Zero-Shot Soft Labeling of Radiology Findings in Chest X-ray Reports Using a Large Language Model

Jungwoo Nam
Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan
smskawjddn11@gmail.com

Po-Chih Kuo
Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan
kuopc@cs.nthu.edu.tw

Abstract

Automated radiology report labeling is essential for scaling medical image analysis, yet existing frameworks like rule-based CheXpert and BERT-based CheXbert suffer from limited scalability and high reliance on expert-annotated data. To address these bottlenecks, we propose a flexible zero-shot labeling approach utilizing the Llama-3-8b model. By employing generative prompting, our method classifies 14 major thoracic diseases without explicit task-specific training. Furthermore, we introduce a soft-labeling paradigm that captures clinical ambiguity through continuous probability outputs (0 to 1), moving beyond rigid binary hard-labeling. Experimental results demonstrate that our zero-shot labeling approach maintains competitive classification performance while offering superior flexibility for sub-class expansion. Notably, downstream image classification models trained on our soft-labeled datasets exhibit improved generalization and precision, suggesting that preserving clinical uncertainty is critical for robust medical AI development.

Keywords: Chest X-ray, Radiology Report Labeling, Large Language Models, Zero-Shot Learning, Soft-Labeling

1 Introduction

Chest X-ray (CXR) imaging combined with radiology reports is critical for diagnosing thoracic diseases. While recent deep learning architectures (e.g., ResNet, DenseNet, ViT (Huang et al., 2017; Dosovitskiy, 2020; He et al., 2016)) have significantly improved CXR image classification (Manzari et al., 2023), training these models requires vast amounts of high-quality annotated data (Quigley et al., 2022).

Expert manual annotation is prohibitively expensive and time-consuming (Sivarajkumar and Wang, 2022). To scale text-to-label conversion, automated extractors like CheXpert (rule-based) (Irvin et al., 2019) and CheXbert (fine-tuned BERT) (Smit et al., 2020) were developed. However, these frameworks lack scalability adding new clinical categories or subclasses requires rebuilding rules or labor-intensive re-annotation, limiting their utility in clinical data expansions (Horng et al., 2021).

To overcome this rigid bottleneck, Large Language Models (LLMs) paired with Zero-Shot Learning (ZSL) (Annadani and Biswas, 2018) offer a flexible alternative without expert labels. Furthermore, traditional approaches rely on rigid hard-labeling (e.g., mapping uncertain terms into binary values) (Irvin et al., 2019; Smit et al., 2020), which distorts nuance. We propose a soft-labeling approach where an LLM continuous probability outputs (0 to 1) preserve clinical ambiguity and improve vision model generalization (Müller et al., 2019).

In this study, we leverage an open-source Llama-3-8b(Touvron et al., 2023) to build a prompt labeling approach for CXR reports (Johnson et al., 2019a). Our contributions are twofold: (1) we compare the classification performance of 14 major diseases against existing tools and evaluate scalability for limited sub-classes in radiology reports, and (2) we conduct a comparative analysis between traditional hard-labeling and our soft-labeling approach to demonstrate how preserving continuous clinical metrics enhances downstream vision model generalization.

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 293–301
November 6-7, 2026 ©2026 Association for Computational Linguistics

2 Related Work

2.1 Automated Radiology Report Labeling and LLMs

Automated radiology report labeling tools are critical for enhancing medical image analysis efficiency while minimizing time and resource constraints (Monshi et al., 2020). Traditional approaches include rule-based methods like CheXpert (Irvin et al., 2019; Peng et al., 2018), which lack flexibility for diverse expressions in free-text reports, and fine-tuned architectures like CheXbert (Smit et al., 2020), which require labor-intensive expert annotations. In contrast, zero-shot learning addresses these data scarcity bottlenecks by performing clinical tasks without explicit training (Annadani and Biswas, 2018). Specifically, Meta's open-source Llama-3-8b (Meta, 2024) serves as a powerful framework to flexibly adapt to subclass expansions without additional labeling costs, leveraging its vastly expanded training dataset and an 8,192-token context window to effectively process complex clinical contexts.

