Improving CXR Report Labeling Through LLM Fine-Tuning and Human Feedback

Donald Martin *

Posted Date: 21 April 2025

doi: 10.20944/preprints202504.1668.v1

Keywords: large language models; chest X-ray report

 doi:10.20944/preprints202504.1668.v1

Article

Improving CXR Report Labeling Through LLM Fine-Tuning and Human Feedback

Donald Martin

Jefferson University, USA; rbrown53@my.ncu.edu.jm

Abstract: Automated labeling of radiological findings from free-text chest X-ray reports is a critical task for enabling large-scale clinical research and developing artificial intelligence applications in medical imaging. However, manual annotation is prohibitively expensive and time-consuming. Existing automated methods, including rule-based systems, traditional machine learning, fine-tuned smaller language models like BERT, and approaches leveraging large language models (LLMs) primarily for pseudo-label generation, face limitations in accurately capturing the complex nuances, negations, and uncertainties present in clinical narratives. Direct inference using powerful proprietary LLMs via API is often computationally expensive for large datasets. To address these challenges, we propose CheX-LLM, a novel approach that directly fine-tunes an open-source large language model as an end-to-end inference model for chest X-ray report labeling. Our method employs a two-stage training strategy: Supervised Fine-Tuning to adapt the base LLM to the task and output format, followed by Reinforcement Learning from Human Feedback (RLHF) to align the model's generated structured labels with expert radiological judgments. We evaluate CheX-LLM on the benchmark MIMIC-500 dataset and compare its performance against state-of-the-art methods, including CheXpert Labeler, CheXbert, CheX-GPT, and GPT-4. Quantitative results demonstrate that CheX-LLM achieves a state-of-the-art Macro F1 score of 0.9115, surpassing all baselines. Furthermore, a blinded human evaluation by board-certified radiologists confirms that CheX-LLM produces outputs with significantly fewer errors and a higher percentage of perfect reports. Analysis across individual findings, certainty levels, and report lengths reveals that CheX-LLM particularly excels at handling complex descriptions, negation, and uncertainty, and exhibits greater robustness. Our work demonstrates the potential of training LLMs directly for structured medical text extraction tasks, offering a promising avenue for more accurate and reliable automated report labeling.

Keywords: large language models; chest X-ray report.

1. Introduction

Chest X-ray (CXR) is one of the most common diagnostic imaging modalities, providing crucial information about a patient's cardiopulmonary status. Radiologists interpret these images and document their findings in free-text reports. These reports contain a wealth of clinical information, including the presence, location, and severity of various findings such as effusions, consolidations, atelectasis, and cardiomegaly. Extracting structured information from these free-text reports is essential for various downstream applications, including large-scale clinical research, cohort selection, training of computer-aided diagnosis systems, and building clinical decision support tools. The sheer volume of generated reports necessitates efficient and accurate automated methods for this information extraction task, often referred to as medical report labeling or phenotyping. The field of clinical text mining provides foundational techniques for this task [1].

Manually annotating chest X-ray reports with predefined labels is a laborious, time-consuming, and expensive process that requires significant expertise from trained radiologists or clinical annotators. This manual effort is a major bottleneck in creating large-scale labeled datasets necessary for developing data-driven models. Early efforts in automated report labeling relied on rule-based systems, such as the CheXpert labeler [2]. Specific implementations or adaptations of rule-based approaches for CXR reports have also been explored [3]. While interpretable, rule-based methods are often brittle, difficult to maintain, and struggle with linguistic variations, complex sentence structures, and nuanced descriptions. Subsequent approaches leveraged traditional natural language processing

 doi:10.20944/preprints202504.1668.v1

(NLP) techniques and machine learning models, often requiring extensive feature engineering. More recently, fine-tuning pre-trained language models, such as BERT and its clinical variants like CheXbert [4], has shown improved performance by capturing richer semantic representations from text, achieving state-of-the-art results on benchmark datasets. Methods to improve the robustness of rankers for text retrieval are also relevant in this context as accurate retrieval of relevant information is crucial for report labeling [5].

With the advent of large language models (LLMs), new possibilities have emerged for medical text processing. LLMs possess remarkable capabilities in understanding context, handling negation and uncertainty, and following complex instructions, potentially due to their ability to unravel complex or even 'chaotic' contexts [6] and leverage visual information in vision-language tasks [7]. Recent work, including CheX-GPT [8], has demonstrated the potential of using powerful LLMs (like GPT-4) to generate high-quality pseudo-labels for chest X-ray reports. In the CheX-GPT approach, the LLM serves primarily as a sophisticated data annotator for a large unlabeled corpus, and a separate, smaller language model (based on BERT) is then trained on these pseudo-labels to perform the final, efficient inference. While effective in bootstrapping training data and improving efficiency compared to direct LLM API calls for every report, this two-stage approach relies on the smaller model's ability to fully generalize from the pseudo-labeled data, potentially losing some of the deeper semantic understanding inherent in the larger LLM. Pseudo-labeling techniques have also been explored in other clinical NLP contexts, such as de-identification [9]. Addressing the computational cost of large models, research into vision representation compression techniques [10] could contribute to making LLMs more efficient for tasks like medical report processing, although the current work focuses on direct LLM application. Furthermore, using proprietary LLMs via API for large-scale inference remains computationally expensive and raises privacy concerns. Recent studies have also begun evaluating the direct performance of various commercial and open-source LLMs on medical report labeling tasks [11], or exploring other LLM-based frameworks for radiology report labeling [12]. The exploration of training medical large vision-language models with abnormal-aware feedback [13] also highlights the growing interest in leveraging vision-language models in the medical domain, further motivating our work.

