

Privacy-Preserving Information Extraction Framework for Diverse Imaging Reports using Large Language Models

Dabin Min

Interdisciplinary Program in Bioengineering, Seoul National University Graduate School, Seoul

Soyeon Kim

Interdisciplinary Program in Bioengineering, Seoul National University Graduate School, Seoul

Sangheum Hwang

Seoul National University of Science and Technology

Kwang Nam Jin

Boramae Medical Center, Seoul National University College of Medicine

SangHeum Bang

Boramae Medical Center, Seoul National University College of Medicine

Won Gi Jeong

Chonnam National University Hwasun Hospital, Chonnam National University Medical School

Jinwook Choi

Seoul National University College of Medicine

Jung Min Chang

Seoul National University College of Medicine

Chang Min Park

`morphius@snu.ac.kr`

Seoul National University College of Medicine

Article

Keywords: Radiology Report Information Extraction, Large Language Models, Privacy-Preserving Framework

Posted Date: May 6th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-6267208/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Efficient extraction of structured information from unstructured radiology reports remains a critical challenge in healthcare. We introduce the Radiology Report Information Extraction Framework (RRIEF), a privacy-preserving approach utilizing parameter-efficient fine-tuning of open-source large language models (LLMs). We validated RRIEF across chest X-ray (CXR), mammography, and coronary CT angiography (CCTA) reports, evaluating its performance against specialized methods and proprietary LLMs (GPT-4o, Gemini-1.5-Flash, Claude-3.5-Sonnet). For CXR, RRIEF-LLaMA1-65B achieved F1 scores of 0.87 and 0.85 in internal and external tests, significantly outperforming CheXpert Labeler (0.70 and 0.69, $P < .001$), CheXbert (0.72 and 0.69, $P < .001$), and all proprietary LLMs (Claude-3.5-Sonnet: 0.69 and 0.62, $P < .001$). For mammography, RRIEF-LLaMA1-30B/65B reached F1 scores of 0.91 and 0.99 in internal and external tests, exceeding all proprietary LLMs (0.86 and 0.92, $P = .002$). For CCTA, using only 100 training reports, RRIEF-LLaMA3-8B significantly outperformed Gemini-1.5-Flash in stenosis severity (0.87 vs 0.83, $P = .02$), GPT-4o in external testing (0.83 vs 0.68, $P < .001$), and all proprietary models for modifiers in external testing (1.00 vs 0.93, $P = .004$). Notably, RRIEF-LLaMA3-8B achieved superior performance on CXR with only 200 training samples compared to all baselines including CheXbert and proprietary LLMs ($P < .001$). Our locally deployable framework enables high-performance information extraction from different types of radiology reports, facilitating large-scale research and clinical practice. We provide our complete implementation code publicly to promote accessibility and adoption.

Introduction

Radiology reports represent a vital component of healthcare data, containing detailed interpretations of medical images that are crucial for patient care, clinical research, and quality assurance. However, an estimated 80% of this clinical information exists in unstructured free-text format, making it challenging to leverage for systematic analysis and clinical decision support (1). Automated information extraction from these reports—converting unstructured text into structured, quantifiable data—has emerged as a critical need across various imaging modalities (2, 3). Making these unstructured reports computationally accessible could advance clinical research, enhance quality monitoring, and streamline billing processes while reducing the manual annotation burden on healthcare professionals (4–7).

Efforts for automated information extraction from radiology reports have evolved significantly over the past decades. Early approaches relied heavily on rule-based systems and traditional machine learning methods, which often required extensive manual feature engineering and were highly sensitive to variations in reporting styles and terminology (2, 5, 8). More recent deep learning-based methods have shown promise in specific settings, particularly for chest X-ray (CXR) reports, achieving notable accuracy in identifying common findings (6). However, these solutions typically require large annotated datasets for training and are often confined to specific imaging modalities or institutional reporting formats (3). For example, CheXpert and CheXbert have demonstrated success in information extraction from CXR reports but their application remains limited to this single modality (5, 6). Similar constraints exist in other imaging domains, where the development of automated labeling systems has been hampered by

the need for substantial manual annotations and the challenge of accommodating diverse reporting patterns across institutions and radiologists (7–11). Furthermore, these specialized approaches often lack the flexibility to adapt to evolving medical terminology and reporting practices, necessitating frequent retraining and updates (3).