2.2 Uncertainty Handling and Soft-labeling Methods

Uncertainty in radiology reports naturally reflects clinical variations and is vital for accurate diagnosis (Johnson et al., 2019a,b). However, traditional tools rely on hard-labeling (Irvin et al., 2019; Smit et al., 2020), compressing findings into discrete categories like Positive, Negative, Uncertain, and Unmentioned (Irvin et al., 2019). Standard binary mapping approaches simplify training but distort clinical nuances, degrading downstream performance. To mitigate this, soft labeling outputs continuous values between 0 and 1 to preserve rich contextual information. Prior CXR classification research demonstrated that replacing hard-labeling with soft-labeling combined with Label Smoothing Regularization (LSR) significantly enhanced prediction performance, outperforming traditional binary mapping baselines by a clear margin. While existing studies rely on artificial numeric adjustments, our approach leverages LLM prompts to extract continuous probability metrics directly from textual clinical context.

3 Methodology

3.1 Data set

The MIMIC-CXR dataset(Johnson et al., 2019a) was large publicly available dataset consisting of chest X-ray and free-text radiology reports. This dataset comprised 337,110 chest X-ray images collected from 75,379 patients, along with 227,827 associated free-text reports.

Crucially, rather than focusing solely on major disease categories Table 1, our evaluation leverages a long-tailed CXR(Holste et al., 2024) distribution to additionally test and evaluate relatively rare sub-diseases and nuanced clinical findings. This poses a significant challenge for traditional rule-based or rigid binary classification systems, which typically fail due to sparse sample sizes or fixed keyword definitions.

The test set consisted of 675 images and radiology reports from 295 patients. There were a total of 14 labels. Each label was classified into positive, negative, uncertain and unmentioned.

3.2 Llama-3-8b Prompt Labeling

A crucial component of our methodology was the generation of prompts for radiology report labeling using the **Llama-3-8B-Instruct** model. To address privacy concerns regarding sensitive medical data, all inferences were executed locally using open-weights, ensuring complete data security without external API dependency. We established 14 categories comprising key findings observed in radiology reports. To classify these conditions, we utilized two distinct approaches: hard labeling and soft labeling. The hard labeling method categorized each disease into four definitive classes—Positive, Negative, Uncertain, and Unmentioned—whereas the soft labeling method outputted continuous probability values ranging within $[0, 1]$. This soft approach effectively captured the inherent uncertainty of ambiguous or hedging expressions found in radiology reports.

3.3 Normalization for Soft-ZSL

To transform the raw probability value x generated from the prompt, we apply a linear normalization. This adjustment shifts the original

Disease Category	Positive	Negative	Total
Calcification of the Aorta	12	595	607
Emphysema	21	586	607
Pneumomediastinum	0	607	607
Tortuous Aorta	9	598	607

Table 1: Distribution of evaluated rare sub-disease classes in the test long-tailed CXR dataset, reporting the number of positive instances across the four target conditions.

values into a specific range, yielding a more balanced distribution:

$$\hat{x} = x \cdot \frac{5}{10} + 0.45, \quad \text{for } x > 0 \tag{1}$$

Next, we applied a Sigmoid function to the linearly normalized values $\hat{x}$:

$$\text{Sigmoid}(\hat{x}),\ \hat{x} > 0 \tag{2}$$

While the resulting values remain within the range of $[0, 1]$, this non-linear transformation smooths out extreme variances and yields more stable and consistent probability distributions.

3.4 Image Model Training

To validate the efficacy of the automated annotators, the label information generated from the radiology reports is systematically mapped onto the corresponding CXR images. For image classification, we employ three backbone architectures initialized with ImageNet-1K pre-trained weights. Input images are converted to grayscale and resized to 224×224 pixels, with data augmentation applied, including random rotation ($\pm 5°$), horizontal flipping, and standard normalization. The models are trained for 5 epochs using the Adam optimizer with a fixed learning rate of 1×10^{-4} and Binary Cross-Entropy (BCE) as the loss function.