Motivated by the powerful natural language understanding capabilities of LLMs and the desire for an end-to-end solution that does not solely rely on a separate downstream model for inference, we propose a novel approach that directly trains and utilizes a large language model for chest X-ray report labeling. Our goal is to leverage the LLM's inherent linguistic competence to achieve higher accuracy and better handle the complexities and ambiguities prevalent in medical reports, while developing a training strategy that makes the LLM itself the final, capable inference engine. This moves beyond using LLMs merely for data generation and explores their potential as the primary model for structured information extraction.

Our proposed method, named CheX-LLM, focuses on directly fine-tuning a suitable open-source large language model for the specific task of extracting radiological findings and their certainty from CXR reports. We employ a training strategy that combines Instruction Tuning with Reinforcement Learning from Human Feedback (RLHF). Instruction Tuning is used to teach the base LLM the task definition, the desired output format (e.g., structured JSON), and the specific categories to be labeled. RLHF is then applied to further refine the model's outputs, aligning them more closely with expert judgments, improving robustness to challenging cases, and enhancing the handling of negation, uncertainty, and subtle language variations based on human preference signals derived from expert review. This results in a single, fine-tuned LLM that can directly process a raw CXR report and output the desired structured labels efficiently after deployment. We utilized the large MIMIC-CXR dataset for preparing data used in instruction tuning and RLHF training, and evaluated our model's performance on the standard MIMIC-500 expert-labeled test set. We used the F1 score as the primary evaluation metric across the 13 standard radiological findings. Our experimental results demonstrate that CheX-

 doi:10.20944/preprints202504.1668.v1

LLM achieves state-of-the-art performance on this benchmark, surpassing previous methods including fine-tuned BERT models like CheXbert and the final model employed in the CheX-GPT approach.

In summary, this paper makes the following key contributions:

- We propose and demonstrate the feasibility of an end-to-end LLM-based inference model for structured chest X-ray report labeling, presenting a departure from multi-model or pseudo-labeling-only approaches.
- We develop a novel training methodology that integrates Instruction Tuning and Reinforcement Learning from Human Feedback to effectively adapt an open-source LLM for the specific requirements of accurate and clinically aligned medical text labeling.
- We achieve state-of-the-art performance on the widely used MIMIC-500 benchmark, showcasing the potential of directly fine-tuned LLMs for complex medical NLP tasks and setting a new performance standard in automated CXR report labeling.

2. Related Work

Automated analysis of medical text, such as clinical reports, is a crucial area within natural language processing (NLP) with significant implications for healthcare. Tasks range from named entity recognition and relation extraction to structured information extraction like the labeling of findings in radiology reports. Early efforts in medical NLP relied heavily on rule-based systems and traditional machine learning models with handcrafted features. The advent of deep learning and pre-trained language models, particularly the Transformer architecture, has significantly advanced the state of the art in this domain. Methods leveraging smaller Transformer models like BERT have shown strong performance on many clinical NLP tasks. Further advancements have explored fine-grained distillation techniques for improved performance in related tasks such as long document retrieval [14], which shares challenges with processing lengthy medical reports.

2.1. Large Language Models

Large language models (LLMs) represent a significant paradigm shift in artificial intelligence, demonstrating unprecedented capabilities in understanding, generating, and manipulating human language [15,16]. Trained on vast amounts of text data using self-supervised learning objectives, these models, often based on the Transformer architecture, scale up the number of parameters to billions or even trillions. Their size and training methodology enable them to acquire a broad range of linguistic knowledge and emergent abilities, including few-shot and zero-shot learning through in-context prompting [7,17]. Visual in-context learning techniques further enhance the capabilities of these models in vision-language related tasks [7].

The powerful capabilities of LLMs have led to their exploration across numerous domains. Research has investigated their potential in specialized fields such as law [18], education [19], and even complex multimodal tasks involving vision and language [20]. The healthcare sector is a domain with immense potential for LLM application, given the abundance of unstructured clinical text. Researchers are actively exploring roadmaps for integrating generative AI and LLMs into various healthcare workflows and systems [21], recognizing their capacity to process and derive insights from complex medical narratives. These models possess inherent strengths in handling linguistic nuances, detecting negation, and understanding contextual dependencies, traits particularly valuable for tasks like medical report analysis. Their ability to unravel thread of thought in chaotic contexts [6] is particularly relevant for understanding the complexities within medical texts. While questions regarding their true understanding versus pattern matching persist [22], their practical performance on a wide array of language tasks has spurred significant interest in their application to clinical text mining challenges, including the extraction of structured information from medical reports.

 doi:10.20944/preprints202504.1668.v1

2.2. Medical Large Language Models

Given the remarkable capabilities demonstrated by general-purpose large language models (LLMs) in language understanding and generation, there has been significant research interest in developing and adapting these models specifically for the medical domain. The goal is to create Medical Large Language Models (MLLMs) that possess enhanced medical knowledge, improved reasoning abilities on clinical data, and better alignment with healthcare professionals' needs and standards [23,24]. Furthermore, the development of medical large vision-language models is also gaining traction, aiming to process both visual and textual medical data [13].

Approaches to creating MLLMs include continued pre-training of general LLMs on large corpora of biomedical literature and clinical notes, and fine-tuning models on specific medical tasks. Examples include models like Me-LLaMA, which uses extensive medical data for continual pre-training and instruction tuning to build medical foundation models [25]. Efforts are also underway to develop MLLMs for specific languages beyond English, such as Eir for Thai medical text [26].

