Robustness of Fine-Tuned CXR Foundation Models to Adversarial Perturbations

Assignee Research

June 13, 2026

Abstract

Radiologists play a crucial role in translating medical images into actionable reports. However, the field faces staffing shortages and increasing workloads. While automated approaches using vision-language models (VLMs) show promise as assistants, they require exceptionally high accuracy. Most current VLMs in radiology rely solely on supervised fine-tuning. Meanwhile, additional preference fine-tuning in the post-training pipeline has become standard practice in the general domain. The challenge in radiology lies in the prohibitive cost of obtaining radiologist feedback at scale. To address t

1 Introduction

This paper examines: CheXalign: Preference fine-tuning in chest X-ray interpretation models without human feedback. Research question: What is the effect of fine-tuning CXR foundation models with human feedback on their robustness to adversarial perturbations, measured by comparing model confidence and accuracy on perturbed and original MIMIC-CXR/NIH-CXR14 test sets?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.2/10.

3 Results

15 papers retrieved. 27 claims extracted; 19 independently verified. Quality review score: 7.2/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Expert human feedback from radiologists is the gold standard for preference data generation in Radiology Report Generati	✓	0.20
Large-scale annotation tasks by radiologists are impractical or unfeasible due to limited availability.	×	0.13
In the general domain, Large Language Models (LLMs) are commonly leveraged for cost-effective preference data generation	×	0.14
Zheng et al. (2023) categorized 'LLM-as-a-Judge' evaluation methods into pairwise, single answer, and reference-guided g	✓	0.29
Pairwise grading is the most common method in the general domain for both preference data generation and evaluation.	✓	0.23
Existing LLM-as-a-Judge methods are tailored for uni-modal, general-domain LLMs and do not directly apply to multi-modal	✓	0.24
Factual grounding is essential in Radiology Report Generation (RRG) to ensure clinical reliability.	×	0.11
Publicly available datasets exist that contain paired prompts (including images) and radiologist-written reference repor	✓	0.25
Reference-based grading allows for factually grounded annotations without the need for a multi-modal 'Judge' metric.	✓	0.22
Preference pairs for a given Judge can be obtained by repeat sampling from the Supervised Fine-Tuning (SFT) baseline.	×	0.14
Canonical alignment algorithms such as Direct Preference Optimization (DPO) can be used to preference fine-tune models.	×	0.14
Human (radiologist) feedback is multi-modal in nature.	✓	0.20
Obtaining human radiologist feedback at scale is prohibitively expensive.	×	0.11
Using LLM or metrics-based Judges is highly scalable.	✓	0.20
Obtaining a high-quality multi-modal Judge is difficult.	✓	0.21
Standard NLG metrics such as BLEU, ROUGE, and BERTScore may not differentiate between subtle nuances that are clinically	✓	0.20
GREEN is a state-of-the-art metric for radiology report evaluation based on a single-answer reference-guided LLM-as-a-Ju	✓	0.26
CheXbert scores are clinical efficacy metrics based on extracting 14 labels using the CheXbert labeler	✓	0.22

References

- http://arxiv.org/abs/2311.11096v1
- http://arxiv.org/abs/2410.07025v3
- http://arxiv.org/abs/2212.08228v2