The advent of large language models (LLMs) like ChatGPT and LLaMA has introduced new possibilities for medical information extraction, potentially addressing the limitations of traditional approaches (12–19). In healthcare applications, these models have demonstrated remarkable capabilities across tasks including medical licensing examinations, clinical decision support, and extraction of clinical information from various text sources (20–28). Specifically for extracting information from radiology reports, recent studies have explored zero-shot and few-shot approaches using these models, leveraging their pre-trained knowledge without requiring extensive task-specific training data (13, 14, 26, 29, 30). However, while these attempts highlight the potential of LLMs, their performance has generally remained comparable to or slightly below that of existing specialized methods, primarily due to reliance on zero/few-shot learning approaches (13, 29, 31). Moreover, proprietary LLMs raise significant privacy concerns when processing healthcare data, spurring interest in locally-deployable open-source alternatives, though optimal training and implementation approaches for medical report labeling remain unclear (13, 14, 32–35).

Thus, the purpose of our study is to develop and validate an efficient framework for information extraction from radiology reports. Our approach focuses on three key objectives: First, to establish a privacy-preserving framework that outperforms existing methodologies. Second, to demonstrate the framework's effectiveness across different imaging modalities with distinct characteristics and reporting styles. Third, to facilitate widespread adoption in diverse clinical settings by analyzing training data requirements and providing our complete implementation as an open-source resource.

Material and Methods

This retrospective study was approved by the Institutional Review Board of Seoul National University Hospital (IRB No. 2303-155-1417) and the requirement for written informed consent was waived. To promote accessibility and facilitate adoption in diverse clinical settings, our complete implementation code is publicly available at https://github.com/reonaledo/report_labeler (currently under development).

Data

In our study, we utilized the following datasets: MIMIC-CXR (36) and Open-i (37) for CXR reports, a private dataset from Seoul National University Hospital (SiteA-MMG) and CDD-CESM (38) for mammography reports, and two private datasets from Seoul National University Bundang Hospital (SiteB-CCTA) and Chonnam National University Hwasun Hospital (SiteC-CCTA) for coronary computed tomography angiography (CCTA) reports.

For CXR, 2,000 reports were sampled from MIMIC-CXR for training (n = 1,000) and internal testing (n = 1,000), with an additional 751 reports from Open-i for external test. For mammography, 1,500 reports sampled from 224,136 screening cases collected by SiteA-MMG between January 2001 and December 2022 were used for training (n = 500) and internal testing (n = 1,000), with the entire CDD-CESM dataset (n = 326) used for external test. For CCTA, we collected 150 reports from SiteB-CCTA for training (n = 100) and internal testing (n = 50), with an additional 51 reports from SiteC-CCTA for external testing. All CCTA reports were synthetically generated by cardiothoracic radiologists following their institutional reporting formats. Detailed descriptions of dataset selection and preparation are provided in **Supplemental Text 1**. Examples from each dataset, illustrating the variations in reporting styles, are shown in Table 1.

Table 1
Examples of Each Dataset.

Dataset	Modality	Example 1	Example 2
MIMIC-CXR	CXR	The lungs are clear. There is no focal consolidation, pleural effusion, or pneumothorax. The cardiomediastinal silhouette is normal. There is no free air under the hemidiaphragms. No pancreatic calcifications visualized. Osseous structures are intact.	The cardiomediastinal and hilar contours are within normal limits. The lungs are well expanded and clear. There is no large pleural effusion, pneumothorax or focal consolidation concerning for pneumonia. There is no evidence of free air.
Open-i		Heart is mildly enlarged stable. Mediastinal contour is normal. Pulmonary vascularity is normal. Lungs are hyperexpanded but clear. No pleural effusions or pneumothoraces.	The cardiomediastinal silhouette is normal in size and contour. Stable right lower lobe calcified granuloma. No focal consolidation, pneumothorax or large pleural effusion. Spurring of the thoracic spine.
SiteA-MMG	Mammography	Gr 3 Suspicious malignant microcal, RUO: Segmental/ fine linear pleomorphic/ductal extension to SA/ nipple retraction(+) --> C5 Suspicious malignant mass with calcifications, Rt inner --> C4c r/o metastatic LN, Rt axilla --> C4c	Gr 2 and a few benign calci on Rt.
CDD-CESM		Right Breast: ACR C: Heterogeneously dense breasts. Upper outer quadrant benign macrocalcifications. No suspicious microcalcifications. Multiple lower and central inner equal density rounded and oval shaped masses are seen, some of them show circumscribed margin and others show obscured margin. Normal skin thickness and contour of breast.	Right Breast: Diffuse edematous changes evidenced by increased skin thickness and coarsened trabeculae. Associated upper outer clusters of pleomorphic microcalcifications are also seen. Left Breast: Central benign macrocalcifications are noted. No speculated mass lesions

