate, however, was insufficient to bridge the gap between Human+AI teams and AI significantly. We do not notice a significant improvement in human+AI performance (green distribution in Fig. 3-6). This suggests a sizable gap between realized and best-attainable team performance. *Radiologists still miss opportunities to choose the AI answer when it would help most.*

For complete results with more metrics, refer to Table B.6 in the Appendix.

Changes in reliance loss vary across radiologists, which is not predictable from their reliance levels. Among 10 radiologists, we observe that 8 of them (P1-2, P4-5, P7-8) achieve less reliance loss on AI in Study 2 than Study 1, i.e., higher mis-reliant benchmarks. Only one radiologist (P3) achieves more reliance loss in Study 2 than in Study 1 and one radiologist's reliance loss (P6) remains the same. Moreover, we find that change in reliance loss is not predictable from the change on reliance level. For example, P6 achieves a reliance level in Study 2 (0.56) that is farther from the appropriate reliance (0.24) than their reliance level in Study 1 (0.19) (Table 3.2), but their reliance loss remains the same (Fig. 3-7). This suggests that the evaluation of reliance also depends on the distinguishability of the human-alone decisions and the AI decisions, as it is not necessarily reflected by the distance to the appropriate reliance.

3.5 Conclusion

While there is a growing interest in evaluating AI assistance with human decision makers, only a handful of previous works have attempted to evaluate AI systems directly with domain experts, and even fewer have achieved complementary performance or investigated

human behavior. We contribute a comprehensive study with domain experts about how a clinical AI tools might be integrated in practice with two realistic design of workflows. Our findings suggest that while human-AI teams consistently outperform humans alone, they still underperform compared to AI due to under-reliance. More importantly, we look beyond performance and investigate human behavioral patterns in human-AI interaction. Even when domain experts are made aware of their diagnostic performance, the performance gap relative to AI, and their prior AI-assisted outcomes, they still struggle to effectively calibrate their trust and reliance on AI tools. As in prior studies, our work finds that complementary performance—where Human+AI outperforms both human and AI alone—is *difficult* to achieve in practice. Most importantly, we go a step further by uncovering actionable strategies that can help human decision-makers move closer to this ideal. These insights highlight a critical opportunity: rather than solely characterizing human behavior, we can begin to design human-AI systems and workflows that actively support effective reliance. Our findings lay the groundwork for more impactful, behaviorally-informed research directions to improve human-AI decision-making in high-stakes domains like medical diagnosis.

Chapter 4

Evaluation of Frontier Models in Radiology Report Generation

4.1 Introduction

Large language models (LLMs) are becoming multimodal, and GPT-4 model series represent the state-of-the-art. [1] Similar to the claimed general-purpose capabilities in LLMs [17, 84], large multimodal models are supposed to possess advanced skills across a wide range of domains, including high-stakes scenarios such as medicine [123]. However, in the field of radiology report generation, where relatively rich datasets are available, there has been *inconclusive* and even *contradictory* evidence regarding the performance of LMMs. Some studies [72, 123] claimed that GPT-4V performs well to some extent based on case studies and qualitative analysis. In contrast, Brin et al. [16] found that the model is not yet a reliable tool for radiological image interpretation on a small private dataset. Wu et al. [121] observed that GPT-4V can generate structured reports with incorrect content, as evidenced by case studies and qualitative analysis. Moreover, existing evaluation works tend to work with either very small size of samples [123] or limited evaluation metrics [66]. Our work distinguishes itself by providing an in-depth evaluation and analysis on *why* GPT-4V fails at this task. [2]

[1] `https://huggingface.co/spaces/WildVision/vision-arena`.

[2] We access GPT-4 vision model series (inlcuding gpt-4-vision-preview and gpt-4o) through Azure OpenAI service to prevent sharing data with third parties. Due to limited space, we mainly show evaluation results of GPT-4o in the main paper. Throughout the paper, we use GPT-4V to refer to GPT-4o, unless otherwise specified. Full evaluation results of GPT-4o, GPT-4-vision-preview, and the open sourced Llama3.2-90B-vision-instruct

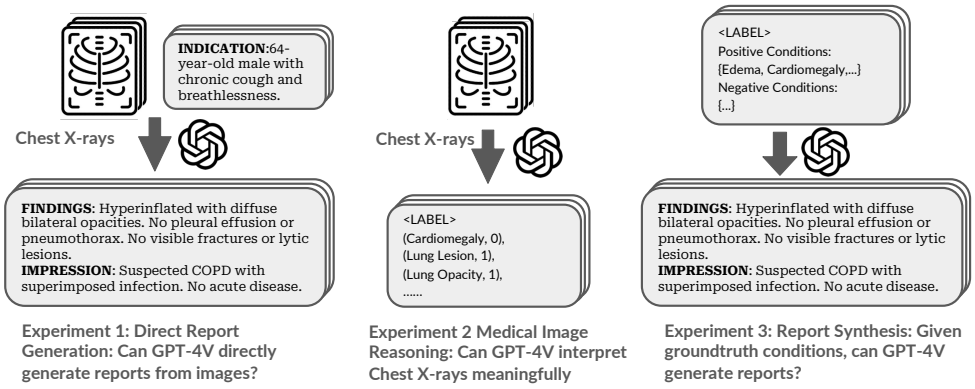


Figure 4-1: Evaluation overview. In Experiment 1, we evaluate the out-of-box capability of GPT-4V. We further decompose the task into medical image reasoning (Experiment 2) and report synthesis (Experiment 3).

To do that, we perform three experiments as shown in Fig. 4-1 on three main radiology report generation benchmarks: MIMIC-CXR, CheXpert Plus, and IU X-Ray. Our evaluation starts with Experiment 1: **direct report generation**. Different from previous works [72, 123], we conduct a thorough evaluation of GPT-4V's capability to directly generate reports from chest X-rays, utilizing different prompting strategies and assessing both lexical metrics, which measure how textually similar a generated report is to a reference report, and clinical efficacy metrics, which measure how clinically accurate it is. We experiment with various prompting strategies, including zero-shot, contextual enhancement, chain-of-thought (CoT) [120], and few-shot in-context learning. Despite our various attempts, the performance of GPT-4V is consistently low in both metrics.

To further investigate the reason for GPT-4V's poor performance, we break down report generation into two steps, **medical image reasoning** and **report synthesis given medical conditions**. For Experiment 2 (medical image reasoning), we first test whether GPT-4V can identify medical conditions from X-rays. Our findings indicate that GPT-4's performance in identifying medical conditions from images is unsatisfactory across different prompts. Based on limited capability results, we further compare the difference between distributions of predicted medical condition labels conditioned on different groundtruth image labels. We find that GPT-4V cannot interpret medical images meaningfully as the distribution of predicted labels does not vary depend on the groundtruth label.

---

can be found in Appendix C.3.

Finally, in Experiment 3 (report synthesis), we explore whether bypassing the image reasoning bottleneck by providing groundtruth conditions enables GPT-4V to generate clinically usable reports. As expected, reports generated by GPT-4V achieve higher clinical efficacy; however, the limited improvement in lexical metrics suggests that GPT-4V-generated reports remain dissimilar to human-written reports in style. Most importantly, GPT-4V underperforms a finetuned Llama-2 in both lexical metrics and clinical efficacy metrics, calling into question its utility. We further validate our findings by conducting an additional human reader study with two radiologist to evaluate the clinical viability of GPT-4V generated reports.

In summary, our key contributions and conclusions are as follows:

- We perform the first systematic and in-depth evaluation to benchmark GPT-4V in radiology report generation. Our main conclusion is that GPT-4V cannot generate radiology reports yet.
- To understand the poor performance, we decompose the task into medical image reasoning and report synthesis. We find that GPT-4V cannot interpret chest X-ray images meaningfully in the image reasoning step, and further validate this finding through rigorous hypothesis testing.
- During report synthesis, we address the image reasoning bottleneck by providing groundtruth conditions. Nonetheless, both experimental results and human evaluations consistently show that GPT-4V performs worse than a finetuned Llama-2 baseline.

4.2 Experimental Setup

Method. In Experiment 1 (Section 4.3), we evaluate GPT-4V's ability to directly generate radiology report given chest X-ray images. We consider five variations of prompts as outlined in Table 4.1. Prompt 1.1 (Basic generation) is a prompt to test the out-of-the-box capability of GPT-4V. We implement three additional prompting strategies leveraging insights in prompt engineering: (1) inspired by [82], we add relevant contextual information (i.e., the INDICATION) to derive Prompt 1.2 as "Indication enhancement", and add instructions on

Experiment 1: Direct Report Generation	
Prompt 1.1 Basic generation	Direct report generation based on chest X-ray images
Prompt 1.2 +Indication	Contextual enhancement by providing the indication section
Prompt 1.3 +Instruction	Contextual enhancement by providing instructions on medical conditions
Prompt 1.4 Chain-of-Thought (CoT)	Step 1 - medical condition labeling; Step 2 - report synthesis
Prompt 1.5 Few-shot	Few-shot: in-context learning given a few examples
Experiment 2: Medical Image Reasoning Capability	
Prompt 2.1 Image reasoning	Medical condition labeling directly from chest X-ray images
Experiment 3: Report Synthesis Given Medical Conditions	
Prompt 3.1 Report synthesis	Report generation using provided positive and negative conditions

Table 4.1: An index to prompts used in all of our experiments.

medical condition labels to Prompt 1.3 as "+instruction" enhancement; (2) we use a chain-of-thought (CoT) strategy in Prompt 1.4, eliciting the model with two steps: medical condition label prediction based on images followed by report synthesis based on the predicted labels; (3) We adopt few-shot in-context learning by adding a few example image-report pairs in Prompt 1.5. We compare these results with the state-of-the-art (SOTA) models.

In addition to evaluation of the end-to-end radiology report generation capability, we further evaluate on the decomposed tasks: Experiment 2 (Section 4.3): chest X-ray image reasoning; and Experiment 3 (Section 4.3): synthesizing a radiology report from given conditions. This decomposition allows us to look into the bottlenecks in the current generation performance. In Experiment 2, we prompt the model to directly output medical condition labels from images (Prompt 2.1). In Experiment 3, we bypass image reasoning to test GPT-4V's textual synthesis ability and provide groundtruth conditions to evaluate the model's report composition capability independently (Prompt 3.1). To contextualize the performance of GPT-4V, we also report the performance of a finetuned Llama-2-7B on groundtruth labels and groundtruth impressions following Alpaca [112].

Dataset and pre-processing. We use three chest X-ray datasets: MIMIC-CXR, IU X-ray and CheXpert Plus. The MIMIC-CXR dataset [49] contains chest X-ray images and their corresponding free-text radiology reports. The dataset includes 377,110 images from 227,835 studies. Each study has one radiology report and one or more chest X-rays. The IU X-raydataset [29] (also known as "Open-i") includes 3996 de-identified radiology reports and 8121 associated images from the Indiana University hospital network. CheXpert Plus

dataset [23] is a newly-released and comprehensive radiology data collection, featuring a total of 223,462 unique pairs of radiology reports and chest X-rays across 187,711 studies from 64,725 patients. For our evaluation, we use the whole validation set of CheXpert Plus (200 samples) and randomly sample 300 studies from each of the MIMIC-CXR and IU X-RAY datasets after removing studies with empty impression or indication sections. This results in a total of 800 samples. More details about data processing can be checked in Appendix C.2.

Evaluation metrics. We evaluate the generated reports from two aspects:

- Lexical metrics. Lexical metrics focus on the surface form and the exact word matches between the generated and reference texts. We adopt common lexical metrics: BLEU [89] (1-gram and 4-gram), ROUGE-L [68], and METEOR [8].

- Clinical efficacy metrics. We first evaluate on **clinical correctness** based on labeler results on generated reports. Following existing works [43, 115, 82], we use the CheXbert automatic labeler [109] to extract labels for each of 14 Chexpert medical conditions [44]. We compute both positive F1 and negative F1, where each condition has four labels: `present, absent, uncertain, unmentioned`. Positive F1 considers only positive labels against all others, while negative F1 considers negative labels as 1 and all other labels as 0. We report the macro-averaged F1 on all 14 conditions and on top 5 conditions (which only reports on the five most common conditions[3]). We also report RadGraph F1 [47], which captures the overlap in clinical entities and relations between a generated report and a reference report.

  Additionally, from a **pragmatic** viewpoint, commenting on negative observations is essential in radiology reports. Following Nguyen et al. [82], we compute Negative F1 and Negative F1-5, to evaluate whether the model can accurately identify negative conditions and include that in the generated reports. All F1 scores are macro-averaged. We also use the **hallucination** metric to quantify the proportion of uninferable information. Following Nguyen et al. [82], we define uninferable information to include previous studies, previous

[3]Top five conditions are Pneumothorax, Pneumonia, Edema, Pleural Effusion, and Consolidation.

Experiment	Lexical metrics				Clinical Efficacy Metrics					
	BLEU-1	**BLEU-4**	**ROUGE**	**METEOR**	**Pos F1**	**Pos F1@5**	**Rad. F1**	**Neg F1***	**Neg F1@5***	**Hall.*↓**
MIMIC-CXR										
Basic	0.331	0.031	0.225	0.282	0.134	0.153	0.163	0.038	0.081	0.587
+Indication	0.333	0.044	0.234	0.290	**0.273**	0.255	0.182	**0.048**	**0.126**	0.547
+Instruction	0.276	0.028	0.203	0.264	0.155	0.230	0.154	0.042	0.101	**0.359**
CoT	0.242	0.017	0.186	0.201	0.172	**0.272**	0.119	0.008	0.020	0.397
Few-shot	**0.337**	**0.055**	**0.257**	**0.301**	0.170	0.203	**0.188**	0.035	0.091	0.477
Llama-3.2[1]	0.258	0.026	0.249	0.216	0.172	0.300	0.144	0.045	0.117	0.049
SOTA	0.402	0.142	0.291	0.333	0.473	0.516	0.267	0.077	0.156	0.158
[ref.]	Liu et al.	Hyland et al.	Liu et al.	Hyland et al.	Liu et al.	Tu et al.	Tu et al.	Nguyen et al.	Nguyen et al.	Nguyen et al.
Δ(to SOTA)	-16.17%	-61.27%	-11.68%	-9.61%	-42.28%	-47.29%	-29.59%	-37.66%	-19.23%	20.10%
IU X-RAY										
Basic	0.316	0.045	0.238	0.311	0.059	0.045	0.203	0.000	0.000	0.303
+Indication	**0.330**	**0.049**	0.242	**0.323**	0.077	0.098	**0.214**	**0.071**	**0.051**	0.307
+Instruction	0.238	0.030	0.207	0.283	**0.081**	0.146	0.174	0.000	0.000	**0.177**
CoT	0.239	0.024	0.194	0.231	0.077	**0.161**	0.144	0.000	0.000	0.197
Few-Shot	0.279	0.044	**0.243**	0.250	0.037	0.031	0.187	0.010	0.025	0.211
Llama-3.2	0.248	0.027	0.239	0.231	0.104	0.214	0.158	0.015	0.040	0.011
SOTA	0.499	0.184	0.390	0.208	-	-	-	-	-	-
[ref.]	Liu et al.	Liu et al.	Liu et al.	Liu et al.						
Δ(to SOTA)	-36.27%	-75.82%	-38.72%	+49.52%	-	-	-	-	-	-
CHEXPERT PLUS										
Basic	**0.237**	**0.015**	**0.176**	**0.191**	0.228	0.191	**0.112**	0.013	0.035	0.680
+Instruction	0.191	0.007	0.159	0.172	0.210	0.325	0.101	0.042	0.085	0.377
CoT	0.166	0.011	0.155	0.139	**0.234**	**0.339**	0.077	0.000	0.000	0.400
Few-shot	0.171	0.007	0.158	0.149	0.188	0.224	0.094	**0.043**	**0.111**	**0.370**
Llama-3.2	0.166	0.006	0.175	0.147	0.261	0.355	0.092	0.031	0.081	0.058
SOTA	-	0.069	0.279	-	0.366	0.495	0.285	-	-	-
[ref.]		Chambon et al.				Chambon et al.				
Δ(to SOTA)	-	-78.26%	-36.92%	-	-36.07%	-31.52%	-60.70%	-	-	-

* To compare with SOTA numbers, all metrics, except for those marked with * (Neg F1, Neg F1@5, and Hall), are evaluated on the findings section. * columns are based on the impression section.

CheXpert Plus doesn't have indication section in reports, thus we skip experiment with +Indication prompt.

[1] Due to the space limit, we only show the best results of prompt 1.1-1.5 for Llama-3.2-90B-Vision-Instruct. It is noted that the low hallucination rate is likely because it only outputs medical conditions (mostly wrong but not uninferable).

Full details of the performance of GPT-4o, GPT-4-vision-preview, and Llama3.2-90B-Vision-Instruct, including results for both the findings and impression sections, are provided in the Appendix C.3.

Table 4.2: Direct report generation performance comparison. GPT-4V shows a significant performance gap compared to SOTA, and the results are consistent across the five prompting strategies. Open sourced Llama3.2 performs similarly compared with GPT-4V. Examples of generated reports across different prompts can be found in Appendix C.3.5.

treatment details, recommendations, doctor communications, and image view descriptions.

4.3 Results

Experiment 1: Can GPT-4V directly generate reports from images?

We first evaluate the out-of-the-box capability of GPT-4V in generating radiology reports from chest X-ray images using basic generation (Prompt 1.1). Table 4.2 shows the results compared with existing state-of-the-art (SOTA) models. Overall, GPT-4V significantly underperforms the state-of-the-art models on both lexical and clinical efficacy metrics, with the exception of the METEOR score on the IU X-RAY dataset. The relatively better METEOR performance is due to its comprehensive evaluation criteria, which include synonymy and paraphrasing, not just exact word matches like BLEU and ROUGE. This allows METEOR to recognize semantic equivalents, even if the word choice differs. In other words, the generated report somewhat resembles a radiology report, although it fails at the exact word-level matching. For clinical efficacy metrics, the gaps to SOTA are consistently large. This suggests that GPT-4V struggles to accurately identify conditions in its generated reports from images alone.

Our results are consistent across prompting strategies. Our prompting strategies include adding contextual information, chain-of-thought reasoning, and few-shot prompting. While indication enhancement (Prompt 1.2) provides indication section as input in addition to chest X-rays and improves many metrics for both MIMIC-CXR and IU X-RAY, it remains within the same range and does not significantly reduce the gap compared to SOTA. Instruction enhancement (Prompt 1.3) provides medical condition descriptions and makes a moderate yet still limited difference to SOTA in Positive F1 scores and Hallucination. Following the same labeling instructions, Chain-of-Thought (Prompt 1.4) similarly increases Positive F1-5 by 11.9% in MIMIC-CXR, 11.6% in IU X-RAY and 14.8% in CheXpert Plus, marking the most effective advances so far. However, it still faces a substantial gap to SOTA, with 47.29% in MIMIC-CXR and 31.52% in CheXpert Plus. Few-Shot (Prompt 1.5) provides image-report pairs as context and generally improves lexical metrics, RadGraph F1, and Hallucination, while clinical correctness, particularly in identifying positive conditions,

Metric	Chain-of-Thought (1st Step)	Image Reasoning
Positive F1	0.195	0.161
Positive F1@5	0.298	0.242

Table 4.3: Image reasoning performance of GPT-4V on MIMIC-CXR. The model performs poorly in identifying medical conditions from chest X-ray images.

remains consistently low across three datasets. This indicates that while few-shot prompting might help GPT-4V mimic the format of groundtruth reports, it still falls short in generating accurate reports.

Experiment 2: Can GPT-4V interpret chest X-rays meaningfully?

In this section, we probe GPT-4V's ability to reason about chest X-ray images alone. Specifically, we evaluate whether the model can meaningfully interpret chest X-ray images by measuring how accurately GPT-4V can label medical conditions present (positive F1). Table 4.3 provides an overview of GPT-4V's labeling performance under different prompting strategies.

We can see that GPT-4V cannot accurately specify positive conditions from given chest X-rays. This can be highlighted by consistently poor Positive F1 scores observed from various prompting strategies. Furthermore, this inability to accurately interpret images may directly contribute to GPT-4V's failure in generating high-quality reports, as supported by similar Positive F1 score of 0.172 and Positive F1-5 score of 0.272 from the report synthesis phase of Chain-of-Thought (see Table 4.2), compared to 0.195 (Positive F1) and 0.298 (Positive F1-5) from the initial label generation phase of Chain-of-Thought.

Overall, these results indicate GPT-4V's limited ability in identifying medical conditions from chest X-ray images, regardless of whether labels are derived from CoT 1st step or direct prompting.

Testing whether GPT-4V generates labels based on given chest X-rays. Considering the failure of GPT-4V to accurately label medical conditions, we would like to investigate to what extent can GPT-4V predict meaningful labels given a specific chest X-ray image.

To test this, we group chest X-rays by their groundtruth conditions and then analyze the generated label distribution for each group. If the label distributions are similar across different condition groups, it would suggest that GPT-4V is not meaningfully identifying labels from the chest X-rays but rather assigning labels randomly without proper image interpretation. For example, if the model's generated label probabilities are roughly the same regardless of whether the groundtruth condition of the given image is Edema or Cardiomegaly, it indicates a limited capability in medical image reasoning.

Formally, let X_{ij} be a binary random variable that takes the value 1 if GPT-4V labels the j-th condition as positive for the chest X-ray image associated with the i-th study, and 0 otherwise, where $i = 1, 2, \ldots, 300$ and $j = 1, 2, \ldots, 13$. We exclude the "No Findings" condition from this study. We define $Y_j = \sum_{i=1}^{300} X_{ij}$ as the sum of positive mentions for the j-th condition across all 300 studies, and $\mathbf{Y} = [Y_1, \ldots, Y_{13}]$ as the count vector. Next, we categorize the study pool into 13 condition groups, where group k consists all studies that are ground truth positive for the k-th condition based on the associated radiology report. Note that there might be overlaps between these groups, as a single study can be positive for multiple conditions. For each group k, GPT-4V's labeling process given the chest X-ray image from i-th study can be modeled as:

$$\begin{cases} X_{ij}^{(k)} \sim \text{Bernoulli}(P_j^{(k)}) \\ \qquad\qquad \text{for } i \in \text{group } k \text{ and } j = 1, \ldots, 13 \\ \mathbf{Y}_k \sim \text{Multinomial}(n_k; \mathbf{P}_k) \\ \qquad\qquad \text{with } \mathbf{P}_k = [P_1^{(k)}, \ldots, P_{13}^{(k)}] \end{cases} \tag{4.1}$$

where n_k is the number of studies in group k, and $P_j^{(k)}$ is the probability that GPT-4V labels the j-th condition as positive for the chest X-ray images associated with the studies in group k.

We first use a χ^2**-test** to test if GPT-4V follows the same label distribution across different groups, i.e., testing the null hypothesis (H_0) that $\mathbf{P}_k = \mathbf{P}_{k'}$ for any groups k and k'. Additionally, we use **bootstrap confidence interval** [27] to test if GPT-4V labels one certain condition independently of the groundtruth condition group. Specifically, we test the

Statistics	Overall		Top 5 Conditions	
	Groundtruth	GPT-4o	Groundtruth	GPT-4o
χ^2 statistic	1770.38	66.05	243.51	5.30
p-value	p < 1e-4	1.00	p < 1e-4	0.994
df.	144	144	16	16

Table 4.4: χ^2-test for homogeneity of label distribution across different condition groups. When p-value is smaller than 0.0001, at 0.01% significance level, we can reject the null hypothesis that different groups follow the same label distribution.

null hypothesis (H_0) that $P_j^{(k)} = P_j$ for any condition j and group k. More test details and robustness check can be found in Appendix C.3.1.

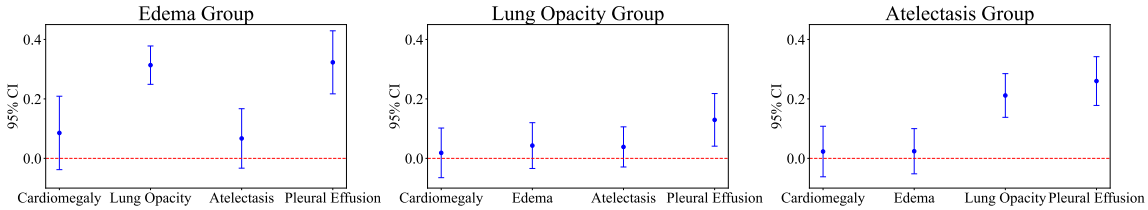


Figure 4-2: 95% Bootstrap confidence interval of example 3 conditions for MIMIC-CXR. When zero falls into the interval, at 95% confidence level, we cannot reject the null hypothesis that GPT-4V labels j-th condition independent of which condition group this study belongs to.

Experiment	Lexical metrics				Clinic Efficacy Metrics					
	BLEU-1	BLEU-4	ROUGE	METEOR	Pos F1	Pos F1@5	Rad. F1	Neg F1	Neg F1@5	Hall.↓
GPT-4o	0.159	0.006	0.142	0.185	0.123	0.152	0.077	0.038	0.081	0.587
GPT-4o (gt)	0.175	0.009	0.187	0.183	0.879	0.972	0.105	0.639	0.956	**0.427**
GPT-4-vision-preview (gt)	0.176	0.007	0.185	0.179	0.885	**0.977**	0.103	0.584	**0.958**	0.431
Finetuned Llama-2 (gt)	**0.301**	**0.094**	**0.330**	**0.348**	**0.923**	0.957	**0.286**	**0.703**	0.941	0.710

Table 4.5: Performance in report generation with groundtruth conditions. Although GPT-4V's performance improves significantly, it still underperforms finetuned Llama-2, especially in matching the writing style of groundtruth reports.

