Can Rule-Based Insights Enhance LLMs for Radiology Report Classification? Introducing the RadPrompt Methodology.

Panagiotis Fytas[1] Anna Breger[*,2,3] Ian Selby[*,4,5]
Simon Baker[1] Shahab Shahipasand[5] Anna Korhonen[1]

[1]Language Technology Lab, University of Cambridge
[2]Department of Applied Mathematics and Theoretical Physics, University of Cambridge
[3]Center of Medical Physics and Biomedical Engineering, Medical University of Vienna
[4]Department of Radiology, University of Cambridge
[5]Cambridge University Hospitals, NHS Foundation Trust
pf376@cam.ac.uk

Abstract

Developing imaging models capable of detecting pathologies from chest X-rays can be cost and time-prohibitive for large datasets as it requires supervision to attain state-of-the-art performance. Instead, labels extracted from radiology reports may serve as distant supervision since these are routinely generated as part of clinical practice. Despite their widespread use, current rule-based methods for label extraction rely on extensive rule sets that are limited in their robustness to syntactic variability. To alleviate these limitations, we introduce RadPert, a rule-based system that integrates an uncertainty-aware information schema with a streamlined set of rules, enhancing performance. Additionally, we have developed RadPrompt, a multi-turn prompting strategy that leverages RadPert to bolster the zero-shot predictive capabilities of large language models, achieving a statistically significant improvement in weighted average F1 score over GPT-4 Turbo. Most notably, RadPrompt surpasses both its underlying models, showcasing the synergistic potential of LLMs with rule-based models. We have evaluated our methods on two English Corpora: the MIMIC-CXR gold-standard test set and a gold-standard dataset collected from the Cambridge University Hospitals.

1 Introduction

Supervised deep learning for medical imaging classification has accomplished significant milestones. In the chest X-ray (CXR) domain, such models have exhibited predictive capabilities on par with expert physicians (Rajpurkar et al., 2018; Tang et al., 2020) and are being utilized in collaborative settings to increase clinician accuracy (Rajpurkar et al., 2020).

Annotating medical images, however, is expensive and arduous: it requires a committee of expert radiologists to resolve the inherently high degree of annotator variance and subjectivity (Razzak et al., 2018). This issue is particularly problematic considering the global shortage of radiologists (Jeganathan, 2023; Kalidindi and Gandhi, 2023; Konstantinidis, 2023). Instead, we often have access to a form of distant supervision: the radiology report. Radiology reports are semi-structured free-text interpretations of an X-ray image and are generated as a routine part of clinical practice to communicate findings.

In the past, rule-based models (Irvin et al., 2019; Peng et al., 2017) have been used to extract structured labels from radiology reports in various imaging datasets, including ChestX-ray14 (Wang et al., 2017), CheXpert (Irvin et al., 2019), MIMIC-CXR (Johnson et al., 2019) and BRAX (Reis et al., 2022). However, those rule-based methods are often based on elementary techniques and, thus, exhibit limited robustness to syntactic variation. Naturally, supervised deep learning models offer superior performance through their robustness to syntactic variability (Smit et al., 2020; Jain et al., 2021b). In contrast, Large Language Models (LLMs) represent a significant improvement over rule-based models in an unsupervised setting and have achieved impressive performance in the field of radiology (Infante et al., 2024; Adams et al., 2023; Liu et al., 2023).

In this paper, we present RadPert, a rule-based model built on the RadGraph knowledge graph (Jain et al., 2021a). RadPert leverages entity-level uncertainty labels from RadGraph, reducing the

[*]Equal contribution.

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 212–235
August 16, 2024. ©2024 Association for Computational Linguistics

need for a comprehensive rule set and enhancing its resilience to syntactic variations. We have evaluated RadPert internally on MIMIC-CXR and externally on a dataset collected from the Cambridge University Hospitals (CUH). RadPert surpasses CheXpert, the former rule-based state-of-the-art (SOTA), by achieving statistically significant improvement in weighted average F1 score.

Furthermore, we explore the collaborative potential of LLMs with rule-based models through RadPrompt. RadPrompt is a multi-turn prompting strategy that employs RadPert as an implicit means of encoding medical knowledge (Figure 1). In fact, RadPrompt, based on GPT-4 Turbo, manages to outperform both its underlying models in a zero-shot setting.

2 Related Work

Numerous natural language processing methods have been developed to derive structured predictions from radiology reports (Peng et al., 2017; Hassanpour et al., 2017; Pons et al., 2016; Bozkurt et al., 2019; Wang et al., 2018). Many of those approaches are designed for the multitask classification of radiology reports, written in English, into labels representing prevalent pathologies from CXRs. Each such label can exhibit one of four output classes: *Null, Positive, Negative* and *Uncertain*. CheXpert (Irvin et al., 2019), the rule-based SOTA, follows an approach based on regular expression matching and the Universal Dependency Graph (UDG) of a radiology report. Due to the rudimentary regular expression matching, however, CheXpert is sensitive to syntactic variation. Thus, multiple over-generalized rules are used in an attempt to alleviate these shortcomings. Furthermore, the UDG is a type of information extraction that does not explicitly identify negation and uncertainty. Therefore, its ability to detect uncertainty in complex phrases is hampered despite the extensive rule set. Extensions of CheXpert have been developed for Brazilian Portuguese (Reis et al., 2022) and German (Wollek et al., 2024). CheXbert (Smit et al., 2020) is a semi-supervised model pretrained on automatically extracted labels from the CheXpert model, fine-tuned on manually annotated reports, and evaluated on 687 MIMIC-CXR gold-standard test set reports. However, the published model weights[1] of CheXbert differ from the original model. This discrepancy complicates comparisons on the MIMIC-CXR dataset as the published model is fine-tuned on unspecified MIMIC-CXR manually annotated reports, which can potentially overlap with the MIMIC-CXR gold-standard test set.

Recent work has also explored the adoption of LLMs for radiology report classification. Specifically, Dorfner et al. (2024) examine the zero and few-shot capabilities of LLMs. However, they mainly treat the task as a binary classification for each pathology. Namely, for multitask classification, they only report the few-shot results on an unpublished institutional dataset. CheX-GPT (Gu et al., 2024) utilizes zero-shot GPT-4 labels as a distant supervision to fine-tune a BERT-based model. Nonetheless, they also simplify the task into binary classification.

Alternative approaches to the classification of chest X-rays (CXRs) explore moving away from the distantly supervised paradigm of training unimodal vision models on classifying structured labels extracted from radiology reports. In lieu of structured prediction, Vision-Language (VL) models are trained to align the embedding representations of CXRs with the representations of the corresponding radiology reports via self-supervised contrastive learning objectives (Huang et al., 2021; Boecking et al., 2022; Tiu et al., 2022; Wang et al., 2022; Bannur et al., 2023). This alignment task is transformed into CXR classification through the cosine similarity of CXR embeddings to the embeddings of textual prompts representing the existence or absence of pathologies. However, vision models trained with the structured prediction paradigm outperform VL models such as CheXzero (Tiu et al., 2022), even when the latter utilizes an expert-annotated validation set for selecting optimal classification thresholds.

In this paper, we will focus on improving the unsupervised SOTA for the multitask classification of radiology reports.

3 Methods

3.1 Task

Similar to CheXpert and CheXbert, we will focus on the multitask classification of CXR radiology reports. Specifically, our models classify thirteen labels that correspond to pathologies (Atelectasis, Edema, Cardiomegaly, Consolidation, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, Pleural Effusion, Pleural Other, Pneumoth-

[1] https://github.com/stanfordmlgroup/CheXbert

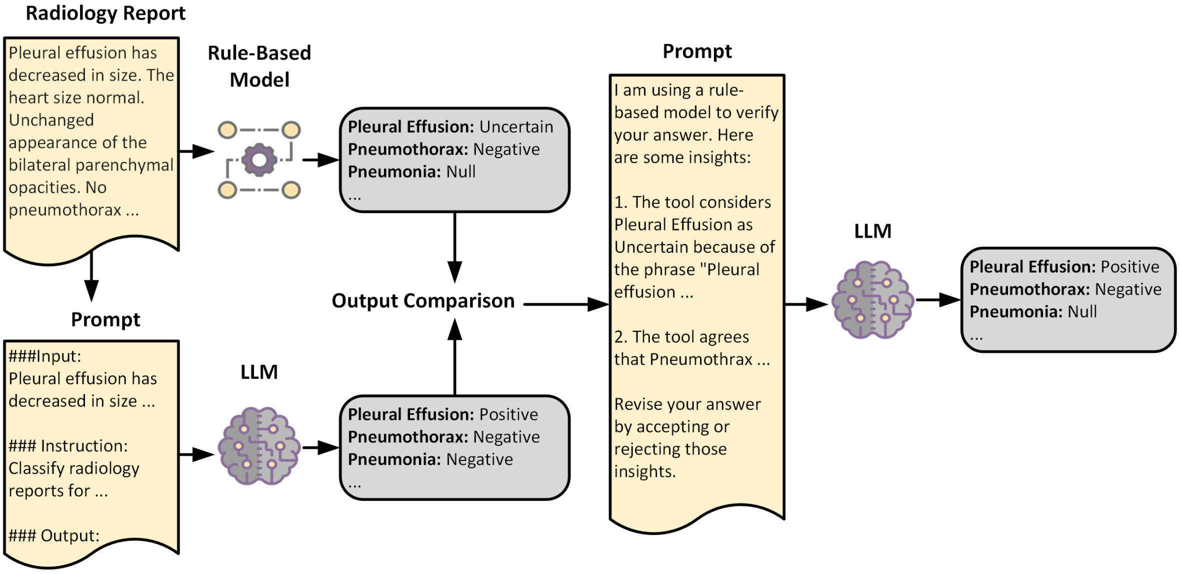

Radiology Report
Pleural effusion has decreased in size. The heart size normal. Unchanged appearance of the bilateral parenchymal opacities. No pneumothorax ...
Rule-Based Model
Pleural Effusion: Uncertain
Pneumothorax: Negative
Pneumonia: Null
...
Prompt
###Input:
Pleural effusion has decreased in size ...
### Instruction:
Classify radiology reports for ...
### Output:
LLM
Pleural Effusion: Positive
Pneumothorax: Negative
Pneumonia: Negative
...
Output Comparison
Prompt
I am using a rule-based model to verify your answer. Here are some insights:
1. The tool considers Pleural Effusion as Uncertain because of the phrase "Pleural effusion ...
2. The tool agrees that Pneumothorax ...
Revise your answer by accepting or rejecting those insights.
LLM
Pleural Effusion: Positive
Pneumothorax: Negative
Pneumonia: Null
...

Figure 1: Overview of the RadPrompt methodology. RadPrompt utilizes the rule-based RadPert model to detect potential errors in the original (first-turn) LLM classification decision. A second-turn prompt is then constructed, offering evidence that may cause the LLM to revise its original classification outcome.

orax, Support Devices and Pneumonia), with each label having four possible output classes: *Null, Positive, Negative* and Uncertain. A pathology is classified as *Null* if there are no references to it in the radiology report. It is considered *Negative* when its absence is explicitly mentioned. *Positive* classes entail that the existence of the corresponding pathology is specified in the report. Finally, *Uncertain* classes imply that while the pathology is discussed in the report, its existence cannot be determined.

3.2 RadPert

In order to overcome the limitations of existing tools, we have designed RadPert. RadPert incorporates hand-crafted rules with the RadGraph (Jain et al., 2021a) knowledge graph.

3.2.1 RadGraph Information Schema

RadGraph (Jain et al., 2021a) defines an information schema specifically designed for radiology reports. It contains two top-level entity types: *Anatomy (ANAT)* and *Observation (OBS)*. *Anatomy* entities describe bodily anatomical structures (e.g. “lobe”) and their spatial characteristics (e.g. “left”). *Observation* entities include pathological abnormalities (e.g. “opacities”), diagnosed diseases (e.g. “pneumonia”) and various other characteristics (e.g. “acute”). It is important to note that *Observation* entities are further categorized into three second-level attributes: *Definitely Present (DP)*, *Definitely Absent (DA)* and *Uncertain (U)*.

Additionally, RadGraph defines three types of directed relations between entities. Firstly, the *suggestive of* relation indicates that some *Observation* implies the existence of another *Observation*. Secondly, *located at* relations account for *Observations* relating to specific *Anatomies*. Finally, *modify* relations can exist only between the same type of entity and describe the characteristics relating to a specific entity (e.g., *modify*(“left”, “lung”)).