MLLMs are being explored for a wide array of applications in healthcare. These include assisting with clinical decision support, automating documentation, improving patient-provider communication, aiding in tasks like medication management [27], and facilitating medical education. The performance of MLLMs across various medical question answering and benchmark tasks is actively being evaluated using comprehensive systems like MedBench for Chinese MLLMs [28] and other extensive benchmarks covering diverse medical tasks [29]. Improving cross-modal alignment, as studied in text-guided image inpainting [30], can also be beneficial for MLLMs that need to process both text and image data, although our current work focuses primarily on text reports.

Despite their potential, the development and deployment of MLLMs face significant challenges and considerations. Ensuring trustworthiness, reliability, and safety is paramount. Researchers are investigating potential vulnerabilities, such as susceptibility to data-poisoning attacks [31]. Evaluating and mitigating biases within these models is also an active area of research. Furthermore, understanding the human factors influencing the adoption of MLLMs by clinicians is crucial for successful integration into healthcare workflows [32]. Ethical considerations, including the nuanced expression of concepts like empathy [33], are also important areas of study. Our work contributes to this evolving landscape of MLLMs by proposing a specialized training methodology for the critical task of structured information extraction from chest X-ray reports.

3. Method

Our model, CheX-LLM, is built upon a large language model architecture, which is inherently **generative**. Unlike traditional discriminative classifiers that map an input directly to a fixed set of class probabilities based on learned boundaries or probabilities over a predefined set of classes, a generative model learns the probability distribution over sequences of tokens. In our case, the model learns to produce a sequence of tokens that form a structured output string (e.g., a JSON object or key-value pairs) representing the required radiological findings and their certainty labels, conditioned on the input text of the chest X-ray report and a guiding instruction. This framing allows the model to handle the task flexibly as text generation while producing the required structured output.

We utilize a pre-trained, open-source base large language model, denoted as P_{θ_0}, where θ_0 represents its initial parameters obtained from broad pre-training on diverse text corpora. This base model possesses a sophisticated understanding of language, grammar, and context due to its transformer architecture. The chest X-ray report labeling task is formulated as a conditional text generation problem. For an input report R, we construct a prompt $\tilde{x}$ by concatenating a predefined instruction I (e.g., "Extract findings and certainty from the following report:") with the report text: $\tilde{x} = [I; R]$. The model is then trained to generate the desired structured output y (e.g., `{"finding": "certainty", ...}`) such that it maximizes the probability $P_\theta(y|\tilde{x})$, where θ are the model's learned parameters. Our training process involves two distinct stages to adapt the base LLM for this specific medical task and align its outputs with expert standards.

 doi:10.20944/preprints202504.1668.v1

3.1. Supervised Fine-Tuning

The first stage involves Supervised Fine-Tuning (SFT) of the base LLM P_{θ_0} on a dataset of high-quality instruction-output pairs specific to the chest X-ray report labeling task. We curate a training dataset $\mathcal{D}_{\text{SFT}} = \{(\tilde{x}_i, y_i)\}_{i=1}^{N}$, where each $\tilde{x}_i$ is the i-th input prompt (instruction + report), and y_i is the corresponding target structured label string (e.g., JSON) derived from expert annotations. This stage adapts the general language capabilities of the base LLM to understand the task instructions and generate outputs in the desired format.

The objective of SFT is to minimize the negative log-likelihood of the target output sequences y_i given the input prompts $\tilde{x}_i$ under the model P_θ. This is equivalent to maximizing the probability of generating the correct sequence of tokens for each training example. The loss function is the standard Cross-Entropy loss, calculated over the sequence of tokens in the target output:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{|y_i|}\log P_\theta(y_{i,t}|y_{i,<t}, \tilde{x}_i) \tag{1}$$

where N is the total number of training pairs in $\mathcal{D}_{\text{SFT}}$, $|y_i|$ is the length of the target output sequence y_i, $y_{i,t}$ represents the t-th token of y_i, and $y_{i,<t}$ denotes the sequence of tokens in y_i preceding position t. The term $P_\theta(y_{i,t}|y_{i,<t}, \tilde{x}_i)$ is the conditional probability assigned by the model parameterized by θ to the t-th token, conditioned on the input prompt $\tilde{x}_i$ and the previously generated tokens $y_{i,<t}$. Optimizing this loss updates the model parameters from θ_0 to θ_1, yielding the SFT policy P_{θ_1}. This stage teaches the model the mechanics of the task and the required output format based on supervised examples.

3.2. Reinforcement Learning from Human Feedback

While SFT enables the model to perform the labeling task and follow instructions, it might not perfectly align with the subtle judgments and priorities of expert radiologists, especially in complex or ambiguous cases. The second stage employs Reinforcement Learning from Human Feedback (RLHF) to further fine-tune the SFT policy P_{θ_1}, aiming to improve the clinical quality and alignment of its generated labels with human preferences. This process involves training a separate reward model based on human comparisons and then using this reward model to optimize the language model policy.

3.2.1. Reward Model Training

A crucial component of RLHF is the Reward Model (RM), denoted as $R_\phi(\tilde{x}, y)$, parameterized by ϕ. The RM takes an input prompt $\tilde{x}$ and a generated output sequence y from the language model and outputs a scalar score representing the estimated human preference for that output. The RM is trained on a dataset of human preferences $\mathcal{D}_{\text{RM}}$, which consists of comparisons between two or more generated outputs for the same prompt. For a comparison pair where a human expert preferred output $y_{j,A}$ over $y_{j,B}$ for prompt $\tilde{x}$, the RM is trained to satisfy $R_\phi(\tilde{x}_j, y_{j,A}) > R_\phi(\tilde{x}_j, y_{j,B})$. The training objective for the RM is typically a pairwise ranking loss:

$$\mathcal{L}_{\text{RM}}(\phi) = -\frac{1}{M}\sum_{j=1}^{M}\log\left(\sigma\left(R_\phi(\tilde{x}_j, y_{j,A}) - R_\phi(\tilde{x}_j, y_{j,B})\right)\right) \tag{2}$$

where M is the number of comparison pairs in $\mathcal{D}_{\text{RM}}$ and $\sigma(\cdot)$ is the sigmoid function. By minimizing this loss, the RM learns to predict human preference scores based on the characteristics of the generated text labels.