Table 4.4 presents χ^2-test results for the homogeneity of label distribution across different groups. For both the overall and top 5 conditions[4], at 0.01% significance level, we can both reject the null hypothesis for groundtruth reports that different groups follow the same label distribution, but not for GPT-4V's generated reports.

[4]Due to the sparsity of the original study pool, we report results for two different tables: (1) A modified table with zero elements replaced by 0.001; (2) A reduced table with only five most frequent medical conditions in the subsample.

Figure 4-2 illustrates the 95% bootstrap confidence intervals for top 5 conditions[5]. If zero falls within the interval, we cannot reject the null hypothesis that GPT-4V labels the j-th condition independently of the condition group at 95% confidence level. The figure shows that, in 7 out of 12 cases (58.3%), we cannot reject the null hypothesis.

In summary, the results show that GPT-4V labels conditions independently of the groundtruth condition, and there is no significant difference in label distributions across groups in GPT-4V's generated reports, unlike the groundtruth reports.

Experiment 3: Given groundtruth conditions, can GPT-4V generate reports?

Given that GPT-4V cannot perform image reasoning, we investigate whether GPT-4V can generate high-quality radiology reports given groundtruth medical conditions. We conduct an experiment on report synthesis (Prompt 3.1) and use a finetuned Llama-2 model as a baseline for comparison.

Table 4.5 shows that while using groundtruth conditions significantly enhances GPT-4V's clinical accuracy, it still does not perform as well as the finetuned Llama-2, particularly in matching the content of groundtruth reports. Progress in clinical accuracy is evidenced by large improvements in F1 scores compared to basic generation (Prompt 1.1). However, there are only minor changes in lexical metrics and RadGraph F1, which focus on entity relation matching in groundtruth reports, along with consistently large gaps with finetuned Llama-2, suggest that groundtruth conditions are insufficient to align GPT-4V's writing closely with that of groundtruth reports. The higher scores of the finetuned Llama-2 in lexical metrics also indicate that finetuning open models is an effective way to leverage existing datasets.

Human Evaluation. To further evaluate the quality of GPT-4V-generated reports beyond automatic metrics, we collaborate with two experts: one board-certified and one board-eligible radiologist, to conduct a human evaluation. From our testing set of 300 studies, we randomly select 50 cases for blind human evaluation. Each radiologist is provided

[5]Due to the sparsity of the original study pool, we limit our analysis to the five most frequent medical conditions in our subsample.

	Binary	Likert Scale (1-5)		
	Clinically Usable	**Diagnostic Accuracy**	**Completeness**	**Clarity/Readability**
Groundtruth	50/50 (100%)	**4.86**	**4.87**	4.92
Finetuned Llama-2	46/50 (92%)	4.16	4.42	**4.94**
GPT-4-vision-preview	36/50 (72%)	3.63	3.49	3.80

All metrics are evaluated on the impression section and are averaged based on results from two experts.

Table 4.6: Human evaluation of radiology report quality. From the perspective of radiologists, GPT-4V underperforms relative to the finetuned Llama-2, particularly in fine-grained likert scale metrics of diagnostic accuracy, completeness, and clartiy/readability.

Groundtruth labels	**Model**	**Report**	**Usable**	**Acc.**	**Comp.**	**Clarity**
Positive: Enlarged Cardiomediastinum, Lung Opacity, Atelectasis Negative: Pneumonia	Groundtruth	In comparison with study of ___, there is little change in the cardiomediastinal silhouette and pacer leads. Continued elevation of the right hemidiaphragmatic contour. Opacification medially above the elevated hemidiaphragm most likely represents atelectatic changes. No definite acute focal pneumonia.	✓	5	5	4
	Finetuned Llama-2	1. No evidence of pneumonia. 2. Stable postoperative appearance of the chest. 3. Increased retrocardiac opacity likely reflects atelectasis.	✓	5	5	5
	GPT-4-vision-preview	The chest radiograph demonstrates an enlarged cardiomediastinum, lung opacity which may be due to various etiologies, and atelectasis. There is no clear evidence of pneumonia on the radiograph.	✓	3	4	3

Table 4.7: Comparison of example reports from human annotation.

with anonymized chest X-ray images and randomly ordered IMPRESSION sections from groundtruth reports, as well as reports generated by Llama-2 and GPT-4V. Both Llama-2 and GPT-4V are prompted with groundtruth medical conditions. The evaluation involves a detailed review of three reports per study case, assessing each report's clinical usability with a binary label as the first step. Then, the radiologist rates each report on two dimensions: clinical efficacy (diagnostic accuracy and completeness) and lexical performance (clarity/readability). Reports are rated on a Likert scale, where a score of 5 denotes superior performance and a score of 1 denotes poor performance. We compute and report the average

scores for each metric across different report types.

Table 4.6 shows that, from the perspective of radiologists, GPT-4V still underperforms the finetuned Llama-2. Groundtruth reports are indeed of high quality, rated as clinically usable in 50 out of 50 cases. However, a significant usability gap is observed between Llama-2 and GPT-4V, with Llama-2 being deemed clinically usable in 46 out of 50 cases, compared to 36 out of 50 for GPT-4V. Furthermore, Llama-2 consistently outperforms GPT-4V across all other Likert scale metrics, especially in completeness and clarity/readability.

Table 4.7 presents three example reports. While groundtruth reports offer detailed clinical insights and varied descriptors, GPT-4V tends to provide vague statements, only stating "lung opacity which may be due to various etiologies" without specifying its location, severity, or offering a differential diagnosis. Llama-2 performs slightly better by offering some specific diagnoses, yet still lacks detailed descriptions.

In summary, human annotation corroborates with our findings from Experiment 3. Given groundtruth conditions, GPT-4V generated reports still lack comprehensive coverage of all relevant clinical findings and do not effectively summarize and organize medical conditions, compared with human-written reports.

4.4 Conclusion

We perform the first systematic and in-depth evaluation of the GPT-4V series models in radiology report generation using three chest X-ray benchmarks. We find that GPT-4V cannot generate radiology reports, even with different prompting strategies. Open-sourced models like Llama-3.2 vision perform similarly poorly compared to GPT-4V. Both closed-source and open-source models exhibit a significant gap when compared to specialized SOTA radiology report generation models. To understand the low performance, we decompose the task into image reasoning and report synthesis. The results demonstrate that GPT-4V struggles significantly with interpreting chest X-rays meaningfully, which directly impacts its ability to generate reports. Furthermore, even when we bypass this problem by providing groundtruth conditions, GPT-4V still underperforms a finetuned Llama-2 baseline and

consistently fails to replicate the writing style of groundtruth reports or meet the preferences of radiologists. Overall, our study highlights substantial concerns regarding the feasibility of integrating GPT-4V into real radiology workflows.

Chapter 5

CLEAR: A Clinically-Grounded Tabular Framework for Radiology Report Evaluation

5.1 Introduction

Evaluation is becoming increasingly challenging in the era of large language models (LLMs). While models continue to hill-climb on benchmarks rapidly [75, 86, 5, 116, 77], it remains unclear whether these reported metrics match task-specific needs [34, 98, 12]. In the context of radiology, the pursuit of generalist foundation models achieves promising progress [9, 127], but do these appealing automated metrics truly capture clinically aligned qualities [92]?

In the existing literature, three main types of metrics have been proposed to assess the quality of generated radiology reports, as illustrated in Figure 5-1: (i) **Lexical metrics** measure surface-level similarity between the generated and ground-truth reports [89, 68, 128]. While straightforward and easy to compute, they struggle to capture nuanced semantics and domain-specific terminology, leading to poor sensitivity to clinically significant errors. (ii) **Clinical efficacy metrics** evaluate the correctness of medical entities and their relationships [47, 125, 130], typically through structured extraction-based comparisons. Although more clinically informed than lexical metrics, they lack the resolution to assess fine-grained attributes such as severity, temporal progression, or treatment recommendations. (iii) **LLM-based metrics** [87, 42, 127] represent the latest direction, often leveraging the pipeline of LLM-as-a-Judge [131] with pre-defined taxonomies such as the six error categories from ReXVal dataset [126]. While getting closer to expert judgment compared with previous two

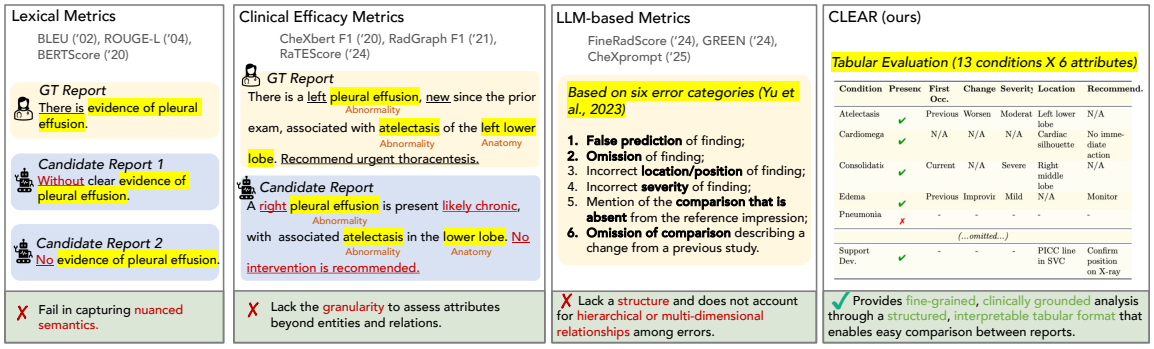


Figure 5-1: A comparison of existing metrics with CLEAR. Yellow highlights indicate the main evaluation mechanism for each type of metric. Red underlining marks an erroneous term in the candidate report, in contrast to the black underlined term in the ground-truth report, which the designed metric fails to evaluate.

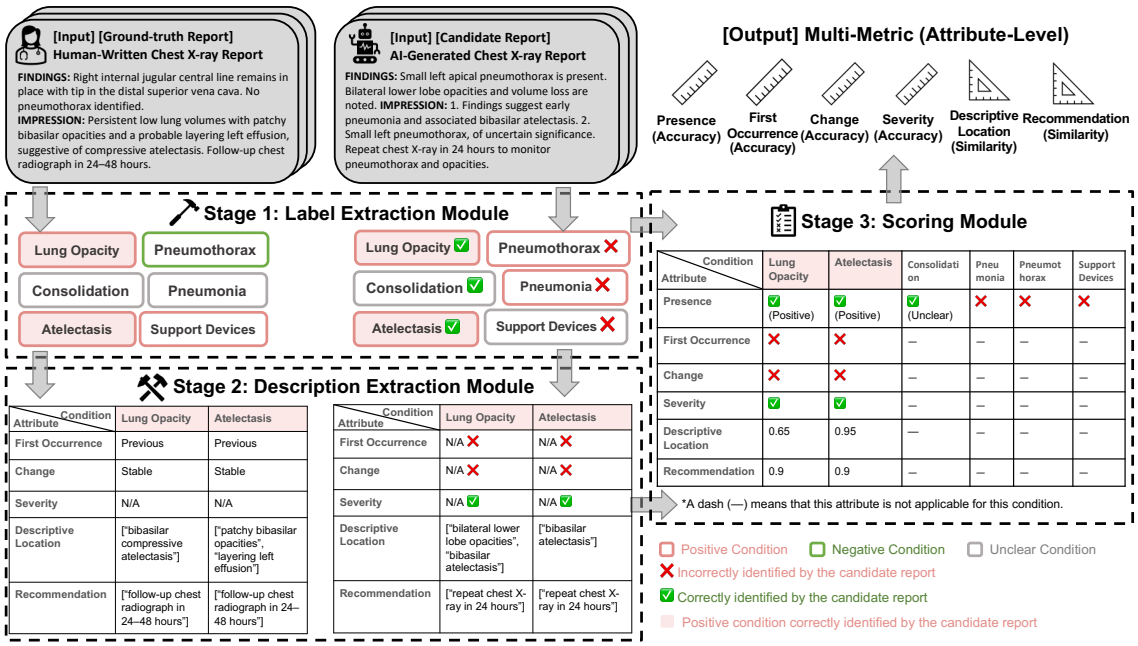


Figure 5-2: **CLEAR Framework.** Given a pair of ground-truth and candidate reports, we first assesses whether the candidate report can accurately identify a set of medical observations in the **label extraction module**. For each correctly identified positive condition, the **description extraction module** further evaluates the report's ability to describe the condition across five attributes: `first occurrence`, `change`, `severity`, `descriptive location`, and `recommendation`. Finally, the **scoring module** compiles and outputs the evaluation metrics.

types, these methods may still lack comprehensive structured attribution and condition-level interpretability.

Therefore, to address the limitations of existing metrics, we introduce **CLEAR** (Section 5.2), the first clinically-grounded attribute-level evaluation framework that leverages

LLMs to map free-text radiology reports to a structured tabular format. Compared to prior work, CLEAR transforms the coarse, single-dimensional taxonomy into a fine-grained, multidimensional structure. Our design not only enables more comprehensive comparisons between candidate and ground-truth reports, but also provides interpretable outputs to assess report quality at the level of condition-attribute pairs. Given the strong adaptability of LLMs across diverse language tasks, they serve as an ideal unified model to operationalize our proposed framework.

Specifically, CLEAR begins with the **Label Extraction Module** (Section 5.2), which evaluates whether the candidate report can precisely identify the presence or absence of specific medical conditions. To ensure robust performance across model scales, we enhance this module using high-quality, expert-curated labels. Next, for each correctly identified positive condition, the **Description Extraction Module** (Section 5.2) assesses whether the candidate report can accurately describe the condition. Jointly established with one research radiologist and reviewed by one clinical radiologist, we define five commonly used attributes in a radiology report (`first occurrence`, `change`, `severity`, `descriptive location`, and `recommendation`), enabling the first systematic evaluation of these critical facets. Finally, the **Scoring Module** (Section 5.2) compiles and outputs metric scores for each attribute. We carefully design automated measurements based on the output type from previous modules: accuracy metrics aim at exact matches for single-label outputs while similarity metrics focus on contextual relevance for multi-phrasing outputs.

Additionally, since no existing datasets [113, 126, 97] are compatible with CLEAR, we work closely with radiologists to create **CLEAR-Bench** (Section 5.3), an expert-curated, attribute-level dataset to assess clinical alignment. CLEAR-Bench consists of 100 studies randomly sampled from MIMIC-CXR-JPG test and validation sets [49, 48]. Each study is annotated and reviewed by at least two radiologists across 6 report attributes and 13 CheXpert conditions[1] [44]. CLEAR-Bench includes two components: (i) **Expert ensemble labels** includes ground-truth labels for `presence` attribute of each condition. These labels are constructed via majority voting among three radiologists, followed by one round of

[1]Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax, and Support Devices.

Attribute	Value Set	NLP Task	Metric
Presence	$S_1 \in$ {"Positive", "Unclear", "Negative"}	Cls (Prompt 1)	Accuracy
Temporal Assessment			
First Occurrence	$S_2 \in$ {"Previous", "Current", "N/A"}	QA (Prompt 2)	Accuracy
Change	$S_3 \in$ {"Improving", "Stable", "Worsening", "Mixed", "N/A"}	QA (Prompt 3)	Accuracy
Description Assessment			
Severity	$S_4 \in$ {"Severe", "Moderate", "Mild", "Mixed", "N/A"}	QA (Prompt 4)	Accuracy
Descriptive Location	$S_5 = \{\text{Entry}_1, ..., \text{Entry}_n\}$ (e.g., Entry_m = "left mid lung atelectasis")	IE (Prompt 5)	Similarity
Treatment Assessment			
Recommendation	$S_6 = \{\text{Entry}_1, ..., \text{Entry}_n\}$ (e.g., Entry_m = "recommend follow-up at 4 weeks")	IE (Prompt 6)	Similarity

* Cls denotes "Classification," QA denotes "Question Answering," and IE denotes "Information Extraction."

Table 5.1: An overview of our expert-curated fine-grained attributes in CLEAR.

consensus discussion. (ii) **Expert curated attributes** contains the remaining five report attributes for each condition positively identified in the ensemble labels. These attributes are first generated by LLMs, then independently curated by two radiologists, and finalized through one round of discussion and resolution. Additionally, during the curation process, we collect expert Likert scores for each model output, contributing to the assessment of how well proposed automated metrics align with clinical judgment.

Finally, we evaluate each component of CLEAR using the CLEAR-Bench. Our experimental results (Section 5.4) show that: (i) the Label Extraction Module achieves high accuracy compared to expert ensemble labels and significantly outperforms existing labelers across all metrics; (ii) the Description Extraction Module can accurately extract attribute-level information according to clinical assessment; (iii) our proposed automated metrics serve as effective proxies for expert scoring.

5.2 CLEAR Framework

We introduce the CLEAR framework, a hierarchical and fine-grained system for evaluating the clinical accuracy of radiology reports. CLEAR addresses both high-level diagnostic correctness and the descriptive quality of positive findings. As shown in Figure 5-2, CLEAR includes three sequential stages: label extraction, description extraction, and structured scoring.

Specifically, given a ground-truth and a candidate report pair, CLEAR first identifies whether the candidate correctly recognizes the presence or absence of specific medical conditions (Stage 1). It then examines, for each positively identified condition, whether the ground-truth and candidate reports are aligned across a set of expert-curated descriptive dimensions (Stage 2). Finally, it aggregates these evaluations into standardized, multi-dimensional metrics (Stage 3).

Stage 1: Label Extraction

This stage determines the presence or absence of 13 pre-defined medical conditions in the candidate report, following the CheXpert structure [44]. Since accurately identifying and describing abnormalities is more clinically significant in radiology reporting, we exclude the "No Findings" label and focus on the remaining 13 conditions. Each condition is labeled as `positive`, `unclear`, or `negative` based on report content.

While existing labelers like CheXbert [108] and CheXpert [44] are available, our pilot analysis (see Table 5.2) showed that their performance was limited. Since label extraction involves understanding and interpreting clinical narratives to assign structured labels, we hypothesized that LLMs could offer significant improvements over existing approaches. In particular, LLMs can handle complex linguistic nuances, such as negation, uncertainty, and context-dependent phrasing, more effectively in free-form radiology reports.

Base model variants and training strategies. We support three model scales: small (fine-tuned Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct), medium (Llama-3.3-70B-Instruct and Llama-3.1-70B-Instruct), and large (GPT-4o). For medium and large models, we apply different prompting strategies, including zero-shot (Prompt 1) and five-shot. For small models, we perform full-parameter fine-tuning using our curated dataset. To avoid overfitting, we first conduct hyperparameter tuning through 5-fold cross-validation and a grid search over learning rate, gradient accumulation steps, and number of epochs, followed by re-training on the full dataset. Full implementation details are provided in Appendix D.3.

Expert-in-the-loop label curation. High-quality labeled data is essential for training our label extraction model. To build a gold training dataset, we implemented a multi-stage annotation refinement with expert in the loop. We began with the test set from MIMIC-CXR-JPG [48], which includes a single radiologist's annotations for 13 CheXpert conditions [44]. Each condition is originally labeled as `positive`, `negative`, `unmentioned`, or `uncertain`. In initial discussions with a radiologist, we identified two major issues with the original annotations: labeling errors (e.g., conditions mentioned in the report but left unlabeled) and category ambiguity (e.g., vague distinctions between `negative` and `unmentioned`). To address these, we used GPT-4o to pre-screen and re-label the reports, prompting it with the original MIMIC labeling guidelines. We then flagged cases with label mismatches between GPT-4o and the original annotations. We then asked an expert to re-annotate the discrepancy cases. To reduce the radiologist's workload, reports with more than five mismatched condition labels are discarded from expert annotation, as such extensive disagreement often signals deeper interpretive ambiguities or quality issues in the original reports. While this introduces potential bias, we prioritized curating a high-quality subset over exhaustively correcting all samples. For the remaining reports, our collaborating radiologist independently re-annotated only the discrepant conditions, reviewing the original report text without seeing prior labels. During human annotation process, we observed that the original labeling schema lacked sufficient granularity to reflect the nuanced certainty levels expressed in radiology. In discussion with our expert radiologist, we expanded the label set to:`{confidently present, likely present, neutral, likely absent, confidently absent}`. In total, we curated 550 studies, each with high-quality labels for all 13 conditions. For consistency with prior work and to simplify downstream modeling, we further merged all labels into three classes `{positive, negative, unclear}`. A detailed description of the annotation process and instructions are provided in Appendix D.2.

Stage 2: Description Extraction

Building on the condition labels from Stage 1, this module extracts fine-grained clinical features that capture essential descriptive information for accurate reporting. The primary

motivation is to transform the narrative text of radiology reports into a comprehensive, structured tabular format that distills all clinically significant attributes. In collaboration with two radiologists, we developed five clinically significant dimensions: `first occurrence` (whether the condition is newly observed), `change` (progression or improvement from prior studies), `severity` (the extent or intensity of the condition), `descriptive location` (specific anatomical site), and `recommendation` (suggested follow-up actions). These expert-developed attributes were specifically designed to reflect the nuanced but essential information radiologists routinely document when interpreting chest X-rays. By extracting these attributes, our approach enables a more comprehensive evaluation beyond simple condition detection.

Implementation details. We use prompt-based methods to extract each of the five attributes from free-text reports. Each attribute can be naturally framed as a standalone language understanding task. To operationalize this, we design custom prompts tailored to the nature of each attribute: we use a Question Answering (QA) template to prompt the model for `first occurrence` (Prompt 2), `change` (Prompt 3), and `severity` (Prompt 4), and an Information Extraction (IE) template for `descriptive location` (Prompt 5) and `recommendation` (Prompt 6). For QA tasks, the model selects the best answer from multiple-choice options based on its understanding of the report. For IE tasks, it extracts relevant phrases guided by condition-specific example terminologies. Our prompt templates and terminology lists are summarized in Appendix D.4, and were reviewed by two radiologists. We use a single model to process all five prompt types, one prompt per query to extract each attribute from a given report. We evaluate two model scales: a smaller Llama-3.1-8B-Instruct and a larger GPT-4o from OpenAI.

Stage 3: Scoring and Metrics

In this module, we process outputs from Stage 1 and Stage 2 into numeric metrics for each attribute. Given the i-th pair of ground-truth and candidate attribute sets, denote the attributes extracted from the ground-truth report as $\{S_j^{(i)}\}_{j=1}^{6}$ and from the candidate report

as $\{\hat{S}_j^{(i)}\}_{j=1}^{6}$. An overview of the attributes is provided in Table 5.1.

For `presence` $(S_1, \hat{S}_1)$, we evaluate the accuracy of identifying `Positive` and `Negative` conditions. We define a target class $c \in \{\texttt{Positive}, \texttt{Negative}\}$, treating all other labels as non-target. The corresponding binary F1 score, $F1_c$, is computed for each target class, resulting a positive-F1 and negative-F1. We report these scores at three levels: micro average, Top-5 condition average[2], and across all 13 conditions.

For `first occurrence` $(S_2, \hat{S}_2)$, `change` $(S_3, \hat{S}_3)$, and `severity` $(S_4, \hat{S}_4)$, we assess the exact match between predictions and ground truth. Considering that these attributes are framed as multiple-choice questions in the prompt, exact match is a natural and appropriate metric. Accuracy is calculated as $\text{Acc.}_j = \frac{\sum_i \mathbb{1}[S_j^{(i)} = \hat{S}_j^{(i)}]}{\sum_i 1}$. We report accuracy at the micro level, as well as averaged across reports and the 13 conditions.

For `descriptive location` $(S_5, \hat{S}_5)$ and `recommendation` $(S_6, \hat{S}_6)$, which involve free-text descriptions, we measure phrase-level similarity against clinically meaningful expressions. To evaluate alignment, we first use optimal matching–based metrics with similarity scores such as BLEU-4 [89] and ROUGE-L [68]:

$$\text{Score}_j^{(i)} = \frac{1}{|S_j^{(i)}|} \sum_{e \in S_j^{(i)}} \max_{\hat{e} \in \hat{S}_j^{(i)}} \text{Similarity}(e, \hat{e}),$$

where $S_j^{(i)} = \{e_k\}_{k=1}^{n}$ and $\hat{S}_j^{(i)} = \{\hat{e}_k\}_{k=1}^{n'}$. Additionally, to better approximate clinical judgment from an expert's perspective, we prompt o1-mini (Prompt 8) to directly compare each attribute pair and return a similarity score in the range $[0, 1]$.

5.3 CLEAR-Bench: Attribute-Level Expert Alignment Dataset

In this section, we introduce CLEAR-Bench, an expert-curated, attribute-level dataset in collaboration with five radiologists. Inspired by recent expert evaluation datasets for chest X-ray reports [113, 126, 97], CLEAR-Bench is specifically designed to assess how

[2]Top five conditions in MIMIC-CXR-JPG are Pneumothorax, Pneumonia, Edema, Pleural Effusion, and Consolidation.