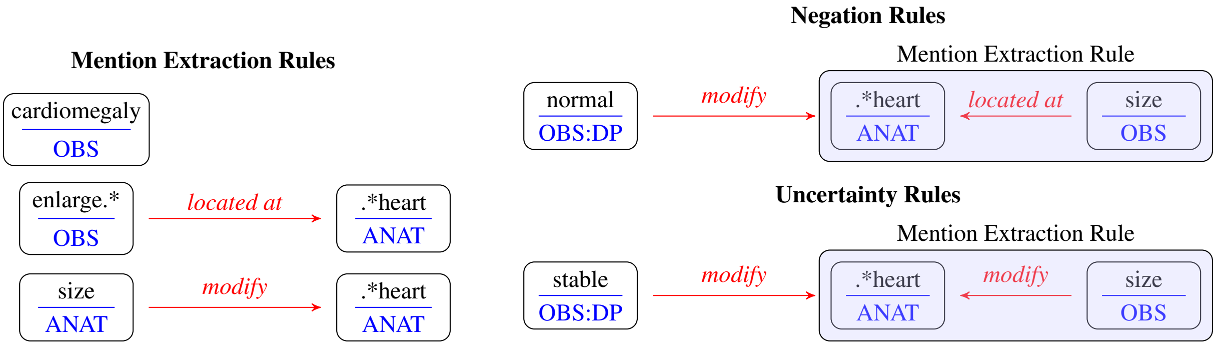
The RadGraph model is based on the DyGIE++ (Wadden et al., 2019) framework initialized with PubMedBERT weights (Gu et al., 2021). The model is fine-tuned on 500 expert-annotated MIMIC-CXR reports based on the RadGraph information schema.

3.2.2 RadPert Pipeline

RadPert employs the following four-stage pipeline:

Knowledge graph extraction. We first extract the RadGraph entities and relations from radiology reports. Utilizing RadGraph instead of the UDG allows uncertainty and negation classes to be extracted at an entity level. Thus, the negation and the uncertainty of various complex phrases can be determined based on those classes, reducing the need for complex negation and uncertainty rules.

Mention extraction. In this stage, for each pathology label, we have adapted and simplified the CheXpert rules (Irvin et al., 2019) so they can be applied to RadGraph entities and relations. Essentially, those rules can be represented as graphs



(a) *Mention extraction* rules.

(b) *Negation/uncertainty detection* rules.

Figure 2: Examples of RadPert rules for Cardiomegaly. The rules take the form of graphs that follow the RadGraph (Jain et al., 2021a) information schema. The ".*" symbolizes allowing the matching of different prefixes and suffixes within the entity span.

based on the RadGraph information schema. Figure 2a includes examples of mention extraction rules in the form of graphs. Checking whether a pathology is mentioned in a radiology report amounts to determining whether any rule-graphs for the specific pathology are subgraphs of the radiology report knowledge graph[2]. If none of the pathology rules match a given radiology report, then the class for that pathology is *Null*.

Negation/uncertainty detection. We next aim to determine whether an extracted mention is *Positive*, *Negative*, or *Uncertain*. For mentions that contain *Observation* entities in their subgraph, the uncertainty quantifier of the *Observation* determines the initial class of that mention. For instance, if a "heart" *Anatomy* is connected with an "enlarged" *Observation*, which is characterized as *Definitely Absent*, then that mention will be labeled as *Negative*. If a mention only possesses *Anatomy* entities, then we consider by default that mention to be *Positive*. However, certain phrases contain implicit negations/uncertainties. In cases such as "normal heart size", the entity "normal" would be considered under RadGraph a *Definitely Present Observation* attached to an *Anatomy*. Thus, in order to detect such implicit negations/uncertainties and determine the final uncertainty class for each pathology, we have developed a negation and an uncertainty rule set. Both rule sets are constructed from hand-crafted rules in the form of graphs. Examples of Cardiomegaly negation/uncertainty rules can be observed in Figure 2b. When a negation rule is activated, the initial class of the mention will be negated (i.e., *Positive* becomes *Negative* and *Negative* becomes *Positive*). However, when an uncertainty rule is matched, RadPert considers the class of the mention to be *Uncertain*.

Mention aggregation. After extracting and classifying all mentions in a radiology report for a specific label, RadPert aggregates them into the final uncertainty class for that label. Similarly to CheXpert (Irvin et al., 2019), we prioritize positive mentions, followed by uncertain ones, while negative mentions have the lowest priority.

3.3 RadPrompt

RadPert, through its rules, implicitly encodes expert knowledge vital to classifying radiology reports. However, as a rule-based system, it is still sensitive to syntactic and lexical variability. To alleviate this limitation, we propose RadPrompt, a zero-shot prompting technique that injects prompts with insights derived from the application of RadPert. RadPrompt, as seen in Figure 1, employs a two-turn prompting strategy.

In the first turn, the zero-shot prompt contains instructions, which define the task, and the radiology report that needs to be classified. After a response is received from the LLM, the first-turn classification outcome is compared with the output of RadPert.

In the second turn, a prompt is constructed by specifying that a rule-based model is used to verify the validity of the LLM's answer. Hints are then added by specifying for each pathology either RadPert's agreement with the LLM or the radiology report sentence that leads RadPert to a disagreement. This is possible since RadPert, as a rule-based sys-

[2] This problem corresponds to the edge-colored and node-colored variant of Induced Subgraph Isomorphism. Exhaustive search with subgraphs of fixed-length has polynomial complexity (Floderus et al., 2015).

tem, allows the detection of the specific mention that leads to the classification decision. Finally, the prompt instructs the LLM to adjust its answer by accepting or rejecting RadPert's hints. In Table 14 of the Appendix, we present the format of our first and second-turn prompts.

3.3.1 Base Model

As a base model for the RadPrompt strategy, we explore various LLMs, including API-based models such as Gemini-1.5 Pro (Reid et al., 2024), Claude-3 Sonnet, GPT-4 Turbo (OpenAI, 2023), and Llama-2 (Touvron et al., 2023). In the case of Llama-2, we are using the 70 billion parameter chat variant, quantized with the Int 4 AWQ method (Lin et al., 2024), which we run locally with a single NVIDIA RTX 6000 Ada GPU.

4 Results and Discussion

4.1 Evaluation

To allow comparison with previous work (Irvin et al., 2019; Smit et al., 2020), for each pathology, we evaluate our methodology based on the weighted average F1 score across three aspects of the task: negation detection, positive mention detection, and uncertainty detection. We report the F1 scores of the sub-tasks in the Appendix. Each of those sub-tasks amounts to binary classification. For instance, *Negative* classes are transformed into positive in negation detection, while the other classes are transformed into negative. Positive mention detection and uncertainty detection are constructed with an analogous logic. The reported scores correspond to the averages across 1000 bootstrap replicates (Efron and Tibshirani, 1986), reported along the 95% Confidence Intervals (CI).

4.2 Data

For internal evaluation, we are evaluating the models on the gold-standard test set of annotated radiology reports used in the MIMIC-CXR paper (Johnson et al., 2019). MIMIC-CXR is considered an internal dataset for methods based on RadPert since RadGraph is trained on MIMIC-CXR radiology reports. The MIMIC-CXR gold-standard test set contains 687 radiology reports that do not overlap with the training and validation set of RadGraph.

For external evaluation, we have collected a private dataset from the Cambridge University Hospitals in Cambridge, UK. The CUH dataset consists of 650 radiology reports annotated by a single consultant radiologist with six years of experience, using the same annotation guidelines as MIMIC-CXR[3]. Details regarding the label distribution of both datasets are attached in Table 15 of the Appendix.

4.3 RadPert Evaluation

In Table 1, we report the weighted average F1 scores across the sub-tasks of positive mention detection, negation detection, and uncertainty Detection for the MIMIC-CXR and CUH datasets. We are also reporting the improvements over the CheXpert labeler alongside their confidence intervals. Radpert achieves a statistically significant improvement both on average and on the majority of the pathologies. Namely, for MIMIC-CXR, RadPert is 8.0% (95% CI: 5.5%, 10.8%) better than CheXpert, yielding an average F1 score of 0.757 (95% CI: 0.779, 0.800).

In Table 6 of the Appendix, we also report fine-grained results in the distinct sub-tasks. In addition to the sub-tasks of negation, positive mention, and uncertainty detection, we also report the performance improvement in mention detection. Mention detection treats *Null* as the positive class, and *Negative*, *Uncertain*, and *Positive* as the negative class.

4.3.1 Discussion on RadPert's Performance

We observe performance improvement in all subtasks. The strongest improvement is achieved in the uncertainty detection task, showcasing the effectiveness of utilizing the uncertainty labels of RadGraph. However, the improvement in mention detection is marginal. A primary cause of mention detection failure is the reliance on the RadGraph model, which occasionally fails to recall all entities and relations within a radiology report.

Focusing on specific pathologies, RadPert fails to consistently outperform CheXpert for Atelectasis, Edema, and Pleural Effusion. In the case of Atelectasis and Edema, the rule sets are straightforward, and their mentions often lack syntactic variability in practice, offering limited benefit from the uncertainty-aware entity representations of RadGraph. Regarding Pleural Effusion, RadPert is hindered by the divergence between RadGraph annota-

[3] MIMIC-CXR annotation guidelines were provided upon request by the authors of Johnson et al. (2019).

	MIMIC-CXR Gold Standard Test Set				CUH			
Pathologies	Weighted F1		Improvement over		Weighted F1		Improvement over	
	RadPert		CheXpert (%)		RadPert		CheXpert (%)	
Atelectasis	0.782	(0.740, 0.825)	-5.2	(-10.2, 0.2)	0.893	(0.836, 0.941)	-0.8	(-6.3, 4.4)
Cardiomegaly	0.801	(0.749, 0.846)	8.1	(4.2, 12.6)	0.910	(0.872, 0.945)	27.3	(16.4, 41.1)
Consolidation	0.806	(0.731, 0.872)	15.5	(1.9, 33.4)	0.951	(0.928, 0.971)	3.0	(0.4, 5.8)
Edema	0.801	(0.758, 0.843)	0.1	(-5.6, 3.9)	0.625	(0.466, 0.754)	-5.5	(-28.1, 19.1)
Enlarged Card.	0.628	(0.548, 0.702)	23.8	(5.6, 4.7)	0.908	(0.860, 0.950)	0.7	(-2.1, 3.5)
Fracture	0.866	(0.765, 0.946)	30.8	(9.7, 60.9)	0.764	(0.593, 0.898)	12.7	(-8.1, 47.5)
Lung Lesion	0.696	(0.583, 0.797)	4.0	(-5.4, 14.8)	0.816	(0.706, 0.911)	660.4	(210.8, 2700.3)
Lung Opacity	0.783	(0.741, 0.827)	3.2	(-1.3, 8.7)	0.712	(0.652, 0.766)	0.8	(-1.6, 3.5)
Pleur. Effusion	0.873	(0.843, 0.901)	-3.3	(-6.4, -0.2)	0.641	(0.587, 0.689)	0.1	(-2.5, 2.8)
Pleur. Other	0.547	(0.390, 0.692)	16.7	(1.6, 44.0)	0.082	(0.043, 0.127)	189.8	(45.9, 713.3)
Pneumonia	0.757	(0.704, 0.806)	28.1	(15.8, 42.5)	0.656	(0.520, 0.773)	54.5	(9.4, 130.9)
Pneumothorax	0.898	(0.856, 0.934)	5.1	(-0.4, 10.9)	0.626	(0.568, 0.682)	2.1	(-0.9, 5.7)
Sup. Devices	0.886	(0.854, 0.915)	2.1	(-0.4, 5.1)	0.858	(0.825, 0.890)	-2.6	(-4.8, -0.6)
Macro Avg.	0.757	(0.779, 0.800)	8.0	(5.5, 10.8)	0.726	(0.699, 0.752)	14.6	(10.4, 19.1)
Weighted Avg.	0.816	(0.802, 0.830)	3.4	(1.5, 5.3)	0.787	(0.765, 0.808)	5.0	(2.6, 7.3)

Table 1: Weighted average F1 scores for RadPert alongside improvements over the CheXpert model on the MIMIC-CXR gold-standard and CUH test sets. The F1 scores are averaged across the sub-tasks of positive mention detection, negation detection, and uncertainty detection weighted by the support sets. The scores correspond to the averages across 1000 bootstrap replicates and are reported alongside their confidence intervals.

tion guidelines[4] and those of the MIMIC-CXR and CUH datasets concerning uncertainty. Specifically, RadGraph suggests annotating any degree of uncertainty as *OBS:Uncertain* (Jain et al., 2021a) while the MIMIC-CXR guidelines, also used by CUH, permit some degree of uncertainty within *Positive* and *Negative* labels. For instance, "likely representing pneumonia" should be labeled as positive according to MIMIC-CXR guidelines. For Pleural Effusion, uncertain mentions such as "minimal if any pleural effusion" are commonplace and labeled inconsistently by the annotators in MIMIC-CXR. However, due to RadGraph's annotation guidelines, RadPert primarily labels such mentions as *Uncertain*, resulting in low precision in the uncertainty detection task for Pleural Effusion. This behavior can be observed in the Pleural Effusion confusion matrices (Appendix, Figure 3).

Notably, RadPert's performance for Lung Lesion showed a substantial improvement over CheXpert's performance on the CUH dataset compared to MIMIC-CXR. This discrepancy arises because "lung lesion" is a specific term frequently used in the CUH reports, while it rarely appears in MIMIC-CXR reports. The CheXpert labeler treats Lung Lesion as an umbrella term encompassing "masses", "nodular opacities", and "carcinomata", lacking specific rules for "lung lesions" and only identifying the less general terms, leading to inconsistent performance in CUH. Additionally, variations such as "edema" in the US and "oedema" in the UK also illustrate the divergent terminology and spelling conventions between the two corpora, although these spelling differences do not affect the ability of CheXpert to detect Edema mentions.