4 Results

4.1 Capturing Clinical Nuances and Severity Gradients with Soft-ZSL

While traditional rule-based systems (e.g. CheXpert) often struggle with complex linguistic phenomena such as negations, hedging expressions, and resolved historical findings that lead to false-positive over-activations, as illustrated in Table 2, In contrast, our proposed Zero-Shot framework overcomes these limitations of rule-based systems, demonstrating the potential for more robust categorical predictions.

As demonstrated in Table 3, rather than enforcing rigid binary assignments (Hard-ZSL), the soft-labeling paradigm scales output values in response to descriptive modifiers—assigning lower probabilities to tentative pleural effusion (0.2) and mild atelectasis (0.4), and higher weights to severely increasing edema (0.9). By preventing distortion of clinical ambiguity, this capacity allows the model to faithfully capture clinical nuances and severity gradients for subsequent supervision.

4.2 Performance Evaluation on Rare Sub-diseases and Long-Tailed Distribution

To evaluate the model's generalization capabilities on underrepresented conditions beyond major pathologies, we conducted an extended evaluation focusing on rare sub-diseases infrequently found in standard clinical reports. Specifically, the evaluation was performed on four specific rare sub-diseases using 607 overlapping radiology reports from our gold standard dataset: Calcification of the Aorta, Emphysema, Pneumomediastinum, and Tortuous Aorta.

As summarized in Table 4, our framework achieved a Recall of 1.00 for both Calcification of the Aorta and Tortuous Aorta, followed by 0.95 for Emphysema. Furthermore, for Pneumomediastinum—which had no positive instances in the evaluation set—the model correctly predicted zero occurrences, effectively preventing false positives.

Radiology Report Sample	Target Disease	CheXpert	Zero-Shot
(a) ... No focal parenchymal opacities indicating pneumonia.	Pneumonia	P	N
(b) ... Combination of mediastinal fat and possible middle lobe atelectasis should not be mistaken for pneumonia.	Pneumonia	P	UM
(c) ... The atelectatic change at the left base seen previously is essentially cleared.	Atelectasis	P	N
(d) Pulmonary edema on _ has almost entirely cleared. ...	Edema	N	P
(e) ... Heart size is slightly larger but still within normal range.	Cardiomegaly	P	N

Table 2: Comparative analysis between the rule-based CheXpert and the proposed Zero-Shot framework on complex clinical narratives. 'P' denotes positive, 'N' denotes negative, and 'UM' denotes unmentioned.

Radiology Report Sample	Target Disease	Hard-ZSL	Soft-ZSL
(a) ... The left lung is clear aside from mild left base atelectasis.	Atelectasis	P	0.4
(b) ... relatively extensive atelectatic changes in the right lung bases.	Atelectasis	P	0.8
(c) ... Pleural Effusion, if any, is mild on the left side and similar.	Pleural Effusion	P	0.2
(d) ... These are strongly suggestive for severely increasing pulmonary edema.	Edema	P	0.9
(e) ... Pulmonary edema on _ has almost entirely cleared.	Edema	P	0.4

Table 3: Granular comparison between Hard-ZSL and Soft-ZSL outputs. 'P' indicates discrete positive assignments, while numerical values denote calibrated continuous probabilities.

4.3 Image Labeling Evaluation on Test Set

To evaluate the downstream image classification performance, we utilized a dedicated test set comprising 675 images from 295 patients, strictly partitioned to prevent patient-level data leakage. The evaluation focused on five key thoracic pathologies: Atelectasis, Cardiomegaly, Edema, Pleural Effusion, and Lung Opacity. We assessed four annotation configurations—CheXpert, CheXbert, Hard-ZSL, and Soft-ZSL. The final reported results represent the ensemble average of outputs from the three backbone architectures (ResNet-101, DenseNet-121, and ViT-B/16). Classification efficacy is systematically measured across five key performance metrics: Area Under the ROC Curve (AUC), Recall, F1-Score, Precision, and Accuracy.