Experiments	Pos F1@13	Pos F1@5	Pos F1 (micro)	Neg F1@13	Neg F1@5	Neg F1 (micro)
LARGE MODELS						
GPT-4o (base)	**0.805**	0.929	0.934	0.476	0.648	0.815
GPT-4o (5-shot)	0.795	**0.940**	**0.934**	0.510	0.723	0.842
MEDIUM MODELS						
Llama-3.1-70B-Instruct (base)	0.782	0.890	0.924	0.630	0.850	0.920
Llama-3.1-70B-Instruct (5-shot)	0.794	0.916	0.924	**0.744**	0.890	**0.958**
Llama-3.3-70B-Instruct (base)	0.780	0.894	0.925	0.602	0.876	0.926
Llama-3.3-70B-Instruct (5-shot)	0.781	0.907	0.926	0.695	**0.892**	0.953
SMALL MODELS						
Llama-3.1-8B-Instruct (base)	0.736	0.880	0.910	0.418	0.660	0.714
Llama-3.1-8B-Instruct (550 finetune)	0.729	0.806	0.905	0.482	0.803	0.949
Qwen2.5-7B-Instruct (base)	0.694	0.834	0.880	0.413	0.616	0.736
Qwen2.5-7B-Instruct (550 finetune)	0.727	0.800	0.905	0.511	0.849	0.953
BASELINES						
CheXbert [108]	0.695	0.833	0.897	0.498	0.877	0.952
CheXpert [44]	0.674	0.811	0.888	0.522	0.831	0.948
Δ Improvement over SOTA	+15.8%	+12.8%	+4.1%	+42.5%	+1.7%	+0.06%

Table 5.2: Evaluation of the label extraction module. CLEAR outperforms existing labelers across all metrics in identifying both positive and negative conditions. Specifically, larger models perform better at capturing positive conditions, while techniques such as 5-shot prompting and supervised fine-tuning significantly improve the detection of negative conditions.

well automated evaluators like CLEAR align with radiologist judgments. It consists of two annotation subsets: expert ensemble labels and expert-curated attributes. We defer full details of the instruction criteria, interface design, and annotation workflow to Appendix D.2.

Expert ensemble labels. These provide the ground-truth labels for the `Presence` attribute. We randomly selected 100 studies from the validation and test sets of MIMIC-CXR-JPG [48], excluding any training samples and normal studies. Each report was independently annotated from scratch by three board-certified radiologists. During annotation, the radiologists categorized each of 13 CheXpert conditions [44] into one of five categories: `confidently absent`, `likely absent`, `neutral`, `likely present`, and `confidently present`, based on their best interpretation of the report. After the initial round of annotations, we merged `confidently present` and `likely present` into a single category `positive`, while `likely absent` and `confidently absent` into `negative`. We then assessed agreement across annotators. Remaining disagreements were first resolved by majority vote, followed by a consensus discussion for any unresolved conflicts. The finalized dataset serves as the ground truth for evaluating model performance in the Label Extraction Module.

Expert-curated attributes. These cover the remaining five report attributes: `first occurrence`, `change`, `severity`, `descriptive location`, and `recommendation`. We began by preparing two sets of model-generated attributes, one from Llama-3.1-8B-Instruct and the other from GPT-4o, for each positive condition identified in the expert ensemble labels. These two sets were merged and then randomly split into two review sets, each with 50 samples from Llama and 50 from GPT-4o. Each set was independently reviewed by separate radiologists. During curation, each radiologist first rated each attribute as `incorrect`, `partially correct`, or `correct`. For non-`correct` attributes, the radiologist also provided a revised version, which was used to construct the ground-truth attribute set.

5.4 Experiments

Experimental setup. To evaluate the effectiveness and clinical reliability of our proposed CLEAR framework, we conduct experiments using CLEAR-Bench. For the Label Extraction Module, we compare CLEAR's performance against two established baselines: the BERT-based labeler CheXbert [108] and the rule-based labeler CheXpert [44], using the Expert Ensemble Labels from CLEAR-Bench. We report F1 scores as introduced in Section 5.2. For the Description Extraction Module, we evaluate CLEAR using the Expert-Curated Attributes from CLEAR-Bench. As no prior baselines exist for this task, we report expert evaluation scores directly, along with automated metrics defined in Section 5.2.

LLM-based labeler achieves substantial gains over existing labelers. We begin with evaluating the performance of the Label Extraction Module. As shown in Table 5.2, our text generation-based approach (Prompt 1) significantly outperforms the best BERT-based labeler [108] and the top rule-based labeler [44] across all accuracy metrics. In identifying positive conditions, our module achieves a notable improvement in accuracy averaged over all 13 medical conditions (+15.8%), with smaller increase on the Top 5 conditions (+12.8%) and the full label pool (+4.1%). This is likely because text generation models can understand the full sentence and overall report, instead of relying on token-level classification or hard-coded rules. Furthermore, this contextual understanding generalizes across conditions,

	First Occurrence		Change		Severity		Descriptive Location		Recommendation	
Metric	**GPT-4o**	**Llama 8B**	**GPT-4o**	**Llama 8B**	**GPT-4o**	**Llama 8B**	**GPT-4o**	**Llama 8B**	**GPT-4o**	**Llama 8B**
EXPERT EVALUATION SCORES										
Experts (condition averaged)	**0.818**	0.685	0.837	0.685	**0.809**	0.565	0.857	0.761	0.933	**0.474**
Experts (report averaged)	0.783	**0.680**	**0.867**	**0.688**	0.771	**0.583**	0.872	**0.763**	**0.940**	0.416
Experts (micro)	0.777	0.662	0.855	0.663	0.777	0.570	**0.867**	0.757	0.936	0.404
ACCURACY METRICS										
Acc. (condition averaged)	0.740	0.688	0.710	0.589	0.682	0.470	–	–	–	–
Acc. (report averaged)	0.755	0.679	0.759	0.596	0.685	0.532	–	–	–	–
Acc. (micro)	0.737	0.665	0.754	0.575	0.671	0.494	–	–	–	–
SIMILARITY METRICS										
o1-mini (micro)	–	–	–	–	–	–	0.785	0.739	0.888	0.361
ROUGE-L (micro)	–	–	–	–	–	–	0.686	0.672	0.887	0.268
BLEU-4 (micro)	–	–	–	–	–	–	0.500	0.402	0.885	0.263
Average (experts)	0.793	0.676	0.853	0.679	0.786	0.573	0.865	0.760	0.936	0.431
Average (all)	0.768	0.677	0.797	0.633	0.733	0.536	0.761	0.682	0.911	0.364
Δ(GPT-4o − Llama)	+0.091		+0.164		+0.197		+0.079		+0.547	

* A dash (–) indicates the metric is not applicable for this attribute.
* Bold values highlight the highest scores per metric. Colored cells distinguish GPT-4o (green) from Llama 8B (yellow).
* The bottom row shows the difference between GPT-4o and Llama 8B for the "Average (all)" metric.

Table 5.3: Evaluation of the description extraction module. Expert ratings are averaged across all samples (0 = incorrect, 0.5 = partially correct, 1 = correct). According to radiologists' clinical judgment, CLEAR can accurately extract attribute-level information from free-text reports. Additionally, GPT-4o is consistently preferred over Llama-3.1-8B-Instruct, though Llama performs reasonably well, especially on `descriptive location`, and remains a low-cost, open-source option.

especially for rare conditions (e.g., fracture) where BERT-based models struggle due to data imbalance, and unseen patterns (e.g., pleural other) where rule-based systems fail to capture beyond their predefined scope. This advantage is even more evident in negative conditions, which require interpreting implicit cues (e.g., "lungs are clear"). Our module achieves a substantial boost (+42.5%) in average accuracy across all conditions, highlighting once again its strength in semantic understanding beyond explicit mentions.

Ablation study of model scales and adaptation. For identifying positive clinical findings, model scale plays a major role, with GPT-4o achieving the highest performance across all accuracy metrics. In contrast, model adaptation strategies, including both few-shot prompting and supervised fine-tuning, have relatively limited impact compared to each base model. This is likely because the base models already encode sufficient clinical knowledge to accurately identify positive findings, and larger model scales are more strongly related

with the richness of this knowledge. However, when it comes to negative mentions, model adaptation strategies stand out, with all metrics improving notably across scales. The reason is that these strategies effectively incorporate expert-derived "side" information, which is typically not captured by base models during pre-training, through few-shot examples or supervised training data. Specifically, among different strategies, supervised fine-tuning consistently outperforms few-shot prompting, with average gains of 26.8% for small models from fineuning, 7.9% for medium models from few-shot, and 7.3% for large models from few-shot.

LLMs, especially GPT-4o, excel at fine-grained attribute extraction. We next probe our description extraction module to assess how reliably a unified language model can handle all five fine-grained attributes (see Table 5.3). Overall, GPT-4o shows strong performance across all five attributes, achieving the highest average score of 0.911 (`recommendation` average all) and a minimum of 0.733 (`severity`). When analyzing by task type, GPT-4o performs better on IE tasks (`location` and `recommendation`), with an average score of 0.836, particularly for attributes that involve highly formulaic language (e.g., "follow-up imaging recommended to assess the resolution of opacity" for `recommendation`). In contrast, it achieves a relatively lower score of 0.766 on QA tasks (`first occurrence`, `change`, and `severity`), which typically require deeper clinical contextual understanding. In comparison, Llama-3.1-8B-Instruct (a small-scale model) shows mixed performance across attributes. In QA tasks, it captures temporal information reasonably well, scoring 0.677 for `first Occurrence` average all and 0.633 for `change`, though its interpretation of clinical findings is weaker (0.536 for `severity`). As for IE tasks, hallucinations significantly affect performance. But with a customized terminology list (see Table D.3), it achieves 0.682 on `location`, the closest to GPT-4o. However, unrelated descriptive phrases (e.g., "signs of generalized fluid overload") significantly lower `recommendation` score to 0.364.

CLEAR aligns well with expert ratings. Generally, all the implementations of CLEAR are highly correlated with expert scoring, as shown in Table 5.4. However, automated metrics are typically slightly lower than expert scores, as observed in Table 5.3. This is because

Automated Metric	Corr. with Expert Scoring
Accuracy Metrics produced by CLEAR	
Acc. (condition averaged)	0.894
Acc. (report averaged)	0.908
Acc. (micro)	0.915
Similarity Metrics produced by CLEAR	
o1-mini (micro)	0.994
ROUGE-L (micro)	0.977
BLEU-4 (micro)	0.811

Table 5.4: Pearson correlation between CLEAR and expert scores. All of automated metrics generated by CLEAR show strong alignment with expert evaluations.

similarity metrics based on ROUGE-L and BLEU-4 prioritize exact matches against ground truth, whereas expert scoring includes a `Partially Correct` category, allowing some tolerance for clinically reasonable but not perfectly matched responses. This distinction is further supported by the exceptionally high correlation of o1-mini scores with expert ratings, reaching 0.994. Compared to other lexical metrics, o1-mini can more effectively capture semantic and clinical alignment, making it a closer proxy to expert judgment.

5.5 Related Work

Lexical metrics. Traditional word-overlap metrics such as BLEU [89], ROUGE [68],and METEOR [8] are commonly used in natural language generation tasks and are therefore also commonly applied to radiology report generation. However, these metrics fail to capture subtle semantic nuances, such as negations or synonyms, which are critical in the clinical domain. Embedding-based metrics like BERTScore [128] improve on semantic matching but remain inadequate in capturing nuanced semantics and domain-specific medical terms, thereby missing clinically important errors.

Clinical efficacy metrics. To bridge the gap between surface-level fluency and clinical correctness, domain-specific metrics have been introduced. Label-based metrics such as CheXpert [44] map reports to 14 predefined clinical labels and measure classification accuracy, but their rule-based pipelines propagate annotation noise. CheXbert [108]

improves semantic understanding over CheXpert by fine-tuning BERT-based classifiers; however, it still lags behind recent LLMs due to the limited capacity of BERT compared to newer and more powerful language models. More recent entity-centric methods such as RadGraph F1 [47], RadGraph2 [50], MEDCON [124] and RaTEScore [130] capture subject–relation–object triples. Although these approaches effectively identify and compare medical entities and their relationships, they often lack the granularity to evaluate specific attributes such as severity, temporal progression, or treatments. To better align automatic metrics with radiologist judgments, RadCliQ [125] combines BLEU, BERTScore, CheXbert similarity, and RadGraph F1 into a weighted score learned from 160 radiologist-annotated report pairs (ReXVal). These annotations are provided at an aggregate level, quantifying the total number of clinically significant and insignificant errors without distinguishing specific clinical attributes.

LLM-based metrics. More recently, researchers have been using LLMs to assess radiology reports. Several methods, including GREEN and CheXprompt, build on six categories of the clinical-error taxonomy introduced in RadCliQ. GREEN [87] tallies the number of errors and matched findings of each type and then aggregates them into a single report-level score, which limits granularity and makes it difficult to isolate specific mistakes. CheXprompt [127] uses GPT-4 to quantify clinically significant and insignificant errors in radiology reports, categorizing them into six predefined types. Similarly, it focuses primarily on counting these errors without delving into the nuanced contextual attributes of each error instance. FineRadScore [42] takes a different route: it calculates the minimum line-by-line edits required to transform a generated report into a reference report. While this encourages precision, it penalizes semantically equivalent but differently phrased outputs. RadFact [9] decomposes each report into atomic sentences and uses LLM to determine whether each generated sentence is entailed by the reference report, which does not differentiate different types of clinical errors or severity.

5.6 Conclusion

We present CLEAR, the first clinically grounded, attribute-level evaluation framework that leverages LLMs to convert free-text radiology reports into a structured tabular format. CLEAR consists of three components: (1) a label extraction module to assess the accurate identification of medical conditions; (2) a description extraction module to evaluate the precision of condition descriptions; and (3) a scoring module to compile multi-metric evaluation results. We also introduce CLEAR-Bench, an expert-curated alignment dataset covering 6 report attributes and 13 medical conditions. Our experiments show that CLEAR can effectively identify clinical conditions, faithfully extract attribute-level information in line with clinical validation, and provide automated metrics that serve as reliable proxies for expert scoring.

Limitations

While CLEAR provides a clinically grounded framework and demonstrates strong alignment with expert clinical assessment, it has several limitations. First, like all existing evaluation metrics, CLEAR relies solely on ground-truth reports without incorporating image information, overlooking the fact that reference reports may not fully capture all relevant findings present in the image. Future work could explore integrating image-based evaluation to better reflect clinical completeness. Second, CLEAR is built on the CheXpert label structure, which is limited in both granularity and anatomical coverage. Extending the framework to include additional specialties such as breast imaging, cardiology, and gastroenterology in the future could enhance its generalizability. Lastly, although we prioritize high-quality annotations, both the training and evaluation datasets remain relatively small due to the common tradeoff between annotation quality and dataset scale.

Chapter 6

Conclusion

This thesis begins with a theoretical investigation into the foundations of effective human–AI collaboration. We propose a formal framework that models how human intuitions interact with machine explanations. Our analysis demonstrates that task-specific human intuitions are essential for enabling complementary performance, where combined human–AI decisions outperform either alone. This theoretical work provides a principled foundation for designing effective human-AI collaborative systems.

Building on this foundation, we conducted empirical studies involving expert radiologists diagnosing prostate cancer from MRI. These studies allowed us to directly observe how AI tools are used in clinical workflows and how human reliance on AI evolves across different interface and feedback conditions. While AI assistance improved diagnostic accuracy compared to unaided human performance, we found that complementary performance remained elusive due to under-reliance. Using a theoretical reliance framework, we analyzed behavioral patterns and identified actionable strategies that guarantee a complementary performance. One such strategy involves deferring to a positive diagnosis only when both the radiologist and AI model independently agree. Unlike prior work that typically stops at surface-level performance evaluation, our study takes a step further by offering concrete recommendations for effective human-AI collaboration.

Finally, we turned our attention to the evaluation of advanced multimodal large language models (MLLMs) for radiology report generation from chest X-rays. Despite the remarkable progress in general domains, our diagnostic analysis revealed that current MLLMs struggle

with clinical image interpretation. To address the limitations of existing evaluation metrics, we introduced a fine-grained, tabular-based framework grounded in expert-curated annotations. This benchmark enables more comprehensive and clinically meaningful assessment of generated reports and provides practical guidance for improving model performance.

Looking ahead, this work opens several promising avenues for future research. On the human side, designing better interfaces and workflows may further improve appropriate reliance and trust, ultimately enabling complementary performance. On the model side, advancing toward expert-level capabilities remains a critical challenge. More broadly, the findings of this thesis contribute to the growing recognition that as AI tools become increasingly powerful, the key question is no longer just what AI can do, but how we can most effectively integrate these tools with human expertise in high-stakes decision-making. Understanding and shaping this human–AI interaction is essential for building systems that are not only technically advanced, but also safe, trustworthy, and aligned with real-world human needs.

Bibliography

[1] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y Lim. 2020. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[2] Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. 2023. *Combining human expertise with artificial intelligence: Experimental evidence from radiology*. Technical Report. National Bureau of Economic Research.

[3] Saar Alon-Barkat and Madalina Busuioc. 2023. Human–AI interactions in public sector decision making:"automation bias" and "selective adherence" to algorithmic advice. *Journal of Public Administration Research and Theory* 33, 1 (2023), 153–169.

[4] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 275–285.

[5] Anthropic. 2025. Claude 3.7 Sonnet System Card. `https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf` Accessed: 2025-05-19.

[6] Lucrezia Greta Armando, Gianluca Miglio, Pierluigi de Cosmo, and Clara Cena. 2023. Clinical decision support systems to improve drug prescription and therapy

optimisation in clinical practice: a scoping review. *BMJ Health & Care Informatics* 30, 1 (2023).

[7] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10, 7 (2015), e0130140.

[8] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 65–72.

[9] 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 T. A. Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. 2024. MAIRA-2: Grounded Radiology Report Generation. *arXiv* abs/2406.04449 (2024). `https://arxiv.org/abs/2406.04449`

[10] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.

[11] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. 2020. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. *arXiv preprint arXiv:2006.14779* (2020).

[12] Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A. Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, Hyo Jung Hong, Mehr Kashyap, Akash R. Chaurasia,

Nirav R. Shah, Karandeep Singh, Troy Tazbaz, Arnold Milstein, Michael A. Pfeffer, and Nigam H. Shah. 2025. Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. *JAMA* 333, 4 (01 2025), 319–328. https://doi.org/10.1001/jama.2024.21700 arXiv:https://jamanetwork.com/journals/jama/articlepdf/2825147/jama_bedi_2024_oi_240124_1737130

[13] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

[14] Or Biran and Kathleen R McKeown. 2017. Human-Centric Justification of Machine Learning Predictions.. In *IJCAI*, Vol. 2017. 1461–1467.

[15] Joeran S Bosma, Anindo Saha, Matin Hosseinzadeh, Ilse Slootweg, Maarten de Rooij, and Henkjan Huisman. 2021. Annotation-efficient cancer detection with report-guided lesion annotation for deep learning-based prostate cancer detection in bpMRI. *arXiv preprint arXiv:2112.05151* (2021).

[16] Dana Brin, Vera Sorin, Yiftach Barash, Eli Konen, Benjamin S. Glicksberg, Girish Nadkarni, and Eyal Klang. 2023. Assessing GPT-4 Multimodal Performance in Radiological Image Analysis. *medRxiv* (nov 2023).

[17] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).

[18] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 454–464.

[19] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *arXiv preprint arXiv:2102.09692* (2021).

[20] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*. IEEE, 160–169.

[21] Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2020. Feature-Based Explanations Don't Help People Detect Misclassifications of Online Toxicity. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 95–106.

[22] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 1721–1730.

[23] Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P. Langlotz. 2024. CheXpert Plus: Augmenting a Large Chest X-ray Dataset with Text Radiology Reports, Patient Demographics and Additional Image Formats. arXiv:2405.19538 [cs.CL] `https://arxiv.org/abs/2405.19538`

[24] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. 2018. Do explanations make VQA models more predictable to a human? *arXiv preprint arXiv:1810.12366* (2018).

[25] Aritrick Chatterjee, Ambereen N Yousuf, Roger Engelmann, Carla Harmath, Grace Lee, Milica Medved, Ernest B Jamison, Abel Lorente Campos, Batuhan Gundogdu, Glenn Gerber, et al. 2025. Prospective Validation of an Automated Hybrid Multidimensional MRI Tool for Prostate Cancer Detection Using Targeted Biopsy: Comparison with PI-RADS-based Assessment. *Radiology: Imaging Cancer* 7, 1 (2025), e240156.

[26] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *26th International Conference on Intelligent User Interfaces*. 307–317.

[27] Anthony Christopher Davison and David Victor Hinkley. 1997. *Bootstrap methods and their application*. Number 1. Cambridge university press.

[28] Maarten De Rooij, Esther HJ Hamoen, Jurgen J Fütterer, Jelle O Barentsz, and Maroeska M Rovers. 2014. Accuracy of multiparametric MRI for prostate cancer detection: a meta-analysis. *American Journal of Roentgenology* 202, 2 (2014), 343–351.

[29] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association* 23, 2 (2016), 304–310.

[30] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[31] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[32] Shi Feng and Jordan Boyd-Graber. 2018. What can AI do for me: Evaluating Machine Learning Interpretations in Cooperative Play. *arXiv preprint arXiv:1810.09648* (2018).

[33] Sorelle A Friedler, Chitradeep Dutta Roy, Carlos Scheidegger, and Dylan Slack. 2019. Assessing the Local Interpretability of Machine Learning Models. *arXiv preprint arXiv:1902.03501* (2019).

[34] Deep Ganguli, Nicholas Schiefer, Marina Favaro, and Jack Clark. 2023.

Challenges in evaluating AI systems. https://www.anthropic.com/index/evaluating-ai-systems

[35] Ana Valeria Gonzalez, Gagan Bansal, Angela Fan, Robin Jia, Yashar Mehdad, and Srinivasan Iyer. 2020. Human Evaluation of Spoken vs. Visual Explanations for Open-Domain QA. *arXiv preprint arXiv:2012.15075* (2020).

[36] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 90–99.

[37] Shunan Guo, Fan Du, Sana Malik, Eunyee Koh, Sungchul Kim, Zhicheng Liu, Donghyun Kim, Hongyuan Zha, and Nan Cao. 2019. Visualizing uncertainty and alternatives in event sequence predictions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[38] Ziyang Guo, Yifan Wu, Jason D Hartline, and Jessica Hullman. 2024. A decision theoretic framework for measuring AI reliance. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 221–236.

[39] H Benjamin Harvey and Vrushab Gowda. 2020. How the FDA regulates AI. *Academic radiology* 27, 1 (2020), 58–61.

[40] Peter Hase and Mohit Bansal. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? *arXiv preprint arXiv:2005.01831* (2020).

[41] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H Schwartz, and Hugo JWL Aerts. 2018. Artificial intelligence in radiology. *Nature Reviews Cancer* 18, 8 (2018), 500–510.

[42] Alyssa Huang, Oishi Banerjee, Kay Wu, Eduardo Pontes Reis, and Pranav Rajpurkar. 2024. FineRadScore: A Radiology Report Line-by-Line Evaluation Technique Generating Corrections with Severity Scores. In *Machine Learning for Healthcare Conference*. PMLR.

[43] Stephanie L Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, et al. 2023. MAIRA-1: A specialised large multimodal model for radiology report generation. *arXiv preprint arXiv:2311.13668* (2023).

[44] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 590–597.

[45] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods* 18, 2 (2021), 203–211.

[46] Ayush Jain, David Way, Vishakha Gupta, Yi Gao, Guilherme de Oliveira Marinho, Jay Hartford, Rory Sayres, Kimberly Kanada, Clara Eng, Kunal Nagpal, et al. 2021. Development and assessment of an artificial intelligence–based tool for skin condition diagnosis by primary care physicians and nurse practitioners in teledermatology practices. *JAMA network open* 4, 4 (2021), e217249–e217249.

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

[48] Alistair Johnson, Matthew Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. 2024. MIMIC-CXR-JPG - Chest Radiographs with Structured Labels (version 2.1.0). `https://doi.org/10.13026/jsn5-t979`.

[49] A Johnson, T Pollard, R Mark, S Berkowitz, and Steven Horng. 2019. MIMIC-CXR database (version 2.0. 0). *physionet* 2 (2019), 5.

[50] Sameer Khanna, Adam Dejl, Kibo Yoon, Steven QH Truong, Hanh Duong, Agustina Saenz, and Pranav Rajpurkar. 2023. Radgraph2: Modeling disease progression in radiology reports via hierarchical information extraction. In *Machine Learning for Healthcare Conference*. PMLR, 381–402.

[51] Amirhossein Kiani, Bora Uyumazturk, Pranav Rajpurkar, Alex Wang, Rebecca Gao, Erik Jones, Yifan Yu, Curtis P Langlotz, Robyn L Ball, Thomas J Montine, et al. 2020. Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ digital medicine* 3, 1 (2020), 1–8.

[52] Hyo-Eun Kim, Hak Hee Kim, Boo-Kyung Han, Ki Hwan Kim, Kyunghwa Han, Hyeonseob Nam, Eun Hye Lee, and Eun-Kyung Kim. 2020. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *The Lancet Digital Health* 2, 3 (2020), e138–e148.

[53] Bobby Kleinberg, Renato Paes Leme, Jon Schneider, and Yifeng Teng. 2023. U-calibration: Forecasting for an unknown agent. In *The Thirty Sixth Annual Conference on Learning Theory*. PMLR, 5143–5145.

[54] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

[55] Ramaravind Kommiya Mothilal, Divyat Mahajan, Chenhao Tan, and Amit Sharma. 2021. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 652–663.

[56] Marie-Luise Kromrey, Laura Steiner, Felix Schön, Julie Gamain, Christian Roller, and Carolin Malsch. 2024. Navigating the Spectrum: Assessing the Concordance of ML-Based AI Findings with Radiology in Chest X-Rays in Clinical Settings. In *Healthcare*, Vol. 12. MDPI, 2225.

[57] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

[58] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006* (2019).

[59] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. *arXiv preprint arXiv:2112.11471* (2021).

[60] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2023. Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. In *Proceedings of FAccT*.

[61] Vivian Lai, Han Liu, and Chenhao Tan. 2020. " Why is' Chicago'deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[62] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 29–38.

[63] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1675–1684.

[64] Curtis P. Langlotz. 2015. *The Radiology Report: A Guide to Thoughtful Communication for Radiologists and Other Medical Professionals*. CreateSpace Independent Publishing Platform, North Charleston, SC.