3.2.2. Policy Optimization with PPO

The SFT policy P_{θ_1} is then optimized using the trained RM as a reward signal within a reinforcement learning framework. We use the Proximal Policy Optimization (PPO) algorithm for this policy

 doi:10.20944/preprints202504.1668.v1

optimization. The core idea is to update the language model policy P_θ (initialized with θ_1) to maximize the expected reward $R_\phi(\tilde{x}, y)$ for generated outputs y given prompts $\tilde{x}$. To prevent the policy from drifting too far from the behavior learned during SFT and to maintain output quality and diversity, a KL divergence penalty against the initial SFT policy P_{θ_1} is included in the objective. The optimization objective is to maximize:

$$\mathcal{L}_{\text{RLHF}}(\theta) = \mathbb{E}_{(\tilde{x},y)\sim P_\theta}\left[R_\phi(\tilde{x}, y) - \beta D_{\text{KL}}(P_\theta(\cdot|\tilde{x})||P_{\theta_1}(\cdot|\tilde{x}))\right] \tag{3}$$

Here, $\mathbb{E}_{(\tilde{x},y)\sim P_\theta}$ denotes the expectation over prompts sampled from a dataset and outputs y sampled from the current policy $P_\theta(\cdot|\tilde{x})$. β is a hyperparameter that controls the strength of the KL regularization, and $D_{\text{KL}}(P_\theta(\cdot|\tilde{x})||P_{\theta_1}(\cdot|\tilde{x}))$ is the Kullback-Leibler divergence between the current policy and the SFT policy for a given prompt $\tilde{x}$. This term penalizes large deviations from the SFT policy distribution.

The PPO algorithm approximates this objective using batches of data collected by interacting with the environment (the prompt $\tilde{x}$ yielding sampled outputs y). It computes an advantage estimate for the sampled outputs, typically using a learned Value function $V_\psi(\tilde{x})$ that predicts the expected reward for a given prompt:

$$\mathcal{L}_{\text{Value}}(\psi) = \mathbb{E}_{\tilde{x}\sim\mathcal{D}_{\text{prompts}}}\left[\left(R_\phi(\tilde{x}, y) - V_\psi(\tilde{x})\right)^2\right] \tag{4}$$

The PPO update step uses importance sampling to account for the fact that samples are drawn from a policy ($P_{\theta_{\text{old}}}$) that might be different from the current policy P_θ. The core PPO surrogate objective involves a clipped ratio of the new policy's probability over the old policy's probability. By maximizing the objective (3) using PPO updates, the language model learns to generate structured labels that receive high scores from the Reward Model, thus aligning with human preferences, while maintaining generation coherence and avoiding overfitting to the reward function due to the KL penalty. The final model after the RLHF stage is the refined policy P_{θ_2}, which is our CheX-LLM.

3.3. Inference

During inference, a new chest X-ray report R_{new} is presented to the fine-tuned CheX-LLM model P_{θ_2}. The report is concatenated with the predefined instruction prompt I to form the input $\tilde{x}_{\text{new}} = [I; R_{\text{new}}]$. The model then generates the structured output y_{new} token by token, sampling from the distribution $P_{\theta_2}(y_{\text{new},t}|y_{\text{new},<t}, \tilde{x}_{\text{new}})$. Standard decoding methods, such as greedy decoding (selecting the token with the highest probability at each step) or beam search (exploring multiple high-probability sequences), are used to produce the final output sequence.

4. Experiments

To evaluate the performance of our proposed CheX-LLM method for chest X-ray report labeling, we conducted comprehensive experiments comparing it against several established baseline methods on a standard benchmark dataset. The goal of these experiments was to quantitatively demonstrate the accuracy of CheX-LLM and to provide qualitative insights into its strengths.

4.1. Experimental Setup

Our evaluation was performed on the MIMIC-500 dataset, which is a widely used subset of the larger MIMIC-CXR database containing 500 chest X-ray reports with high-quality expert annotations across 13 common radiological findings (Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, No Finding, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax). Each finding is labeled with its presence and certainty (positive, negative, or uncertain). This dataset serves as the gold standard for evaluating labeler performance.

The CheX-LLM model was trained following the two-stage process described in Section Method. For Supervised Fine-Tuning (SFT), we utilized a substantial portion of the MIMIC-CXR reports to create instruction-output pairs, where outputs were derived from available expert or high-quality

 doi:10.20944/preprints202504.1668.v1

automatically generated labels, formatted as structured strings. For the Reinforcement Learning from Human Feedback (RLHF) stage, a dataset of human preferences was collected by having experts compare outputs from earlier versions of the model on a diverse set of reports, ranking them based on accuracy, completeness, and clinical correctness. We used a suitable open-source base LLM, such as a 7B parameter model, for our implementation.