Note: In CXR reports, both MIMIC-CXR and Open-i maintain a structured approach with minor differences in detail expression. For mammography reports, SiteA-MMG utilizes both common abbreviations (e.g., 'LN' for lymph node, 'Rt' for right) and unconventional shorthand (e.g., 'microcal', for microcalcification, 'calci' for calcification) in a list-like format, catering to specialist readers. In contrast, CDD-CESM adopts a narrative style, using minimal abbreviations and favoring standard terminology, providing detailed descriptions for each breast followed by an opinion section. For CCTA, SiteB-CCTA employs a concise narrative format focusing on core findings with minimal structural divisions, while SiteC-CCTA uses a comprehensive starred-section format with detailed anatomical categorization, explicit measurement values, and a separate conclusion section that systematically summarizes all findings. CXR = Chest X-ray, CCTA = Coronary CT Angiography.

Dataset	Modality	Example 1	Example 2
		Left Breast: Status postoperative with flap reconstruction showing no speculated mass lesions or suspicious microcalcifications. Normal skin thickness. OPINION: Right Breast: Multiple lower and central inner rounded and oval shaped benign looking homogenously enhancing masses with circumscribed margin (BIRADS 3). Upper outer quadrant benign macrocalifications (BIRADS 2). Left Breast: Status postoperative with flap reconstruction showing no evidence of recurrent lesions (BIRADS 2).	or suspicious microcalcifications. Normal skin thickness and contour of breast. OPINION: Right Breast: Diffuse edematous changes associated with upper outer suspicious microcalcifications (BIRADS 5). Left Breast: Central benign macrocalcifications (BIRADS 2).
SiteB-CCTA	CCTA	Mild atherosclerosis with no significant stenosis in the coronary arteries. LM and LCx, unremarkable pLAD and pRCA, < 20–30% stenosis with calcified plaques No remarkable finding in the LV myocardium. Total coronary calcium score = 205.68	Atherosclerosis with significant stenosis in the coronary arteries. LM RI LCx RCA < 20% stenosis with calcified plaques. pLAD focal 70–80% stenosis with mixed plaque. No remarkable finding in the LV myocardium. Total coronary calcium score = 550.04. Ascending aorta dilatation: 45 mm.
SiteC-CCTA		* Image quality: good quality * Calcium Scoring Total Agatston Score: 245.0 Coronary calcium volume: 210.0	* Image quality: limitation due to severe coronary calcification. * Calcium Scoring Total Agatston Score: 996.0
Note: In CXR reports, both MIMIC-CXR and Open-i maintain a structured approach with minor differences in detail expression. For mammography reports, SiteA-MMG utilizes both common abbreviations (e.g., ‘LN’ for lymph node, ‘Rt’ for right) and unconventional shorthand (e.g., ‘microcal’, for microcalcification, ‘calci’ for calcification) in a list-like format, catering to specialist readers. In contrast, CDD-CESM adopts a narrative style, using minimal abbreviations and favoring standard terminology, providing detailed descriptions for each breast followed by an opinion section. For CCTA, SiteB-CCTA employs a concise narrative format focusing on core findings with minimal structural divisions, while SiteC-CCTA uses a comprehensive starred-section format with detailed anatomical categorization, explicit measurement values, and a separate conclusion section that systematically summarizes all findings. CXR = Chest X-ray, CCTA = Coronary CT Angiography.			

Dataset	Modality	Example 1	Example 2
		* Dominancy: Right dominancy * Coronary Anomaly or variant; Absent. Coronary artery stenosis 1) LM : No plaque, no stenosis 2) LAD with branches : pLAD ; mixed plaque with mild stenosis mLCA: small calcified plaque with mild stenosis 3) LCX with branches : Hypoplastic LCX, No plaque, no stenosis 4) RCA with branches : pRCA, mRCA, dRCA ; small calcified plaque with mild stenosis 5) Ramus Intermedius : Absent Vulnerable plaque : None. Other cardiac finding : S/P MVR, AVR Hypertrabeculation at LV apex. Extracardiac finding : Within normal limits. ""1. Moderate coronary calcification (Total Agatston Score: 245.0) (1) pRCA, mRCA, dRCA ; small calcified plaque with mild stenosis (2) pLAD ; mixed plaque with mild stenosis	Coronary calcium volume: 802.1 * Dominancy: Right dominancy * Coronary Anomaly or variant; Absent. Coronary artery stenosis (1) proximal to distal RCA ; multifocal calcified plaque with mild stenosis (2) Proximal LAD ; calcified/mixed plaque with mild to moderate stenosis ...suspicious moderate to severe stenosis at pLAD near 2nd diagonal branch/proximal 2nd diagonal. mid LAD ; shallow myocardial bridging distal LAD ; calcified plaque with mild stenosis (3) proximal LCX ; calcified plaque with mild stenosis Vulnerable plaque : None. Other cardiac finding : Within normal limits. Extracardiac finding : Within normal limits Conclusion 1. severe coronary calcification (Total Agatston Score: 996.0).