[65] Curtis P Langlotz. 2019. Will artificial intelligence replace radiologists? , e190058 pages.

[66] Yingshu Li, Yunyi Liu, Zhanyu Wang, Xinyu Liang, Lingqiao Liu, Lei Wang, Leyang Cui, Zhaopeng Tu, Longyue Wang, and Luping Zhou. 2023. A Comprehensive Study of GPT-4V's Multimodal Capabilities in Medical Imaging. *medRxiv* (2023), 2023–11.

[67] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2119–2128.

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

[69] Zachary C Lipton. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490* (2016).

[70] Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. 2024. Bootstrapping Large Language Models for Radiology Report Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 18635–18643.

[71] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. *arXiv preprint arXiv:2101.05303* (2021).

[72] Zhengliang Liu, Hanqi Jiang, Tianyang Zhong, Zihao Wu, Chong Ma, Yiwei Li, Xiaowei Yu, Yutong Zhang, Yi Pan, Peng Shu, et al. 2023. Holistic evaluation of gpt-4v for biomedical imaging. *arXiv preprint arXiv:2312.05256* (2023).

[73] Ana Lucic, Hinda Haned, and Maarten de Rijke. 2020. Why does my model fail? contrastive local explanations for retail forecasting. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 90–98.

[74] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of NIPS*.

[75] Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Tobi Walsh, Armin Hamrah, Lapo Santarlasci, Julia Betts Lotufo, Alexandra Rome, Andrew Shi, and Sukrut Oak. 2025. *The AI Index 2025 Annual Report*. Technical Report. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA. `https://hai.stanford.edu/ai-index/2025-ai-index-report`.

[76] Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, et al. 2023. Towards accurate differential diagnosis with large language models. *arXiv preprint arXiv:2312.00164* (2023).

[77] Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, et al. 2025. Towards accurate differential diagnosis with large language models. *Nature* (2025), 1–7.

[78] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. 2020. International evaluation of an AI system for breast cancer screening. *Nature* 577, 7788 (2020), 89–94.

[79] Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. 2019. Explaining machine learning classifiers through diverse counterfactual explanations. *arXiv preprint arXiv:1905.07697* (2019).

[80] John Ashworth Nelder and Robert WM Wedderburn. 1972. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)* 135, 3 (1972), 370–384.

[81] Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 1069–1078.

[82] Dang Nguyen, Chacha Chen, He He, and Chenhao Tan. 2023. Pragmatic Radiology Report Generation. In *Machine Learning for Health (ML4H)*. PMLR, 385–402.

[83] Justin G Norden and Nirav R Shah. 2022. What AI in health care can learn from the long road to autonomous vehicles. *NEJM Catalyst Innovations in Care Delivery* 3, 2 (2022).

[84] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375* (2023).

[85] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *26th International Conference on Intelligent User Interfaces*. 340–350.

[86] OpenAI. 2025. Introducing GPT-4.5. `https://openai.com/index/introducing-gpt-4-5/` Accessed: 2025-05-19.

[87] Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Md, Michael Moseley, Curtis Langlotz, Akshay Chaudhari, et al. 2024. GREEN: Generative Radiology Report Evaluation and Error Notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 374–390.

[88] Khaled Ouanes and Nesren Farhah. 2024. Effectiveness of artificial intelligence (AI) in clinical decision support systems and care delivery. *Journal of Medical Systems* 48, 1 (2024), 74.

[89] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a

method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[90] Allison Park, Chris Chute, Pranav Rajpurkar, Joe Lou, Robyn L Ball, Katie Shpanskaya, Rashad Jabarkheel, Lily H Kim, Emily McKenna, Joe Tseng, et al. 2019. Deep learning–assisted diagnosis of cerebral aneurysms using the HeadXNet model. *JAMA network open* 2, 6 (2019), e195600–e195600.

[91] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–15.

[92] Magdalini Paschali, Zhihong Chen, Louis Blankemeier, Maya Varma, Alaa Youssef, Christian Bluethgen, Curtis Langlotz, Sergios Gatidis, and Akshay Chaudhari. 2025. Foundation Models in Radiology: What, How, Why, and Why Not. *Radiology* 314, 2 (2025), e240597.

[93] Ruben Pauwels, Danieli Moura Brasil, Mayra Cristina Yamasaki, Reinhilde Jacobs, Hilde Bosmans, Deborah Queiroz Freitas, and Francisco Haiter-Neto. 2021. Artificial intelligence for detection of periapical lesions on intraoral radiographs: Comparison between convolutional neural networks and human observers. *Oral surgery, oral medicine, oral pathology and oral radiology* 131, 5 (2021), 610–616.

[94] Judea Pearl. 2009. *Causality*. Cambridge university press.

[95] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–52.

[96] Pranav Rajpurkar, Chloe O'Connell, Amit Schechter, Nishit Asnani, Jason Li, Amirhossein Kiani, Robyn L Ball, Marc Mendelson, Gary Maartens, Daniël J van Hoving, et al. 2020. CheXaid: deep learning assistance for physician diagnosis of

tuberculosis using chest x-rays in patients with HIV. *NPJ digital medicine* 3, 1 (2020), 115.

[97] V. Rao, S. Zhang, J. Acosta, S. Adithan, and P. Rajpurkar. 2025. ReXErr-v1: Clinically Meaningful Chest X-Ray Report Errors Derived from MIMIC-CXR. PhysioNet. https://doi.org/10.13026/9dns-vd94

[98] Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Ramona Comanescu, Canfer Akbulut, Tom Stepleton, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, William Isaac, and Laura Weidinger. 2024. Gaps in the Safety Evaluation of Generative AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, 1 (Oct. 2024), 1200–1217. https://doi.org/10.1609/aies.v7i1.31717

[99] Andreas M Rauschecker, Jeffrey D Rudie, Long Xie, Jiancong Wang, Michael Tran Duong, Emmanuel J Botzolakis, Asha M Kovalovich, John Egan, Tessa C Cook, R Nick Bryan, et al. 2020. Artificial intelligence system approaching neuroradiologist-level differential diagnosis accuracy at brain MRI. *Radiology* 295, 3 (2020), 626–637.

[100] Carlo Reverberi, Tommaso Rigon, Aldo Solari, Cesare Hassan, Paolo Cherubini, and Andrea Cherubini. 2022. Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific reports* 12, 1 (2022), 14952.

[101] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of KDD*.

[102] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations.. In *AAAI*, Vol. 18. 1527–1535.

[103] Alejandro Rodriguez-Ruiz, Kristina Lång, Albert Gubern-Merida, Mireille Broeders, Gisella Gennaro, Paola Clauser, Thomas H Helbich, Margarita Chevalier, Tao Tan, Thomas Mertelmeier, et al. 2019. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *JNCI: Journal of the National Cancer Institute* 111, 9 (2019), 916–922.

[104] Anindo Saha, Joeran S Bosma, Jasper J Twilt, Bram van Ginneken, Anders Bjartell, Anwar R Padhani, David Bonekamp, Geert Villeirs, Georg Salomon, Gianluca Giannarini, Jayashree Kalpathy-Cramer, Jelle Barentsz, Klaus H Maier-Hein, Mirabela Rusu, Olivier Rouvière, Roderick van den Bergh, Valeria Panebianco, Veeru Kasivisvanathan, Nancy A Obuchowski, Derya Yakar, Mattijs Elschot, Jeroen Veltman, Jurgen J Fütterer, Constant R. Noordman, Ivan Slootweg, Christian Roest, Stefan J. Fransen, Mohammed R.S. Sunoqrot, Tone F. Bathen, Dennis Rouw, Jos Immerzeel, Jeroen Geerdink, Chris van Run, Miriam Groeneveld, James Meakin, Ahmet Karagöz, Alexandre Bône, Alexandre Routier, Arnaud Marcoux, Clément Abi-Nader, Cynthia Xinran Li, Dagan Feng, Deniz Alis, Ercan Karaarslan, Eui-joon Ahn, François Nicolas, Geoffrey A. Sonn, Indrani Bhattacharya, Jinman Kim, Jun Shi, Hassan Jahanandish, Hong An, Hongyu Kan, Ilkay Oksuz, Liang Qiao, Marc-Michel Rohé, Mert Yergin, Mohamed Khadra, Mustafa E. Şeker, Mustafa S. Kartal, Noëlie Debs, Richard E. Fan, Sara Saunders, Simon J.C. Soerensen, Stefania Moroianu, Sulaiman Vesal, Yuan Yuan, Afsoun Malakoti-Fard, Agnė Mačiūnien, Akira Kawashima, Ana M.M. de M.G. de Sousa Machadov, Ana Sofia L. Moreira, Andrea Ponsiglione, Annelies Rappaport, Arnaldo Stanzione, Arturas Ciuvasovas, Baris Turkbey, Bart de Keyzer, Bodil G. Pedersen, Bram Eijlers, Christine Chen, Ciabattoni Riccardo, Deniz Alis, Ewout F.W. Courrech Staal, Fredrik Jäderling, Fredrik Langkilde, Giacomo Aringhieri, Giorgio Brembilla, Hannah Son, Hans Vanderlelij, Henricus P.J. Raat, Ingrida Pikūnienė, Iva Macova, Ivo Schoots, Iztok Caglic, Jeries P. Zawaideh, Jonas Wallström, Leonardo K. Bittencourt, Misbah Khurram, Moon H. Choi, Naoki Takahashi, Nelly Tan, Paolo N. Franco, Patricia A. Gutierrez, Per Erik Thimansson, Pieter Hanus, Philippe Puech, Philipp R. Rau, Pieter de Visschere, Ramette Guillaume, Renato Cuocolo, Ricardo O. Falcão, Rogier S.A. van Stiphout, Rossano Girometti, Ruta Briediene, Rūta Grigienė, Samuel Gitau, Samuel Withey, Sangeet Ghai, Tobias Penzkofer, Tristan Barrett, Varaha S. Tammisetti, Vibeke B. Løgager, Vladimír Černý, Wulphert Venderink, Yan M. Law, Young J. Lee, Maarten de Rooij, and Henkjan Huisman. 2024. Artificial intelligence and radiologists in prostate cancer detection on MRI (PI-CAI): an international, paired,

non-inferiority, confirmatory study. *The Lancet Oncology* 25, 7 (2024), 879–887. https://doi.org/10.1016/S1470-2045(24)00220-1

[105] Maxi Scherer. 2019. Artificial Intelligence and Legal Decision-Making: The Wide Open? *Journal of international arbitration* 36, 5 (2019).

[106] Jarrel CY Seah, Cyril HM Tang, Quinlan D Buchlak, Xavier G Holt, Jeffrey B Wardman, Anuar Aimoldin, Nazanin Esmaili, Hassan Ahmad, Hung Pham, John F Lambert, et al. 2021. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *The Lancet Digital Health* 3, 8 (2021), e496–e506.

[107] Yongsik Sim, Myung Jin Chung, Elmar Kotter, Sehyo Yune, Myeongchan Kim, Synho Do, Kyunghwa Han, Hanmyoung Kim, Seungwook Yang, Dong-Jae Lee, et al. 2020. Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest radiographs. *Radiology* 294, 1 (2020), 199–209.

[108] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. 2020. Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1500–1519.

[109] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. 2020. CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. *arXiv preprint arXiv:2004.09167* (2020).

[110] David F Steiner, Robert MacDonald, Yun Liu, Peter Truszkowski, Jason D Hipp, Christopher Gammage, Florence Thng, Lily Peng, and Martin C Stumpe. 2018. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *The American journal of surgical pathology* 42, 12 (2018), 1636–1646.

[111] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*. PMLR, 3319–3328.

[112] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html* 3, 6 (2023), 7.

[113] Kevin Tian, S. J. Hartung, A. A. Li, J. Jeong, F. Behzadi, J. Calle-Toro, S. Adithan, M. Pohlen, D. Osayande, and P. Rajpurkar. 2023. ReFiSco: Report Fix and Score Dataset for Radiology Report Generation. PhysioNet. https://doi.org/10.13026/cneg-zk64

[114] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. https://github.com/heartexlabs/label-studio Open source software available from https://github.com/heartexlabs/label-studio.

[115] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. 2024. Towards generalist biomedical ai. *NEJM AI* 1, 3 (2024), AIoa2300138.

[116] Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. 2025. Towards Conversational Diagnostic Artificial Intelligence. *Nature* (April 2025), 1–9. https://doi.org/10.1038/s41586-025-08866-7

[117] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

[118] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26rd International Conference on Intelligent User Interfaces*.

[119] Hilde JP Weerts, Werner van Ipenburg, and Mykola Pechenizkiy. 2019. A human-grounded evaluation of shap for alert processing. *arXiv preprint arXiv:1907.03324* (2019).

[120] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[121] Chaoyi Wu, Jiayu Lei, Qiaoyu Zheng, Weike Zhao, Weixiong Lin, Xiaoman Zhang, Xiao Zhou, Ziheng Zhao, Ya Zhang, Yanfeng Wang, et al. 2023. Can gpt-4v (ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis. *arXiv preprint arXiv:2310.09909* (2023).

[122] Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanisław Jastrzębski, Thibault Févry, Joe Katsnelson, Eric Kim, et al. 2019. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE transactions on medical imaging* 39, 4 (2019), 1184–1194.

[123] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421* 9, 1 (2023), 1.

[124] Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific data* 10, 1 (2023), 586.

[125] Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hos-

sein Abad, Andrew Y Ng, et al. 2023. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns* 4, 9 (2023).

[126] F. Yu, M. Endo, R. Krishnan, I. Pan, A. Tsai, E. P. Reis, E. Kaiser Ururahy Nunes Fonseca, H. Lee, Z. Shakeri, A. Ng, C. Langlotz, V. K. Venugopal, and P. Rajpurkar. 2023. Radiology Report Expert Evaluation (ReXVal) Dataset (version 1.0.0). https://doi.org/10.13026/2fp8-qr71

[127] Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, Tristan Naumann, Muhao Chen, Matthew P. Lungren, Akshay Chaudhari, Serena Yeung-Levy, Curtis P. Langlotz, Sheng Wang, and Hoifung Poon. 2025. A clinically accessible small multimodal radiology model and evaluation metric for chest X-ray findings. *Nature Communications* 16, 1 (01 Apr 2025), 3108. https://doi.org/10.1038/s41467-025-58344-x

[128] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).

[129] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. *arXiv preprint arXiv:2001.02114* (2020).

[130] Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. RaTEScore: A Metric for Radiology Report Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 15004–15019. https://doi.org/10.18653/v1/2024.emnlp-main.836

[131] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging

llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2023), 46595–46623.

[132] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Association for Computational Linguistics, Bangkok, Thailand. http://arxiv.org/abs/2403.13372

Chapter A

Prior Work Details

A.1 A Summary of Recent Empirical Studies

We surveyed literature of recent empirical studies that include quantifiable metric for evaluating human understanding with machine explanations. In Table A.1 we present a summary of recent empirical studies and provide a reinterpretation with the three core concepts: model decision boundary g, model error z, and task decision boundary f, as we defined in the main paper.

Paper	Model	Prediction	Explanations	g	z	f
[1]	Generalized additive models	Shown	Global feature importance (shape function of GAMs)	✓	✗	✗
[67]	Decision trees/random forests	Shown	Rule-based explanations (tree-based explanation); Counterfactual explanations (counterfactual examples)	✓	✗	✗
[24]	Convolution Neural Networks	Hidden	Local feature importance (attention, gradient-based)	✓	✗	✗
[26]	Decision trees/random forests	Shown	Local feature importance (perturbation-based SHAP)	✓	✗	✗
[81]	Logistic regression; Shallow (1- to 2-layer) neural networks	Hidden	Local feature importance (gradient-based, perturbation-based (LIME))	✓	✗	✗
[19]	Wizard of Oz	Shown	Model uncertainty (classification confidence (or probability))	✗	✓	✓

[95]	Linear regression	Shown	Presentation of simple models (linear regression); Information about training data (input features or information the model considers)	✓	✓	✓
[11]	RoBERTa; Generalized additive models (GAMs)	Shown	Model uncertainty (classification confidence (or probability)); Local feature importance (perturbation-based (LIME)); Natural language explanations (expert-generated rationales);	✗	✓	✓
[118]	Logistic regression	Shown	Example-based methods (Nearest neighbor or similar training instances); Counterfactual explanations (counterfactual examples); Global feature importance (permutation-based);	✓	✓	✗
[129]	Decision trees/random forests	Shown	Model uncertainty (classification confidence (or probability)); Local feature importance (perturbation-based SHAP); Information about training data (input features or information the model considers)	✗	✓	✓
[14]	Logistic regression	Shown	Natural language explanations (model-generated rationales)	✗	✓	✓
[71]	Support-vector machines (SVMs)	Shown	Local feature importance (coefficients)	✓	✓	✓
[73]	Decision trees/random forests	Hidden	Counterfactual explanations (contrastive or sensitive features)	✓	✗	✗
[37]	Recurrent Neural Networks	Shown	Model uncertainty (classification confidence (or probability))	✗	✗	✓
[18]	Wizard of Oz	Mixed	Example-based methods (Nearest neighbor or similar training instances)	✓	✓	✓
[102]	VQA model (hybrid LSTM and CNN)	Hidden	Rule-based explanations (anchors)	✓	✗	✗
[33]	Logistic regression; Decision trees/random forests; Shallow (1- to 2-layer) neural networks	Hidden	Counterfactual explanations (counterfactual examples); Presentation of simple models (decision trees, logistic regression, one-layer MLP)	✓	✗	✗
[85]	Other deep learning models	Shown	Local feature importance (video features)	✓	✓	✓
[119]	Decision trees/random forests	Hidden	Model uncertainty (classification confidence (or probability)); Local feature importance (perturbation-based SHAP)	✗	✓	✓
[62]	Support-vector machines (SVMs)	Shown	Example-based methods (Nearest neighbor or similar training instances); Model performance (accuracy)	✗	✓	✓
[51]	Other deep learning models	Shown	Model uncertainty (classification confidence (or probability)); Local feature importance (gradient-based)	✗	✓	✓
[35]	Other deep learning models	Shown	extractive evidence	✗	✓	✓

[32]	Generalized additive models (GAMs)	Shown	Model uncertainty (classification confidence (or probability)); Global example-based explanations (prototypes)	✗	✓	✓
[20]	Wizard of Oz	Shown	Model uncertainty (classification confidence (or probability))	✗	✓	✗
[63]	Bayesian decision lists	Hidden	Rule-based explanations (decision sets)	✓	✗	✗
[61]	BERT; Support-vector machines (SVMs)	Shown	Local feature importance (attention); Model performance (accuracy); Global example-based explanations (model tutorial)	✗	✓	✓
[4]	Convolution Neural Networks	Hidden	Local feature importance (propagation-based (LRP), perturbation-based (LIME))	✓	✗	✗
[21]	Recurrent Neural Networks	Shown	Local feature importance (attention)	✗	✓	✓
[40]	Other deep learning models	Shown	Local feature importance (perturbation-based (LIME)); Rule-based explanations (anchors); Example-based methods (Nearest neighbor or similar training instances); Partial decision boundary (traversing the latent space around a data input)	✓	✗	✗
[101]	Support-vector machines (SVMs) and Inception neural network	Shown	Local feature importance (perturbation-based (LIME))	✓	✗	✗

Table A.1: A summary of recent empirical studies measuring human understanding with machine explanations. The papers are sorted by time, starting form the newest. Note: columns g, z, and f mean model decision boundary, model error, and task decision boundary respectively. ✓(or ✗) means the study measures (or does not measure) the corresponding type of human understanding.

Measuring human understanding of model decision boundary via human simulatability. A straightforward way of model decision boundary evaluation is to measure how well humans can simulate the model predictions, or in other words, the human ability of forward simulation/prediction [31]. Humans are typically asked to simulate model predictions given an input and some explanations [24, 26, 81, 95, 118, 71, 73, 18, 102, 33, 85, 63, 4, 40, 31, 69]. For example, given profiles of criminal defendants and machine explanations, participants are asked to guess what the AI model would predict [118].

Measuring human understanding of model decision boundary via counterfactual reasoning. Sometimes researchers measure human understanding of the decision boundary by evaluating participants' counterfactual reasoning abilities [33, 73]. Counterfactual reasoning investigates the ability to answer the 'what if' question. In practice, participants

are asked to determine the output of a perturbed input applied to the same ML model [33]. Lucic et al. [73] asked participants to manipulate the input to change the model output.

Measuring human understanding of model decision boundary via feature importance. Additionally, Wang and Yin [118] also tested human understanding of model decision boundary via feature importance, specifically by (1) asking the participants to select among a list of features which one was most/least influential on the model's predictions and (2) specifying a feature's marginal effect on predictions. Ribeiro et al. [101] asked participants to perform feature engineering by identifying features to remove, given the LIME explanations. These can be viewed as a coarse inquiry into properties of the model's model decision boundary.

Measuring human understanding of task decision boundary and model error via human+AI performance. Similar to the application-grounded evaluation defined in Doshi-Velez and Kim [31], one of the most well-adopted evaluation measurement of human understanding is to measure human understanding of the task decision boundary through human+AI performance [31, 19, 95, 11, 129, 14, 71, 37, 18, 85, 119, 62, 51, 35, 32, 61, 21]. In those experiments, participants are shown machine predictions and explanations, then they are asked to give a final decision based on the information, with the goal of achieving complementary performance. For example, human decision-makers are asked to predict whether this defendant would re-offend within two years, given a machine prediction and explanations [118]. Note that for binary classification problems, measuring human understanding of the model error is equivalent to measuring human understanding of the task decision boundary if machine predictions are shown.

Measuring human understanding of model error through human trust. In some other cases, trust or reliance is introduced as a criterion reflecting the human understanding of the model error. Explanations are used to guide people to trust an AI model when it is right and not to trust it when it is wrong. Hence, by analyzing when and how often human follows machine predictions, trust can reflect the human understanding of the model error [118, 19, 129]. In other cases, the measure of human understanding of model error can be

used as an intermediate measurement towards measuring task decision boundary [10, 95, 11, 14, 71, 18, 85, 119, 62, 51, 35, 32, 20, 61, 21], where human subjects are asked whether they agree with machine predictions.

A.2 Characterizing Relationship between Core Functions & Human Understandings

Fig. 2-3 visualizes the realizations of the base diagram under the two conditions and organizes them in a two-level decision tree. At the root, we have the base diagram. At the next level, we have two realizations based on whether condition 1 is satisfied: diagram (b) for emulation and diagram (c) for discovery. The branches at the leaf level are determined by condition 2, i.e., whether model prediction $\hat{Y}$ is shown. Next, we unfold the effect of these two conditions.

Effect of $show(Y)$. We observe differences in the diagrams between emulation and discovery tasks. First, human local understanding of task decision boundary $Y^H$is collapsed with Y in emulation tasks (Fig. 2-3b), so no edge goes into Y^H, and Y^Haffects $Z^H$and $\hat{Y}^H$. However, in discovery tasks (Fig. 2-3c), since $Y^H \not\equiv Y$, the edge connections remain the same, i.e., we are unable to rule out any connections for now. Hence, human understanding of task decision boundary is usually not of interest in emulation tasks [24, 81]. In comparison, human understanding of both model decision boundary and task decision boundary is explored in discovery tasks [11, 32, 118, 91].

Effect of $show(\hat{Y})$**.** We start with emulation tasks, where the relationships are relatively straightforward because the human understanding of task decision boundary is perfect ($Y^H \equiv Y$). When $\hat{Y}$ is shown (Fig. 2-3d), human understanding of local predicted label becomes perfect, i.e., $\hat{Y}^H \equiv \hat{Y}$. It follows that $Z^H = I(Y^H \neq \hat{Y}^H) = I(Y \neq \hat{Y}) = Z$. This scenario happens in debugging for emulation tasks, where model developers know the true label, the predicted label, and naturally whether the predicted label is incorrect for the

given instance. It is clear that the desired understanding is not local, but about global model decision boundary.

In comparison, when $\hat{Y}$ is not shown (e.g., an auditor tries to extrapolate the model prediction), recall $Y^H \equiv Y$ in emulation tasks, so Y^H can affect $\hat{Y}^H$ and Z^H. As shown in Fig. 2-3e, the connection between $\hat{Y}^H$ and Z^H remains unclear.

In discovery tasks, when $\hat{Y}$ is shown (Fig. 2-3f), $\hat{Y}^H \equiv \hat{Y}$. The relationships between Y^H and Z^H, however, remain unclear and can be potentially shaped by further information such as machine explanations. When $\hat{Y}$ is not shown (Fig. 2-3g), we do not receive any new information in discovery tasks. Therefore, Fig. 2-3g is the same as the base diagram where all interactions between local understandings are possible, which highlights the fact that no insights about human understandings can be derived without any assumption or intervention.

Implications. Our framework reveals the underlying mechanism of human local understanding with two important conditions: 1) knowing the task decision boundary; and 2) showing machine predictions $\hat{Y}$. Such conditions allow us to rule out connections between human understanding of core variables. For example, in emulation with prediction shown, the relationship between all variables is simplified to a deterministic state.

Another implication is that we need to make explicit assumptions in order to make claims such as human performance improves because human understanding of the model error is better (i.e., humans place appropriate trust in model predictions). Because there exist dashed links between variables, for example, in discovery tasks with prediction shown, we can not tell whether it is $Y^H \rightarrow Z^H$ or $Z^H \rightarrow Y^H$, nor can we tell from observational data without making assumptions. The alternative hypothesis to "appropriate trust $\rightarrow$ improved task performance" is that $\hat{Y}$ directly improves human understanding of the task decision boundary. In these ambiguous cases, explanations can be seen as shaping which scenario is more likely, and it is critical to make the assumptions explicit to support causal claims.