Finally, in Table 5 of the Appendix, we provide carbon estimates for both CheXpert and RadPert. RadPert not only improves upon CheXpert in performance but also demonstrates greater energy efficiency.

4.4 RadPrompt Evaluation

In Table 2, we present the improvement in the weighted average F1 score of RadPrompt for various base LLMs on the MIMIC-CXR gold-standard test set. Specifically, we compare the revised classification outcome of the second-turn prompt, which is infused with RadPert hints, to the first-turn classification outcome. For all tested LLMs, we observe that the RadPrompt strategy leads, on average (across pathologies), to a statistically significant improvement over the baseline zero-shot prompting. For clarity, in Tables 7, 8, 9, 10 and 11 of the Appendix, we also report the task-specific F1 scores of the first and second turns of RadPrompt.

Furthermore, we compare RadPrompt's second-

[4] Available on OpenReview.

	RadPrompt Improvement of Weighted Average F1 Over 1st Turn (%)							
Pathologies	**Gemini-1.5 Pro**		**Llama-2 70B**		**Claude-3 Sonnet**		**GPT-4 Turbo**	
Atelectasis	-0.9	(-4.4, 3.0)	-7.0	(-12.6, -0.2)	-1.4	(-7.1, 5.3)	-3.9	(-7.2, -0.4)
Cardiomegaly	-2.3	(-6.6, 1.9)	14.3	(9.2, 20.2)	2.7	(-2.4, 7.6)	-1.9	(-5.5, 1.5)
Consolidation	26.6	(13.9, 40.5)	70.7	(43.8, 102.6)	31.9	(15.7, 49.7)	2.6	(-3.6, 9.4)
Edema	7.7	(3.2, 12.5)	10.3	(4.5, 16.6)	7.4	(1.9, 13.1)	-3.1	(-5.9, -0.4)
Enlarged Card.	49.7	(22.1, 89.6)	160.2	(75.1, 309.4)	103.0	(55.7, 167.3)	3.9	(-8.6, 17.3)
Fracture	10.7	(1.4, 23.6)	20.1	(4.6, 42.0)	14.8	(0.8, 31.2)	5.2	(0.9, 9.9)
Lung Lesion	65.5	(37.3, 100.6)	24.0	(3.7, 48.0)	3.2	(-11.5, 18.5)	6.5	(-7.0, 20.4)
Lung Opacity	26.9	(18.8, 36.2)	23.5	(15.9, 32.3)	23.6	(14.1, 34.0)	8.1	(2.2, 14.4)
Pleural Effusion	4.1	(1.5, 6.5)	4.9	(1.2, 9.0)	8.3	(5.2, 11.4)	0.3	(-1.8, 2.4)
Pleural Other	21.0	(1.8, 44.6)	158.3	(-0.1, 291.8)	36.8	(8.2, 72.8)	10.8	(-6.9, 29.4)
Pneumonia	15.6	(10.3, 21.4)	-5.3	(-14.1, 4.0)	22.0	(14.2, 30.5)	4.5	(1.2, 8.3)
Pneumothorax	20.5	(14.9, 26.3)	19.3	(12.7, 26.8)	34.9	(28.2, 42.5)	1.0	(-1.3, 3.3)
Support Devices	4.1	(1.8, 6.7)	23.1	(15.7, 31.7)	1.1	(-0.8, 3.3)	0.5	(-0.5, 1.6)
Macro Average	14.8	(12.2, 17.3)	20.8	(16.2, 25.8)	16.2	(13.1, 19.4)	2.1	(0.3, 4.1)
Weighted Average	10.2	(8.4, 12.0)	12.5	(9.7, 15.4)	12.7	(10.7, 15.0)	0.9	(-0.2, 2.1)

Table 2: Improvement of weighted average F1 scores for RadPrompt over the base LLM on MIMIC-CXR gold-standard test set, alongside confidence intervals. Improvement is measured in a multi-turn chat setting by comparing the initial classification decision of the LLM to the revised classification decision after introducing RadPert hints.

	RadPrompt Improvement of Weighted Average F1 Over RadPert (%)							
Pathologies	**Gemini-1.5 Pro**		**Llama-2 70B**		**Claude-3 Sonnet**		**GPT-4 Turbo**	
Atelectasis	6.2	(0.8, 11.7)	-0.0	(-1.6, 1.7)	3.8	(0.6, 7.5)	6.2	(0.7, 11.7)
Cardiomegaly	-1.4	(-4.0, 1.2)	0.7	(-0.9, 2.4)	-0.2	(-1.5, 1.0)	0.7	(-2.8, 4.5)
Consolidation	-7.7	(-16.0, 0.1)	-22.4	(-29.6, -16.2)	-0.6	(-4.2, 3.2)	2.4	(-3.8, 9.1)
Edema	-0.9	(-3.9, 2.3)	0.5	(-0.8, 1.9)	0.1	(-0.8, 1.2)	1.3	(-1.7, 4.7)
Enlarged Card.	-11.6	(-19.1, -5.1)	-1.5	(-4.0, 0.7)	-8.0	(-14.1, -2.4)	-6.5	(-12.8, -0.8)
Fracture	-8.5	(-15.7, -1.2)	-1.2	(-4.0, 1.2)	-2.0	(-5.2, 0.0)	-4.5	(-11.7, 3.3)
Lung Lesion	-2.4	(-9.2, 4.9)	-28.4	(-37.9, -19.4)	2.1	(-5.3, 11.1)	-2.9	(-14.0, 9.2)
Lung Opacity	-5.0	(-8.0, -2.1)	-0.4	(-1.9, 1.1)	-0.4	(-2.7, 1.8)	-0.2	(-3.1, 2.8)
Pleural Effusion	2.0	(0.0, 4.1)	-0.7	(-2.1, 0.9)	3.2	(1.6, 5.0)	2.8	(0.4, 5.4)
Pleural Other	-10.0	(-20.3, 1.8)	-4.0	(-12.5, 0.0)	0.0	(0.0, 0.0)	13.5	(-3.9, 39.7)
Pneumonia	4.2	(-0.1, 9.4)	-14.8	(-19.8, -9.7)	3.0	(0.5, 6.4)	4.4	(-0.4, 9.5)
Pneumothorax	-0.6	(-3.1, 2.1)	-3.0	(-5.0, -1.3)	2.7	(0.3, 5.6)	3.5	(0.8, 7.1)
Support Devices	2.2	(0.5, 4.0)	-0.2	(-1.2, 0.5)	1.2	(-0.0, 2.5)	0.2	(-2.4, 2.8)
Macro Average	-2.2	(-3.8, -0.6)	-5.5	(-6.9, -4.3)	0.5	(-0.4, 1.4)	1.4	(-0.5, 3.2)
Weighted Average	-0.2	(-1.5, 1.2)	-3.5	(-4.4, -2.7)	1.4	(0.7, 2.1)	1.9	(0.7, 3.2)

Table 3: Improvement of weighted average F1 scores for RadPrompt over the rule-based RadPert on the MIMIC-CXR gold-standard test set, alongside confidence intervals.

turn results with RadPert in Table 3 for the MIMIC-CXR gold-standard test set. On average, RadPrompt with Gemini-1.5 Pro and Llama-2 70 B fail to outperform RadPert. However, Claude-3 Sonnet and GPT-4 Turbo-based RadPrompt surpass RadPert.

Regarding the external evaluation of RadPrompt, the current ethical agreement with the Cambridge University Hospitals limits the use of third-party APIs. Thus, we are only able to evaluate RadPrompt with a Llama-2 base. We present the weighted average and the sub-task-specific results in Tables 12 and 13. Similarly to the MIMIC-CXR gold-standard test set, we observe that Llama-2-based RadPrompt enhances the performance of Llama-2 but fails to improve upon RadPert.

4.4.1 Discussion on RadPrompt's Performance

We can observe from Tables 2 and 3 that RadPrompt on Claude-3 Sonnet and on GPT-4 Turbo exceeds, on average, both RadPert and the initial LLM predictions. Namely, RadPrompt with GPT-4 Turbo is 2.1% (CI 0.3%, 4.1%) better than baseline GPT-4 Turbo and 1.4% (CI -0.5%, 3.2%) better than RadPert.

Focusing on individual pathologies, we notice that RadPrompt with a Gemini-1.5 Pro base manages to outperform both of its underlying models for Pleural Effusion, Pneumonia, and Support Devices. Additionally, RadPrompt with Claude-3 Sonnet surpasses its underlying models in the case of Lung Lesion, Pleural Effusion, Pneumonia, Pneumothorax, and Support Devices. For a GPT-4 Turbo base, the same behavior is observed for Consolidation, Pleural Effusion, Pleural Other, Pneumonia, and Pneumothorax. The ability of RadPrompt to boost the performance of both its underlying models demonstrates the potential of combining the language reasoning capabilities of LLMs with the insights encoded in rule-based models.

In Table 4, we present a fine-grained comparison between the first and second turns of RadPrompt. We observe that all models, with the exception of GPT-4 Turbo, initially struggled to understand that we intended to classify only those pathologies explicitly mentioned in the report. This effect disproportionately affects the *Negative* class since *Null* is often conflated with *Negative*. The distinction, however, between those two labels is non-negligible. Inconsistencies often exist between the gold-standard labels extracted directly from chest X-ray Images and the gold-standard labels of their corresponding radiology reports, and thus, pathologies visible within a chest X-ray may be excluded from the radiology report (Jain et al., 2021b). Such observations are also noted in other clinical domains, such as Magnetic Resonance Imaging (MRI), where the clinical context and the referrer physician may bias the observations mentioned within a radiology report (Wood et al., 2020).

5 Limitations

While this study demonstrates promising improvements in radiology report classification using the RadPrompt methodology, several limitations must be considered.

RadPert and RadPrompt are exclusively developed and tested for the English language. The study also centers around a list of pathologies typical of chest X-rays. As such, the extension of our methodologies to other languages, types of medical imaging, and additional pathologies was not verified.

Furthermore, previous studies have highlighted discrepancies between labels from radiology report annotations and those from the corresponding imaging study annotations (Jain et al., 2021b; Wood et al., 2020). The source of such inconsistencies includes incomplete radiology report impressions, hierarchical relationships within labels, and the undeniable uncertainty of the task. In future work, we aim to study this effect within the CUH test set.

Due to ethical considerations, we are currently unable to perform inference for the CUH test set through third-party APIs. Thus, we have not evaluated RadPrompt externally for SOTA LLMs. We expect to overcome this limitation after the planned release of the CUH dataset.

Additionally, we cannot estimate the computational cost and carbon footprint for GPT-4-based RadPrompt due to a lack of specific metrics. In the Appendix, we provide carbon footprint estimates for the Llama-2-based RadPrompt, which is significantly higher than RadPert and CheXpert. Nonetheless, RadPert delivers performance comparable to GPT-4 while operating on a commercial CPU with minimal carbon emissions, underscoring its benefits in resource-limited environments.

Finally, there is an inherent degree of ambiguity in classifying radiology reports, especially as it pertains to the *Uncertainty* labels. We aim to extend current datasets with labels from multiple annota-

	RadPrompt Improvement of Weighted Average F1 Over 1st Turn (%)			
Sub-task	**Gemini-1.5 Pro**	**Llama-2 70B**	**Claude-3 Sonnet**	**GPT-4 Turbo**
Mention Detection	17.8 (15.6, 20.0)	26.7 (23.8, 29.8)	24.6 (21.7, 27.8)	1.9 (0.9, 3.0)
Negation Detection	31.9 (26.4, 37.6)	54.8 (45.8, 64.2)	62.3 (52.1, 73.1)	4.9 (2.3, 8.1)
Pos. Mention Detection	3.8 (2.4, 5.2)	1.7 (-0.5, 4.1)	0.7 (-0.9, 2.4)	-0.4 (-1.6, 0.7)
Uncertainty Detection	2.9 (-5.9, 13.0)	-6.4 (-20.7, 9.8)	-0.5 (-13.0, 14.0)	-2.6 (-10.3, 5.9)
Weighted Average	10.2 (8.4, 12.0)	12.5 (9.7, 15.4)	12.7 (10.7, 15.0)	0.9 (-0.2, 2.1)

Table 4: Improvement of RadPrompt over the base LLM for the different sub-tasks on MIMIC-CXR gold-standard test set. For each sub-task. we report the improvement of the weighted average F1 score across all pathologies, along with confidence intervals. The weighted average refers to averaging over sub-tasks, excluding the mention detection task.

tors in order to measure annotator agreement.

6 Conclusions

This paper introduced RadPert, a rule-based system enhanced by the RadGraph information schema, demonstrating significant improvements in the classification of radiology reports. By leveraging entity-level uncertainty labels, RadPert reduces reliance on comprehensive rule sets. Our evaluations show that RadPert surpasses CheXpert, the previous rule-based SOTA, by achieving an 8.0% (95% CI: 5.5%, 10.8%) increase in F1 score, with confidence intervals strongly supporting this improvement.