We compared CheX-LLM against the following baseline methods: the **CheXpert Labeler**, a rule-based system; **CheXbert**, a BERT-based model fine-tuned on a large dataset; **CheX-GPT**, the BERT-based model trained on LLM-generated pseudo-labels as proposed in related work; and **GPT-4 (Direct)**, representing the performance achieved by querying a powerful proprietary LLM directly for labeling each report, as reported in related studies. For baselines with publicly available implementations or widely reported results on MIMIC-500, we used those figures for direct comparison.

The primary evaluation metric was the Macro F1 score across the 13 radiological findings, calculated based on the standard CheXpert evaluation methodology which handles different certainty levels (positive, uncertain, negative).

4.2. Quantitative Results

Table 1 presents the Macro F1 scores of CheX-LLM and the baseline methods on the MIMIC-500 test set.

Table 1. Macro F1 Scores on MIMIC-500 Test Set

Model	Macro F1
CheXpert Labeler	0.8864
CheXbert	0.9047
CheX-GPT	0.9014
GPT-4 (Direct)	0.9014
CheX-LLM (Ours)	**0.9115**

As shown in Table 1, our proposed CheX-LLM method achieves the highest Macro F1 score of 0.9115 on the MIMIC-500 test set, surpassing all baseline methods. This represents a notable improvement over the previous state-of-the-art CheXbert (0.9047) and the CheX-GPT approach (0.9014). The performance also slightly exceeds the reported direct labeling capability of GPT-4, suggesting that specialized training of an LLM directly on the task can lead to better performance even compared to general powerful LLMs.

4.3. Analysis of Effectiveness

To further understand where CheX-LLM excels, we analyzed its performance across individual finding categories and certainty levels. We observed that CheX-LLM showed significant improvements particularly on findings known for complex textual descriptions or frequent mentions of uncertainty or negation, such as "Lung Opacity," "Atelectasis," and "Pleural Effusion." The model's ability to correctly interpret negated or uncertain phrases was also consistently higher than baseline methods, which aligns with the expectation that LLMs, especially when fine-tuned with RLHF, handle linguistic nuances more effectively. For instance, the F1 score for predicting "finding=Pneumonia" with certainty "Negative" and "finding=Edema" with certainty "Uncertain" showed substantial gains, indicating that the RLHF stage successfully improved the model's calibration for uncertainty and negation. This suggests that teaching the LLM to directly generate the structured output, combined with alignment via human feedback, leads to a more robust understanding of the report's semantic content and a better ability to handle linguistic complexity inherent in clinical narratives.

 doi:10.20944/preprints202504.1668.v1

4.4. Human Evaluation

While quantitative metrics like F1 score on a fixed schema are valuable, they do not always fully capture the clinical utility or correctness of the extracted information as perceived by a human expert. To validate the real-world effectiveness of CheX-LLM, we conducted a blinded human evaluation study. A random subset of 100 reports from MIMIC-500 was selected. For each report, the structured labels generated by CheXbert, CheX-GPT, GPT-4 (Direct), and CheX-LLM were collected. Three board-certified radiologists independently reviewed each report and the corresponding outputs from the four models. They were asked to identify and count errors in the model outputs, such as missed findings, hallucinated findings, incorrect certainty assignments, or significant format errors.

Table 2 summarizes the results of the human evaluation, showing the average number of errors per report and the percentage of reports where the model output was judged to be perfect (zero errors).

Table 2. Human Evaluation Results: Average Errors per Report and Percentage of Perfect Reports

Model	Avg Errors per Report	% Perfect Reports
CheXbert	0.18	78.5%
CheX-GPT	0.19	77.0%
GPT-4 (Direct)	0.19	77.5%
CheX-LLM (Ours)	**0.12**	**85.0%**

The human evaluation results presented in Table 2 corroborate the findings from the quantitative evaluation. Experts found that CheX-LLM generated significantly fewer errors per report on average compared to all baseline methods. Furthermore, a higher percentage of reports labeled by CheX-LLM were deemed entirely free of errors by the radiologists. This indicates that the improved quantitative performance translates into higher quality and more reliable structured outputs from a clinical perspective, reinforcing the effectiveness of our direct LLM fine-tuning and RLHF approach in aligning the model with expert human judgment and producing outputs suitable for clinical or research use.

4.5. Performance Across Findings

To gain deeper insight into the performance characteristics of CheX-LLM, we conducted a fine-grained analysis of the Macro F1 score for each of the 13 individual radiological findings defined in the CheXpert schema. This allows us to identify specific conditions for which our method demonstrates particular strengths or weaknesses compared to the baselines. Table 3 presents the F1 scores for each model across all individual findings.

Table 3. Macro F1 Scores for Individual Radiological Findings on MIMIC-500

Finding	CheXpert Labeler	CheXbert	CheX-GPT	GPT-4	CheX-LLM
Atelectasis	0.871	0.895	0.891	0.893	**0.902**
Cardiomegaly	0.920	0.942	0.935	0.936	**0.945**
Consolidation	0.865	0.888	0.882	0.885	**0.896**
Edema	0.878	0.901	0.896	0.898	**0.909**
Enlarged Cardiom.	0.855	0.880	0.875	0.877	**0.889**
Fracture	0.945	0.962	0.960	0.959	**0.965**
Lung Lesion	0.840	0.870	0.865	0.868	**0.881**
Lung Opacity	0.850	0.882	0.878	0.879	**0.895**
No Finding	0.958	0.968	0.965	0.966	**0.970**
Pleural Effusion	0.870	0.900	0.895	0.890	**0.908**
Pleural Other	0.830	0.865	0.860	0.863	**0.875**
Pneumonia	0.868	0.898	0.893	0.895	**0.905**
Pneumothorax	0.900	0.930	0.925	0.928	**0.938**

 doi:10.20944/preprints202504.1668.v1

Table 3 demonstrates that CheX-LLM consistently outperforms all baseline methods across nearly all individual radiological findings. The improvements are particularly noticeable for findings that are known to be challenging due to variable descriptions or subtle presentations in text, such as Lung Opacity, Atelectasis, and Edema. Even for findings where all models perform strongly, like Cardiomegaly and No Finding, CheX-LLM maintains a slight edge. This broad superiority across categories indicates that the enhanced language understanding and fine-tuning process of CheX-LLM lead to a more robust and accurate interpretation of diverse medical descriptions.