Chapter B

AI Prostate Experiments Details

B.1 Model Impementation Details

Training configurations We use the established nnU-Net implementation[1] for image segmentation. The framework was configured to handle dataset preprocessing, augmentation, and training pipeline generation automatically. The training process utilized a batch size of 8 and a learning rate of 0.001, optimized using the AdamW optimizer. Training was performed over 1000 epochs on one NVIDIA A40 GPU. nnU-Net's default data augmentation techniques, such as random cropping, flipping, and intensity scaling, were employed to improve generalization. For lesion-level prediction, we set the threshold to 0.5. The framework's automatic hyperparameter tuning ensured optimal performance, and we monitored model training using AUROC and average precision on the validation set. A detailed performance is shown in table appendix B.1.

	Training (n=1211)				Testing (n=200)			
	AUROC	AP	Accuracy	F1	AUROC	AP	Accuracy	F1
Per-patient	0.910	0.737	0.847	0.725	0.799	0.624	0.735	0.644
Per-lesion	0.940	0.682	0.948	0.664	0.824	0.484	0.911	0.531

Table B.1: AI model performance.

[1] https://github.com/DIAGNijmegen/picai_baseline

B.2 Statistical Test Details

Table B.2: Performance across different scoring rules and conditions. Values are reported as the mean with 95% confidence intervals (CI).

Scoring Rule	Human-alone	Human+AI (Study 1)	Human+AI (Study 2)	AI
$\mu = 0.0$ specificity	0.587 (0.566, 0.608)	0.593 (0.572, 0.614)	0.606 (0.588, 0.624)	0.611 (0.598, 0.625)
$\mu = 0.125$	0.598 (0.576, 0.620)	0.610 (0.588, 0.632)	0.629 (0.610, 0.648)	0.641 (0.628, 0.655)
$\mu = 0.25$	0.609 (0.583, 0.636)	0.627 (0.602, 0.653)	0.651 (0.629, 0.673)	0.671 (0.656, 0.687)
$\mu = 0.375$	0.620 (0.588, 0.652)	0.645 (0.613, 0.676)	0.674 (0.647, 0.700)	0.701 (0.682, 0.721)
$\mu = 0.5$ accuracy	0.632 (0.593, 0.670)	0.662 (0.624, 0.700)	0.696 (0.664, 0.728)	0.731 (0.708, 0.755)
$\mu = 0.625$	0.610 (0.574, 0.646)	0.638 (0.603, 0.674)	0.670 (0.640, 0.700)	0.704 (0.682, 0.725)
$\mu = 0.75$	0.588 (0.554, 0.623)	0.615 (0.581, 0.649)	0.643 (0.615, 0.672)	0.676 (0.655, 0.697)
$\mu = 0.875$	0.567 (0.533, 0.600)	0.592 (0.559, 0.625)	0.617 (0.589, 0.644)	0.648 (0.627, 0.668)
$\mu = 1.0$ sensitivity	0.545 (0.512, 0.578)	0.568 (0.535, 0.601)	0.590 (0.562, 0.618)	0.620 (0.599, 0.641)

Table B.3: Statistical comparisons across scoring rules (μ values) for different contrasts. t test and Benjamini–Hochberg adjusted p-values.

Scoring Rule	H vs AI	H+AI(1) vs AI	H+AI(2) vs AI	H+AI(1) vs H+AI(2)	H vs H+AI(1)	H vs H+AI(2)
$\mu = 0.0$ specificity	0.0007***	0.017*	$0.082^{\dagger}$	0.178	0.178	0.047*
$\mu = 0.125$	0.0003***	0.009**	$0.074^{\dagger}$	0.170	0.170	0.030*
$\mu = 0.25$	0.0002***	0.006**	$0.073^{\dagger}$	0.145	0.170	0.022*
$\mu = 0.375$	0.0002***	0.006**	$0.073^{\dagger}$	0.132	0.170	0.019*
$\mu = 0.5$ accuracy	0.0002***	0.006**	$0.075^{\dagger}$	0.125	0.170	0.019*
$\mu = 0.625$	0.0002***	0.006**	$0.085^{\dagger}$	0.125	0.170	0.019*
$\mu = 0.75$	0.0003***	0.009**	0.121	0.125	0.178	0.022*
$\mu = 0.875$	0.004**	0.025*	0.170	0.145	0.226	0.046*
$\mu = 1.0$ sensitivity	$0.057^{\dagger}$	0.125	0.330	0.191	0.330	0.125

Note: H = Human-alone, H+AI(1) = Human+AI (Study 1), H+AI(2) = Human+AI (Study 2). *p<0.05, **p<0.01, ***p<0.001, $^{\dagger}p$<0.1. All tests are one-sided (less than).

Table B.4: Degrees of freedom for corresponding t-tests in Table B.3.

Scoring Rule	H vs AI	H+AI(1) vs AI	H+AI(2) vs AI	H+AI(1) vs H+AI(2)	H vs H+AI(1)	H vs H+AI(2)
$\mu = 0.0$	1089.6	1097.1	1641.7	1275.9	1197.9	1270.4
$\mu = 0.125$	1075.3	1086.2	1632.6	1272.6	1197.8	1264.3
$\mu = 0.25$	1063.7	1077.5	1624.0	1270.6	1197.7	1259.8
$\mu = 0.375$	1055.8	1071.8	1617.2	1269.9	1197.6	1257.1
$\mu = 0.5$	1051.5	1069.0	1612.4	1270.3	1197.6	1256.2
$\mu = 0.625$	1050.3	1068.8	1609.1	1271.9	1197.5	1257.0
$\mu = 0.75$	1055.4	1074.1	1608.8	1276.1	1197.5	1261.3
$\mu = 0.875$	1075.2	1091.3	1617.3	1284.4	1197.7	1272.4
$\mu = 1.0$	1110.4	1119.2	1636.7	1293.9	1197.9	1288.0

To assess differences in performance across agent configurations, we conducted one-sided Welch's t-tests, which accommodate unequal variances and sample sizes. This approach aligns with our directional hypotheses (e.g., AI-alone outperforms human-alone) and accounts for inter-group variability.

B.2.1 Pairwise Comparisons

We performed six pairwise comparisons between agent conditions:

1. Human-alone vs. AI-alone
2. Human+AI (Study 1) vs. AI-alone
3. Human+AI (Study 2) vs. AI-alone
4. Human+AI (Study 1) vs. Human+AI (Study 2)
5. Human-alone vs. Human+AI (Study 1)
6. Human-alone vs. Human+AI (Study 2)

All p-values were corrected for multiple testing using the Benjamini–Hochberg procedure to control the false discovery rate.

Numerical performance results are reported in Table B.2, with associated p-values and degrees of freedom summarized in Table B.3 and Table B.4 respectively.

B.3 Demographics

We recruit 8 practicing radiologists, aged 29 to 52 years (mean: 38.4 years). Respondents were primarily from the United States (n=4), Turkey (n=3), and Italy (n=1). Most participants (n=5) reported advanced or expert-level experience with prostate MRI, whilte the others reported intermediate (n=2). One participant did not answer this question.

B.4 Exit Survey Results

Study 1 Results

In Study 1, participants were highly familiar with the AI tool (mean familiarity: 5/5), though its accuracy received a lower mean rating of 2.4/5. Usefulness and trust in the system were

rated moderately, both averaging 3/5. In open-ended feedback, practitioners reported that the AI tool was most helpful in ambiguous cases and increased confidence in detecting lesions in challenging locations such as the anterior, apical, and transition zones. Concerns included oversensitivity in non-cancerous areas and missed lesions, with suggestions for improvement focusing on providing malignancy probability scores, separate reporting of T2 and DWI/ADC scores, and better performance in transitional zone lesions.

Study 2 Results

In Study 2, the AI tool's helpfulness was rated moderately (mean: 2.9/5), with accuracy ratings remaining low to moderate (mean: 2.1/5). Trust in the AI also averaged 2.5/5. Despite moderate satisfaction, respondents expressed a high likelihood of future AI use (mean: 3.75/5). In open-ended feedback, the AI was perceived as useful in ambiguous cases, with one practitioner noting it reinforced decisions to call studies negative. They also pointed out key challenges such as poor performance in transitional zone lesions, overreliance on diffusion restriction, and limitations in segmenting prostate versus non-prostate tissue. Participants' recommendations for improvement included adopting the PI-RADS classification system, enhancing segmentation capabilities, and improving detection of small lesions. Image quality issues were a significant limitation, with practitioners noting that humans outperform AI in evaluating non-diagnostic images, particularly for diffusion-weighted imaging.

B.5 Fine-grained analysis

Table B.5 and Table B.6 provide an overview of the subgroup analysis of human-AI agreement and disagreement in Studies 1 and 2, respectively. The results indicate that performance metrics are significantly better in subgroups where human and AI decisions align compared to those with disagreement.

For a detailed breakdown, individual-level performance for the different agreement and disagreement subgroups is presented. In Study 1, the results are available in Table B.7,

Table B.5: Study 1 fine-grained subgroup performance.

Condition	Avg (#)	Total	Correct	TP	FP	TN	FN	Acc (%)	Sen (%)	Spc (%)
Initial=AI, final=AI	52.0	416	304	122	99	182	13	73.1	90.4	64.8
Initial=AI, final≠AI	0.4	3	0	0	0	0	3	0.0	0.0	N/A
Initial≠AI, final=AI	4.6	37	29	10	5	19	3	78.4	76.9	79.2
Initial≠AI, final≠AI	18.0	144	64	16	63	48	17	44.4	48.5	43.2

Table B.6: Study 2 fine-grained subgroup performance.

Condition	Avg (#)	Total	Correct	TP	FP	TN	FN	Acc (%)	Sen (%)	Spc (%)
Human $\neq$ AI prediction	21.6	173	61	15	86	46	26	35.3	36.6	34.8
Human $=$ AI prediction	78.4	627	496	198	114	298	17	79.1	92.1	72.4

Table B.8, Table B.9, and Table B.10, each focusing on specific subcategories of agreement or disagreement. Similarly, Study 2 individual-level results are provided in Table B.11, offering finer granularity of the analysis.

B.6 Ensemble on Common-50 Cases

Table B.12 presents a detailed performance comparison among AI, Human, Human-ensemble, Human+AI, and Human+AI ensemble (Study 1 and Study 2) for the common 50-case subset. While the results highlight that the Human-ensemble consistently outperforms individual human performance, the advantage of any ensemble method over AI alone is less significant.

Table B.7: Study 1: Cases where human agreed with AI and decision was kept.

Username	Total Cases	Correct	TP	FP	TN	FN	Accuracy
P1	53	40	17	12	23	1	75.5%
P2	46	33	15	13	18	0	71.7%
P3	67	47	19	17	28	3	70.1%
P4	51	37	14	12	23	2	72.5%
P5	51	37	18	12	19	2	72.5%
P6	46	36	9	8	27	2	78.3%
P7	50	35	16	14	19	1	70.0%
P8	52	39	14	11	25	2	75.0%

Table B.8: Study 1: Cases where human agreed but AI initially but still changed decision against AI.

Username	Total Cases	Correct	TP	FP	TN	FN	Accuracy
P6	3	0	0	0	0	3	0.00%

Table B.9: Study 1: cases where human disagreed with AI but kept original decision.

Username	Total Cases	Correct	TP	FP	TN	FN	Accuracy
P1	20	9	2	10	7	1	45.0%
P2	23	10	4	10	6	3	43.5%
P3	2	1	0	1	1	0	50.0%
P4	18	8	2	7	6	3	44.4%
P5	20	9	2	11	7	0	45.0%
P6	22	12	2	5	10	5	54.5%
P7	18	6	2	11	4	1	33.3%
P8	21	9	2	8	7	4	42.9%

Table B.10: Study 1: cases where human disagreed with AI but followed AI advice.

Username	Total Cases	Correct	TP	FP	TN	FN	Accuracy
P1	2	1	1	0	0	1	50.0%
P2	6	6	1	0	5	0	100.0%
P3	6	4	0	1	4	1	66.7%
P4	6	5	2	1	3	0	83.3%
P5	4	4	1	0	3	0	100.0%
P6	4	3	2	1	1	0	75.0%
P7	7	5	2	1	3	1	71.4%
P8	2	1	1	1	0	0	50.0%

Table B.11: Finegrained analysis for Study 2: (1) When Human disagrees with AI, human are prone to errors (accuracy is lower than 50%); (2) Human is better at identifying AI false positives than identifying false negatives, i.e., humans are better at catching AI's false alarms than its missed cases.

Username	#Disagreements	Correct	TP	FP	TN	FN	Accuracy
P1	28	11	1	10	10	7	39.3%
P2	27	6	2	19	4	2	22.2%
P3	11	3	3	8	0	0	27.3%
P4	26	11	1	11	10	4	42.3%
P5	18	7	1	9	6	2	38.9%
P6	20	8	2	6	6	6	40.0%
P7	20	6	2	11	4	3	30.0%
P8	23	9	3	12	6	2	39.1%

Table B.12: Performance comparison between AI, Human, Human-ensemble, Human+AI, and human+AI ensemble (study 1 and 2) for the common 50-case subset.

		Study 1					Study 2		
	AI	Human	Human-ensemble	Human+AI	H+AI ensemble	P (Human-ensemble >Human) P (H+AI ensemble>AI)	Human+AI	H+AI ensemble	P (H+AI ensemble>AI)
AUROC	0.763 [0.727, 0.797]	0.675 [0.630, 0.719]	0.732 [0.690, 0.771]	0.711 [0.668, 0.752]	0.778 [0.741, 0.812]	0.004*/0.265	0.708 [0.666, 0.748]	0.763 [0.726, 0.798]	0.112
Accuracy	70.0% [0.657, 0.745] 35/50	62.5% [0.578, 0.672] 31/50	68.0% [0.635, 0.725] 34/50	65.7% [0.610, 0.703] 33/50	72.0% [0.675, 0.762] 36/50	0.004*/0.216	64.7% [0.600, 0.693] 32/50	70.0% [0.655, 0.745] 35/50	0.229
Sensitivity (Recall)	93.8% [0.892, 0.976] 15/16	81.2% [0.741, 0.878] 13/16	87.5% [0.814, 0.929] 14/16	85.9% [0.797, 0.917] 14/16	93.8% [0.892, 0.976] 15/16	0.028*/0.495	87.5% [0.815, 0.929] 14/16	93.8% [0.892, 0.976] 15/16	0.050
Specificity	58.8% [0.530, 0.646] 20/34	53.7% [0.477, 0.595] 18/34	58.8% [0.529, 0.646] 20/34	56.2% [0.504, 0.620] 19/34	61.8% [0.559, 0.675] 21/34	0.027*/0.197	54.0% [0.482, 0.599] 18/34	58.8% [0.528, 0.647] 20/34	0.498
NPV	95.2% [0.918, 0.982] 20/21	87.0% [0.804, 0.909] 18/21	90.9% [0.864, 0.949] 20/22	90.8% [0.846, 0.938] 19/21	95.5% [0.921, 0.983] 21/22	0.012*/0.467	91.4% [0.854, 0.945] 18/20	95.2% [0.919, 0.982] 20/21	0.051
PPV (Precision)	51.7% [0.453, 0.581] 15/29	45.5% [0.389, 0.517] 13/29	50.0% [0.435, 0.566] 14/28	48.2% [0.416, 0.545] 14/29	53.6% [0.470, 0.602] 15/28	0.005*/0.214	47.4% [0.410, 0.537] 14/30	51.7% [0.452, 0.582] 15/29	0.236

B.7 More Screenshots on User Interface Design

We show screenshots of a login page (Fig. B-1), a consent form (Fig. B-2), a toy demonstration example page (Fig. B-3), and two exit surveys (Fig. B-4, Fig. B-5) for study 1 and study 2 respectively.

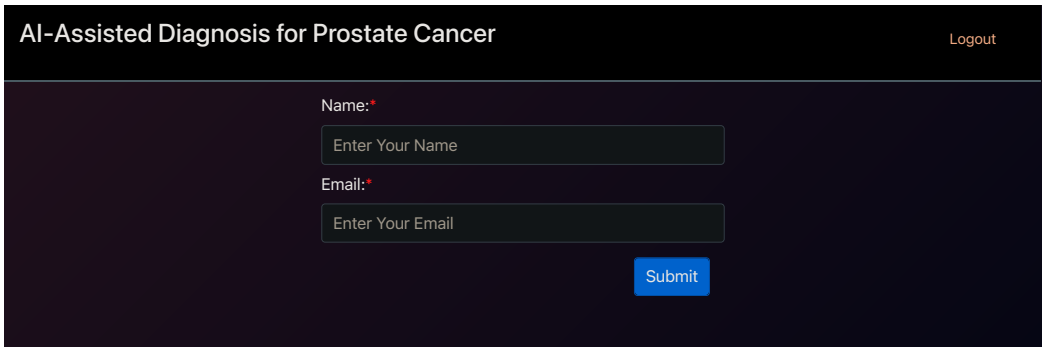


Figure B-1: Login page.

Table B.13: Rational agent's decision rule on the common 50-case subset under the evaluation of different scoring rules.

	Human-alone (d^H)	AI (d^{AI})	$\mathbf{E}_{\theta\sim\hat{\pi}(\theta\mid d^H,d^{AI})}\left[S(d^H,\theta)\right]$	$\mathbf{E}_{\theta\sim\hat{\pi}(\theta\mid d^H,d^{AI})}\left[S(d^{AI},\theta)\right]$	Rational agent (d^r)
$\mu = 1$	0	0	0.995	0.995	$d^H = 0$
	0	1	0.808	0.192	$d^H = 0$
	1	0	0.060	0.940	$d^{AI} = 0$
	1	1	0.282	0.282	$d^H = 1$
$\mu = 0.75$	0	0	0.993	0.993	$d^H = 0$
	0	1	0.713	0.288	$d^H = 0$
	1	0	0.091	0.909	$d^{AI} = 0$
	1	1	0.423	0.423	$d^H = 1$
$\mu = 0.6$	0	0	0.992	0.992	$d^H = 0$
	0	1	0.655	0.345	$d^H = 0$
	1	0	0.109	0.891	$d^{AI} = 0$
	1	1	0.508	0.508	$d^H = 1$
$\mu = 0.5$	0	0	0.991	0.991	$d^H = 0$
	0	1	0.617	0.383	$d^H = 0$
	1	0	0.121	0.879	$d^{AI} = 0$
	1	1	0.564	0.564	$d^H = 1$
$\mu = 0.4$	0	0	0.892	0.892	$d^H = 0$
	0	1	0.555	0.445	$d^H = 0$
	1	0	0.209	0.791	$d^{AI} = 0$
	1	1	0.608	0.608	$d^H = 1$
$\mu = 0.25$	0	0	0.743	0.743	$d^H = 0$
	0	1	0.463	0.538	$d^{AI} = 1$
	1	0	0.341	0.659	$d^{AI} = 0$
	1	1	0.673	0.673	$d^H = 1$
$\mu = 0$	0	0	0.495	0.495	$d^H = 0$
	0	1	0.308	0.692	$d^{AI} = 1$
	1	0	0.560	0.440	$d^H = 1$
	1	1	0.782	0.782	$d^H = 1$

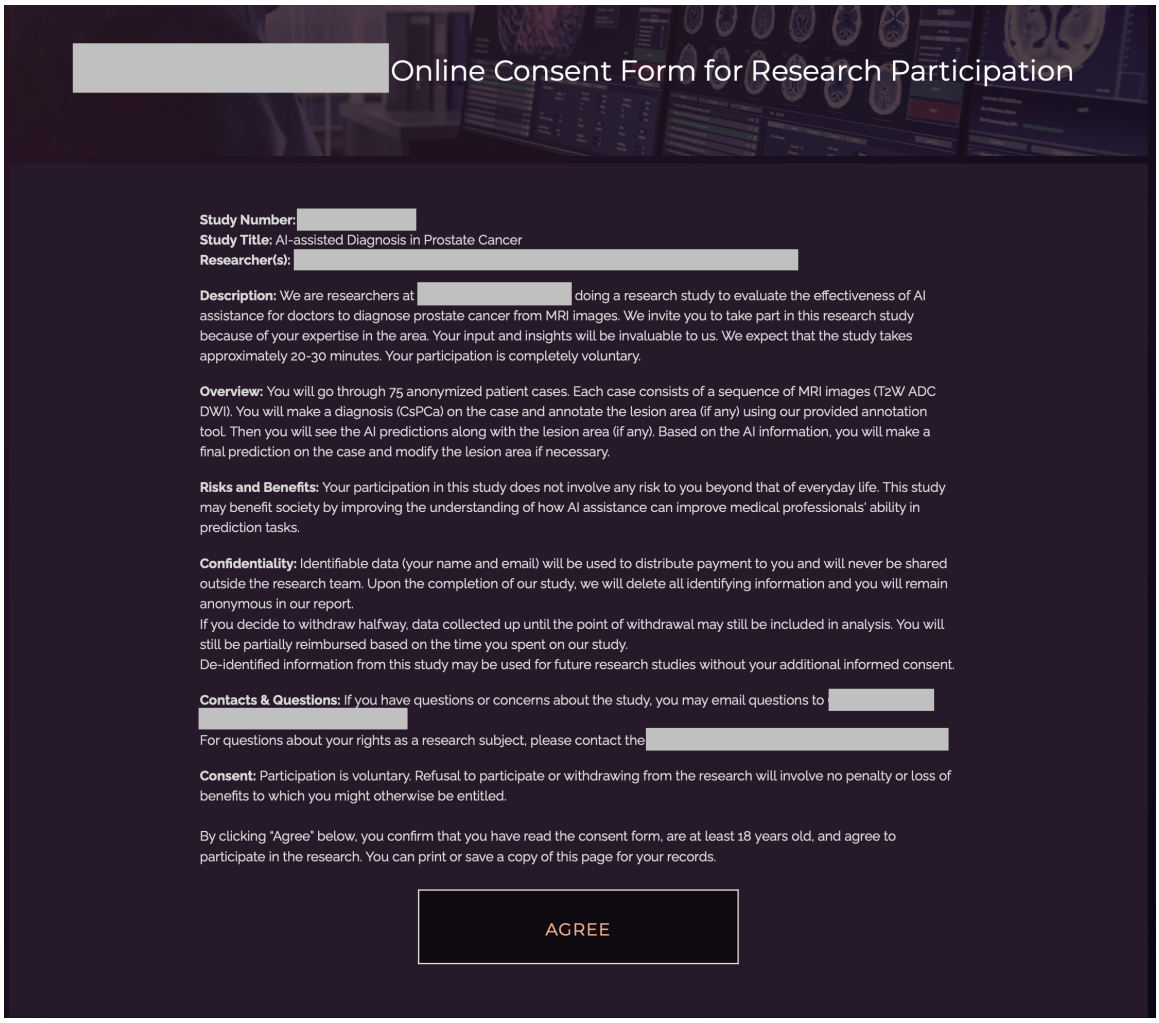


Figure B-2: Consent page.

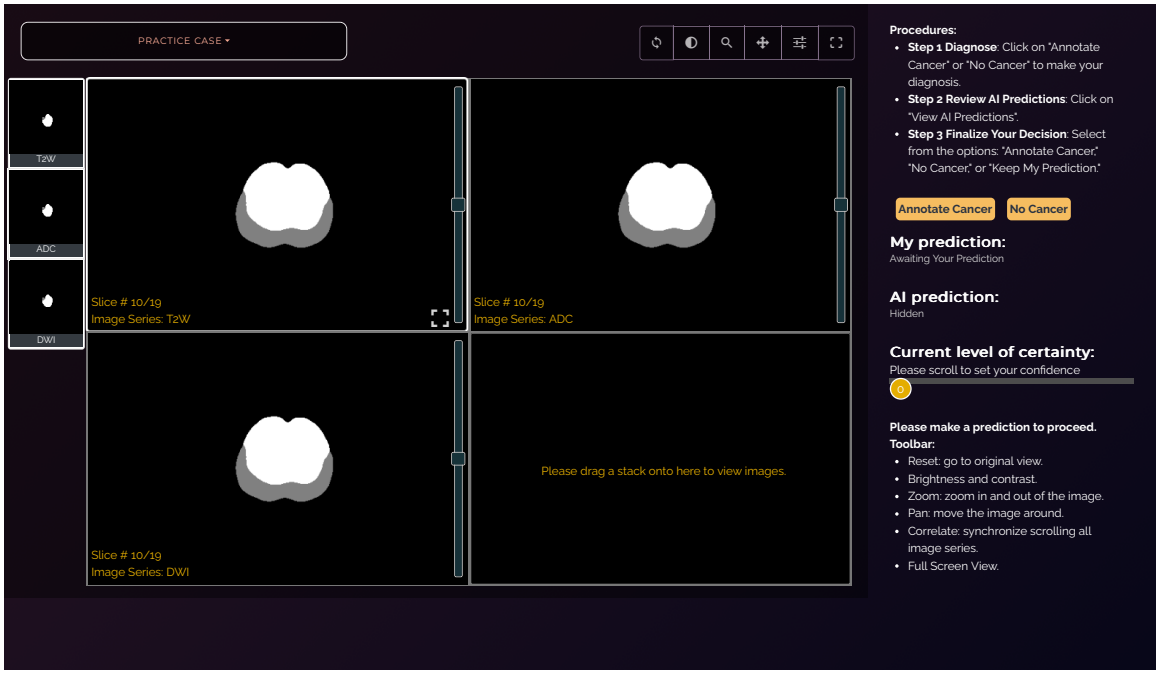


Figure B-3: Toy demonstration example page.