Further extending the application of RadPert, we developed RadPrompt, a multi-turn prompting strategy that utilizes insights from RadPert to enhance the zero-shot prediction capabilities of large language models. RadPrompt demonstrated a 2.1% (95% CI: 0.3%, 4.1%) improvement in F1 score over GPT-4 Turbo, indicating its potential to refine predictions in clinical settings. These results highlight the growing synergy between structured rule-based systems and large language models, offering a promising direction for future research in biomedical Natural Language Processing.

As we continue to refine these tools, future work will focus on expanding the existing datasets and addressing the discrepancies between gold-standard image labels and those extracted from radiology reports.

Code and Data Availability

Code for RadPert and RadPrompt is available on GitHub[5]. The CUH dataset is planned to be released in the following months while managed and made available through the hospital's clinical informatics unit.

Ethical Considerations

For the MIMIC-CXR gold-standard test set, access to LLMs through APIs conforms to the PhysioNet responsible use guidelines[6].

This ethical agreement with Cambridge University Hospitals currently limits the use of third-party APIs, but it is being revised prior to the dataset's release.

Acknowledgments

Panagiotis wishes to thank the David and Claudia Harding Foundation for their support through the University of Cambridge Harding Distinguished Postgraduate Scholars Programme. Anna Breger and Ian wish to acknowledge the EU/EFPIA Innovative Medicines Initiative 2 Joint Undertaking - DRAGON (101005122). Anna Breger is also supported by the Austrian Science Fund (FWF) through project P33217 and Ian by the National Institute for Health and Care Research (NIHR) Cambridge Biomedical Research Centre (BRC-1215-20014). Anna Korhonen wishes to acknowledge the UK Research and Innovation (UKRI) Frontier Research Grant (EP/Y031350/1).

[5] https://github.com/PanagiotisFytas/RadPert-RadPrompt.

[6] https://physionet.org/news/post/gpt-responsible-use

References

Lisa C Adams, Daniel Truhn, Felix Busch, Avan Kader, Stefan M Niehues, Marcus R Makowski, and Keno K Bressem. 2023. Leveraging gpt-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. *Radiology*, 307(4):e230725.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. 2023. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15016–15027.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. 2022. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer.

Selen Bozkurt, Emel Alkim, Imon Banerjee, and Daniel L Rubin. 2019. Automated detection of measurements and their descriptors in radiology reports using a hybrid natural language processing algorithm. *Journal of digital imaging*, 32:544–553.

Felix J Dorfner, Liv Jürgensen, Leonhard Donle, Fares Al Mohamad, Tobias R Bodenmann, Mason C Cleveland, Felix Busch, Lisa C Adams, James Sato, Thomas Schultz, et al. 2024. Is open-source there yet? a comparative study on commercial and open-source llms in their ability to label chest x-ray reports. *arXiv preprint arXiv:2402.12298*.

Bradley Efron and Robert Tibshirani. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, pages 54–75.

Peter Floderus, Mirosław Kowaluk, Andrzej Lingas, and Eva-Marta Lundell. 2015. Induced subgraph isomorphism: Are some patterns substantially easier than others? *Theoretical Computer Science*, 605:119–128.

Jawook Gu, Han-Cheol Cho, Jiho Kim, Kihyun You, Eun Kyoung Hong, and Byungseok Roh. 2024. Chex-gpt: Harnessing large language models for enhanced chest x-ray report labeling. *arXiv preprint arXiv:2401.11505*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Saeed Hassanpour, Curtis P Langlotz, Timothy J Amrhein, Nicholas T Befera, and Matthew P Lungren. 2017. Performance of a machine learning classifier of knee mri reports in two large academic radiology practices: a tool to estimate diagnostic yield. *American Journal of Roentgenology*, 208(4):750–753.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. 2021. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951.

A Infante, S Gaudino, F Orsini, A Del Ciello, C Gullì, B Merlino, L Natale, R Iezzi, and E Sala. 2024. Large language models (llms) in the evaluation of emergency radiology reports: performance of chatgpt-4, perplexity, and bard. *Clinical radiology*, 79(2):102–106.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *CoRR*, abs/1901.07031.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew Lungren, Andrew Ng, Curtis Langlotz, Pranav Rajpurkar, and Pranav Rajpurkar. 2021a. Radgraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Saahil Jain, Akshay Smit, Steven QH Truong, Chanh DT Nguyen, Minh-Thanh Huynh, Mudit Jain, Victoria A. Young, Andrew Y. Ng, Matthew P. Lungren, and Pranav Rajpurkar. 2021b. Visualchexbert: addressing the discrepancy between radiology report labels and image labels. In *Proceedings of the Conference on Health, Inference, and Learning*, CHIL '21, page 105–115, New York, NY, USA. Association for Computing Machinery.

Sanjay Jeganathan. 2023. The growing problem of radiologist shortages: Australia and new zealand's perspective. *Korean Journal of Radiology*, 24(11):1043.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Sadhana Kalidindi and Sanjay Gandhi. 2023. Workforce crisis in radiology in the uk and the strategies to deal with it: Is artificial intelligence the saviour? *Cureus*, 15(8).

Kleanthis Konstantinidis. 2023. The shortage of radiographers: A global crisis in healthcare. *Journal of Medical Imaging and Radiation Sciences*.

Loïc Lannelongue, Jason Grealey, and Michael Inouye. 2021. Green algorithms: quantifying the carbon footprint of computation. *Advanced science*, 8(12):2100707.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. In *Proceedings of Machine Learning and Systems*, volume 6, pages 87–100.

Qianchu Liu, Stephanie Hyland, Shruthi Bannur, Kenza Bouzid, Daniel Castro, Maria Wetscherek, Robert Tinn, Harshita Sharma, Fernando Pérez-García, Anton Schwaighofer, Pranav Rajpurkar, Sameer Khanna, Hoifung Poon, Naoto Usuyama, Anja Thieme, Aditya Nori, Matthew Lungren, Ozan Oktay, and Javier Alvarez-Valle. 2023. Exploring the boundaries of GPT-4 in radiology. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14414–14445, Singapore. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. 2017. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2017:188–196.

Ewoud Pons, Loes MM Braun, MG Myriam Hunink, and Jan A Kors. 2016. Natural language processing in radiology: a systematic review. *Radiology*, 279(2):329–343.

Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, et al. 2018. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS medicine*, 15(11):e1002686.

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):115.

Muhammad Imran Razzak, Saeeda Naz, and Ahmad Zaib. 2018. Deep learning for medical image processing: Overview, challenges and the future. *Classification in BioApps: Automation of decision making*, pages 323–350.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Eduardo P Reis, Joselisa PQ De Paiva, Maria CB Da Silva, Guilherme AS Ribeiro, Victor F Paiva, Lucas Bulgarelli, Henrique MH Lee, Paulo V Santos, Vanessa M Brito, Lucas TW Amaral, et al. 2022. Brax, brazilian labeled chest x-ray dataset. *Scientific Data*, 9(1):487.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew 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)*, pages 1500–1519, Online. Association for Computational Linguistics.

Yu-Xing Tang, You-Bao Tang, Yifan Peng, Ke Yan, Mohammadhadi Bagheri, Bernadette A Redd, Catherine J Brandon, Zhiyong Lu, Mei Han, Jing Xiao, et al. 2020. Automated abnormality classification of chest radiographs using deep convolutional neural networks. *NPJ digital medicine*, 3(1):70.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. 2022. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, Hong Kong, China. Association for Computational Linguistics.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *CoRR*, abs/1705.02315.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, et al. 2018. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, 77:34–49.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alessandro Wollek, Sardi Hyska, Thomas Sedlmeyr, Philip Haitzer, Johannes Rueckel, Bastian O Sabel, Michael Ingrisch, and Tobias Lasser. 2024. German chexpert chest x-ray radiology report labeler. In *RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*. Georg Thieme Verlag KG.

David A Wood, Sina Kafiabadi, Aisha Al Busaidi, Emily Guilhem, Jeremy Lynch, Matthew Townend, Antanas Montvila, Juveria Siddiqui, Naveen Gadapa, Matthew Benger, et al. 2020. Labelling imaging datasets on the basis of neuroradiology reports: a validation study. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing: Third International Workshop, iMIMIC 2020, Second International Workshop, MIL3ID 2020, and 5th International Workshop, LABELS 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings 3*, pages 254–265. Springer.

Appendix

	CheXpert	RadPert	Llama-2 70B	RadPrompt /w Llama-2 70B
Runtime (min)	7.1	4.8	41.8	43.6
CO_2e (g)	5.44	3.68	85.48	89.16
Device	CPU	CPU	GPU	GPU
Model	Core i7-6700k	Core i7-6700k	NVIDIA RTX 6000 Ada	NVIDIA RTX 6000 Ada

Table 5: Carbon footprint for inference on both MIMIC-CXR gold-standard and CUH test sets, as estimated utilizing the tools from Lannelongue et al. (2021). For RadPert, calculations include the extraction of the RadGraph knowledge graph. Notably, we are not able to provide estimates for GPT-4 Turbo, Gemini-1.5 Pro, and Claude-3 Sonnet since this information is not provided by the respective API providers.

	Negation Detection				**Uncertainty Detection**			
Pathologies	**F1 Score**		**Improvement over**		**F1 Score**		**Improvement over**	
	RadPert		**CheXpert (%)**		**RadPert**		**CheXpert (%)**	
Atelectasis	0.581	(0.000, 0.909)	61.6	(-41.8, 340.2)	0.386	(0.256, 0.511)	0.1	(-29.1, 44.7)
Cardiomegaly	0.834	(0.769, 0.892)	7.1	(0.6, 14.8)	0.093	(0.000, 0.227)	Inf.	(0.0, Inf.)
Consolidation	0.877	(0.762, 0.960)	-6.2	(-17.4, 2.8)	0.665	(0.488, 0.818)	269.7	(0.0, 909.7)
Edema	0.832	(0.773, 0.886)	8.7	(2.6, 16.5)	0.395	(0.160, 0.600)	104.2	(3.4, 275.3)
Enlarged Card.	0.916	(0.836, 0.982)	49.5	(21.9, 96.6)	0.062	(0.000, 0.207)	-3.3	(-28.6, 23.1)
Fracture	0.733	(0.444, 0.947)	0.0	(0.0, 0.0)	0.498	(0.000, 1.000)	Inf.	(0.0, Inf.)
Lung Lesion	0.422	(0.000, 0.800)	-5.1	(-50.0, 55.6)	0.128	(0.000, 0.400)	Inf.	(0.0, Inf.)
Lung Opacity	0.513	(0.353, 0.674)	32.2	(-17.3, 128.6)	0.000	(0.000, 0.000)	0.0	(0.0, 0.0)
Pler. Effusion	0.916	(0.871, 0.956)	-2.6	(-6.3, 1.3)	0.422	(0.267, 0.561)	-14.5	(-42.6, 22.8)
Pler. Other	0.000	(0.000, 0.000)	0.0	(0.0, 0.0)	0.000	(0.000, 0.000)	0.0	(0.0, 0.0)
Pneumonia	0.915	(0.867, 0.955)	17.3	(8.6, 29.0)	0.671	(0.582, 0.743)	43.8	(19.1, 76.5)
Pneumothorax	0.937	(0.912, 0.960)	2.1	(-0.7, 5.2)	0.645	(0.307, 0.909)	125.6	(-7.7, 540.1)
Sup. Devices	0.283	(0.000, 0.545)	0.0	(0.0, 0.0)	0.000	(0.000, 0.000)	0.0	(0.0, 0.0)
Macro Avg.	0.743	(0.686, 0.810)	4.1	(-4.9, 14.5)	0.453	(0.369, 0.554)	40.4	(9.5, 79.8)
Weighted Avg.	0.872	(0.852, 0.893)	5.9	(3.3, 8.7)	0.530	(0.460, 0.607)	31.4	(8.2, 61.3)
	Positive Mention Detection				**Mention Detection**			
Pathologies	**F1 Score**		**Improvement over**		**F1 Score**		**Improvement over**	
	RadPert		**CheXpert (%)**		**RadPert**		**CheXpert (%)**	
Atelectasis	0.819	(0.776, 0.859)	-5.8	(-10.2, -1.5)	0.944	(0.920, 0.965)	0.0	(0.0, 0.0)
Cardiomegaly	0.851	(0.806, 0.893)	7.5	(3.4, 12.0)	0.858	(0.826, 0.890)	-0.0	(-2.8, 3.0)
Consolidation	0.815	(0.724, 0.885)	8.6	(-0.6, 19.8)	0.930	(0.888, 0.963)	0.0	(0.0, 0.0)
Edema	0.809	(0.759, 0.859)	-8.2	(-12.9, -3.3)	0.887	(0.859, 0.916)	-0.2	(-0.9, 0.4)
Enlarged Card.	0.442	(0.336, 0.551)	1.3	(-21.3, 28.8)	0.529	(0.454, 0.609)	16.1	(-0.7, 36.7)
Fracture	0.902	(0.831, 0.964)	8.1	(0.9, 17.5)	0.952	(0.907, 0.990)	5.4	(-0.1, 12.4)
Lung Lesion	0.796	(0.702, 0.878)	1.9	(-6.4, 10.2)	0.834	(0.752, 0.901)	-2.4	(-7.0, 1.7)
Lung Opacity	0.819	(0.774, 0.859)	1.6	(-1.3, 4.5)	0.800	(0.757, 0.840)	-0.0	(-1.2, 1.2)
Pler. Effusion	0.889	(0.859, 0.916)	-3.1	(-6.3, 0.0)	0.979	(0.968, 0.989)	0.6	(-0.1, 1.4)
Pler. Other	0.592	(0.441, 0.727)	16.7	(1.6, 44.0)	0.592	(0.459, 0.709)	1.1	(-5.3, 11.3)
Pneumonia	0.654	(0.550, 0.744)	36.9	(8.5, 75.6)	0.952	(0.931, 0.971)	-0.5	(-1.5, 0.4)
Pneumothorax	0.765	(0.630, 0.870)	9.6	(-7.5, 30.4)	0.963	(0.945, 0.980)	-0.7	(-1.5, 0.0)
Sup. Devices	0.898	(0.869, 0.926)	1.3	(-0.7, 3.5)	0.893	(0.862, 0.918)	1.4	(-0.2, 3.1)
Macro Avg.	0.773	(0.749, 0.796)	4.1	(1.5, 6.8)	0.855	(0.839, 0.870)	1.0	(-0.0, 2.0)
Weighted Avg.	0.824	(0.809, 0.839)	0.9	(-0.8, 2.6)	0.899	(0.890, 0.908)	0.4	(-0.1, 0.9)