Note: In CXR reports, both MIMIC-CXR and Open-i maintain a structured approach with minor differences in detail expression. For mammography reports, SiteA-MMG utilizes both common abbreviations (e.g., 'LN' for lymph node, 'Rt' for right) and unconventional shorthand (e.g., 'microcal', for microcalcification, 'calci' for calcification) in a list-like format, catering to specialist readers. In contrast, CDD-CESM adopts a narrative style, using minimal abbreviations and favoring standard terminology, providing detailed descriptions for each breast followed by an opinion section. For CCTA, SiteB-CCTA employs a concise narrative format focusing on core findings with minimal structural divisions, while SiteC-CCTA uses a comprehensive starred-section format with detailed anatomical categorization, explicit measurement values, and a separate conclusion section that systematically summarizes all findings. CXR = Chest X-ray, CCTA = Coronary CT Angiography.

Dataset	Modality	Example 1	Example 2
		mLAD: small calcified plaque with mild stenosis 2. S/P MVR, AVR Hypertrabeculation at LV apex.	(1) proximal to distal RCA ; multifocal calcified plaque with mild stenosis (2) Proximal LAD ; calcified/mixed plaque with mild to moderate stenosis ...suspicious moderate to severe stenosis at LAD near 2nd diagonal branch/proximal 2nd diagonal. mid LAD ; shallow myocardial bridging distal LAD ; calcified plaque with mild stenosis (3) proximal LCX ; calcified plaque with mild stenosis
Note: In CXR reports, both MIMIC-CXR and Open-i maintain a structured approach with minor differences in detail expression. For mammography reports, SiteA-MMG utilizes both common abbreviations (e.g., 'LN' for lymph node, 'Rt' for right) and unconventional shorthand (e.g., 'microcal', for microcalcification, 'calci' for calcification) in a list-like format, catering to specialist readers. In contrast, CDD-CESM adopts a narrative style, using minimal abbreviations and favoring standard terminology, providing detailed descriptions for each breast followed by an opinion section. For CCTA, SiteB-CCTA employs a concise narrative format focusing on core findings with minimal structural divisions, while SiteC-CCTA uses a comprehensive starred-section format with detailed anatomical categorization, explicit measurement values, and a separate conclusion section that systematically summarizes all findings. CXR = Chest X-ray, CCTA = Coronary CT Angiography.			

Radiology Report Information Extraction Framework (RRIEF)

Figure 1 presents an overview of our framework, **Radiology Report Information Extraction Framework (RRIEF)**, for extracting structured information from radiology reports. The framework consists of four sequential steps: target definition, data annotation, prompt design, and LLM training. We detail each component below.

Target Definition

This step involves specifying the classes of interest and their possible label categories for the target dataset. Classes represent medical findings (e.g., mass, pneumothorax) or inferable information (e.g., severity ratings) from reports, while label categories define the possible values each class can take. These can be freely configured based on the intended use of the extracted information.

For CXR, we used 13 finding classes following the CheXpert Labeler’s protocol, excluding the 'no finding' class as it's inferable from the others (5). The categorization of labels was based on the presence of

these findings in the reports, classified as 'positive', 'negative', 'unsure', or 'not mentioned'. For mammography, given the absence of prior research, we established 11 finding classes derived from the Breast Imaging Lexicon (39), categorized according to their presence in the reports and specified laterality as 'right breast', 'left breast', 'bilateral breasts', 'unsure', or 'not mentioned'. For CCTA, following Coronary Artery Disease-Reporting and Data System (CAD-RADS) 2.0 guidelines (40), we defined classes and labels for stenosis severity categories (0, 1, 2, 3, 4A, 4B, or 5), plaque burden scores (None, P1, P2, P3, or P4), and six modifiers (E, I, N, G, HRP, and S, each labeled as 0 or 1). Details regarding the target findings and labels are provided in **Supplemental Text 2**.