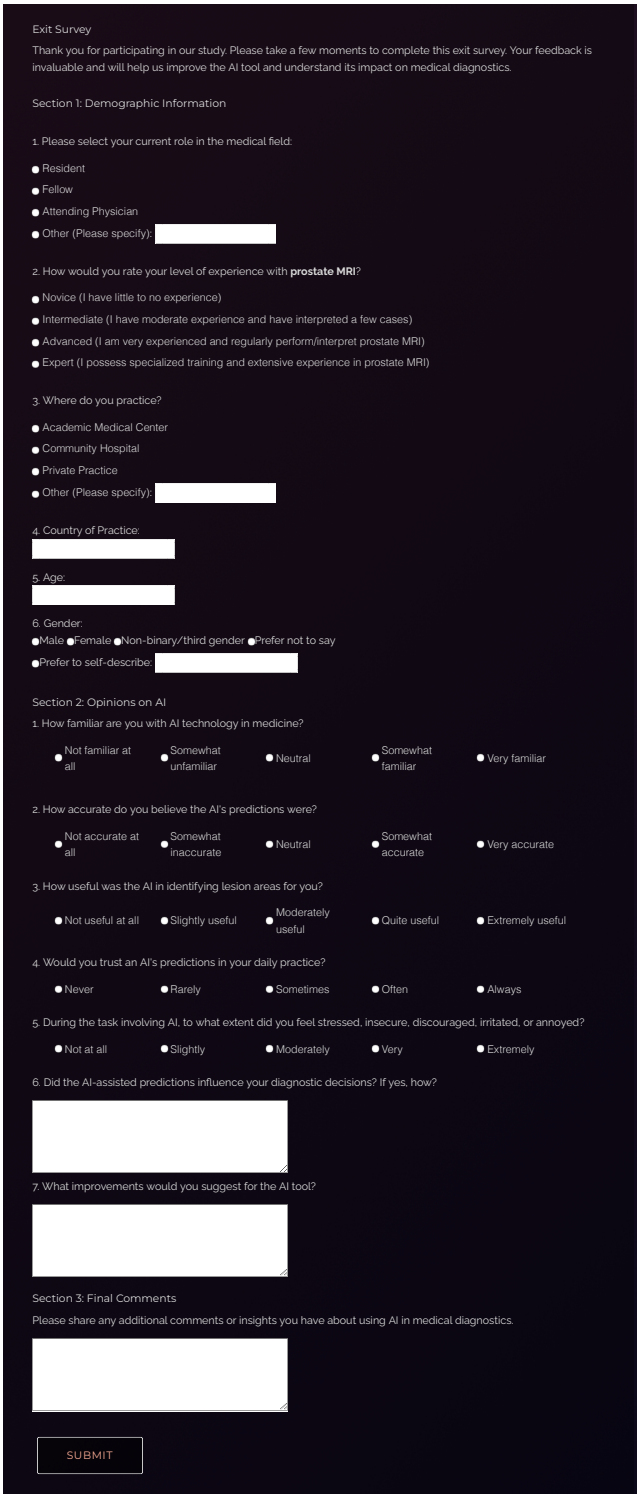


Figure B-4: Exit survey for study 1.

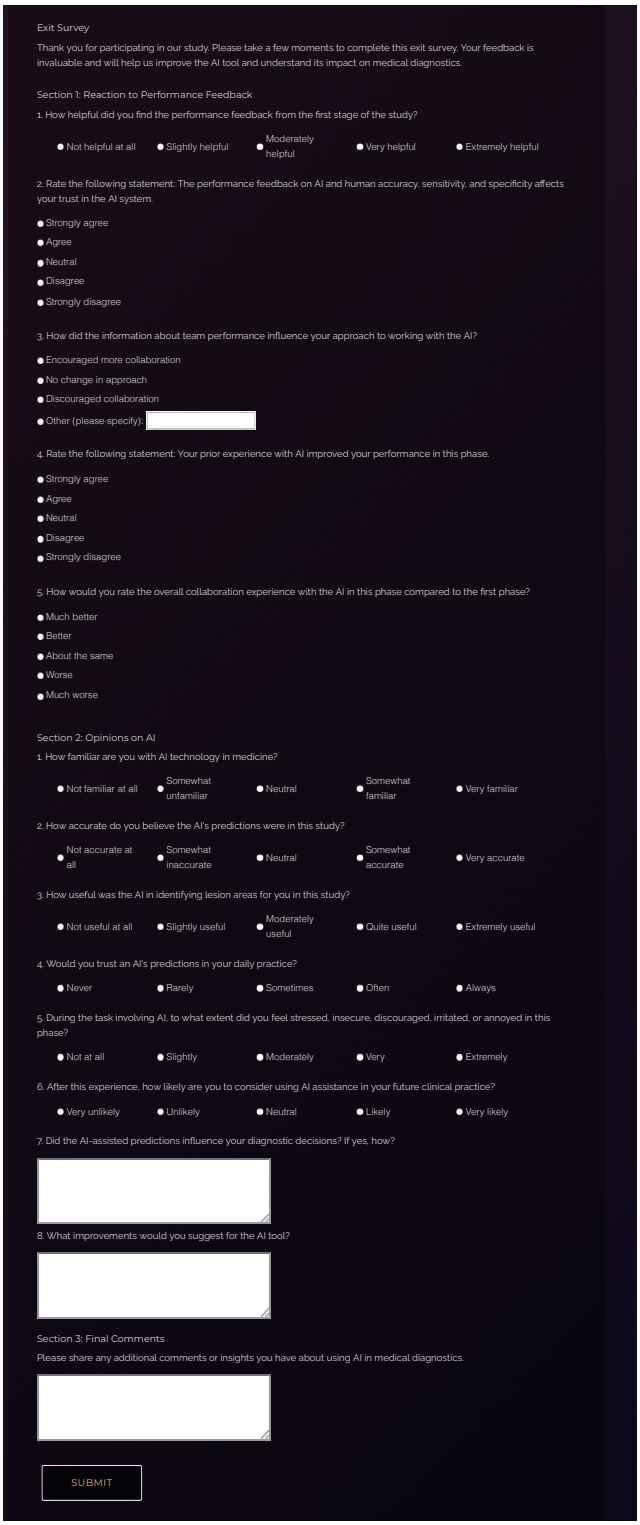


Figure B-5: Exit survey for study 2.

Chapter C

Radiology Report Generation Evaluation Experiments Details

C.1 Model Impementation Details

OpenAI API: We evaluate three datasets using Microsoft's Azure OpenAI service with GPT-4-vision-preview (11/15/2023) and GPT-4o.

Finetuning Llama-2: In the case of the MIMIC dataset, we selectively sample 10% of the official training data, carefully ensuring there is no overlap with the 300-sample test set. For the IU X-ray dataset, we utilize the entire training set, which comprises 3,655 studies, and confirm that these too do not overlap with the test set. The fine-tuning process largely adheres to the default hyperparameters established by Stanford Alpaca [112]. Our hardware includes four A100 GPUs, each equipped with 80GiB of memory, and operates on CUDA version 12.4.

Llama-3.2-90B-Vision-Instruct: Llama-3.2-90B-Vision is the newest flagship open-source multimodal LLM. We choose it to provide a baseline reference. We access the model weights through huggingface and Meta. Our hardware for inference includes four A100 GPUs, each equipped with 80GiB of memory, and operates on CUDA version 12.4.

Code Availability: The source code for our project is publicly accessible on GitHub, enabling users and fellow researchers to review, utilize, or extend our implementations. You can find our repository at https://github.com/ChicagoHAI/cxr-eval-gpt-4v.git.

C.2 Data

Data licenses: MIMIC-CXR license can be found at https://physionet.org/content/mimic-cxr/view-license/2.0.0/. IU X-RAY images are distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (http://creativecommons.org/licenses/by-nc-nd/4.0/).

For MIMIC-CXR, we accessed the data by following the required steps on https://physionet.org/content/mimic-cxr/2.0.0/. We first registered and applied to be a credentialed user, and then completed the required training of CITI Data or Specimens Only Research. We also signed the data use agreement for the project before we get access to the dataset. We downloaded IU X-RAY dataset from https://openi.nlm.nih.gov/faq.

Chexpert Plus dataset is available to download online and the license can be found at https://stanfordaimi.azurewebsites.net/datasets/5158c524-d3ab-4e02-96e9-6ee9efc110a1.

Preprocessing: To prepare the data for the OpenAI API, we first convert the DICOM images to JPEG format, which is required for compatibility with GPT-4V. Then we use base64 encoding to transform the binary image data into its corresponding UTF-8 string.

Ethical consideration of data: There is no substantial concerns around the data, since dataset are de-identified and do not contain harmful or offensive contents.

C.3 Evaluation Experiment Results

C.3.1 Hypothesis Test

Bootstrap Confidence Interval We use **bootstrap confidence interval** [27] to test if GPT-4V labels one certain condition independently of the groundtruth condition group. For this test, for each condition i and group j, we define test statistic θ_{ij} as $P_i^{(j)} - P_i$ and null hypothesis H_0 as $\theta_{ij} = 0$. We construct a 95% confidence interval as $[\hat{\theta}^{(B)}_{ij,\,0.025},\ \hat{\theta}^{(B)}_{ij,\,0.975}]$ with 1000 bootstrap samples for each θ_{ij}. Considering the sparsity of original study pool, we limit our choice of condition i and group j in six most frequent conditions in our subsample.

χ^2 Test Specifically, we use a **χ^2-test** to test if GPT-4V follows the same label distribution across different groups, i.e., testing the null hypothesis (H_0) that $\mathbf{P}_k = \mathbf{P}_{k'}$ for any groups k and k'. For the overall pool, we can construct a 13×13 contingency table with each entry equal to $Y_i^{(j)}$ and then calculate expected count $E_i^{(j)}$ for each entry. Finally, report $\chi^2 = \sum_i \sum_j \frac{(Y_i^{(j)} - E_i^{(j)})^2}{E_i^{(j)}}$. Considering the sparsity of original study pool, we report results of two different tables: (1) A modified table that replaces zero elements with 0.001; (2) A reduced table with only six most frequent conditions in subsample.

Pearson Correlation Coefficient We approximate $P_j^{(k)}$ using $\Pr(X_{ij}^{(k)} = 1)$ to obtain an estimator $\widehat{\mathbf{P}}_\mathbf{k}$ of $\mathbf{P}_\mathbf{k}$ for each group k. Furthermore, we illustrate the correlation $\mathrm{Corr}(\widehat{\mathbf{P}}_\mathbf{m}, \widehat{\mathbf{P}}_\mathbf{n})$ for all groups m and n in Figure C-3 and Figure C-4. It is noted that the condition "Pleural Other" doesn't seem to be highly correlated with other groups. However, considering that "Pleural Other" only has one positive mention in groundtruth conditions and this can be treated as an outlier.

Robustness Check We look into overlap issue between any two groups to further verify our results' robustness. We find out that it does not compromise the assumptions of the multinomial distribution or the robustness of Bootstrap CI results, but it could potentially inflate the Chi-square statistic, leading to incorrect conclusion about "no significance". To

Experiment	IU X-RAY			MIMIC-CXR		
	IMPRESSION	FINDINGS	Labels	IMPRESSION	FINDINGS	Labels
1.1	298/300	259/260	-	300/300	183/183	-
1.2	295/300	259/260	-	300/300	183/183	-
1.3	278/300	241/260	-	300/300	183/183	-
1.4	258/300	223/260	-	300/300	183/183	-
1.5	118/300	101/260	-	83/300	61/183	-
2.1	-	-	237/300	-	-	300/300
3.1	293/300	253/260	-	297/300	182/183	-

Table C.1: Summary of actual sample size across different experiments.

Condition	GT					2.1			2.2				
	Pos	Neg	Unc	Unmnt	Pr(Pos)	Pos	Other	Pr(Pos)	Pos	Neg	Unc	Unmnt	Pr(Pos)
Edema	35	42	15	208	0.117	46	254	0.153	76	174	0	50	0.253
Consolidation	10	17	5	268	0.033	18	282	0.060	30	234	0	36	0.100
Pneumonia	7	37	24	232	0.023	6	294	0.020	14	242	0	44	0.047
Pneumothorax	7	45	3	245	0.023	6	294	0.020	5	272	0	23	0.017
Pleural Effusion	65	30	3	202	0.217	190	110	0.633	212	77	0	11	0.707

Table C.2: Label distribution of top 5 conditions (MIMIC-CXR).

affirm the integrity of our Chi-square test results, we highlight two points: first, comparison with groundtruth studies in Table 4.4 shows a distinct difference in p-values (small for groundtruth and large for generated studies), suggesting minimal impact of overlap on our conclusion that "GPT-4V generates reports following the same label distribution." Additionally, we also check the overlapped samples between groups and find them to be relatively small to original group sizes, with an average ratio of 13.1%, as detailed in Figure C-5.

C.3.2 GPT-4o Evaluation

Overall performance of direct report generation across three datasets is shown in Table C.3, Table C.4 and Table C.5. Complete image reasoning results can be seen in Table C.9. Complete Chi square test can be seen in Table C.10.

Table C.3: Direct report generation performance comparison for IU X-ray findings and impressions (GPT-4o).

Experiment	Lexical metrics				Clinical Efficacy Metrics					
	BLEU-1	BLEU-4	ROUGE-L	METEOR	Pos F1	Pos F1@5	RadGraph	Neg F1	Neg F1@5	Hall.↓
IU X-RAY (FINDINGS)										
1.1	0.316	0.045	0.238	0.311	0.059	0.045	0.203	0.272	0.397	0.354
1.2	0.330	0.049	0.242	0.323	0.077	0.098	0.214	0.285	0.419	0.269
1.3	0.238	0.030	0.207	0.283	0.081	0.146	0.174	0.298	0.458	0.169
1.4	0.239	0.024	0.194	0.231	0.077	0.161	0.144	0.285	0.440	0.138
1.5	0.279	0.044	0.243	0.250	0.037	0.031	0.187	0.240	0.379	0.097
IU X-RAY (IMPRESSION)										
1.1	0.212	0.012	0.194	0.269	0.087	0.105	0.104	0.000	0.000	0.303
1.2	0.248	0.014	0.251	0.341	0.173	0.135	0.153	0.071	0.051	0.307
1.3	0.209	0.012	0.205	0.281	0.111	0.206	0.106	0.000	0.000	0.177
1.4	0.181	0.007	0.143	0.177	0.115	0.255	0.060	0.000	0.000	0.197
1.5	0.138	0.001	0.077	0.124	0.067	0.107	0.034	0.010	0.025	0.211

Table C.4: Performance comparison for MIMIC-CXR findings and impressions (GPT-4o).

Experiment	Lexical metrics				Clinical Efficacy Metrics					
	BLEU-1	BLEU-4	ROUGE-L	METEOR	Pos F1	Pos F1@5	RadGraph	Neg F1	Neg F1@5	Hall.↓
MIMIC-CXR (FINDINGS)										
1.1	0.331	0.031	0.225	0.282	0.134	0.153	0.163	0.244	0.418	0.421
1.2	0.333	0.044	0.234	0.290	0.273	0.255	0.182	0.259	0.431	0.383
1.3	0.276	0.028	0.203	0.264	0.155	0.230	0.154	0.290	0.503	0.298
1.4	0.242	0.017	0.186	0.201	0.172	0.272	0.119	0.249	0.443	0.180
1.5	0.337	0.055	0.257	0.301	0.170	0.203	0.188	0.279	0.481	0.934
MIMIC-CXR (IMPRESSION)										
1.1	0.159	0.006	0.142	0.185	0.123	0.152	0.077	0.038	0.081	0.587
1.2	0.203	0.014	0.195	0.238	0.277	0.258	0.124	0.048	0.126	0.547
1.3	0.168	0.009	0.162	0.200	0.156	0.243	0.095	0.042	0.101	0.359
1.4	0.120	0.004	0.112	0.130	0.177	0.288	0.055	0.008	0.020	0.397
1.5	0.155	0.007	0.153	0.186	0.144	0.232	0.073	0.035	0.091	0.477

C.3.3 GPT-4-vision-preview Evaluation

Medical Restriction in Usage of GPT-4V API Since OpenAI itself will restrict the medical use of GPT-4V, the actual sample size of our expriments will be smaller than 300. Details can be checked in Table C.1.

Misspecification of Negative Mentions in GPT-4V Besides, it is noted that the "negative" category in 2-class labeling actually includes negative, uncertain, and unmentioned mentions. However, GPT-4V shows very similar Negative F1 scores across different labeling types.

Table C.5: Performance comparison for ChexPert Plus findings and impressions (GPT-4o).

Experiment	Lexical metrics				Clinical Efficacy Metrics					
	BLEU-1	BLEU-4	ROUGE-L	METEOR	Pos F1	Pos F1@5	RadGraph	Neg F1	Neg F1@5	Hall.↓
CHEXPERT PLUS (FINDINGS)										
1.1	0.237	0.015	0.176	0.191	0.228	0.191	0.112	0.118	0.174	0.516
1.3	0.191	0.007	0.159	0.172	0.210	0.325	0.101	0.158	0.253	0.371
1.4	0.166	0.011	0.155	0.139	0.234	0.339	0.077	0.160	0.262	0.339
1.5	0.171	0.007	0.158	0.149	0.188	0.224	0.094	0.104	0.135	0.411
3.1	0.085	0.003	0.150	0.108	0.483	0.724	0.080	0.148	0.250	0.226
CHEXPERT PLUS (IMPRESSION)										
1.1	0.105	0.001	0.106	0.090	0.163	0.198	0.008	0.013	0.035	0.680
1.3	0.109	0.002	0.105	0.095	0.172	0.261	0.008	0.042	0.085	0.377
1.4	0.066	0.000	0.093	0.069	0.192	0.242	0.012	0.000	0.000	0.400
1.5	0.084	0.002	0.131	0.088	0.134	0.161	0.029	0.043	0.111	0.370
3.1	0.106	0.003	0.164	0.116	0.854	0.947	0.022	0.662	0.924	0.480

Experiment	Lexical metrics				Clinic Efficacy Metrics					
	BLEU-1	BLEU-4	ROUGE	METEOR	Pos F1	Pos F1@5	Rad. F1	Neg F1	Neg F1@5	Hall.↓
MIMIC-CXR (FINDINGS)										
1.1	0.299	0.035	0.214	0.279	0.117	0.124	0.135	0.241	0.396	0.563
1.2	0.323	0.042	0.227	0.294	0.181	0.194	0.159	0.272	0.464	0.410
1.3	0.265	0.019	0.186	0.262	0.134	0.236	0.109	0.237	0.437	0.607
1.4	0.236	0.008	0.176	0.202	0.151	0.233	0.080	0.151	0.328	0.563
1.5	0.294	0.053	0.223	0.293	0.085	0.036	0.149	0.251	0.462	1.000
MIMIC-CXR (IMPRESSION)										
1.1	0.135	0.018	0.119	0.161	0.118	0.160	0.071	0.004	0.001	0.687
1.2	0.176	0.021	0.163	0.200	0.185	0.200	0.101	0.037	0.096	0.610
1.3	0.141	0.009	0.120	0.174	0.141	0.228	0.068	0.026	0.067	0.593
1.4	0.113	0.002	0.107	0.133	0.150	0.255	0.058	0.023	0.061	0.607
1.5	0.163	0.011	0.160	0.242	0.070	0.072	0.088	0.000	0.000	0.578

Table C.6: Direct report generation performance comparison for MIMIC-CXR findings and impressions (GPT-4-vision-preview).

This suggests that GPT-4V often incorrectly labels uncertain and unmentioned conditions as negative, contributing to its poor performance in accurately identifying negative mentions. More evidence on label distribution is available in Table C.2.

Overall Results of GPT-4V Table C.6 and Table C.7 show all results on both findings section and impression section. Complete image reasoning results can be seen in Table C.8. Complete Chi square test can be seen in Table C.11. Table C.12 show the generated results with groundtruth conditions.

Experiment	Lexical metrics				Clinic Efficacy Metrics					
	BLEU-1	BLEU-4	ROUGE	METEOR	Pos F1	Pos F1@5	Rad. F1	Neg F1	Neg F1@5	Hall.↓
IU X-RAY (FINDINGS)										
1.1	0.278	0.038	0.218	0.326	0.030	0.024	0.178	0.284	0.429	0.494
1.2	0.282	0.042	0.216	0.328	0.023	0.010	0.174	0.308	0.475	0.614
1.3	0.237	0.027	0.189	0.281	0.053	0.052	0.140	0.265	0.429	0.523
1.4	0.233	0.016	0.179	0.235	0.072	0.119	0.105	0.226	0.402	0.619
1.5	0.325	0.037	0.247	0.318	0.061	0.080	0.191	0.290	0.455	0.287
IU X-RAY (IMPRESSION)										
1.1	0.219	0.019	0.232	0.295	0.036	0.041	0.155	0.000	0.000	0.275
1.2	0.209	0.021	0.215	0.295	0.058	0.060	0.169	0.020	0.052	0.410
1.3	0.202	0.013	0.205	0.287	0.041	0.051	0.142	0.041	0.106	0.435
1.4	0.172	0.009	0.155	0.216	0.052	0.100	0.097	0.000	0.000	0.628
1.5	0.247	0.026	0.243	0.279	0.036	0.047	0.138	0.026	0.067	0.263

Table C.7: Direct report generation performance comparison for IU X-RAY findings and impressions (GPT-4-vision-preview).

Metric	MIMIC-CXR		IU X-RAY	
	Chain-of-Thought (1st Step)	Image Reasoning	Chain-of-Thought (1st Step)	Image Reasoning
Positive F1	0.166	0.146	0.072	0.049
Positive F1@5	0.261	0.208	0.095	0.056

Table C.8: Image reasoning performance of GPT-4-vision-preview on chest X-ray images.

C.3.4 Llama-3.2-90B-Vision Evaluation

Overall performance of direct report generation across three datasets is shown in Table C.13, Table C.14 and Table C.15. It is noted that low hallucination of Llama-generated reports is probably due to its straightforward and assertive writing style, exemplified by sentences such as "The heart is enlarged," "There is a left pleural effusion," and "There is a right lung opacity."

C.3.5 Examples of Generated Reports

The generated examples by different prompting strategies are shown in table C.16.

C.4 Human Reader Study

Instructions for human study are shown in Figure C-6.

Metric	MIMIC-CXR		CheXpert Plus	
	Chain-of-Thought (1st Step)	Image Reasoning	Chain-of-Thought (1st Step)	Image Reasoning
Positive F1	0.195	0.161	0.221	0.189
Positive F1@5	0.298	0.242	0.257	0.247

Table C.9: Image reasoning performance of GPT-4o on chest X-ray images.

	MIMIC				CheXpert Plus			
	Overall (GT)	Overall (Gen)	Top 6 (GT)	Top 6 (Gen)	Overall (GT)	Overall (Gen)	Top 6 (GT)	Top 6 (Gen)
χ^2 statistic	1770.38	66.05	317.86	8.71	1546.86	53.24	309.31	3.88
p-value	p<1e-1	1	p<1e-1	0.9989	p<1e-1	1	p<1e-1	1
df.	144	144	25	25	144	144	25	25

Table C.10: χ^2-test for homogeneity of label distribution across different condition groups (GPT-4o).

Statistics	Overall		Top 6 Conditions	
	Groundtruth	GPT-4V	Groundtruth	GPT-4V
χ^2 statistic	1770.38	74.25	317.86	6.11
p-value	p < 1e-4	1.00	p < 1e-4	1.00
df.	144	144	25	25

Table C.11: χ^2-test for homogeneity of label distribution across different condition groups (GPT-4-vision-preview).

Experiment	Lexical metrics				Clinic Efficacy Metrics					
	BLEU-1	BLEU-4	ROUGE	METEOR	Pos F1	Pos F1@5	Rad. F1	Neg F1	Neg F1@5	Hall.↓
MIMIC-CXR										
GPT-4V	0.135	0.018	0.119	0.161	0.118	0.160	0.071	0.004	0.001	0.687
GPT-4V (gt)	0.176	0.007	0.185	0.179	0.885	**0.977**	0.103	0.584	**0.958**	**0.431**
LLaMA-2 (gt)	**0.301**	**0.094**	**0.330**	**0.348**	**0.923**	0.957	**0.286**	**0.703**	0.941	0.710
IU X-RAY										
GPT-4V	0.219	0.019	0.232	0.295	0.036	0.041	0.155	0.000	0.000	0.275
GPT-4V (gt)	0.216	0.003	0.229	0.207	0.852	0.919	0.089	**0.630**	0.868	0.235
LLaMA-2 (gt)	**0.454**	**0.124**	**0.460**	**0.441**	**0.871**	**0.928**	**0.297**	0.627	**0.963**	**0.110**

All metrics are evaluated on the impression section.

Table C.12: Performance in report generation with groundtruth conditions (GPT-4-vision-preiew).

Experiment	Lexical metrics				Clinical Efficacy Metrics					
	BLEU-1	BLEU-4	ROUGE-L	METEOR	Pos F1	Pos F1@5	RadGraph	Neg F1	Neg F1@5	Hall.
MIMIC-CXR (FINDINGS)										
1.1	0.258	0.026	0.236	0.216	0.159	0.266	0.123	0.200	0.267	0.607
1.2	0.251	0.020	0.249	0.215	0.123	0.217	0.144	0.214	0.305	0.710
1.3	0.110	0.012	0.114	0.122	0.118	0.221	0.073	0.054	0.111	0.262
1.4	0.015	0.000	0.051	0.032	0.172	0.300	0.010	0.024	0.045	0.197
1.5	0.233	0.023	0.233	0.202	0.144	0.249	0.120	0.197	0.270	0.579
MIMIC-CXR (IMPRESSION)										
1.1	0.089	0.001	0.085	0.094	0.121	0.231	0.036	0.000	0.000	0.161
1.2	0.085	0.000	0.076	0.088	0.132	0.186	0.028	0.000	0.000	0.200
1.3	0.089	0.001	0.089	0.100	0.141	0.215	0.046	0.045	0.117	0.049
1.4	0.095	0.001	0.068	0.081	0.194	0.334	0.022	0.000	0.000	0.990
1.5	0.086	0.000	0.081	0.088	0.111	0.216	0.035	0.000	0.000	0.160

Table C.13: Direct report generation performance comparison for MIMIC-CXR findings and impressions (Llama-3.2-90B-Vision-Instruct).