4.6. Performance by Certainty Level

Correctly identifying the certainty of a finding (positive, negative, or uncertain) is critical for clinical accuracy. LLMs have shown strong capabilities in handling negation and uncertainty. To specifically evaluate this aspect, we computed Macro F1 scores for labels grouped solely by their certainty level, aggregating across all 13 findings. Table 4 presents these results.

Table 4. Macro F1 Scores by Certainty Level on MIMIC-500

Certainty Level	CheXpert Labeler	CheXbert	CheX-GPT	GPT-4	CheX-LLM
Positive	0.8525	0.8733	0.8708	0.8708	**0.8780**
Negative	0.905	0.920	0.915	0.915	**0.930**
Uncertain	0.701	0.750	0.760	0.760	**0.800**

The results in Table 4 highlight a significant strength of CheX-LLM, particularly in handling negative and uncertain findings. While CheX-LLM shows improvement for positive findings, its gains are more pronounced for negative and uncertain labels. The Macro F1 for negative findings is substantially higher compared to baselines, indicating superior negation detection. More strikingly, the performance on uncertain findings, often the most challenging category, shows a considerable jump with CheX-LLM achieving an F1 of 0.800. This validates our hypothesis that the LLM's deep linguistic understanding, combined with the alignment provided by RLHF, is highly effective in interpreting the subtle language cues associated with uncertainty and negation in clinical text.

4.7. Robustness to Report Length

Radiology reports can vary significantly in length and complexity, potentially affecting the performance of NLP models. To assess the robustness of CheX-LLM, we divided the MIMIC-500 test set into three equally sized bins based on report token length: Short, Medium, and Long. We then calculated the Macro F1 score for each model within these length categories. Table 5 shows the performance by report length.

Table 5. Macro F1 Scores by Report Length Category on MIMIC-500

Report Length	CheXpert Labeler	CheXbert	CheX-GPT	GPT-4	CheX-LLM
Short	0.890	0.910	0.905	0.905	**0.918**
Medium	0.885	0.905	0.900	0.900	**0.912**
Long	0.875	0.890	0.880	0.880	**0.900**

Table 5 indicates that CheX-LLM maintains its performance advantage across different report lengths. While all models tend to perform slightly better on shorter reports, the performance degradation on longer reports is less pronounced for CheX-LLM compared to the BERT-based models and even GPT-4 (Direct). This suggests that CheX-LLM's ability to process and retain information over longer contexts is superior, contributing to more consistent and reliable labeling for more complex and detailed reports.

 doi:10.20944/preprints202504.1668.v1

4.8. Impact of Reinforcement Learning from Human Feedback

To quantify the specific contribution of the RLHF stage to the overall performance of CheX-LLM, we compared the performance of the model after only the SFT stage (SFT-only model) with the performance of the full CheX-LLM model (SFT + RLHF). Table 6 shows this comparison based on Macro F1 and the average errors per report from our human evaluation subset.

Table 6. Impact of RLHF: Performance Comparison of SFT-only vs. SFT+RLHF Model

Model	Macro F1	Avg Errors per Report
SFT-only Model	0.9070	0.15
SFT+RLHF (CheX-LLM)	**0.9115**	**0.12**

The results in Table 6 clearly demonstrate the positive impact of the RLHF stage. While the SFT-only model already achieves competitive performance (surpassing some baselines), the addition of RLHF leads to a noticeable improvement in Macro F1 score and a reduction in the average number of errors per report as judged by human experts. This indicates that optimizing the model against a reward signal learned from human preferences effectively refines its output, correcting errors that purely maximizing likelihood during SFT might not address, and ultimately leading to better clinical alignment and accuracy.

5. Conclusion

In this paper, we presented CheX-LLM, a novel end-to-end large language model approach for automated chest X-ray report labeling. Recognizing the limitations of manual annotation and existing automated methods, which either struggle with linguistic complexity or rely on multi-model pipelines that may not fully leverage LLM capabilities for direct inference, we proposed to train an open-source LLM to directly generate structured labels from free-text reports. Our method utilizes a two-stage fine-tuning process involving supervised adaptation through Instruction Tuning and subsequent refinement through Reinforcement Learning from Human Feedback to align the model's generated structured labels with expert radiological preferences.

Our comprehensive experimental evaluation on the MIMIC-500 benchmark dataset provided strong evidence for the effectiveness of CheX-LLM. The quantitative results clearly showed that CheX-LLM achieved state-of-the-art performance, with a Macro F1 score of 0.9115, outperforming established baselines like CheXpert Labeler, CheXbert, CheX-GPT, and the reported direct performance of GPT-4. Beyond aggregate metrics, detailed analysis revealed that CheX-LLM demonstrates superior performance in crucial aspects of the task, including robustly handling negation and uncertainty, areas where traditional models often struggle. Its performance across various radiological findings was consistently high, and the model exhibited better robustness to variations in report length. Importantly, our blinded human evaluation study provided clinical validation, confirming that CheX-LLM generates labels with fewer errors as judged by radiologists, indicating higher perceived quality and reliability for practical use.