4.4 Image Labeling Result for Test-set

As summarized in Table 5, training models exhibit nuanced performance variations depending on the upstream labeling methodology. For *Atelectasis*, the AUC values remain tightly coupled across all models, with Soft-ZSL achieving a marginal edge at 0.7135, while Hard-ZSL secures the highest Recall (0.6531) and F1-Score (0.5317). In the case of *Cardiomegaly*, Hard-ZSL yields the highest AUC at 0.8198 alongside an Accuracy of 0.8000, closely followed by Soft-ZSL (AUC: 0.8188). Both zero-shot frameworks outperform the rule-based CheXpert and CheXbert in terms of AUC, underscoring the benefits of leveraging fine-grained text context.

Conversely, for *Edema* classification, CheXbert establishes the upper bound with an AUC of 0.8574 and a Precision of 0.6416, whereas Hard-ZSL achieves a notably higher Recall (0.6590) and F1-Score (0.5596). Evaluation on *Pleural Effusion* reveals that Soft-ZSL delivers the strongest performance in AUC (0.8721) and Precision (0.8107), immediately followed by Hard-ZSL in AUC (0.8713) and Recall (0.8041), effectively eclipsing the baseline established by CheXpert and CheXbert. Finally, for *Lung Opacity*, Hard-ZSL achieves the highest AUC (0.6694) and F1-Score (0.4725), whereas CheXbert yields the highest Accuracy (0.7195), indicating that this category presents structural complexities across all paradigms.

Beyond raw discrimination (AUC), the proposed zero-shot architectures frequently outperform traditional labeling extractors across operational thresholds. Models trained via Hard-ZSL consistently exhibit strong Recall performance, particularly in complex conditions. In contrast, Soft-ZSL demonstrates a pronounced superiority in Precision (e.g., 0.8107 in Pleural Effusion and 0.4956 in Atelectasis), implying its practical utility in clinical workflows where minimizing false positives is

Disease	Gold Standard	Zero-Shot	Recall
Calcification of the Aorta	12	12	1.00
Emphysema	21	20	0.95
Pneumomediastinum	0	0	0.00
Tortuous Aorta	9	9	1.00

Table 4: Long-Tail Radiology Report Labeling Results.