Table 6: F1 scores of RadPert and improvement over CheXpert on MIMIC-CXR gold-standard test set. We report results for the sub-tasks of negation detection, uncertainty detection, positive mention detection and mention detection. The scores correspond to the averages across 1000 bootstrap replicates and are reported alongside their confidence intervals.

	Weighted Average F1 Across Tasks							
	Gemini-1.5 Pro				Llama-2 70B			
Pathologies	**Base LLM**		**RadPrompt**		**Base LLM**		**RadPrompt**	
Atelectasis	**0.838**	**(0.792, 0.878)**	0.830	(0.787, 0.870)	**0.842**	**(0.790, 0.884)**	0.782	(0.739, 0.822)
Cardiomegaly	**0.809**	**(0.771, 0.842)**	0.790	(0.740, 0.835)	0.706	(0.657, 0.755)	**0.807**	**(0.756, 0.853)**
Consolidation	0.588	(0.507, 0.662)	**0.743**	**(0.665, 0.815)**	0.368	(0.302, 0.430)	**0.625**	**(0.544, 0.700)**
Edema	0.737	(0.695, 0.778)	**0.794**	**(0.752, 0.834)**	0.729	(0.686, 0.766)	**0.804**	**(0.762, 0.845)**
Enlarged Card.	0.376	(0.275, 0.468)	**0.556**	**(0.464, 0.643)**	0.248	(0.158, 0.343)	**0.619**	**(0.537, 0.695)**
Fracture	0.718	(0.602, 0.820)	**0.792**	**(0.696, 0.874)**	0.717	(0.583, 0.839)	**0.855**	**(0.759, 0.932)**
Lung Lesion	0.413	(0.321, 0.508)	**0.678**	**(0.575, 0.776)**	0.404	(0.313, 0.498)	**0.498**	**(0.397, 0.595)**
Lung Opacity	0.587	(0.532, 0.638)	**0.744**	**(0.700, 0.791)**	0.632	(0.583, 0.681)	**0.780**	**(0.737, 0.824)**
Pleural Effusion	0.856	(0.829, 0.880)	**0.891**	**(0.863, 0.916)**	0.827	(0.798, 0.853)	**0.867**	**(0.837, 0.895)**
Pleural Other	0.409	(0.281, 0.535)	**0.492**	**(0.346, 0.626)**	0.312	(0.129, 0.490)	**0.525**	**(0.363, 0.669)**
Pneumonia	0.683	(0.635, 0.734)	**0.789**	**(0.740, 0.836)**	**0.682**	**(0.638, 0.724)**	0.645	(0.587, 0.698)
Pneumothorax	0.741	(0.699, 0.781)	**0.893**	**(0.855, 0.926)**	0.730	(0.687, 0.773)	**0.871**	**(0.825, 0.908)**
Support Devices	0.870	(0.836, 0.903)	**0.905**	**(0.877, 0.932)**	0.718	(0.669, 0.767)	**0.883**	**(0.851, 0.913)**
Macro Average	0.664	(0.642, 0.685)	**0.761**	**(0.739, 0.783)**	0.609	(0.585, 0.631)	**0.736**	**(0.714, 0.756)**
Weighted Average	0.740	(0.724, 0.755)	**0.815**	**(0.799, 0.829)**	0.700	(0.684, 0.717)	**0.788**	**(0.772, 0.803)**
	Claude-3 Sonnet				**GPT-4 Turbo**			
Pathologies	**Base LLM**		**RadPrompt**		**Base LLM**		**RadPrompt**	
Atelectasis	**0.823**	**(0.774, 0.868)**	0.812	(0.769, 0.850)	**0.864**	**(0.819, 0.902)**	0.830	(0.785, 0.870)
Cardiomegaly	0.778	(0.742, 0.813)	**0.799**	**(0.746, 0.845)**	**0.822**	**(0.777, 0.858)**	0.806	(0.754, 0.849)
Consolidation	0.609	(0.530, 0.679)	**0.801**	**(0.729, 0.865)**	0.804	(0.729, 0.864)	**0.825**	**(0.752, 0.892)**
Edema	0.747	(0.702, 0.788)	**0.802**	**(0.758, 0.846)**	**0.837**	**(0.801, 0.875)**	0.811	(0.771, 0.853)
Enlarged Card.	0.289	(0.211, 0.369)	**0.578**	**(0.494, 0.658)**	0.567	(0.474, 0.650)	**0.587**	**(0.502, 0.667)**
Fracture	0.742	(0.622, 0.856)	**0.849**	**(0.751, 0.929)**	0.785	(0.692, 0.868)	**0.826**	**(0.732, 0.908)**
Lung Lesion	0.689	(0.586, 0.784)	**0.709**	**(0.596, 0.808)**	0.634	(0.534, 0.725)	**0.675**	**(0.565, 0.780)**
Lung Opacity	0.632	(0.574, 0.686)	**0.780**	**(0.737, 0.823)**	0.724	(0.674, 0.769)	**0.782**	**(0.738, 0.823)**
Pleural Effusion	0.832	(0.806, 0.858)	**0.901**	**(0.875, 0.925)**	0.895	(0.869, 0.920)	**0.898**	**(0.872, 0.923)**
Pleural Other	0.404	(0.278, 0.530)	**0.547**	**(0.390, 0.692)**	0.558	(0.418, 0.680)	**0.616**	**(0.462, 0.737)**
Pneumonia	0.640	(0.588, 0.688)	**0.780**	**(0.727, 0.828)**	0.756	(0.699, 0.807)	**0.790**	**(0.738, 0.838)**
Pneumothorax	0.684	(0.638, 0.723)	**0.922**	**(0.887, 0.953)**	0.920	(0.890, 0.948)	**0.929**	**(0.895, 0.960)**
Support Devices	0.886	(0.856, 0.914)	**0.896**	**(0.867, 0.924)**	0.883	(0.849, 0.913)	**0.887**	**(0.855, 0.918)**
Macro Average	0.674	(0.653, 0.693)	**0.783**	**(0.761, 0.804)**	0.773	(0.752, 0.792)	**0.789**	**(0.768, 0.808)**
Weighted Average	0.734	(0.718, 0.750)	**0.827**	**(0.813, 0.841)**	0.824	(0.808, 0.838)	**0.832**	**(0.818, 0.845)**

Table 7: Weighted F1 Scores across positive mention detection, negation detection, and uncertainty detection for RadPrompt on MIMIC-CXR gold-standard test set. The “Base LLM” column refers to the first-turn prediction of the LLM, and the “RadPrompt” column to the second-turn prediction. The scores correspond to the averages across 1000 bootstrap replicates and are reported alongside their confidence intervals.

	Mention Detection F1			
	Gemini-1.5 Pro		Llama-2 70B	
Pathologies	**Base LLM**	**RadPrompt**	**Base LLM**	**RadPrompt**
Atelectasis	0.785 (0.748, 0.819)	**0.926 (0.901, 0.949)**	0.746 (0.706, 0.785)	**0.939 (0.914, 0.961)**
Cardiomegaly	0.827 (0.793, 0.861)	**0.848 (0.815, 0.882)**	0.767 (0.731, 0.803)	**0.865 (0.832, 0.896)**
Consolidation	0.516 (0.450, 0.579)	**0.869 (0.817, 0.912)**	0.374 (0.318, 0.428)	**0.738 (0.672, 0.797)**
Edema	0.781 (0.745, 0.819)	**0.869 (0.838, 0.898)**	0.744 (0.704, 0.782)	**0.883 (0.854, 0.911)**
Enlarged Card.	0.409 (0.344, 0.480)	**0.504 (0.427, 0.582)**	0.326 (0.244, 0.414)	**0.534 (0.458, 0.614)**
Fracture	0.469 (0.385, 0.548)	**0.840 (0.767, 0.906)**	0.324 (0.258, 0.390)	**0.934 (0.881, 0.976)**
Lung Lesion	0.293 (0.236, 0.346)	**0.708 (0.625, 0.788)**	0.283 (0.229, 0.335)	**0.589 (0.495, 0.664)**
Lung Opacity	0.573 (0.526, 0.620)	**0.765 (0.724, 0.807)**	0.592 (0.545, 0.634)	**0.783 (0.743, 0.823)**
Pleural Effusion	0.913 (0.892, 0.932)	**0.966 (0.952, 0.978)**	0.883 (0.860, 0.907)	**0.964 (0.950, 0.977)**
Pleural Other	0.227 (0.158, 0.297)	**0.448 (0.323, 0.560)**	0.149 (0.091, 0.210)	**0.577 (0.444, 0.699)**
Pneumonia	0.802 (0.767, 0.838)	**0.940 (0.917, 0.960)**	0.714 (0.674, 0.753)	**0.890 (0.861, 0.915)**
Pneumothorax	0.760 (0.719, 0.797)	**0.941 (0.919, 0.961)**	0.758 (0.716, 0.795)	**0.943 (0.919, 0.964)**
Support Devices	0.804 (0.767, 0.837)	**0.888 (0.858, 0.913)**	0.655 (0.606, 0.701)	**0.892 (0.862, 0.917)**
Macro Average	0.627 (0.611, 0.647)	**0.809 (0.788, 0.829)**	0.563 (0.547, 0.580)	**0.810 (0.793, 0.827)**
Weighted Average	0.742 (0.724, 0.759)	**0.874 (0.861, 0.887)**	0.687 (0.670, 0.705)	**0.871 (0.860, 0.881)**
	Claude-3 Sonnet		GPT-4 Turbo	
Pathologies	**Base LLM**	**RadPrompt**	**Base LLM**	**RadPrompt**
Atelectasis	0.802 (0.767, 0.837)	**0.936 (0.911, 0.959)**	0.928 (0.901, 0.950)	**0.942 (0.918, 0.962)**
Cardiomegaly	0.777 (0.740, 0.813)	**0.858 (0.826, 0.890)**	0.858 (0.826, 0.889)	**0.859 (0.826, 0.892)**
Consolidation	0.50 (0.437, 0.561)	**0.921 (0.879, 0.956)**	0.882 (0.829, 0.922)	**0.930 (0.888, 0.963)**
Edema	0.789 (0.750, 0.826)	**0.884 (0.855, 0.911)**	**0.895 (0.868, 0.921)**	0.872 (0.842, 0.902)
Enlarged Card.	0.270 (0.222, 0.322)	**0.530 (0.453, 0.610)**	0.585 (0.505, 0.655)	**0.559 (0.478, 0.637)**
Fracture	0.442 (0.360, 0.522)	**0.933 (0.880, 0.976)**	0.811 (0.736, 0.883)	**0.885 (0.821, 0.942)**
Lung Lesion	0.398 (0.330, 0.464)	**0.847 (0.774, 0.908)**	0.701 (0.618, 0.776)	**0.799 (0.722, 0.868)**
Lung Opacity	0.564 (0.514, 0.612)	**0.790 (0.749, 0.830)**	0.742 (0.695, 0.786)	**0.795 (0.754, 0.834)**
Pleural Effusion	0.856 (0.830, 0.881)	**0.977 (0.966, 0.988)**	0.966 (0.953, 0.978)	**0.976 (0.965, 0.987)**
Pleural Other	0.211 (0.141, 0.278)	**0.592 (0.459, 0.709)**	0.560 (0.429, 0.674)	**0.630 (0.50, 0.744)**
Pneumonia	0.748 (0.708, 0.787)	**0.950 (0.928, 0.969)**	0.928 (0.905, 0.950)	**0.953 (0.932, 0.971)**
Pneumothorax	0.693 (0.651, 0.731)	**0.970 (0.953, 0.985)**	0.953 (0.934, 0.973)	**0.970 (0.953, 0.985)**
Support Devices	0.862 (0.831, 0.890)	**0.895 (0.866, 0.920)**	0.897 (0.868, 0.922)	**0.901 (0.875, 0.926)**
Macro Average	0.608 (0.591, 0.628)	**0.853 (0.837, 0.868)**	0.824 (0.804, 0.843)	**0.852 (0.836, 0.867)**
Weighted Average	0.720 (0.701, 0.739)	**0.897 (0.889, 0.906)**	0.881 (0.868, 0.892)	**0.897 (0.888, 0.907)**

Table 8: Mention detection F1 Scores for RadPrompt on MIMIC-CXR gold-standard test set. The "Base LLM" column refers to the first-turn prediction of the LLM, and the "RadPrompt" column to the second-turn prediction. The scores correspond to the averages across 1000 bootstrap replicates and are reported alongside their confidence intervals.