Data Annotation

To expedite the annotation process, we first generate initial annotations using LLMs and then have human review these preliminary labels, significantly reducing the manual annotation effort compared to starting from scratch. For the initial annotation of CXR and mammography datasets, we utilized LLaMA-13B (18) to generate initial annotations based on a small number of annotated examples (we used three fixed examples). These preliminary results were then refined by two graduate students under the guidance of a chest radiologist (D.H.K.; >5 years of experience) and a breast radiologist (J.M.J.; >20 years of experience). For CCTA reports, two board-certified cardiovascular radiologists (W.G.J.; >6 years of experience and B.S.H.; >5 years of experience) independently reviewed all reports and established ground truth annotations through consensus. The labels for all test datasets additionally underwent a comprehensive review by the radiologists, each specializing in their respective fields.

Prompt Design

Our prompts consist of three components: instruction, input, and output. The instruction provides detailed descriptions of the task, including definitions of target classes and their possible labels. The input comprises the radiology report text, while the output is structured as a JSON format containing the extracted information. We designed prompts to be explicit about task requirements while maintaining flexibility for different reporting styles. Examples of prompt design and corresponding outputs are shown in Fig. 2, with complete prompt templates detailed in **Supplemental Text 3**.

LLM Training

We trained LLaMA models (version 1 and 3) independently for each imaging modality using different model sizes ranging from 7B to 70B parameters (16, 18). To minimize computational resources while maintaining performance, we employed Quantized Low-Rank Adaptation (QLoRA) (41), a parameter-efficient fine-tuning technique. This approach reduced the trainable parameters to less than 0.1% of the total model parameters, enabling efficient training of even the largest model (LLaMA3-70B) on a single 48GB A6000 GPU within 6 hours. To ensure reproducibility and consistent label generation across

different runs, we employed deterministic decoding by setting the temperature parameter to 0 (13). Detailed training configurations and hyperparameters are provided in **Supplemental Text 4**.

Evaluation

For comparative analysis, we tested RRIEF with various LLaMA models (LLaMA1: 7B, 13B, 30B and 65B; LLaMA3: 8B and 70B) against three comparison groups: their base versions without additional training, the open-source DeepSeek-R1-Distill-Qwen-14B model (a specialized model optimized for reasoning tasks) (42), and leading proprietary LLMs (GPT-4o, Gemini-1.5-flash, Claude-3.5-sonnet), all evaluated under zero/few-shot settings (minimal example-based inference where models perform tasks with no or few examples). We conducted these comparative experiments across all three modalities: CXR, mammography, and CCTA. For CXR reports, we additionally compared against specialized methods including CheXpert Labeler and CheXbert. Implementation details of zero/few-shot learnings are provided in **Supplemental Text 5**.

To investigate the relationship between training data size and model performance, we conducted experiments using LLaMA3-8B on the MIMIC-CXR dataset. Training data sizes ranged from 50 to 1,000 reports (50, 100, 200, 500 and 1,000), all randomly sampled from the original training set of 1,000 reports without replacement. To ensure statistical reliability and reproducibility, we repeated each experiment five times using different random seeds.

Performance was evaluated using the Macro F1 score (rationale detailed in **Supplemental Text 6**), with results averaged across 10 bootstrap repetitions and presented with 95% confidence intervals. Statistical comparisons between methods used the one-sided Wilcoxon Signed Rank test ($P < .05$ for significance). Cases failing to produce valid JSON outputs were excluded from analysis (**Supplemental Text 7**).

Result

Patient Characteristics

A flowchart illustrating the data selection process is provided in Fig. 3. For CXR, we utilized the MIMIC-CXR dataset, including 1,000 reports from 631 patients (median age 62 years [IQR 50–74 years]; 50% female) for training and 1,000 reports from 807 patients (median age 63 years [IQR 49–72 years]; 53% female) for internal testing. The Open-i dataset provided 751 reports from 751 patients (demographics unavailable) for external testing. For mammography, the SiteA-MMG dataset contributed 500 reports from 486 patients (median age 51 years [IQR 47–56 years]; all female) for training and 1,000 reports from 995 patients (median age 53 years [IQR 47–60 years]; all female) for internal testing. The CDD-CESM dataset provided 326 reports from 326 patients (median age 50 years [IQR 43–58 years]; all female) for external testing. For CCTA, we employed the SiteB-CCTA dataset with 100 reports for training

and 50 reports for internal testing. The SiteC-CCTA dataset contributed 51 reports for external testing. As synthetic data was used for CCTA, conventional demographic distributions could not be represented.