Experiment	Lexical metrics				Clinical Efficacy Metrics					
	BLEU-1	BLEU-4	ROUGE-L	METEOR	Pos F1	Pos F1@5	RadGraph	Neg F1	Neg F1@5	Hall.
IU X-RAY (FINDINGS)										
1.1	0.248	0.018	0.231	0.231	0.064	0.075	0.146	0.220	0.255	0.404
1.2	0.244	0.019	0.239	0.216	0.021	0.056	0.156	0.221	0.265	0.342
1.3	0.203	0.027	0.190	0.187	0.068	0.082	0.138	0.170	0.308	0.142
1.4	0.015	0.001	0.053	0.037	0.104	0.214	0.008	0.010	0.012	0.677
1.5	0.231	0.018	0.236	0.221	0.061	0.094	0.158	0.227	0.278	0.579
IU X-RAY (IMPRESSION)										
1.1	0.117	0.000	0.063	0.102	0.054	0.087	0.019	0.000	0.000	0.057
1.2	0.117	0.000	0.058	0.094	0.080	0.197	0.015	0.000	0.000	0.043
1.3	0.096	0.000	0.059	0.104	0.058	0.072	0.023	0.012	0.030	0.011
1.4	0.049	0.000	0.054	0.053	0.082	0.186	0.003	0.015	0.040	0.983
1.5	0.114	0.000	0.065	0.100	0.063	0.102	0.018	0.000	0.000	0.037

Table C.14: Direct report generation performance comparison for IU X-RAY findings and impressions (Llama-3.2-90B-Vision-Instruct).

Experiment	Lexical metrics				Clinical Efficacy Metrics					
	BLEU-1	BLEU-4	ROUGE-L	METEOR	Pos F1	Pos F1@5	RadGraph	Neg F1	Neg F1@5	Hall. Prop.
CHEXPERT PLUS (FINDINGS)										
1.1	0.166	0.006	0.175	0.147	0.214	0.290	0.092	0.096	0.153	0.419
1.3	0.079	0.004	0.100	0.088	0.230	0.266	0.069	0.069	0.180	0.339
1.4	0.022	0.003	0.062	0.029	0.247	0.355	0.006	0.042	0.059	0.097
1.5	0.152	0.005	0.173	0.131	0.261	0.285	0.092	0.109	0.170	0.355
CHEXPERT PLUS (IMPRESSION)										
1.1	0.073	0.002	0.106	0.072	0.152	0.221	0.013	0.003	0.007	0.173
1.3	0.086	0.002	0.100	0.084	0.174	0.238	0.013	0.031	0.081	0.058
1.4	0.091	0.000	0.087	0.071	0.190	0.281	0.005	0.000	0.000	0.985
1.5	0.067	0.001	0.102	0.069	0.154	0.209	0.012	0.003	0.007	0.157

Table C.15: Performance comparison for Chexpert Plus findings and impressions (Llama-3.2-90B-Vision-Instruct).

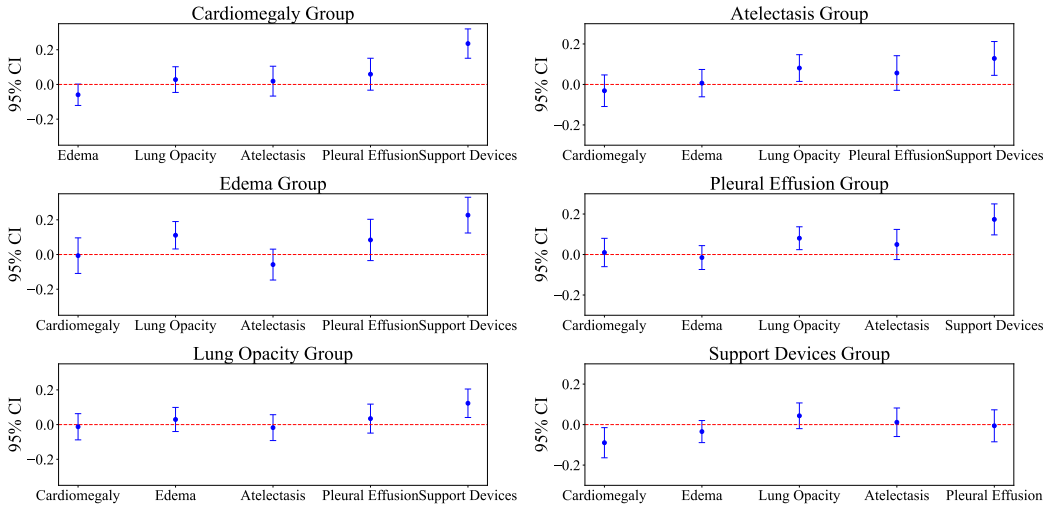


Figure C-1: 95% Bootstrap confidence interval of top 6 conditions in our sample for GPT-4-vision-preview.

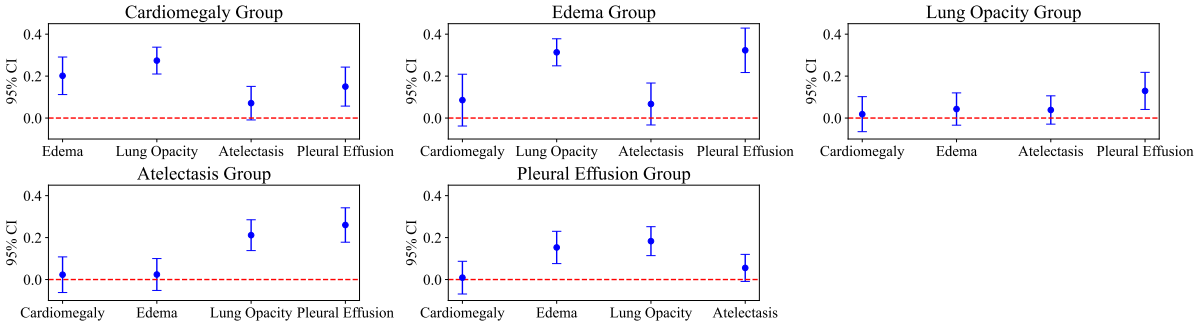


Figure C-2: 95% Bootstrap confidence interval of top 5 conditions in our sample for GPT-4o.

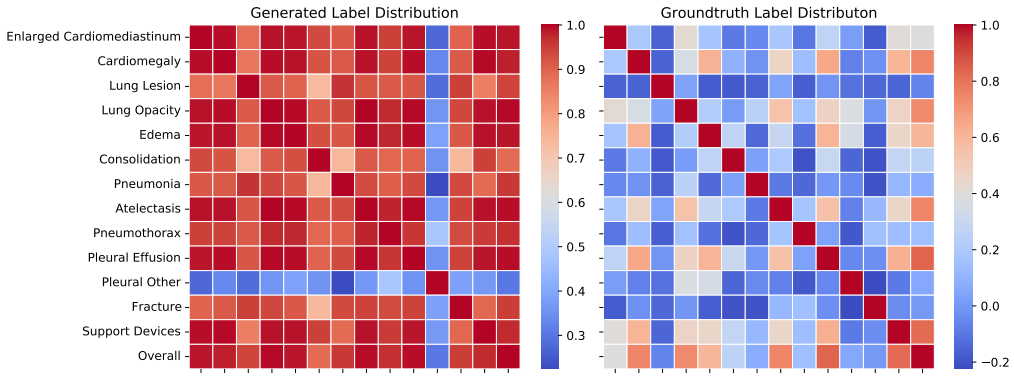


Figure C-3: Correlations between distributions of Pr(Pos) for different condition groups (GPT-4-vision-preview).

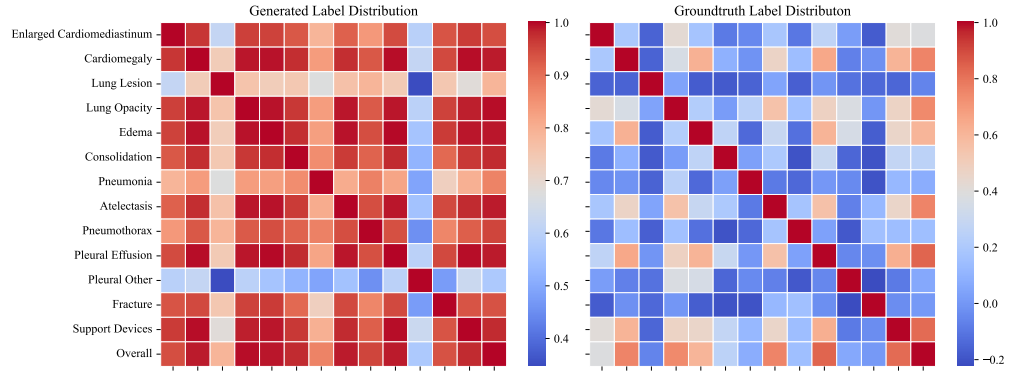


Figure C-4: Correlations between distributions of Pr(Pos) for different condition groups (GPT-4o).

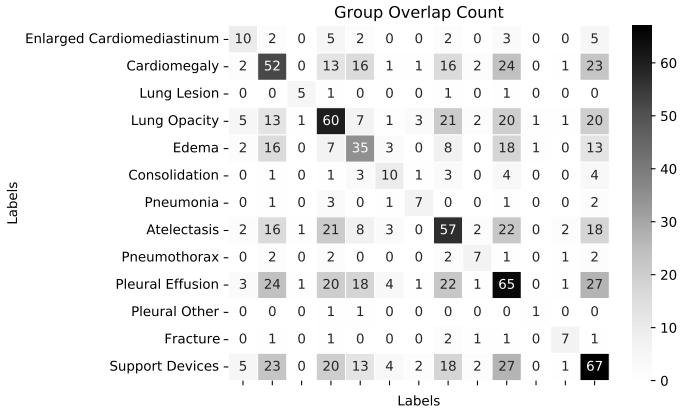


Figure C-5: Overlap count heatmap within groudtruth condition groups. The overlap between any two groups is relatively small compared to the original group sizes, which relate to the counts on the diagonal, ensuring that this overlap does not compromise the robustness of our results.

Experiment	Findings	Impression
GroundTruth	Single AP view of the chest provided. Interval increase in opacification of the right hemithorax with unchanged rightward shift of midline structures. The left lung is clear. Hilar and cardiomediastinal contours are obscured.	Worsening collapse of the right lung from ___ with unchanged rightward shift of mediastinal structures.
Basic	The cardiomediastinal silhouette is severely enlarged. There is opacification of the right hemithorax with loss of visualization of the right heart border; this may be attributable to a large pleural effusion, consolidation, or a combination of both. The left lung field remains clear with normal aeration, without evidence of focal consolidation. The trachea is midline. There is no evidence of pneumothorax on the left. The osseous structures of the visualized portion of the thorax appear intact. Nasogastric tube is in place with its tip projecting over the gastric bubble.	1. Severe cardiomegaly. 2. Opacification of the right hemithorax consistent with a large pleural effusion and/or consolidation. 3. No pneumothorax or acute bony abnormality identified. 4. The presence of a nasogastric tube with an appropriately situated tip. Further investigation with a CT scan may be considered to differentiate between effusion and consolidation and to assess for any additional intrathoracic pathology. Clinical correlation is recommended.
+Indication	The cardiac silhouette appears enlarged. There is an opacity in the right upper lobe consistent with lobar collapse and volume loss as evidenced by crowding of adjacent rib spaces. There are also patchy infiltrates seen in the right lower lobe. The remaining lung fields are clear without evidence of pleural effusion. The bony thorax and soft tissues appear unremarkable.	Right upper lobe and right lower lobe collapse with an enlarged cardiac silhouette. These findings could be consistent with lung atelectasis secondary to obstruction, possibly due to an underlying mass or mucous plugging, and heart failure should be considered as a differential for cardiomegaly. Further diagnostic evaluation and correlation with clinical findings are recommended.
+Instruction	- Enlarged Cardiomediastinum is observed. - Cardiomegaly is present. - Evidence of Lung Opacity is seen. - Extensive consolidation is apparent. - There is Pleural Effusion. - Support devices in the form of a central venous catheter are noted. - No definite pneumothorax, pleural lesions, or fractures are identified.	- Cardiomegaly with associated enlargement of the cardiomediastinum. - Significant lung opacity indicating consolidation, likely due to an infectious or inflammatory process, necessitating further clinical correlation. - Pleural effusion is present. - Central venous catheter in situ. - Absence of pneumothorax or significant pleural disease.
CoT	The cardiomediastinal silhouette is enlarged due to cardiomegaly. There are findings suggesting the presence of pulmonary edema, as evidenced by the perihilar alveolar opacities bilaterally, which may extend to the periphery. No evidence of atelectasis. There is a presence of pleural effusion and support devices, including endotracheal tube above the carina and a gastric tube in the stomach. No lung masses or nodules are identified. No definite pneumothorax or pneumoperitoneum.	Cardiomegaly with signs of pulmonary edema and pleural effusion. Supportive devices in situ. No evidence of atelectasis, pneumothorax, or acute bony injury.
Few-shot	AP supine and lateral decubitus views of the chest demonstrate extensive consolidation involving the right hemithorax with associated volume loss. Evidence of air bronchograms within the consolidation suggests an alveolar filling process, such as pneumonia. There is near-complete opacification of the right lung, with the trachea deviated towards the right side, indicative of volume loss. No visible pneumothorax or significant pleural effusion is noted on the limited view available. The left lung is hyperinflated, likely compensatory. There is a loss of definition of the right heart border and diaphragm silhouette in keeping with the adjacent consolidation. The left hemidiaphragm and heart borders are normal. Osseous structures appear intact but are poorly visualized due to overlying opacity.	Extensive right lung consolidation with volume loss, likely representing a severe pneumonia process. No clear evidence of pleural effusion or pneumothorax on the available projections. Compensatory hyperinflation of the left lung. Clinical correlation and possibly further imaging, such as a CT scan, are recommended for a comprehensive assessment.

Table C.16: Comparison of generated reports across different prompting strategies for one study.

Thanks for participating in our human evaluation study! Feel free to reach me at [redacted] if you encounter any issue or any questions in the process.

To begin, please register at https://physionet.org/content/mimic-cxr/2.0.0/ and sign the data agreement before the study.

You will be provided with a zipped folder containing an Excel sheet ***reader_study_{YOUR_NAME}.csv***, which is the main file you will work on, accompanied by chest X-ray images. If you are a Windows user, no additional action is required. However, if you are using a MacBook, please upload the entire folder to OneDrive and open the Excel sheet from there.

We would like you to complete the rating for 50 patient cases. Each case will include a chest x-ray image, and three reports generated by different methods. ***Please note that the reports we provided are only the impression section.***

Please simulate the real clinical workflow, and give your ratings to the following dimensions:

1. First, take a look at the chest x-ray, and provide your decision on whether each of the reports (impression section) are clinically usable or not, as you would in real practice. (1 means usable and 0 means not usable).

And then, please use likert scale to evaluate each of them in terms of

2. Clinical Efficacy -- Diagnostic Accuracy (Likert 1-5)
   This dimension evaluates how accurate the report is in diagnosing the patient's condition based on the provided chest X-ray image.
3. Clinical Efficacy -- Completeness (Likert 1-5)
   This dimension assesses whether the report covers all relevant clinical findings and provides a comprehensive summary of the patient's condition.
4. Lexical Metrics -- Clarity/Readability (Likert 1-5)
   This dimension evaluates the linguistic quality of the report, focusing on how clear, concise, and easily understandable the text is.

5 is the best score and 1 is the lowest score.

A few notes:

1. Study might contain multiple images. Please click the link_1, link_2, … to view them.
2. Some reports might have masked measurements (xxx cm), and dates, doctor names. Those are de-identified information, please do not consider them as errors.
3. If a report has mention of prior study, please ignore that part and provide your rating only based on the current study's information.

Additionally, please feel free to leave any free-form notes or comments regarding each case. Once you have completed the evaluation for all 50 cases, please return only the Excel sheet to [redacted].

Figure C-6: Human reader study instructions.

C.5 Prompts

Table C.17: Prompt 1.1 Basic generation: direct report generation based on chest X-ray images.

System	You are a professional chest radiologist that reads chest X-ray image(s).
User	Write a report that contains only the FINDINGS and IMPRESSION sections based on the attached images. Provide only your generated report, without any additional explanation and special format. Your answer is for reference only and is not used for actual diagnosis.

Table C.18: Prompt 1.2 Indication enhancement: providing the indication section.

System	You are a professional chest radiologist that reads chest X-ray image(s).
User	Below is INDICATION related to chest X-ray images. INDICATION: {} Write a report that contains only the FINDINGS and IMPRESSION sections based on the attached images and INDICATION. Provide only your generated report, without any additional explanation and special format. Your answer is for reference only and is not used for actual diagnosis.

Table C.19: Prompt 1.3 Instruction enhancement: providing information on medical condition labels.

System	You are a professional chest radiologist that reads chest X-ray image(s).
User	Below is an observation plan consisting of 14 conditions: “No Finding”, “Enlarged Cardiomediastinum”, “Cardiomegaly”, “Lung Lesion”, “Lung Opacity”, “Edema”, “Consolidation”, “Pneumonia”, “Atelectasis”, “Pneumothorax”, “Pleural Effusion”, “Pleural Other”, “Fracture”, “Support Devices”. Based on attached images, assign labels for each condition except “No Finding”: “1”, “0”, “-1”, “2”. It is noted that “No Finding” is either “2” or “1”. These labels have the following interpretation: 1 - The observation was clearly present on the chest X-ray image. 0 - The observation was absent on the chest X-ray image and was mentioned as negative. -1 - The observation was unclear if it exists. 2 - The observation was absent but not explicitly mentioned. Based on labels you choose for each condition, write a report that contains only the FINDINGS and IMPRESSION sections. Don’t return any of your assigned labels. Provide only your generated report, without any additional explanation and special format. Your answer is for reference only and is not used for actual diagnosis.

Table C.20: Prompt 1.4 Chain-of-Thought: step 1 - medical condition labeling; step 2 - report synthesis.

System	You are a professional chest radiologist that reads chest X-ray image(s).
User	Below is an observation plan consisting of 14 conditions: “No Finding”, “Enlarged Cardiomediastinum”, “Cardiomegaly”, “Lung Lesion”, “Lung Opacity”, “Edema”, “Consolidation”, “Pneumonia”, “Atelectasis”, “Pneumothorax”, “Pleural Effusion”, “Pleural Other”, “Fracture”, “Support Devices”. Based on attached images, assign labels for each condition except “No Finding”: “1”, “0”, “-1”, “2”. It is noted that “No Finding” is either “2” or “1”. These labels have the following interpretation: 1 - The observation was clearly present on the chest X-ray image. 0 - The observation was absent on the chest X-ray image and was mentioned as negative. -1 - The observation was unclear if it exists. 2 - The observation was absent but not explicitly mentioned. The first step is to return one list of your assigned labels. For multiple images, assign the labels based on all images and return only one list of labels for the given 14 conditions. The second step is to write a report that contains only the FINDINGS and IMPRESSION sections based on labels you choose for each condition. Your answer is for reference only and is not used for actual diagnosis. Strictly follow the format below to provide your output. <LABEL> [(“No Finding”, “1”|“2”), (“Enlarged Cardiomediastinum”, “0”|“1”|“2”|“-1”), (“Cardiomegaly”, “0”|“1”|“2”|“-1”), (“Lung Lesion”, “0”|“1”|“2”|“-1”), (“Lung Opacity”, “0”|“1”|“2”|“-1”), (“Edema”, “0”|“1”|“2”|“-1”), (“Consolidation”, “0”|“1”|“2”|“-1”), (“Pneumonia”, “0”|“1”|“2”|“-1”), (“Atelectasis”, “0”|“1”|“2”|“-1”), (“Pneumothorax”, “0”|“1”|“2”|“-1”), (“Pleural Effusion”, “0”|“1”|“2”|“-1”), (“Pleural Other”, “0”|“1”|“2”|“-1”), (“Fracture”, “0”|“1”|“2”|“-1”), (“Support Devices”, “0”|“1”|“2”|“-1”)] </LABEL> <REPORT> FINDINGS: <findings>

Table C.21: Prompt 1.5 Few-shot: few-shot in-context learning given a few examples (MIMIC).

System	You are a professional chest radiologist that reads chest X-ray image(s).
User	Write a report that contains only the FINDINGS and IMPRESSION sections based on the attached images. Provide only your generated report, without any additional explanation and special format. Your answer is for reference only and is not used for actual diagnosis. [.JPEG] FINDINGS: Single portable view of the chest is compared to previous exam from ___. Enteric tube is seen with tip off the inferior field of view. Left PICC is seen; however, tip is not clearly delineated. Persistent bibasilar effusions and a right pigtail catheter projecting over the lower chest. There is possible right apical pneumothorax. Superiorly, the lungs are clear of consolidation. Cardiac silhouette is within normal limits. Osseous and soft tissue structures are unremarkable. IMPRESSION: No significant interval change with bilateral **pleural effusions** with right pigtail catheter in the lower chest. Possible small right apical pneumothorax. [.JPEG] FINDINGS: Frontal and lateral radiographs of the chest show hyperinflated lungs with flattened diaphragm, consistent with emphysema. Asymmetric opacity in the right middle lobe is concerning for pneumonia. No pleural effusion or pneumothorax is seen. The cardiomediastinal contours are within normal limits aside from a tortuous aorta. IMPRESSION: Right middle lobe opacity concerning for **pneumonia**. [.JPEG] FINDINGS: PA and lateral views of the chest provided. Midline sternotomy wires and mediastinal clips again noted. Suture is again noted in the right lower lung with adjacent rib resection. There is mild scarring in the right lower lung as on prior. There is no focal consolidation, large effusion or pneumothorax. No signs of congestion or edema. The heart remains moderately enlarged. The mediastinal contour is stable. IMPRESSION: Postsurgical changes in the right hemithorax. Mild **cardiomegaly** unchanged. No edema or pneumonia. [.JPEG] FINDINGS: PA and lateral views of the chest provided. Biapical pleural parenchymal scarring noted. No focal consolidation concerning for pneumonia. No effusion or pneumothorax. No signs of congestion or edema. Cardiomediastinal silhouette is stable with an unfolded thoracic aorta and top-normal heart size. Bony structures are intact. IMPRESSION: **No acute findings.** Top-normal heart size. [.JPEG]

Table C.22: Prompt 1.5 Few-shot: few-shot in-context learning given a few examples (IU X-RAY).

System	You are a professional chest radiologist that reads chest X-ray image(s).
User	Write a report that contains only the FINDINGS and IMPRESSION sections based on the attached images. Provide only your generated report, without any additional explanation and special format. Your answer is for reference only and is not used for actual diagnosis. [.PNG] FINDINGS: 2 images. Heart size upper limits of normal. Mediastinal contours are maintained. The patient is mildly rotated. There is a small to moderate sized right apical pneumothorax which measures approximately 2.0 cm. No focal airspace consolidation is seen. Left chest is clear. No definite displaced bony injury is seen. Results called XXXX. XXXX XXXX p.m. XXXX, XXXX. IMPRESSION: Small to moderate right apical **pneumothorax**. [.PNG] FINDINGS: The heart is normal in size and contour. There is focal airspace disease in the right middle lobe. There is no pneumothorax or effusion. IMPRESSION: Focal airspace disease in the right middle lobe. This is most concerning for **pneumonia**. Recommend follow up to ensure resolution. [.PNG] FINDINGS: Stable cardiomegaly with vascular prominence without overt edema. No focal airspace disease. No large pleural effusion or pneumothorax. The XXXX are intact. IMPRESSION: Stable **cardiomegaly** without overt pulmonary edema. [.PNG] FINDINGS: Heart is enlarged. There is prominence of the central pulmonary vasculature. Mild diffuse interstitial opacities bilaterally, predominantly in the bases, with no focal consolidation, pleural effusion, or pneumothoraces. XXXX and soft tissues are unremarkable. IMPRESSION: Cardiomegaly with pulmonary interstitial edema and XXXX bilateral **pleural effusions**. [.PNG] FINDINGS: The cardiac silhouette and mediastinum size are within normal limits. There is no pulmonary edema. There is no focal consolidation. There are no XXXX of a pleural effusion. There is no evidence of pneumothorax. IMPRESSION: **Normal chest x-XXXX.**

[.PNG]
FINDINGS:
IMPRESSION: **Presumed closure device** at the level of the ligamentum

Table C.23: Prompt 2.1 Image reasoning: medical condition labeling from chest X-ray images (2-class).

System	You are a professional chest radiologist that reads chest X-ray image(s).
User	Below is an observation plan consisting of 14 conditions: “No Finding”, “Enlarged Cardiomediastinum”, “Cardiomegaly”, “Lung Lesion”, “Lung Opacity”, “Edema”, “Consolidation”, “Pneumonia”, “Atelectasis”, “Pneumothorax”, “Pleural Effusion”, “Pleural Other”, “Fracture”, “Support Devices”. Based on attached images, assign labels for each condition: “1”, “0”. If the observation was clearly present on the chest X-ray image, assign “1” to the condition. Otherwise, assign “0” to the condition. For multiple images, assign the labels based on all images and return only one list of labels for the given 14 conditions. Your answer is for reference only and is not used for actual diagnosis. Strictly follow the format below to provide your output. <LABEL> [(“No Finding”, “0”|“1”), (“Enlarged Cardiomediastinum”, “0”|“1”), (“Cardiomegaly”, “0”|“1”), (“Lung Lesion”, “0”|“1”), (“Lung Opacity”, “0”|“1”), (“Edema”, “0”|“1”), (“Consolidation”, “0”|“1”), (“Pneumonia”, “0”|“1”), (“Atelectasis”, “0”|“1”), (“Pneumothorax”, “0”|“1”), (“Pleural Effusion”, “0”|“1”), (“Pleural Other”, “0”|“1”), (“Fracture”, “0”|“1”), (“Support Devices”, “0”|“1”)] </LABEL>

Table C.24: Prompt 2.2 Image reasoning: medical condition labeling from chest X-ray images (4-class).