	Negation Detection F1							
	Gemini-1.5 Pro				Llama-2 70B			
Pathologies	**Base LLM**		**RadPrompt**		**Base LLM**		**RadPrompt**	
Atelectasis	0.066	(0.000, 0.143)	**0.340**	**(0.000, 0.616)**	0.015	(0.000, 0.047)	**0.579**	**(0.000, 0.909)**
Cardiomegaly	0.673	(0.594, 0.739)	**0.742**	**(0.667, 0.811)**	0.546	(0.477, 0.617)	**0.852**	**(0.789, 0.906)**
Consolidation	0.286	(0.188, 0.379)	**0.790**	**(0.654, 0.893)**	0.210	(0.136, 0.293)	**0.739**	**(0.591, 0.857)**
Edema	0.656	(0.583, 0.721)	**0.801**	**(0.737, 0.857)**	0.555	(0.483, 0.621)	**0.833**	**(0.769, 0.890)**
Enlarged Card.	0.455	(0.344, 0.561)	**0.696**	**(0.581, 0.804)**	0.125	(0.000, 0.294)	**0.887**	**(0.800, 0.962)**
Fracture	0.122	(0.043, 0.206)	**0.474**	**(0.235, 0.688)**	0.058	(0.020, 0.105)	**0.688**	**(0.400, 0.909)**
Lung Lesion	0.036	(0.000, 0.089)	**0.225**	**(0.000, 0.500)**	0.028	(0.000, 0.068)	**0.263**	**(0.000, 0.750)**
Lung Opacity	0.203	(0.101, 0.306)	**0.428**	**(0.256, 0.600)**	0.235	(0.142, 0.327)	**0.539**	**(0.373, 0.692)**
Pleural Effusion	0.733	(0.660, 0.798)	**0.888**	**(0.839, 0.932)**	0.625	(0.545, 0.698)	**0.870**	**(0.816, 0.923)**
Pleural Other	**0.035**	**(0.000, 0.087)**	0.000	(0.000, 0.000)	**0.023**	**(0.000, 0.057)**	0.000	(0.000, 0.000)
Pneumonia	0.624	(0.553, 0.692)	**0.887**	**(0.833, 0.933)**	0.498	(0.428, 0.567)	**0.823**	**(0.753, 0.888)**
Pneumothorax	0.714	(0.665, 0.756)	**0.922**	**(0.894, 0.948)**	0.710	(0.664, 0.753)	**0.909**	**(0.878, 0.937)**
Support Devices	0.059	(0.000, 0.143)	**0.207**	**(0.000, 0.414)**	0.026	(0.000, 0.065)	**0.295**	**(0.000, 0.556)**
Macro Average	0.371	(0.340, 0.416)	**0.628**	**(0.569, 0.693)**	0.302	(0.268, 0.350)	**0.717**	**(0.648, 0.787)**
Weighted Average	0.622	(0.590, 0.655)	**0.820**	**(0.788, 0.850)**	0.544	(0.511, 0.579)	**0.841**	**(0.818, 0.864)**
	Claude-3 Sonnet				**GPT-4 Turbo**			
Pathologies	**Base LLM**		**RadPrompt**		**Base LLM**		**RadPrompt**	
Atelectasis	0.099	(0.025, 0.198)	**0.581**	**(0.000, 0.909)**	0.515	(0.167, 0.800)	**0.868**	**(0.500, 1.000)**
Cardiomegaly	0.554	(0.478, 0.621)	**0.796**	**(0.721, 0.857)**	0.717	(0.640, 0.785)	**0.761**	**(0.684, 0.829)**
Consolidation	0.227	(0.151, 0.305)	**0.896**	**(0.788, 0.973)**	0.752	(0.615, 0.857)	**0.899**	**(0.800, 0.978)**
Edema	0.661	(0.589, 0.731)	**0.827**	**(0.768, 0.884)**	**0.871**	**(0.814, 0.921)**	0.836	(0.775, 0.893)
Enlarged Card.	0.148	(0.102, 0.199)	**0.713**	**(0.590, 0.818)**	0.620	(0.488, 0.736)	**0.741**	**(0.625, 0.841)**
Fracture	0.090	(0.031, 0.160)	**0.733**	**(0.444, 0.947)**	0.627	(0.333, 0.857)	**0.811**	**(0.545, 1.000)**
Lung Lesion	0.023	(0.000, 0.058)	**0.530**	**(0.000, 1.000)**	0.239	(0.000, 0.500)	**0.412**	**(0.000, 0.800)**
Lung Opacity	0.117	(0.059, 0.177)	**0.495**	**(0.326, 0.647)**	0.382	(0.217, 0.540)	**0.494**	**(0.318, 0.653)**
Pleural Effusion	0.572	(0.500, 0.644)	**0.937**	**(0.897, 0.973)**	0.903	(0.855, 0.946)	**0.948**	**(0.910, 0.981)**
Pleural Other	**0.032**	**(0.000, 0.076)**	0.000	(0.000, 0.000)	0.305	(0.000, 0.632)	**0.425**	**(0.000, 1.000)**
Pneumonia	0.537	(0.466, 0.604)	**0.909**	**(0.859, 0.951)**	0.862	(0.802, 0.910)	**0.914**	**(0.866, 0.954)**
Pneumothorax	0.638	(0.591, 0.683)	**0.947**	**(0.924, 0.969)**	0.937	(0.913, 0.959)	**0.955**	**(0.934, 0.973)**
Support Devices	0.174	(0.000, 0.350)	**0.259**	**(0.000, 0.500)**	**0.182**	**(0.000, 0.545)**	0.123	(0.000, 0.375)
Macro Average	0.305	(0.276, 0.340)	**0.731**	**(0.673, 0.795)**	0.640	(0.571, 0.715)	**0.757**	**(0.697, 0.832)**
Weighted Average	0.532	(0.495, 0.571)	**0.862**	**(0.842, 0.882)**	0.827	(0.796, 0.855)	**0.867**	**(0.847, 0.888)**

Table 9: Negation detection F1 Scores for RadPrompt on MIMIC-CXR gold-standard test set. The "Base LLM" column refers to the first-turn prediction of the LLM, and the "RadPrompt" column to the second-turn prediction. The scores correspond to the averages across 1000 bootstrap replicates and are reported alongside their confidence intervals.

	Uncertainty Detection F1							
	Gemini-1.5 Pro				**Llama-2 70B**			
Pathologies	**Base LLM**		**RadPrompt**		**Base LLM**		**RadPrompt**	
Atelectasis	0.301	(0.136, 0.464)	**0.376**	**(0.208, 0.536)**	0.364	(0.143, 0.560)	**0.386**	**(0.256, 0.515)**
Cardiomegaly	**0.385**	**(0.235, 0.529)**	0.170	(0.044, 0.320)	0.000	(0.000, 0.000)	**0.095**	**(0.000, 0.227)**
Consolidation	0.258	(0.138, 0.386)	**0.448**	**(0.250, 0.643)**	0.236	(0.133, 0.341)	**0.542**	**(0.367, 0.706)**
Edema	0.253	(0.082, 0.410)	**0.317**	**(0.087, 0.522)**	0.382	(0.154, 0.571)	**0.382**	**(0.148, 0.585)**
Enlarged Card.	0.000	(0.000, 0.000)	**0.045**	**(0.000, 0.150)**	0.000	(0.000, 0.000)	**0.068**	**(0.000, 0.229)**
Fracture	0.292	(0.000, 0.800)	**0.341**	**(0.000, 1.000)**	0.000	(0.000, 0.000)	**0.405**	**(0.000, 1.000)**
Lung Lesion	0.041	(0.000, 0.092)	**0.136**	**(0.000, 0.324)**	0.035	(0.000, 0.086)	**0.085**	**(0.000, 0.276)**
Lung Opacity	0.000	(0.000, 0.000)	**0.000**	**(0.000, 0.000)**	0.000	(0.000, 0.000)	0.000	(0.000, 0.000)
Pleural Effusion	**0.483**	**(0.333, 0.619)**	0.466	(0.296, 0.606)	0.488	(0.308, 0.654)	**0.434**	**(0.276, 0.571)**
Pleural Other	0.000	(0.000, 0.000)	0.000	(0.000, 0.000)	0.000	(0.000, 0.000)	0.000	(0.000, 0.000)
Pneumonia	0.705	(0.621, 0.776)	**0.710**	**(0.624, 0.781)**	**0.704**	**(0.614, 0.788)**	0.592	(0.497, 0.678)
Pneumothorax	**0.652**	**(0.353, 0.870)**	0.568	(0.222, 0.834)	0.475	(0.000, 0.800)	**0.566**	**(0.250, 0.846)**
Support Devices	0.000	(0.000, 0.000)	0.000	(0.000, 0.000)	0.000	(0.000, 0.000)	0.000	(0.000, 0.000)
Macro Average	0.393	(0.332, 0.462)	0.393	(0.322, 0.473)	0.394	(0.316, 0.476)	**0.407**	**(0.329, 0.496)**
Weighted Average	0.498	(0.432, 0.560)	**0.512**	**(0.445, 0.577)**	**0.513**	**(0.439, 0.584)**	0.478	(0.414, 0.548)
	Claude-3 Sonnet				**GPT-4 Turbo**			
Pathologies	**Base LLM**		**RadPrompt**		**Base LLM**		**RadPrompt**	
Atelectasis	0.310	(0.095, 0.522)	**0.398**	**(0.254, 0.530)**	**0.406**	**(0.200, 0.583)**	0.337	(0.182, 0.491)
Cardiomegaly	**0.476**	**(0.300, 0.634)**	0.096	(0.000, 0.227)	**0.474**	**(0.293, 0.644)**	0.248	(0.074, 0.429)
Consolidation	0.454	(0.286, 0.607)	**0.634**	**(0.444, 0.783)**	**0.651**	**(0.455, 0.815)**	0.648	(0.461, 0.810)
Edema	0.202	(0.048, 0.344)	**0.395**	**(0.154, 0.606)**	**0.498**	**(0.222, 0.714)**	0.496	(0.222, 0.706)
Enlarged Card.	0.000	(0.000, 0.000)	**0.062**	**(0.000, 0.207)**	0.000	(0.000, 0.000)	0.000	(0.000, 0.000)
Fracture	**0.722**	**(0.000, 1.000)**	0.498	(0.000, 1.000)	0.342	(0.000, 1.000)	**0.498**	**(0.000, 1.000)**
Lung Lesion	**0.275**	**(0.062, 0.500)**	0.120	(0.000, 0.375)	**0.287**	**(0.000, 0.526)**	0.186	(0.000, 0.545)
Lung Opacity	0.000	(0.000, 0.000)	0.000	(0.000, 0.000)	0.000	(0.000, 0.000)	0.000	(0.000, 0.000)
Pleural Effusion	**0.583**	**(0.390, 0.735)**	0.490	(0.318, 0.633)	**0.469**	**(0.292, 0.625)**	0.464	(0.300, 0.615)
Pleural Other	0.000	(0.000, 0.000)	0.000	(0.000, 0.000)	0.000	(0.000, 0.000)	0.000	(0.000, 0.000)
Pneumonia	**0.688**	**(0.595, 0.765)**	0.687	(0.598, 0.761)	0.683	(0.599, 0.758)	**0.708**	**(0.626, 0.780)**
Pneumothorax	**0.652**	**(0.363, 0.880)**	0.645	(0.307, 0.909)	**0.651**	**(0.308, 0.889)**	0.645	(0.307, 0.909)
Support Devices	0.000	(0.000, 0.000)	0.000	(0.000, 0.000)	0.000	(0.000, 0.000)	**0.000**	**(0.000, 0.000)**
Macro Average	**0.493**	**(0.418, 0.567)**	0.460	(0.374, 0.569)	**0.521**	**(0.448, 0.599)**	0.511	(0.425, 0.592)
Weighted Average	**0.546**	**(0.484, 0.606)**	0.543	(0.469, 0.618)	**0.579**	**(0.516, 0.639)**	0.564	(0.497, 0.627)

Table 10: Uncertainty detection F1 Scores for RadPrompt on MIMIC-CXR gold-standard test set. The “Base LLM” column refers to the first-turn prediction of the LLM, and the “RadPrompt” column to the second-turn prediction. The scores correspond to the averages across 1000 bootstrap replicates and are reported alongside their confidence intervals.