Overall Performance in Chest X-ray and Mammography Report Labeling

The overall performance results for CXR and mammography report labeling are summarized in Table 2. In our internal MIMIC-CXR test, RRIEF-LLaMA1-65B achieved a mean F1 score of 0.87 (95% CI: 0.86, 0.88), significantly outperforming CheXpert Labeler (0.70, 95% CI: 0.70, 0.71; $P < .001$) and CheXbert (0.72, 95% CI: 0.71, 0.72; $P < .001$). This performance surpassed all zero/few-shot LLMs including Claude-3.5-Sonnet (0.69, 95% CI: 0.68, 0.70; $P < .001$). Even the smallest model, RRIEF-LLaMA1-7B, demonstrated superior performance (0.83, 95% CI: 0.81, 0.84; $P < .001$) compared to these benchmarks. On the external Open-i dataset, RRIEF-LLaMA3-70B scored highest at 0.85 (95% CI: 0.83, 0.87), surpassing CheXpert Labeler (0.69, 95% CI: 0.68, 0.70; $P < .001$) and CheXbert (0.69, 95% CI: 0.68, 0.71; $P < .001$). RRIEF-LLaMA3-70B showed better generalizability than RRIEF-LLaMA1-65B (0.80, 95% CI: 0.77, 0.82; $P = .002$). Notably, smaller models such as RRIEF-LLaMA1-7B (0.80, 95% CI: 0.79, 0.81) and RRIEF-LLaMA3-8B (0.84, 95% CI: 0.82, 0.85) outperformed all zero/few-shot benchmarks including Claude-3.5-Sonnet (0.62, 95% CI: 0.60, 0.63; $P < .001$), aligning with their internal test results.

For mammography reports, internal testing on SiteA-MMG showed RRIEF-LLaMA1-30B and RRIEF-LLaMA1-65B both achieved 0.91 (95% CI: 0.89, 0.92 and 0.90, 0.92, respectively), exceeding zero/few-shot performances of LLaMA models ($P < .001$ for all comparisons) and proprietary LLMs including Gemini-1.5-Flash (0.86, 95% CI: 0.85, 0.87; $P = .002$). Even the smallest model, RRIEF-LLaMA1-7B (0.88, 95% CI: 0.86, 0.90), outperformed GPT-4o and Claude-3.5-Sonnet in few-shot settings ($P < .001$ for all comparisons). In the external CDD-CESM test, RRIEF-LLaMA3-70B achieved the highest score of 0.99 (95% CI: 0.98, 1.00), confirming its superior generalizability over RRIEF-LLaMA1-65B ($P = .02$) and outperforming all zero/few-shot configurations, including Claude-3.5-Sonnet (0.92, 95% CI: 0.90, 0.94; $P = .002$).

Table 2
Overall Performance on Chest X-ray and Mammography Reports.

Method	Model	Chest X-ray		Mammography	
		MIMIC-CXR (n = 1,000)	Open-i (n = 751)	SiteA-MMG (n = 1,000)	CDD-CESM (n = 326)
CheXpert Labeler	-	0.70 (0.70, 0.71)	0.69 (0.68, 0.70)	-	-
CheXbert	-	0.72 (0.71, 0.72)	0.69 (0.68, 0.71)	-	-
0-shot	LLaMA1-65B	0.47 (0.47, 0.47)	0.45 (0.44, 0.46)	0.40 (0.35, 0.44)	0.32 (0.31, 0.34)
	LLaMA3-70B	0.36 (0.35, 0.37)	0.31 (0.30, 0.32)	0.65 (0.64, 0.66)	0.84 (0.83, 0.86)
	DeepSeek-R1-Distill-Qwen-14B	0.54 (0.53, 0.55)	0.47 (0.46, 0.48)	0.71 (0.70, 0.72)	0.45 (0.44, 0.47)
	GPT-4o	0.59 (0.58, 0.59)	0.47 (0.45, 0.49)	0.71 (0.71, 0.72)	0.72 (0.71, 0.73)
	Gemini-1.5-Flash	0.57 (0.56, 0.58)	0.60 (0.59, 0.62)	0.71 (0.70, 0.72)	0.72 (0.70, 0.75)
	Claude-3.5-Sonnet	0.67 (0.66, 0.68)	0.47 (0.45, 0.49)	0.72 (0.71, 0.72)	0.79 (0.78, 0.81)
3-shot	LLaMA1-65B	0.47 (0.45, 0.48)	0.45 (0.44, 0.47)	0.64 (0.63, 0.66)	0.60 (0.58, 0.61)

Note: Data represent the average Macro F1 Score and 95% Confidence Interval from 10 Bootstrap Tests. For each dataset, the highest performance is shown in bold and the second-highest is underlined. RRIEF = Radiology Report Information Extraction Framework, B = Billion, LLaMA = Large language model Meta AI.