User	Below is an observation plan consisting of 14 conditions: “No Finding”, “Enlarged Cardiomediastinum”, “Cardiomegaly”, “Lung Lesion”, “Lung Opacity”, “Edema”, “Consolidation”, “Pneumonia”, “Atelectasis”, “Pneumothorax”, “Pleural Effusion”, “Pleural Other”, “Fracture”, “Support Devices”. Based on attached images, assign labels for each condition except “No Finding”: “1”, “0”, “-1”, “2”. It is noted that “No Finding” is either “2” or “1”. These labels have the following interpretation: 1 - The observation was clearly present on the chest X-ray image. 0 - The observation was absent on the chest X-ray image and was mentioned as negative. -1 - The observation was unclear if it exists. 2 - The observation was absent but not explicitly mentioned. For multiple images, assign the labels based on all images and return only one list of labels for the given 14 conditions. Your answer is for reference only and is not used for actual diagnosis. Strictly follow the format below to provide your output. <LABEL> [(“No Finding”, “1”|“2”), (“Enlarged Cardiomediastinum”, “0”|“1”|“2”|“-1”), (“Cardiomegaly”, “0”|“1”|“2”|“-1”), (“Lung Lesion”, “0”|“1”|“2”|“-1”), (“Lung Opacity”, “0”|“1”|“2”|“-1”), (“Edema”, “0”|“1”|“2”|“-1”), (“Consolidation”, “0”|“1”|“2”|“-1”), (“Pneumonia”, “0”|“1”|“2”|“-1”), (“Atelectasis”, “0”|“1”|“2”|“-1”), (“Pneumothorax”, “0”|“1”|“2”|“-1”), (“Pleural Effusion”, “0”|“1”|“2”|“-1”), (“Pleural Other”, “0”|“1”|“2”|“-1”), (“Fracture”, “0”|“1”|“2”|“-1”), (“Support Devices”, “0”|“1”|“2”|“-1”)] </LABEL>

Table C.25: Prompt 3.1 Report synthesis: report generation using provided positive and negative conditions.

System	You are a professional chest radiologist that reads chest X-ray image(s).
User	Below is a given observation plan: <LABEL> Positive Conditions: {} Negative Conditions: {} </LABEL> Write a report that contains only the FINDINGS and IMPRESSION sections based on given labels rather than images. For positive conditions, you should clearly mention it in the report. For negative conditions, you should clearly mention in the report that there is no clear evidence of this condition. You should not mention any other conditions not listed above. Your answer is for reference only and is not used for actual diagnosis. Strictly follow the format below to provide your output. <REPORT> FINDINGS: <findings> IMPRESSION: <impression> </REPORT>

Table C.26: Prompt of finetuned LLaMA-2 report synthesis given groundtruth labels

System	Write a radiology report that includes all given positive labels and negative labels.
User	**Input:** Positive labels: {positive_labels} Negative labels: {negative_labels} **Output:** {output}

Chapter D

CLEAR Evaluation Framework Details

D.1 Open-sourced Artifacts

We will formally release code package for CLEAR on GitHub at the camera-ready stage. The current version supports both open-source models via the vLLM backend and closed-source models through the Azure OpenAI API.

Our collected ground-truth dataset, CLEAR-Bench, and related data documentation will also be made publicly available on Physionet to support future research in this area.

D.2 Data Annotation and Curation

We accessed MIMIC-CXR-JPG data by following the required steps on `https://physionet.org/content/mimic-cxr-jpg/2.1.0/`. We first registered and applied to be a credentialed user, and then completed the required training of CITI Data or Specimens Only Research. Data license can be found at `https://physionet.org/content/mimic-cxr-jpg/view-license/2.1.0/`.

During each human annotation process, we follow a traditional paradigm: initial pilot rounds are conducted to gather user feedback, followed by formal, independent large-scale annotation, data analysis for quality control and final resolution via consensus discussion. Our annotation platform is built upon an open source data labeling tool, Label Studio [114].

D.2.1 Label Structure Refinement

MIMIC-CXR-JPG Labeling Criteria

Positive (1.0): The label is positively mentioned in the report and present in one or more associated images.
Example: "A large pleural effusion"

Negative (0.0): The label is negatively mentioned in the report and should not be present in any associated image.
Example: "No pneumothorax."

Uncertain (-1.0): The label is either: (1) mentioned with uncertainty, so presence in the image is unclear; or (2) described ambiguously, with uncertain existence.
Explicit uncertainty: "The cardiac size cannot be evaluated."
Ambiguous language: "The cardiac contours are stable."

Unmentioned (Missing): The label is not mentioned in the report at all.

Figure D-1: 4-type labeling criteria in MIMIC.

During the interaction of pilot training, we closely work with all involved radiologists and collect a lot of valuable feedback for user experience with designed interfaces and task instruction.

After summarizing input feedback, we recognize some shared and repeatedly mentioned issues in the 4-type label structure of MIMIC-CXR-JPG (see Figure D-1): (1) The "unmentioned" category has a high degree of overlap with other categories, particularly with "negative" labels. This is because radiologists often do not explicitly state negative findings in the report. However, indirect phrases such as "Lungs are clear" can implicitly negate a wide range of lung-related abnormalities. (2) Additionally, different radiologists have varying tendencies in labeling conditions. More conservative radiologists may lean toward assigning "uncertain" rather than "positive" labels, even when the evidence suggests a likely presence. This inconsistency introduces label noise and ambiguity, particularly when these labels are used for supervised training or evaluation purposes.

Therefore, we refined the original MIMIC label structure into a "5+1" annotation

Our Refined Labeling Criteria

Confidently Absent: The condition is clearly stated as not present in the report.
Example: "No pneumothorax."

Likely Absent: The report implies the condition is likely absent, but the language is ambiguous or uncertain.
Example: "Heart size is normal though increased."

Neutral: The report does not clearly indicate presence or absence.
Explicit uncertainty: "The cardiac size cannot be evaluated."
Ambiguous language: "The cardiac contours are stable."

Likely Present: The report suggests the condition may be present, but uses uncertain or ambiguous language.
Example: "Likely reflecting compressive atelectasis."

Confidently Present: The condition is clearly stated as present in the report.
Example: "A small right pleural effusion."

Figure D-2: Our refined 5-type labeling criteria during expert annotation.

framework. The "5" refers to an extension of MIMIC's original "Positive," "Negative," and "Uncertain" categories into five more nuanced types, as shown in Figure D-2. The "+1" refers to retaining the "Unmentioned" label as a separate flag. Specifically, radiologists are asked to select one of the five labels for each condition and additionally indicate whether this label is explicitly mentioned in the report or not.

After collecting radiologist responses, we map the five types into a final three-type scheme for downstream use: "Confidently Present" and "Likely Present" are merged into "Positive," "Confidently Absent" and "Likely Absent" into "Negative," and "Neutral" is renamed as "Unclear." We then proceed with inter-rater alignment checks for quality control. Notably, the "mentioned" flag is not incorporated into the final label itself but serves as a supporting indicator for data managers to differentiate between labeling disagreements due to quality issues versus differences in individual clinical interpretation. This overall process enables us to accommodate variability in radiologist judgment while maintaining

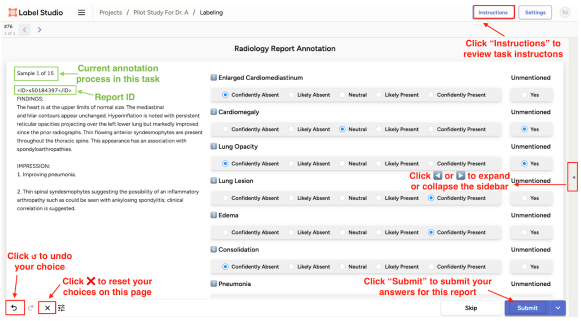


Figure D-3: Interface for Label Annotation.

high annotation quality.

D.2.2 Expert-in-the-loop Dataset Curation

We first exclude 2 cases without any "FINDINGS" or "IMPRESSION" and 30 cases labeled as "No Finding" in the radiologist annotation dataset from MIMIC-CXR-JPG (containing 687 studies in total). Then, we randomly select 20 cases to serve as a pilot set for initial review and refinement of the process.

We then prompt GPT-4o to generate condition labels following the same guidelines used in the original MIMIC documentation for remaining studies excluded 20 pilot cases. After identifying discrepancies between the model-generated labels and the original dataset annotations, we isolate the suspected noisy labels for further review.

For each case, we extract only the relevant report sections (FINDINGS and IMPRESSION), with no images involved, and present them to a board-certified radiologist. The radiologist independently re-annotates the report from scratch based on their clinical judgment.

During the curation, we discard 5 cases due to GPT-4o generation failures. To manage the annotation workload, we limit each review to reports with one to five mismatched conditions per case.

The full curation process took approximately one month, resulting in 550 finalized reports, each annotated with 13 condition labels.

Task instruction can be checked in Figure D-5 and interface can be checked in Figure D-

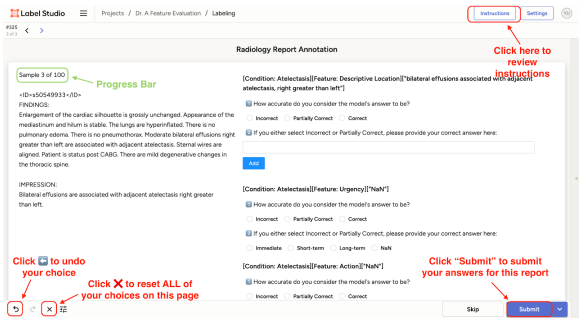


Figure D-4: Interface for Attribute Curation.

3.

D.2.3 CLEAR-Bench: Expert Ensemble

After excluding "No Finding" cases and those already annotated in the curation stage, we selected 5 cases for pilot training and randomly sampled 100 reports from the test and validation sets of MIMIC-CXR-JPG to construct our final evaluation dataset.

Following a brief onboarding process using 5 pilot cases, we collected independent annotations from three radiologists, each labeling the 100 reports from scratch. After an initial round of majority voting, 25 reports with 32 individual condition labels in total remained unresolved. These were finalized through a single round of discussion and consensus among the experts.

The full expert ensemble workflow was completed over the course of three months, resulting in 100 fully annotated reports, each with 13 condition labels.

Task instruction can be checked in Figure D-5 and interface can be checked in Figure D-3.

D.2.4 CLEAR-Bench: Attribute Curation

The blueprint for attribute design was initially inspired by the concept of an “Attribute-Value Format” proposed by Dr. Langlotz in his practical guide to writing radiology reports [64, 207]. Driven by this concept, we generated a list of commonly used report attributes with the assistance of GPT-4o, and refined it through discussion with our collaborating research

radiologist, who is also a co-author. Together, we determined which attributes to include, revise, or remove. During this process, we not only developed a concise yet comprehensive attribute structure but also collected useful example phrases and sentences for each attribute. These examples were later incorporated into the prompts used in the Description Extraction Module (see Appendix D.4). The final version of the prompt set and word list was also reviewed and approved by a clinical radiologist.

We curated attributes using the same 100 studies described earlier, excluding 2 cases that lacked any positively identified conditions in expert ensemble labels. Following a round of pilot training, the formal curation process proceeded as detailed in Section 5.3. After collecting radiologist responses, we conducted a second round of quality control to finalize the ground-truth attributes. The full human curation process took approximately one month.

Task instructions are shown in Figure D-6, and the annotation interface is illustrated in Figure D-4.

D.3 CLEAR: Implementation Details

Base Model	GAS	LR	Epochs
Llama3.1-8B-Instruct	1	7.0×10^{-6}	4
Qwen2.5-7B-Instruct	1	9.0×10^{-6}	5

Table D.1: Hyperparameter search results. GAS denotes the number of gradient-accumulation steps, LR the learning rate, and Epochs the total training epochs.

Supervised finetuning details. All fine-tuned models were obtained through supervised fine-tuning with LLaMA-Factory [132]. To identify an optimal configuration, we developed an automated hyperparameter optimization (HPO) framework that combines five-fold cross-validation with a grid search. Learning rate, number of epoch, and gradient accumulation steps are three objects to be optimized. For learning rate, searching space is $[3.0e^{-6}, 3.0e^{-5}]$, with an interval of $2.0e^{-6}$. For epoch, searching space is $\{2,3,4,5\}$. For gradient accumulation steps, searching target is $\{1,2,4\}$. We conduct extensive experiments to assess hyperparameters' influence. A total of 360 models are finetuned for one base model to

determine the best hyperparameter setting. The best-performing settings, summarized in Table D.1, are used for all experiments reported in Table 5.2. Hyperparameter optimization and model training are performed on NVIDIA A100 80G and NVIDIA H100 94G GPUs. The HPO stage takes 93 h 51 m 20 s on four A100s and 14 h 39 m 36 s on four H100s.

Inference details for local models. We serve the models locally with vLLM (0.8.5.post1) [57]. Inference runs with a temperature of 1e-5 and a max_tokens of 4,096; all other sampling parameters remain at their default settings. A single NVDIA A100 80G is sufficient for inference under this setting.

Model	**Standard Pricing (per 1M Tokens)**
`GPT-4o-2024-1120 (Global)`	Input: $2.50 Cached: $1.25 Output: $10.00
`o1-mini-2024-09-12 (Global)`	Input: $1.10 Cached: $0.55 Output: $4.40

Table D.2: Standard API pricing per 1M tokens for GPT-4o and o1-mini models, based on Azure OpenAI pricing: `https://azure.microsoft.com/en-us/pricing/details/cognitive-services/openai-service/#pricing`.

API Details We access OpenAI's GPT-4o (2024-11-20) and o1-mini (2024-09-12) via Microsoft's Azure. Pricing details can be checked in Table D.2.

D.4 Template & Terminology List

Thank you very much for your support in our human annotation process! To begin with, please register at https://physionet.org/content/mimic-cxr-jpg/2.1.0/ and sign the data agreement before the study. Feel free to reach us at {EMAIL} if you encounter any issue or any questions during the process.

Overview: Task Description
In this task, you will be extracting clinical information from {NUM} radiology reports in total. You will not be shown the corresponding images, so you are being asked to interpret each report, as written, for the extent to which the presence of {NUM} conditions is captured. It is important to note that some reports may have empty FINDINGS or IMPRESSION sections due to limitations in the original MIMIC-CXR-JPG database. Please follow the labeling instructions as below.

INSTRUCTIONS:
For each case, you will be presented with a single radiology report. Your objective is to choose the single most appropriate criterion among 5 options (see below) for each of the {NUM} conditions AND note whether each condition is explicitly mentioned in the report. Please base your decisions solely on the provided report.

CRITERIA:
{See Figure D-2}

Interface User Guide
{Account Information and Usage Tips}

Figure D-5: Instruction Template for Label Annotation Task

Thank you very much for your support in our human annotation process! To begin with, please register at https://physionet.org/content/mimic-cxr-jpg/2.1.0/ and sign the data agreement before the study. Feel free to reach us at {EMAIL} if you encounter any issue or any questions during the process.

Overview: Task Description
This curation task is to identify fine-grained features—such as location, severity, and treatment—related to specific medical conditions (e.g., edema, atelectasis, support devices) in radiology reports. You will review {NUM} text-only reports (no X-ray images) and assess the accuracy of feature annotations generated by an AI model.

Each report includes 13 predefined medical conditions, but you will only see those that were positively labeled by human annotators. As a result, the number of conditions shown per report may vary. For each positive condition, the AI extracts fine-grained details (e.g., location, severity), which you need to review. Start by marking the model's answer as correct, partially correct, or incorrect. If it's incorrect, enter the corrected version in the provided text box.

[optional] If you'd like to understand how the AI generated its responses, you can review the prompts we used at {See section D.4}.

Interface User Guide
{Account Information and Usage Tips}

Figure D-6: Instruction Template for Attribute Curation Task

Prompt 1: Presence

System Instruction:
You are a radiologist reviewing a piece of radiology report to assess the presence of 13 specific medical conditions.

Conditions to evaluate: Cardiomegaly, Enlarged Cardiomediastinum, Atelectasis, Consolidation, Edema, Lung Lesion, Lung Opacity, Pneumonia, Pleural Effusion, Pneumothorax, Pleural Other, Fracture, Support Devices.

Each medical condition in the radiology report must be categorized using one of the following labels: "positive", "negative" or "unclear". The criteria for each label are:
• "positive": The condition is indicated as present in the report.
• "negative": The condition is indicated as not present in the report.
• "unclear": The report does not indicate a clear presence or absence of the condition.

The user will provide you with a piece of radiology report as input. Return your results in the following JSON format:
<TASK1>{
   "Cardiomegaly": "positive"|"negative"|"unclear",
   "Enlarged Cardiomediastinum": "positive"|"negative"|"unclear",
   "Atelectasis": "positive"|"negative"|"unclear",
   "Consolidation": "positive"|"negative"|"unclear",
   "Edema": "positive"|"negative"|"unclear",
   "Lung Lesion": "positive"|"negative"|"unclear",
   "Lung Opacity": "positive"|"negative"|"unclear",
   "Pneumonia": "positive"|"negative"|"unclear",
   "Pleural Effusion": "positive"|"negative"|"unclear",
   "Pneumothorax": "positive"|"negative"|"unclear",
   "Pleural Other": "positive"|"negative"|"unclear",
   "Fracture": "positive"|"negative"|"unclear",
   "Support Devices": "positive"|"negative"|"unclear"
} </TASK1>

User Input:
FINDINGS: {findings}
IMPRESSION: {impression}

Prompt 1

Prompt 2: First Occurrence

System Instruction:
You are a radiologist reviewing a piece of radiology report to extract features for a specific condition, which was already marked as positive during the initial read of this same report.

Please determine from the given report (i.e., current study) whether {condition} is being identified for the first time in current study ["current"], or if the report indicates it was already present or noted in a prior study ["previous"]. If unmentioned, respond with ["N/A"]. Only choose one of the following: ["current"], ["previous"], or ["N/A"].

Example answer: ["current"]

User Input:
FINDINGS: {findings}
IMPRESSION: {impression}

Prompt 2

Prompt 3: Change

System Instruction:
You are a radiologist reviewing a piece of radiology report to extract features for a specific condition, which was already marked as positive during the initial read of this same report.

Please determine from the given report whether {condition} is improving, stable, or worsening according to the given report. If the status is not mentioned, respond with ["N/A"]. If the report describes multiple statuses, respond with ["mixed"]. Only choose one of the following: ["improving"], ["stable"], ["worsening"], ["mixed"] or ["N/A"].

Example answer: ["stable"]

User Input:
FINDINGS: {findings}
IMPRESSION: {impression}

Prompt 3

Prompt 4: Severity

System Instruction:
You are a radiologist reviewing a piece of radiology report to extract features for a specific condition, which was already marked as positive during the initial read of this same report.

Please determine from the given report whether {condition} is mild, moderate, or severe according to the given report. If the status is not mentioned, respond with ["N/A"]. If the report describes multiple statuses, respond with ["mixed"]. Only choose one of the following: ["mild"], ["moderate"], ["severe"], ["mixed"] or ["N/A"].

Example answer: ["mild"]

User Input:
FINDINGS: {findings}
IMPRESSION: {impression}

Prompt 4

Prompt 5: Descriptive Location

System Instruction:
You are a radiologist reviewing a piece of radiology report to extract features for a specific condition, which was already marked as positive during the initial read of this same report.

Please identify the location(s) of {condition} described in the given report. Extract and return a list of phrases that mention the anatomical location(s) {location} specifically related to {condition}. For each location, include any relevant descriptors descriptor and any associated status {status}. {note} If multiple phrases refer to the same location, merge them into one single entry using the most complete, informative, and non-redundant phrasing for that unique area. Format your output as one single list in the following format: ["entry-1","entry-2",...,"entry-n"]. If nothing is mentioned, return ["N/A"].

Example answer:
["left lower lobe compressive atelectasis","right middle lobe bibasilar atelectasis"]

User Input:
FINDINGS: {findings}
IMPRESSION: {impression}

Prompt 5: Additional Notes: location/descriptor/status/note are a list of example key words or phrases for each condition collected from radiologists, such as (e.g., compressive, segmental, focal, terminal, peripheral, etc.).

Condition	Location	Descriptor	Status	Note
Atelectasis	(e.g., left upper, right lower, whole lung, etc.)	(e.g., compressive, segmental, focal, terminal, peripheral, etc.)	(e.g., improving, worsening, stable, unchanged, new, etc.)	
Cardiomegaly		(e.g., mild, moderate, severe, etc.)	(e.g., improving, worsening, stable, unchanged, new, etc.)	
Consolidation	(e.g., left upper, right lower, whole lung, etc.)	(e.g., segmental, focal, terminal, etc.)	(e.g., improving, worsening, stable, unchanged, new, etc.)	
Edema	(e.g., medial (near hilum), middle, lateral (peripheral), etc.)	(e.g., interstitial, alveolar, minimal, mild, moderate, severe, etc.)	(e.g., improving, worsening, stable, unchanged, new, etc.)	
Enlarged Cardiomediastinum		(e.g., mild, moderate, severe, etc.)	(e.g., improving, worsening, stable, unchanged, new, etc.)	
Fracture	(e.g., ribs, cervicothoracic vertebra, etc.)	(e.g., simple or closed, compound or open, incomplete or partial, complete, etc.)	(e.g., improving, worsening, stable, unchanged, new, etc.)	
Lung Lesion	(e.g., central, peripheral, sub-pleural, entire pleural space, etc.)	(e.g., density, internal composition, shape, margin, etc.)	(e.g., improving, worsening, stable, unchanged, new, etc.)	Explicitly refer to a lung lesion (e.g., nodules, masses, infiltrates, metastases, etc.) and ignore findings unrelated to lung lesions.
Lung Opacity	(e.g., left upper, right lower, perihilar, etc.)	(e.g., interstitial, alveolar, diffuse, focal, dense, ill-defined, faint, etc.)	(e.g., improving, worsening, stable, unchanged, new, etc.)	
Pleural Effusion	(e.g., left, right, entire pleural space, etc.)	(e.g., subpulmonic, posterior, loculated, lobular, small, moderate, large, etc.)	(e.g., improving, worsening, stable, unchanged, new, etc.)	
Pneumonia	(e.g., left upper, right lower, whole lung, etc.)	(e.g., segmental, focal, terminal, etc.)	(e.g., improving, worsening, stable, unchanged, new, etc.)	
Pneumothorax	(e.g., left upper, right lower, etc.)	(e.g., simple, tension, open, etc.)	(e.g., improving, worsening, stable, unchanged, new, etc.)	
Pleural Other	(e.g., left upper, right lower, entire pleural space, etc.)	(e.g., subpulmonic, posterior, loculated, lobular, diffuse, focal, etc.)	(e.g., improving, worsening, stable, unchanged, new, etc.)	Do not include findings that pertain solely to Pleural Effusion; only include findings related to other pleural abnormalities (e.g., thickening, plaques, etc.).
Support Devices				Exclude any mention of device removal. Only include information related to existing or currently present devices.

Table D.3: Key Words List for Location Prompt (extracted using GPT-4o, then discussed and confirmed by two radiologists)

Prompt 6: Recommendation

System Instruction:
You are a radiologist reviewing a piece of radiology report to extract features for a specific condition, which was already marked as positive during the initial read of this same report.

Please identify treatment(s)/follow-up(s) associated with {condition} in the given report. Extract and return a list of phrases that only describe specific treatment(s)/follow-up(s) recommended in relation to condition. Do not include any phrase that merely describes the condition without any treatment/follow-up. Each treatment/follow-up should be a single entry. Format your output as a single list in the following format: ["entry-1","entry-2",...,"entry-n"]. If no action is mentioned, return ["N/A"].

Example answer:
["follow-up CT scheduled in 3 months","routine annual imaging advised"]

User Input:
FINDINGS: {findings}
IMPRESSION: {impression}

Prompt 6

Prompt 7: Urgency

System Instruction:
You are a radiologist reviewing a piece of radiology report to extract features for a specific condition, which was already marked as positive during the initial read of this same report.

Please determine from the given report whether {condition} requires immediate, short-term, or long-term treatment/follow-up (e.g., Immediate: Urgent chest tube placement recommended; Short-term: Recommend follow-up chest X-ray in 1-2 weeks; Long-term: Routine annual imaging advised). If unmentioned, answer ["N/A"]. Only choose one of the following: ["immediate"], ["short-term"], ["long-term"], or ["N/A"].

Example answer: ["long-term"]

User Input:
FINDINGS: {findings}
IMPRESSION: {impression}

Prompt 7

o1-mini Scoring

System Instruction:
You are a radiology report comparison assistant. You will be given two lists of findings: one is the ground truth (GT), and the other is a candidate prediction (GEN).

Your task is to compare them and return a similarity score between 0 and 1.
1. A score of 1.0 means they are clinically and semantically identical.
2. A score of 0.0 means they are completely different or unrelated.
3. Partial matches should get a score in between.

Do not explain the score. Just output a float between 0 and 1.
Example answer: </SCORE>"0.8"</SCORE>

User Input:
GT: {groundtruth}
GEN: {candidate}

o1-mini prompt