	Positive Mention Detection F1			
	Gemini-1.5 Pro		**Llama-2 70B**	
Pathologies	**Base LLM**	**RadPrompt**	**Base LLM**	**RadPrompt**
Atelectasis	**0.900** **(0.869, 0.928)**	0.878 (0.843, 0.910)	**0.899** **(0.866, 0.927)**	0.819 (0.775, 0.859)
Cardiomegaly	**0.928** **(0.895, 0.957)**	0.879 (0.840, 0.914)	0.869 (0.828, 0.905)	**0.850** **(0.805, 0.892)**
Consolidation	**0.811** **(0.729, 0.881)**	0.810 (0.733, 0.883)	0.469 (0.382, 0.550)	**0.597** **(0.496, 0.689)**
Edema	**0.824** **(0.775, 0.867)**	0.822 (0.773, 0.867)	**0.866** **(0.822, 0.906)**	0.816 (0.767, 0.866)
Enlarged Card.	0.327 (0.185, 0.472)	**0.468** **(0.347, 0.583)**	0.336 (0.242, 0.437)	**0.446** **(0.336, 0.557)**
Fracture	0.851 (0.762, 0.921)	**0.868** **(0.787, 0.939)**	0.883 (0.800, 0.950)	**0.902** **(0.831, 0.964)**
Lung Lesion	0.495 (0.409, 0.578)	**0.788** **(0.701, 0.865)**	0.485 (0.396, 0.568)	**0.573** **(0.479, 0.662)**
Lung Opacity	0.638 (0.584, 0.688)	**0.786** **(0.743, 0.827)**	0.684 (0.636, 0.731)	**0.811** **(0.769, 0.852)**
Pleural Effusion	0.917 (0.891, 0.939)	**0.918** **(0.893, 0.941)**	**0.909** **(0.883, 0.933)**	0.894 (0.864, 0.920)
Pleural Other	0.438 (0.312, 0.561)	**0.532** **(0.387, 0.659)**	0.335 (0.136, 0.526)	**0.569** **(0.393, 0.704)**
Pneumonia	0.723 (0.634, 0.806)	**0.744** **(0.649, 0.821)**	**0.856** **(0.792, 0.912)**	0.493 (0.398, 0.584)
Pneumothorax	**0.880** **(0.792, 0.950)**	0.817 (0.704, 0.906)	**0.872** **(0.781, 0.945)**	0.754 (0.625, 0.864)
Support Devices	0.887 (0.857, 0.915)	**0.920** **(0.894, 0.945)**	0.733 (0.683, 0.780)	**0.896** **(0.865, 0.922)**
Macro Average	0.740 (0.719, 0.762)	**0.787** **(0.763, 0.809)**	0.708 (0.683, 0.731)	**0.725** **(0.701, 0.749)**
Weighted Average	0.814 (0.800, 0.827)	**0.845** **(0.831, 0.858)**	0.785 (0.770, 0.800)	**0.799** **(0.784, 0.814)**
	Claude-3 Sonnet		**GPT-4 Turbo**	
Pathologies	**Base LLM**	**RadPrompt**	**Base LLM**	**RadPrompt**
Atelectasis	**0.882** **(0.849, 0.913)**	0.850 (0.810, 0.887)	**0.909** **(0.876, 0.936)**	0.871 (0.833, 0.903)
Cardiomegaly	**0.937** **(0.908, 0.964)**	0.870 (0.828, 0.907)	**0.915** **(0.881, 0.947)**	0.887 (0.845, 0.923)
Consolidation	**0.811** **(0.724, 0.880)**	0.808 (0.720, 0.882)	**0.868** **(0.793, 0.930)**	0.844 (0.760, 0.914)
Edema	**0.841** **(0.790, 0.885)**	0.813 (0.762, 0.861)	**0.839** **(0.793, 0.883)**	0.816 (0.765, 0.864)
Enlarged Card.	0.389 (0.271, 0.504)	**0.493** **(0.379, 0.603)**	**0.539** **(0.415, 0.652)**	0.492 (0.378, 0.602)
Fracture	0.864 (0.775, 0.938)	**0.881** **(0.805, 0.946)**	0.829 (0.742, 0.908)	**0.837** **(0.750, 0.913)**
Lung Lesion	0.796 (0.716, 0.871)	**0.806** **(0.727, 0.877)**	0.710 (0.621, 0.789)	**0.762** **(0.673, 0.841)**
Lung Opacity	0.700 (0.644, 0.753)	**0.817** **(0.773, 0.858)**	0.769 (0.717, 0.817)	**0.820** **(0.775, 0.859)**
Pleural Effusion	**0.926** **(0.902, 0.947)**	0.917 (0.892, 0.941)	**0.920** **(0.896, 0.943)**	0.910 (0.885, 0.934)
Pleural Other	0.434 (0.309, 0.557)	**0.592** **(0.441, 0.727)**	0.571 (0.419, 0.693)	**0.624** **(0.480, 0.747)**
Pneumonia	0.706 (0.607, 0.791)	**0.713** **(0.611, 0.800)**	0.698 (0.595, 0.786)	**0.720** **(0.621, 0.803)**
Pneumothorax	**0.895** **(0.812, 0.963)**	0.858 (0.767, 0.938)	**0.891** **(0.800, 0.958)**	0.867 (0.766, 0.949)
Support Devices	0.901 (0.873, 0.927)	**0.910** **(0.882, 0.936)**	0.898 (0.870, 0.924)	**0.904** **(0.876, 0.928)**
Macro Average	0.776 (0.755, 0.795)	**0.795** **(0.772, 0.816)**	**0.797** **(0.773, 0.817)**	0.796 (0.773, 0.816)
Weighted Average	0.837 (0.823, 0.850)	**0.843** **(0.828, 0.857)**	**0.849** **(0.834, 0.863)**	0.846 (0.831, 0.860)

Table 11: Positive mention detection F1 Scores for RadPrompt on MIMIC-CXR gold-standard test set. The “Base LLM” column refers to the first-turn prediction of the LLM, and the “RadPrompt” column to the second-turn prediction. The scores correspond to the averages across 1000 bootstrap replicates and are reported alongside their confidence intervals.

Pathologies	Weighted F1		Improvement over		Improvement over	
	Llama-2 RadPrompt		1st Turn Llama-2 (%)		RadPert (%)	
Atelectasis	0.830	(0.767, 0.888)	37.9	(22.9, 56.8)	-7.1	(-11.6, -3.0)
Cardiomegaly	0.810	(0.747, 0.867)	41.3	(28.6, 57.9)	-11.0	(-16.6, -6.0)
Consolidation	0.929	(0.903, 0.953)	27.7	(21.7, 34.8)	-2.3	(-4.2, -0.7)
Edema	0.529	(0.381, 0.639)	41.5	(-4.1, 99.8)	-15.1	(-27.3, -0.8)
Enlarged Card.	0.844	(0.790, 0.894)	Inf.	(Inf., Inf.)	-7.0	(-10.6, -3.3)
Fracture	0.684	(0.531, 0.817)	12.6	(-5.8, 38.8)	-10.3	(-20.0, -0.6)
Lung Lesion	0.699	(0.577, 0.817)	191.8	(132.3, 268.1)	-14.3	(-23.3, -5.7)
Lung Opacity	0.692	(0.636, 0.748)	2.9	(-6.5, 13.4)	-2.8	(-5.7, -0.1)
Pleur. Effusion	0.615	(0.562, 0.665)	-22.6	(-29.5, -16.1)	-3.9	(-7.6, -0.9)
Pleur. Other	0.106	(0.059, 0.163)	-80.9	(-89.5, -70.8)	34.2	(-8.5, 104.7)
Pneumonia	0.519	(0.374, 0.654)	259.1	(144.6, 433.0)	-21.0	(-33.4, -10.7)
Pneumothorax	0.606	(0.550, 0.661)	-16.0	(-23.4, -8.2)	-3.3	(-6.0, -0.2)
Sup. Devices	0.822	(0.785, 0.857)	6.2	(-0.3, 13.5)	-4.2	(-6.3, -2.3)
Macro Avg.	0.668	(0.639, 0.694)	27.4	(21.9, 32.6)	-8.0	(-10.2, -5.7)
Weighted Avg.	0.695	(0.668, 0.748)	3.0	(-1.3, 10.4)	-11.7	(-13.3, -5.3)

Table 12: Weighted average F1 scores for Llama-2-based RadPrompt on the CUH test set, alongside improvements over 1st turn Llama-2 and RadPert predictions. The scores correspond to the averages across 1000 bootstrap replicates and are reported alongside their confidence intervals.

	Sub-Task F1 Scores			
	Negation Detection		**Uncertainty Detection**	
Pathologies	**Base Llama-2**	**RadPrompt**	**Base Llama-2**	**RadPrompt**
Atelectasis	0.175 (0.111, 0.238)	**0.853 (0.766, 0.923)**	0.000 (0.000, 0.000)	**0.126 (0.000, 0.400)**
Cardiomegaly	0.579 (0.515, 0.639)	**0.835 (0.779, 0.884)**	0.000 (0.000, 0.000)	**0.412 (0.000, 0.727)**
Consolidation	0.665 (0.608, 0.720)	**0.923 (0.887, 0.953)**	0.145 (0.000, 0.298)	**0.490 (0.154, 0.769)**
Edema	0.160 (0.102, 0.223)	**0.444 (0.278, 0.597)**	0.322 (0.000, 1.000)	**0.408 (0.000, 0.800)**
Enlarged Card.	0.000 (0.000, 0.000)	**0.904 (0.854, 0.947)**	0.000 (0.000, 0.000)	**0.639 (0.471, 0.791)**
Fracture	0.052 (0.018, 0.098)	**0.269 (0.071, 0.483)**	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)
Lung Lesion	0.285 (0.215, 0.359)	**0.790 (0.684, 0.884)**	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)
Lung Opacity	0.022 (0.000, 0.056)	**0.187 (0.000, 0.421)**	**0.228 (0.000, 0.667)**	0.000 (0.000, 0.000)
Pleural Effusion	**0.758 (0.717, 0.797)**	0.532 (0.468, 0.592)	**0.414 (0.000, 0.800)**	0.319 (0.000, 0.600)
Pleural Other	**0.556 (0.490, 0.615)**	0.035 (0.000, 0.077)	0.000 (0.000, 0.000)	**0.515 (0.000, 1.000)**
Pneumonia	0.113 (0.063, 0.171)	**0.642 (0.424, 0.813)**	0.128 (0.028, 0.237)	**0.310 (0.087, 0.522)**
Pneumothorax	**0.730 (0.683, 0.771)**	0.610 (0.551, 0.663)	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)
Support Devices	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)
Macro Average	0.377 (0.356, 0.413)	**0.596 (0.550, 0.664)**	0.296 (0.149, 0.441)	**0.459 (0.335, 0.585)**
Weighted Average	0.617 (0.588, 0.643)	**0.607 (0.568, 0.705)**	0.263 (0.127, 0.413)	**0.506 (0.387, 0.616)**
	Positive Mention Detection		**Mention Detection**	
Pathologies	**Base Llama-2**	**RadPrompt**	**Base Llama-2**	**RadPrompt**
Atelectasis	**0.889 (0.826, 0.938)**	0.843 (0.779, 0.902)	0.454 (0.394, 0.510)	**0.868 (0.822, 0.908)**
Cardiomegaly	0.885 (0.750, 0.976)	**0.888 (0.765, 0.977)**	0.625 (0.567, 0.679)	**0.851 (0.805, 0.895)**
Consolidation	0.806 (0.761, 0.851)	**0.950 (0.924, 0.975)**	0.737 (0.704, 0.771)	**0.966 (0.951, 0.980)**
Edema	**0.828 (0.631, 0.960)**	0.697 (0.400, 0.917)	0.230 (0.167, 0.295)	**0.546 (0.405, 0.659)**
Enlarged Card.	0.000 (0.000, 0.000)	0.000 (0.000, 0.000)	0.026 (0.000, 0.065)	**0.876 (0.829, 0.919)**
Fracture	0.843 (0.711, 0.947)	**0.847 (0.722, 0.955)**	0.196 (0.136, 0.257)	**0.637 (0.500, 0.761)**
Lung Lesion	0.094 (0.021, 0.172)	**0.434 (0.000, 0.750)**	0.204 (0.154, 0.257)	**0.742 (0.635, 0.837)**
Lung Opacity	0.701 (0.655, 0.746)	**0.716 (0.659, 0.772)**	0.523 (0.476, 0.566)	**0.696 (0.638, 0.755)**
Pleural Effusion	**0.916 (0.875, 0.953)**	0.848 (0.789, 0.901)	**0.831 (0.801, 0.859)**	0.711 (0.665, 0.752)
Pleural Other	0.687 (0.451, 0.875)	**0.836 (0.640, 0.968)**	**0.577 (0.517, 0.635)**	0.162 (0.096, 0.239)
Pneumonia	0.187 (0.077, 0.298)	**0.479 (0.240, 0.684)**	0.147 (0.099, 0.193)	**0.580 (0.454, 0.693)**
Pneumothorax	**0.618 (0.333, 0.833)**	0.607 (0.286, 0.857)	**0.734 (0.691, 0.777)**	0.625 (0.568, 0.678)
Support Devices	0.780 (0.736, 0.819)	**0.828 (0.792, 0.863)**	0.646 (0.602, 0.688)	**0.818 (0.782, 0.852)**
Macro Average	0.687 (0.647, 0.723)	**0.752 (0.696, 0.798)**	0.461 (0.443, 0.499)	**0.698 (0.671, 0.723)**
Weighted Average	0.781 (0.763, 0.801)	**0.822 (0.799, 0.842)**	0.612 (0.590, 0.652)	**0.726 (0.704, 0.748)**

Table 13: F1 Scores for all sub-tasks for Llama-2-based RadPrompt on the CUH dataset. The “Base Llama-2” column refers to the first-turn prediction of the LLM, and the “RadPrompt” column to the second-turn prediction. The scores correspond to the averages across 1000 bootstrap replicates and are reported alongside their confidence intervals.