Method	Model	Chest X-ray		Mammography	
		MIMIC-CXR (n = 1,000)	Open-i (n = 751)	SiteA-MMG (n = 1,000)	CDD-CESM (n = 326)
	LLaMA3-70B	0.60 (0.59, 0.61)	0.56 (0.54, 0.58)	0.87 (0.85, 0.88)	0.87 (0.86, 0.88)
	DeepSeek-R1-Distill-Qwen-14B	0.57 (0.57, 0.58)	0.53 (0.52, 0.53)	0.73 (0.71, 0.74)	0.85 (0.83, 0.88)
	GPT-4o	0.67 (0.66, 0.68)	0.57 (0.55, 0.59)	0.75 (0.74, 0.75)	0.89 (0.88, 0.90)
	Gemini-1.5-Flash	0.62 (0.61, 0.63)	0.55 (0.53, 0.56)	0.86 (0.85, 0.87)	0.86 (0.84, 0.88)
	Claude-3.5-Sonnet	0.69 (0.68, 0.70)	0.62 (0.60, 0.63)	0.75 (0.74, 0.75)	0.92 (0.90, 0.94)
RRIEF (Ours)	LLaMA1-7B	0.83 (0.81, 0.84)	0.80 (0.79, 0.81)	0.88 (0.86, 0.90)	0.74 (0.72, 0.75)
	LLaMA1-13B	0.83 (0.82, 0.83)	0.79 (0.78, 0.81)	0.90 (0.88, 0.91)	0.79 (0.77, 0.80)
	LLaMA1-30B	0.85 (0.85, 0.86)	0.81 (0.79, 0.82)	0.91 (0.89, 0.92)	0.93 (0.91, 0.94)
	LLaMA1-65B	0.87 (0.86, 0.88)	0.80 (0.77, 0.82)	0.91 (0.90, 0.92)	0.97 (0.97, 0.98)
	LLaMA3-8B	0.85	0.84	0.88	0.85

Note: Data represent the average Macro F1 Score and 95% Confidence Interval from 10 Bootstrap Tests. For each dataset, the highest performance is shown in bold and the second-highest is underlined. RRIEF = Radiology Report Information Extraction Framework, B = Billion, LLaMA = Large language model Meta AI.

Method	Model	Chest X-ray		Mammography	
		MIMIC-CXR (n = 1,000)	Open-i (n = 751)	SiteA-MMG (n = 1,000)	CDD-CESM (n = 326)
		(0.85, 0.86)	(0.82, 0.85)	(0.86, 0.90)	(0.83, 0.86)
	LLaMA3-70B	0.85	0.85	0.90	0.99
		(0.84, 0.86)	(0.83, 0.87)	(0.88, 0.92)	(0.98, 1.00)
Note: Data represent the average Macro F1 Score and 95% Confidence Interval from 10 Bootstrap Tests. For each dataset, the highest performance is shown in bold and the second-highest is underlined. RRIEF = Radiology Report Information Extraction Framework, B = Billion, LLaMA = Large language model Meta AI.					

Finding-specific Performance Advantages of RRIEF

Table 3 highlights RRIEF's average performance advantages in CXR report labeling across individual findings, comparing the average performance of all RRIEF models (Avg. RRIEF: RRIEF-LLaMA1-7B/13B/30B/65B, RRIEF-LLaMA3-8B/70B) against other methods. In labeling 'Support Devices', RRIEF models outperformed traditional methods (Avg. CheXpert & CheXbert) by F1 score differences of 0.32 (MIMIC-CXR) and 0.41 (Open-i). Compared to zero-shot and few-shot performances of LLaMA1-65B and LLaMA3-70B (Avg. LLaMA 0-shot and Avg. LLaMA 3-shot), RRIEF models demonstrated substantial improvements, particularly in 'Enlarged Cardiomedastinum' where F1 score differences reached 0.52–0.64 in both test sets. When compared against proprietary LLMs (Avg. Proprietary LLMs 0-shot and 3-shot: GPT-4o, Gemini-1.5-Flash, Claude-3.5-Sonnet), RRIEF models showed notable improvements (0.23–0.41 F1 score difference) in 'Enlarged Cardiomedastinum', 'Cardiomegaly', and 'Lung Lesion' across both internal and external test sets.