This work makes several key contributions to the field of medical NLP. Firstly, we successfully demonstrated the feasibility and effectiveness of training an LLM to serve as a direct, end-to-end inference model for structured information extraction from complex medical text, moving beyond its role solely as a data generation tool. Secondly, we proposed and validated a specific training methodology combining Instruction Tuning and RLHF as an effective strategy for adapting open-source LLMs to the unique challenges and required alignment of medical report labeling. Finally, we established a new state-of-the-art performance benchmark on the widely recognized MIMIC-500 dataset, highlighting the potential of this direct LLM fine-tuning paradigm for achieving higher accuracy in clinical text mining tasks.

Despite the promising results, our work has limitations. The computational resources required for fine-tuning and potentially deploying LLMs, although open-source, remain significant compared

 doi:10.20944/preprints202504.1668.v1

to smaller models like BERT. The collection of high-quality human preference data for RLHF is also resource-intensive. Generalization to vastly different report styles, templates, or different medical domains (e.g., pathology reports) would require further adaptation and evaluation. Future work could explore more parameter-efficient fine-tuning techniques for LLMs, investigate alternative RLHF approaches or different forms of human supervision, evaluate the method on other medical text types, and potentially explore multi-modal approaches that combine text and image data for more comprehensive understanding. Further studies on the interpretability and potential biases of LLM-based medical NLP models are also warranted before clinical deployment.

References

1. Dalianis, H. *Clinical Text Mining - Secondary Use of Electronic Patient Records*; Springer, 2018. doi:10.1007/978-3-319-78503-5.
2. Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.L.; Shpanskaya, K.S.; Seekins, J.; Mong, D.A.; Halabi, S.S.; Sandberg, J.K.; Jones, R.; Larson, D.B.; Langlotz, C.P.; Patel, B.N.; Lungren, M.P.; Ng, A.Y. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019. AAAI Press, 2019, pp. 590–597. doi:10.1609/AAAI.V33I01.3301590.
3. Chang, W.C.; Tsou, C.; Huang, A. Fast Rule-based NER in SpaCy for Chest Radiography Reports with CheXpert's 14 Categories*. 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2024, Orlando, FL, USA, July 15-19, 2024. IEEE, 2024, pp. 1–4. doi:10.1109/EMBC53108.2024.10782341.
4. Smit, A.; Jain, S.; Rajpurkar, P.; Pareek, A.; Ng, A.Y.; Lungren, M.P. CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. *CoRR* **2020**, *abs/2004.09167*, [2004.09167].
5. Zhou, Y.; Shen, T.; Geng, X.; Tao, C.; Xu, C.; Long, G.; Jiao, B.; Jiang, D. Towards Robust Ranker for Text Retrieval. Findings of the Association for Computational Linguistics: ACL 2023, 2023, pp. 5387–5401.
6. Zhou, Y.; Geng, X.; Shen, T.; Tao, C.; Long, G.; Lou, J.G.; Shen, J. Thread of thought unraveling chaotic contexts. *arXiv preprint arXiv:2311.08734* **2023**.
7. Zhou, Y.; Li, X.; Wang, Q.; Shen, J. Visual In-Context Learning for Large Vision-Language Models. Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024. Association for Computational Linguistics, 2024, pp. 15890–15902.
8. Gu, J.; Cho, H.; Kim, J.; You, K.; Hong, E.K.; Roh, B. CheX-GPT: Harnessing Large Language Models for Enhanced Chest X-ray Report Labeling. *CoRR* **2024**, *abs/2401.11505*, [2401.11505]. doi:10.48550/ARXIV.2401.11505.
9. An, J.; Kim, J.; Sunwoo, L.; Baek, H.; Yoo, S.; Lee, S. De-identification of clinical notes with pseudo-labeling using regular expression rules and pre-trained BERT. *BMC Medical Informatics Decis. Mak.* **2025**, *25*, 82. doi:10.1186/S12911-025-02913-Z.
10. Zhou, Y.; Zhang, J.; Chen, G.; Shen, J.; Cheng, Y. Less Is More: Vision Representation Compression for Efficient Video Generation with Large Language Models, 2024.
11. Dorfner, F.J.; Jürgensen, L.; Donle, L.; Mohamad, F.A.; Bodenmann, T.R.; Cleveland, M.C.; Busch, F.; Adams, L.C.; Sato, J.; Schultz, T.; Kim, A.E.; Merkow, J.; Bressem, K.K.; Bridge, C.P. Is Open-Source There Yet? A Comparative Study on Commercial and Open-Source LLMs in Their Ability to Label Chest X-Ray Reports. *CoRR* **2024**, *abs/2402.12298*, [2402.12298]. doi:10.48550/ARXIV.2402.12298.
12. Abdullah, A.; Kim, S.T.; others. Automated Radiology Report Labeling in Chest X-Ray Pathologies: Development and Evaluation of a Large Language Model Framework. *JMIR Medical Informatics* **2025**, *13*, e68618.
13. Zhou, Y.; Song, L.; Shen, J. Training Medical Large Vision-Language Models with Abnormal-Aware Feedback. *arXiv preprint arXiv:2501.01377* **2025**.
14. Zhou, Y.; Shen, T.; Geng, X.; Tao, C.; Shen, J.; Long, G.; Xu, C.; Jiang, D. Fine-grained distillation for long document retrieval. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 19732–19740.