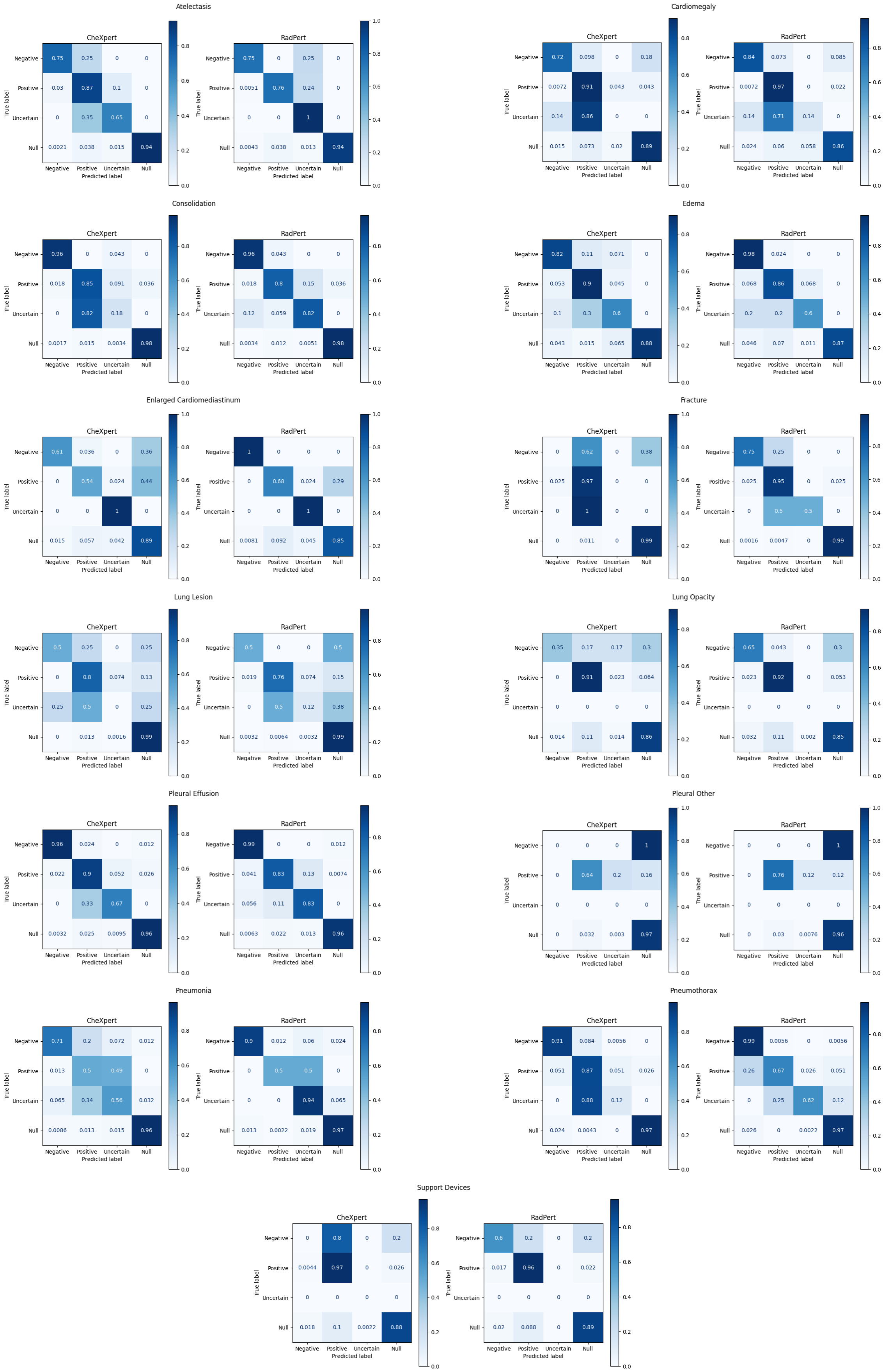
First Turn Prompt	**Second Turn Prompt**

Please accurately classify radiology reports for the presence or absence of findings. For each report, you will classify for the presence or absence of the following findings: Enlarged Cardiomediastinum, Cardiomegaly, ...
 	
Structure your answer like the template I provided to you delimited by triple backticks and return this template and nothing else.
 	
ALWAYS RETURN THE FULL TEMPLATE:
```
 {“Enlarged Cardiomediastinum”:
[ANSWER],
“Cardiomegaly”:
[ANSWER], ...
}
```
 	
If the existence of a finding is mentioned, answer “Yes”.
If a finding is mentioned as not existing, answer “No”.
If it cannot be determined if the patient has the findings, answer “Maybe”.
If a finding is not mentioned in the report, answer ‘Undefined”.
 	
Important steps to consider:
1. Read the radiology report and identify any mentions of Enlarged Cardiomediastinum, Cardiomegaly, ...
2. For every mention, determine if it is a positive, a negative, or an uncertain one.
3. If a finding is not mentioned in the report, answer “Undefined”.
4. For every finding, answer “Yes” if it is mentioned as existing (positive), “Maybe” if it is mentioned as uncertain, and “No” if it is mentioned as not existing (negative).
 	
Classify the following radiology report according to the template. Always output the full template, even if a finding is not mentioned.
 	
<START OF REPORT>
...
<END OF REPORT>
<ANSWER:>
	I am using a rule-based expert model to verify your answer. Here are some insights. However, those suggestions may be wrong. Please give me your new answer after either accepting or rejecting some or all of these suggestions:
	
	1. The tool agrees that the overall report should be classified as “Yes” for Pneumonia.
	2. In agreement with your previous answer, the tool detected no mentions of Enlarged Cardiomediastinum, Cardiomegaly,...
	3. The tool did not detect any explicit mentions for Lung Lesion and, thus, its suggested output is “Undefined” for Lung Lesion.
	4. The tool considers Atelectasis as “Maybe” because of the sentence “...”. However, you previously classified the overall report as “Yes” for Atelectasis.
	
	Please use the same template for your revised answer:
	```
	 {“Enlarged Cardiomediastinum”:
	[ANSWER],
	“Cardiomegaly”:
	[ANSWER], ...
	}
	```

Table 14: Example of RadPrompt first and second-turn prompts. The first-turn prompts are adapted from (Dorfner et al., 2024).

	MIMIC-CXR Gold-Standard				**CUH**			
Pathologies	**Null**	**Negative**	**Uncertain**	**Positive**	**Null**	**Negative**	**Uncertain**	**Positive**
Atelectasis	469	4	17	197	538	41	3	68
Cardiomegaly	452	82	14	139	523	100	10	17
Consolidation	592	23	17	55	355	138	6	151
Edema	460	85	10	132	614	23	2	11
Enlarged Card.	617	28	1	41	536	90	23	1
Fracture	637	8	2	40	623	8	0	19
Lung Lesion	621	4	8	54	607	34	2	7
Lung Opacity	493	23	0	171	471	7	1	171
Pleural Effusion	317	82	18	270	311	240	6	93
Pleural Other	660	2	0	25	476	158	2	14
Pneumonia	464	83	62	78	617	14	8	11
Pneumothorax	461	179	8	39	403	237	2	8
Support Devices	453	5	0	229	369	1	1	279
Total	6696	608	157	1470	6443	1091	66	850

Table 15: Number of output classes per pathology for the MIMIC-CXR gold-standard test set and CUH dataset.



Atelectasis

Cardiomegaly

Consolidation

Edema

Enlarged Cardiomediastinum

Fracture

Lung Lesion

Lung Opacity

Pleural Effusion

Pleural Other

Pneumonia

Pneumothorax

Support Devices

Figure 3: Normalized confusion matrices for MIMIC-CXR gold-standard test set.

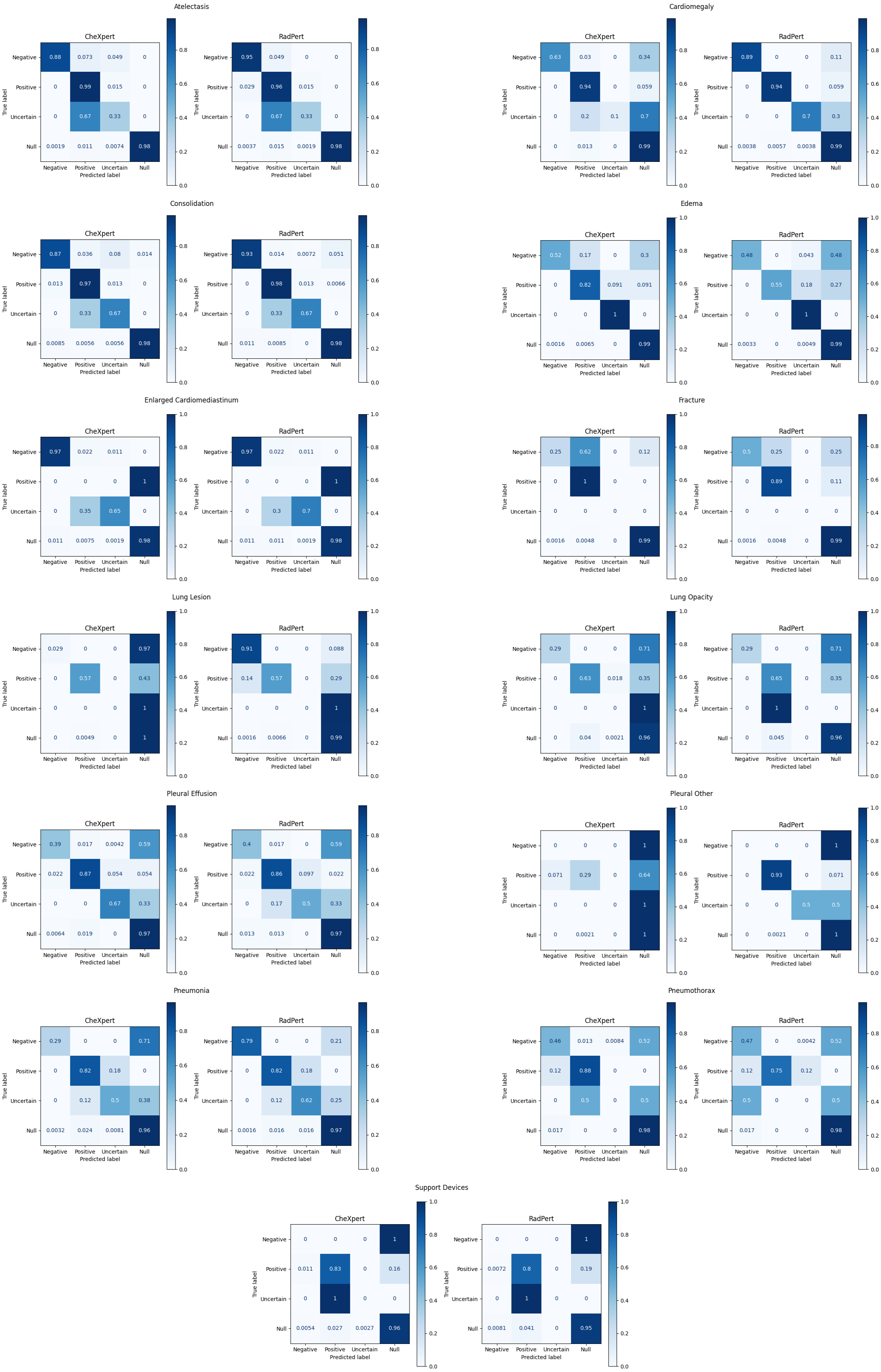


Figure 4: Normalized confusion matrices for CUH test set.