Condition	Metric	CheXpert	CheXbert	Hard-ZSL	Soft-ZSL
Atelectasis	AUC	0.7118 ± 0.0039	0.7107 ± 0.0058	0.7117 ± 0.0125	$\mathbf{0.7135} \pm 0.0163$
	Recall	0.4162 ± 0.1016	0.4839 ± 0.0860	$\mathbf{0.6531} \pm 0.0243$	0.2521 ± 0.1339
	F1-Score	0.4355 ± 0.0574	0.4764 ± 0.0341	$\mathbf{0.5317} \pm 0.0136$	0.3086 ± 0.1363
	Precision	0.4717 ± 0.0075	0.4818 ± 0.0275	0.4486 ± 0.0133	$\mathbf{0.4956} \pm 0.0158$
	Accuracy	0.6933 ± 0.0064	0.6938 ± 0.0204	0.6642 ± 0.0128	$\mathbf{0.7057} \pm 0.0069$
Cardiomegaly	AUC	0.8140 ± 0.0102	0.8126 ± 0.0011	$\mathbf{0.8198} \pm 0.0071$	0.8188 ± 0.0031
	Recall	$\mathbf{0.6272} \pm 0.0522$	0.5351 ± 0.0713	0.5110 ± 0.0764	0.5263 ± 0.0426
	F1-Score	$\mathbf{0.5471} \pm 0.0041$	0.5242 ± 0.0268	0.5322 ± 0.0404	0.5261 ± 0.0211
	Precision	0.4888 ± 0.0240	0.5222 ± 0.0248	$\mathbf{0.5618} \pm 0.0123$	0.5281 ± 0.0125
	Accuracy	0.7664 ± 0.0156	0.7832 ± 0.0091	$\mathbf{0.8000} \pm 0.0055$	0.7872 ± 0.0055
Edema	AUC	0.8501 ± 0.0052	$\mathbf{0.8574} \pm 0.0056$	0.8519 ± 0.0019	0.8499 ± 0.0028
	Recall	0.3384 ± 0.0722	0.3181 ± 0.1441	$\mathbf{0.6590} \pm 0.0908$	0.3893 ± 0.0561
	F1-Score	0.4367 ± 0.0665	0.3995 ± 0.1441	$\mathbf{0.5596} \pm 0.0036$	0.4730 ± 0.0470
	Precision	0.6318 ± 0.0236	$\mathbf{0.6416} \pm 0.0309$	0.4962 ± 0.0414	0.6076 ± 0.0177
	Accuracy	$\mathbf{0.8336} \pm 0.0089$	0.8311 ± 0.0099	0.7990 ± 0.0249	0.8331 ± 0.0073
Pleural Effusion	AUC	0.8656 ± 0.0128	0.8656 ± 0.0075	0.8713 ± 0.0012	$\mathbf{0.8721} \pm 0.0051$
	Recall	0.7457 ± 0.0253	0.7360 ± 0.0426	$\mathbf{0.8041} \pm 0.0124$	0.6582 ± 0.0678
	F1-Score	0.7487 ± 0.0153	0.7340 ± 0.0190	$\mathbf{0.7525} \pm 0.0103$	0.7221 ± 0.0279
	Precision	0.7536 ± 0.0355	0.7350 ± 0.0327	0.7073 ± 0.0152	$\mathbf{0.8107} \pm 0.0455$
	Accuracy	$\mathbf{0.7965} \pm 0.0165$	0.7837 ± 0.0160	0.7852 ± 0.0103	0.7960 ± 0.0103
Lung Opacity	AUC	0.6509 ± 0.0067	0.6670 ± 0.0111	$\mathbf{0.6694} \pm 0.0023$	0.6560 ± 0.0061
	Recall	0.5471 ± 0.1100	0.3745 ± 0.0388	$\mathbf{0.5902} \pm 0.0388$	0.5608 ± 0.0111
	F1-Score	0.4372 ± 0.0156	0.4011 ± 0.0240	$\mathbf{0.4725} \pm 0.0034$	0.4435 ± 0.0044
	Precision	0.3792 ± 0.0421	$\mathbf{0.4341} \pm 0.0090$	0.3968 ± 0.0225	0.3672 ± 0.0104
	Accuracy	0.6489 ± 0.0512	$\mathbf{0.7195} \pm 0.0050$	0.6681 ± 0.0256	0.6454 ± 0.0133

Table 5: Training image model classification test results. Each performance metric represents the average across three distinct model architectures, with the corresponding standard deviations reported adjacent to the mean values.

critical.

5 Discussion

5.1 Performance Evaluation of Labeling Approaches in CXR Images

This study reveals a distinct trade-off between traditional rule-based systems and our zero-shot generative approach. While traditional rule-based baselines excel in standard text-level benchmarks, our ZSL framework achieves competitive performance without requiring manual annotation effort, instead leveraging LLM-based semantic reasoning.

Furthermore, our approach utilizes soft labeling to probabilistically model diagnostic uncertainty. As illustrated in Table 3, this continuous representation offers more granular disease calibration than rigid binary thresholds, overcoming the limitations of rule-based systems like CheXpert in capturing complex semantic nuances (Table 2).

When evaluating the downstream image classification models trained on these automated labels, the intrinsic value of our zero-shot framework becomes more pronounced. Our method achieves the highest Area Under the ROC Curve (AUC) across four major thoracic conditions, excluding Edema. Specifically, the Soft-ZSL framework yields top-tier AUC markers for Atelectasis (0.7135) and Pleural Effusion (0.8721), while Hard-ZSL establishes the upper bound for Cardiomegaly (0.8198) and Lung Opacity (0.6694).