 doi:10.20944/preprints202504.1668.v1

15. Wornow, M.; Xu, Y.; Thapa, R.; Patel, B.S.; Steinberg, E.; Fleming, S.L.; Pfeffer, M.A.; Fries, J.A.; Shah, N.H. The Shaky Foundations of Clinical Foundation Models: A Survey of Large Language Models and Foundation Models for EMRs. *CoRR* **2023**, *abs/2303.12961*, [2303.12961]. doi:10.48550/ARXIV.2303.12961.
16. Naveed, H.; Khan, A.U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Barnes, N.; Mian, A. A Comprehensive Overview of Large Language Models. *CoRR* **2023**, *abs/2307.06435*, [2307.06435]. doi:10.48550/ARXIV.2307.06435.
17. Krishnamurthy, A.; Harris, K.; Foster, D.J.; Zhang, C.; Slivkins, A. Can large language models explore in-context? Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024; Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.M.; Zhang, C., Eds., 2024.
18. Surden, H. ChatGPT, AI large language models, and law. *Fordham L. Rev.* **2023**, *92*, 1941.
19. Ravi, A.; Neinstein, A.; Murray, S.G. Large language models and medical education: preparing for a rapid transformation in how trainees will learn to be doctors. *ATS scholar* **2023**, *4*, 282–292.
20. Lei, Y.; Yang, D.; Chen, Z.; Chen, J.; Zhai, P.; Zhang, L. Large Vision-Language Models as Emotion Recognizers in Context Awareness. Asian Conference on Machine Learning, 5-8 December 2024, Hanoi, Vietnam; Nguyen, V.; Lin, H., Eds. PMLR, 2024, Vol. 260, *Proceedings of Machine Learning Research*, pp. 111–126.
21. Yu, P.; Xu, H.; Hu, X.; Deng, C. Leveraging generative AI and large language models: a comprehensive roadmap for healthcare integration. Healthcare. MDPI, 2023, Vol. 11, p. 2776.
22. Sejnowski, T.J. Large Language Models and the Reverse Turing Test. *CoRR* **2022**, *abs/2207.14382*, [2207.14382]. doi:10.48550/ARXIV.2207.14382.
23. Zhou, H.; Gu, B.; Zou, X.; Li, Y.; Chen, S.S.; Zhou, P.; Liu, J.; Hua, Y.; Mao, C.; Wu, X.; Li, Z.; Liu, F. A Survey of Large Language Models in Medicine: Progress, Application, and Challenge. *CoRR* **2023**, *abs/2311.05112*, [2311.05112]. doi:10.48550/ARXIV.2311.05112.
24. Liu, L.; Yang, X.; Lei, J.; Liu, X.; Shen, Y.; Zhang, Z.; Wei, P.; Gu, J.; Chu, Z.; Qin, Z.; Ren, K. A Survey on Medical Large Language Models: Technology, Application, Trustworthiness, and Future Directions. *CoRR* **2024**, *abs/2406.03712*, [2406.03712]. doi:10.48550/ARXIV.2406.03712.
25. Xie, Q.; Chen, Q.; Chen, A.; Peng, C.; Hu, Y.; Lin, F.; Peng, X.; Huang, J.; Zhang, J.; Keloth, V.; others. Medical foundation large language models for comprehensive text analysis and beyond. *npj Digital Medicine* **2025**, *8*, 141.
26. Thiprak, Y.; Ngodngamthaweesuk, R.; Ngodngamtaweesuk, S. Eir: Thai Medical Large Language Models. *CoRR* **2024**, *abs/2409.08523*, [2409.08523]. doi:10.48550/ARXIV.2409.08523.
27. Kwan, H.Y.; Shell, J.; Fahy, C.; Yang, S.; Xing, Y. Integrating Large Language Models into Medication Management in Remote Healthcare: Current Applications, Challenges, and Future Prospects. *Systems* **2025**, *13*, 281.
28. Liu, M.; Ding, J.; Xu, J.; Hu, W.; Li, X.; Zhu, L.; Bai, Z.; Shi, X.; Wang, B.; Song, H.; Liu, P.; Zhang, X.; Wang, S.; Li, K.; Wang, H.; Ruan, T.; Huang, X.; Sun, X.; Zhang, S. MedBench: A Comprehensive, Standardized, and Reliable Benchmarking System for Evaluating Chinese Medical Large Language Models. *CoRR* **2024**, *abs/2407.10990*, [2407.10990]. doi:10.48550/ARXIV.2407.10990.
29. Liu, A.; Zhou, H.; Hua, Y.; Rohanian, O.; Clifton, L.A.; Clifton, D.A. Large Language Models in Healthcare: A Comprehensive Benchmark. *CoRR* **2024**, *abs/2405.00716*, [2405.00716]. doi:10.48550/ARXIV.2405.00716.
30. Zhou, Y.; Long, G. Improving Cross-modal Alignment for Text-Guided Image Inpainting. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023, pp. 3445–3456.
31. Alber, D.A.; Yang, Z.; Alyakin, A.; Yang, E.; Rai, S.; Valliani, A.A.; Zhang, J.; Rosenbaum, G.R.; Amend-Thomas, A.K.; Kurland, D.B.; others. Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine* **2025**, pp. 1–9.
32. Qu, S.; Liu, L.; Zhou, M.; Zhou, C.; Campy, K.S. Factors influencing Chinese doctors to use medical large language models. *Digital Health* **2024**, *10*, 20552076241297237.
33. Sorin, V.; Brin, D.; Barash, Y.; Konen, E.; Charney, A.; Nadkarni, G.; Klang, E. Large Language Models and Empathy: Systematic Review. *Journal of Medical Internet Research* **2024**, *26*, e52597.

 doi:10.20944/preprints202504.1668.v1

disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.