We observe a critical representation gap: high accuracy in traditional text-level matching does not systematically transfer to downstream CNN performance, whereas our ZSL approach maintains robust cross-modal transferability. The varying efficacy of Soft-ZSL and Hard-ZSL across different pathologies further confirms that the clinical utility of prob-

abilistic labeling is heavily dependent on the inherent visual and structural characteristics of each specific condition.

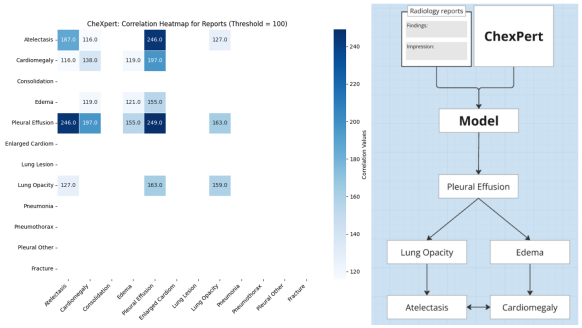

Figure 1: CheXpert: Textual disease correlation heatmap derived from radiology reports.

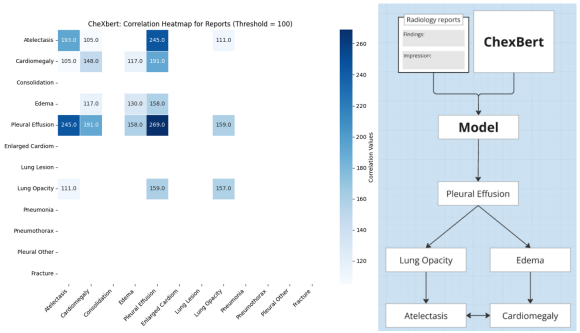

Figure 2: CheXbert: Textual disease correlation heatmap derived from radiology reports.

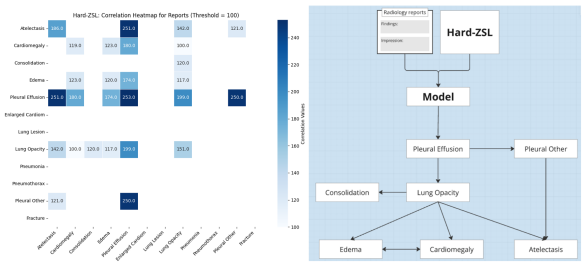

Figure 3: Hard-ZSL: Textual disease correlation heatmap derived from zero-shot report annotations.

5.2 Report: Analysis of Disease Correlations

Figures 1 and 2 display the report-level correlation matrices generated by CheXpert and CheXbert, respectively. Both models isolate strong clinical dependencies among core thoracic observations, with a primary focus on *Pleural Effusion*, alongside tight

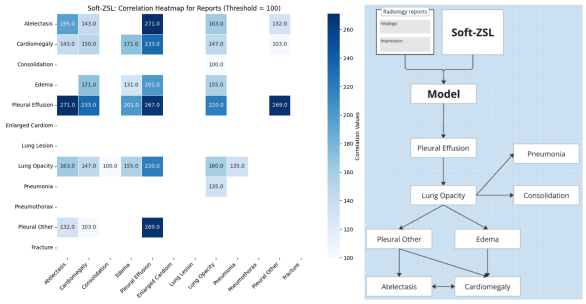

Figure 4: Soft-ZSL: Textual disease correlation heatmap derived from zero-shot report annotations.

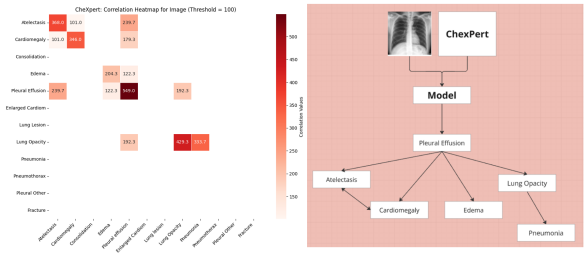

Figure 5: CheXpert: Downstream radiographic image-level disease correlation map.

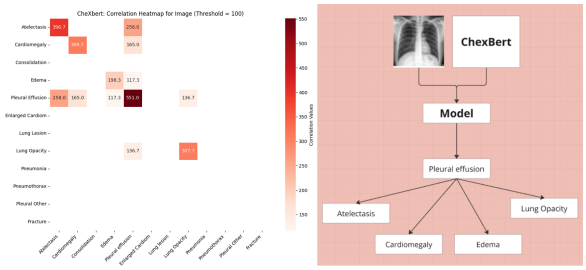

Figure 6: CheXbert: Downstream radiographic image-level disease correlation map.

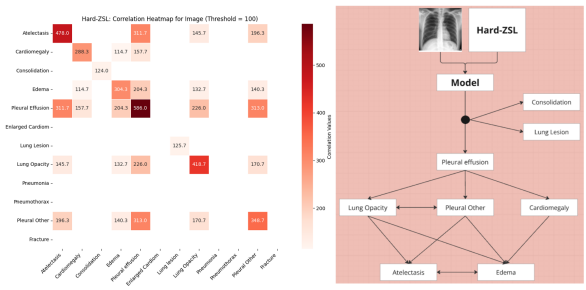

Figure 7: Hard-ZSL: Downstream radiographic image-level disease correlation map.

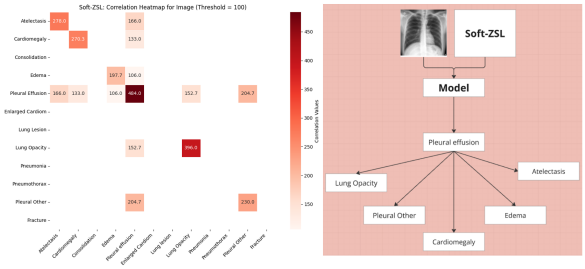

Figure 8: Soft-ZSL: Downstream radiographic image-level disease correlation map.

sub-clusters connecting *Atelectasis* and *Cardiomegaly*. Architecturally, CheXbert leverages a transformer-based vocabulary comprehension refined on 1,000 radiologist-annotated samples from the MIMIC-CXR dataset, which drastically improves sentence-level negation extraction over the rule-based CheXpert. However, because CheXbert's underlying target labels were systematically generated via CheXpert configurations, both frameworks yield highly collinear correlation topologies.

Conversely, the generative LLM prompt pipelines, detailed in Figures 3 and 4, retain the structural anchor around *Pleural Effusion* while mapping extended pathophysiological links. Hard-ZSL demonstrates clear modularity between *Pleural Effusion* and *Lung Opacity*, capturing localized clusters where *Edema* and *Cardiomegaly* co-vary, and linking *Atelectasis* across *Lung Opacity*, *Pleural Effusion*, and *Pleural Other*.

Soft-ZSL mirrors this structural baseline but uniquely uncovers secondary correlations between *Pneumonia* and both the effusion and opacity clusters. This behavior is primarily due to the probabilistic prompt architecture, which prevents discrete multi-label binarization from breaking down overlapping diagnostic signals, thus preserving the clinical intersectionality between infectious processes (*Pneumonia*) and fluid accumulation.

5.3 Image: Analysis of Disease Correlations

When projecting these learned textual structures onto downstream neural networks, the resulting image-level correlation maps display distinct configurations. In the CheXpert pipeline (Figure 5), strong cross-modal correlations are maintained between *Pleural Effusion* and the baseline quartet (*Atelectasis*, *Cardiomegaly*, *Edema*, and *Lung Opacity*), marked by a prominent link between *Atelectasis* and *Cardiomegaly*, alongside an isolated link matching *Pneumonia* to *Lung Opacity*. In comparison, the downstream model trained on CheXbert labels (Figure 6) fails to retain these tight sub-disease links, centering its visual attention almost exclusively on the major *Pleural Effusion* signal.

Finally, evaluating the networks trained via our LLM-driven zero-shot frameworks reveals a more balanced cross-modal transfer. The Soft-ZSL image correlation layout (Figure 8) closely resembles CheXbert's macro-structure but successfully extends its scope to include continuous representations of *Pleural Other*. This indicates that soft-labeled training targets help neural networks preserve fuzzy diagnostic borders that are typically erased by hard labelers.

Crucially, as shown in Figure 7, Hard-ZSL models show distinct, localized couplings between *Edema* and *Atelectasis* within the broader *Cardiomegaly* axis. Furthermore, structurally independent features for *Lung Lesion* and *Consolidation* emerge exclusively within the Hard-ZSL framework. This variation confirms that decoupling rigid label assumptions allows downstream vision models to capture unique radiological signs that traditional rule-based pipelines fail to resolve.

5.4 Advantages of Zero-Shot in Disease Labeling

Our zero-shot approach provides greater flexibility than few-shot learning by eliminating the need for specific examples when data for new conditions is scarce. This method leverages pre-trained knowledge to address diverse clinical scenarios efficiently, significantly reducing expert annotation time and costs. Importantly, this does not replace expert involvement; rather, professional review and refinement of LLM-generated outputs are essential to ensuring high-quality results.

6 Conclusion

In this study, we evaluated the efficacy of a generative zero-shot prompt architecture for automating radiology report annotation,

demonstrating highly promising results in multi-label thoracic disease identification and classification. By leveraging LLM-based reasoning, our framework effectively handles complex sentences and intricate linguistic nuances from radiology reports—such as negations and hedging expressions—that traditionally degrade rule-based extraction.

Furthermore, the soft-labeling methodology naturally bridges textual uncertainty to downstream image training, effectively reflecting the diverse clinical expressions and severity gradients of each disease. Crucially, through the zero-shot paradigm, our approach can efficiently respond not only to primary disease classes but also to emerging sub-disease classes, providing exceptional scalability and clinical flexibility.

Combined with local inference that guarantees complete data privacy, our zero-shot and soft-labeling frameworks offer a scalable, economically viable, and clinically robust alternative to traditional extraction pipelines, laying a solid foundation for flexible multi-modal medical AI architectures.

References

Yashas Annadani and Soma Biswas. 2018. Preserving semantic relations for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7603–7612.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Gregory Holste, Yiliang Zhou, Song Wang, Ajay Jaiswal, Mingquan Lin, Sherry Zhuge, Yuzhe Yang, Dongkyun Kim, Trong-Hieu Nguyen-Mau, Minh-Triet Tran, et al. 2024. Towards long-tailed, multi-label disease classification from chest x-ray: Overview of the cxr-lt challenge. *Medical Image Analysis*, 97:103224.

Steven Horng, Ruizhi Liao, Xin Wang, Sandeep Dalal, Polina Golland, and Seth J Berkowitz. 2021. Deep learning to quantify pulmonary edema in chest radiographs. *Radiology: Artificial Intelligence*, 3(2):e190228.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

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

Alistair Johnson, Matt Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. 2019a. Mimic-cxr-jpg-chest radiographs with structured labels. *PhysioNet*.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019b. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Omid Nejati Manzari, Hamid Ahmadabadi, Hossein Kashiani, Shahriar B Shokouhi, and Ahmad Ayatollahi. 2023. Medvit: a robust vision transformer for generalized medical image classification. *Computers in Biology and Medicine*, 157:106791.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date.

Maram Mahmoud A Monshi, Josiah Poon, and Vera Chung. 2020. Deep learning in generating radiology reports: A survey. *Artificial Intelligence in Medicine*, 106:101878.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. 2018. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2018:188.

Keegan Quigley, Miriam Cha, Ruizhi Liao, Geeticka Chauhan, Steven Horng, Seth Berkowitz, and Polina Golland. 2022. Radtex: Learning efficient radiograph representations from text reports. In *MICCAI Workshop on Resource-Efficient Medical Image Analysis*, pages 22–31. Springer.

Sonish Sivarajkumar and Yanshan Wang. 2022. Healthprompt: a zero-shot learning paradigm for clinical natural language processing. In

AMIA Annual Symposium Proceedings, volume 2022, page 972. American Medical Informatics Association.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. 2020. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.