Table 3
Average Performance Comparison of Methods for Chest X-ray Report Labeling by Findings.

Dataset and finding	Avg. RRIEF (Ours)	Avg. CheXpert & CheXbert	Avg. LLaMA 0-shot	Avg. LLaMA 3-shot	Avg. Proprietary LLMs 0-shot	Avg. Proprietary LLMs 3-shot
<i>MIMIC-CXR (n = 1,000)</i>						
Enlarged Cardiom.	0.91 ± .03	0.74 ± .01 (Δ-0.17)	0.27 ± .04 (Δ-0.64)	0.39 ± .17 (Δ-0.52)	0.60 ± .02 (Δ-0.31)	0.60 ± .02 (Δ-0.31)
Cardiomegaly	0.93 ± .01	0.69 ± .02 (Δ-0.25)	0.50 ± .08 (Δ-0.44)	0.54 ± .07 (Δ-0.39)	0.54 ± .01 (Δ-0.39)	0.59 ± .02 (Δ-0.34)
Lung Lesion	0.84 ± .06	0.70 ± .00 (Δ-0.14)	0.30 ± .00 (Δ-0.54)	0.46 ± .12 (Δ-0.38)	0.51 ± .06 (Δ-0.33)	0.61 ± .05 (Δ-0.23)
Lung Opacity	0.81 ± .03	0.69 ± .09 (Δ-0.12)	0.32 ± .11 (Δ-0.49)	0.44 ± .01 (Δ-0.37)	0.45 ± .06 (Δ-0.36)	0.52 ± .03 (Δ-0.28)
Edema	0.87 ± .02	0.72 ± .00 (Δ-0.15)	0.42 ± .13 (Δ-0.46)	0.55 ± .13 (Δ-0.32)	0.67 ± .03 (Δ-0.20)	0.72 ± .02 (Δ-0.16)
Consolidation	0.89 ± .02	0.88 ± .04 (Δ-0.01)	0.51 ± .04 (Δ-0.38)	0.63 ± .07 (Δ-0.26)	0.65 ± .18 (Δ-0.24)	0.72 ± .12 (Δ-0.17)
Pneumonia	0.82 ± .02	0.72 ± .09 (Δ-0.11)	0.37 ± .07 (Δ-0.45)	0.52 ± .17 (Δ-0.30)	0.60 ± .13 (Δ-0.22)	0.75 ± .04 (Δ-0.07)
Atelectasis	0.75 ± .06	0.65 ± .01 (Δ-0.10)	0.52 ± .35 (Δ-0.23)	0.52 ± .07 (Δ-0.23)	0.55 ± .01 (Δ-0.20)	0.60 ± .01 (Δ-0.15)
Pneumothorax	0.81 ± .05	0.75 ± .04 (Δ-0.06)	0.56 ± .03 (Δ-0.25)	0.69 ± .18 (Δ-0.12)	0.81 ± .06 (Δ-0.00)	0.81 ± .09 (Δ-0.00)
Pleural Effusion	0.93 ± .01	0.87 ± .01 (Δ-0.06)	0.65 ± .01 (Δ-0.28)	0.70 ± .17 (Δ-0.23)	0.78 ± .04 (Δ-0.14)	0.82 ± .04 (Δ-0.10)
Pleural Other	0.83 ± .05	0.62 ± .00 (Δ-0.21)	0.24 ± .08 (Δ-0.59)	0.44 ± .07 (Δ-0.39)	0.62 ± .11 (Δ-0.21)	0.62 ± .12 (Δ-0.21)
Fracture	0.78 ± .04	0.72 ± .10 (Δ-0.06)	0.42 ± .05 (Δ-0.36)	0.58 ± .04 (Δ-0.20)	0.55 ± .03 (Δ-0.23)	0.57 ± .04 (Δ-0.21)

Note: Avg. RRIEF represents the average performance across all RRIEF-LLaMA models (LLaMA1: 7B, 13B, 30B, 65B; LLaMA3: 8B, 70B). Avg. CheXpert & CheXbert is the average performance of CheXpert Labeler and CheXbert. Avg. LLaMA 0-shot and 3-shot represent the average performance of LLaMA1-65B and LLaMA3-70B in 0-shot and 3-shot settings, respectively. Avg. Proprietary LLMs 0-shot and 3-shot represent the average performance of GPT-4o, Gemini-1.5-Flash, and Claude-3.5-Sonnet in 0-shot and 3-shot settings, respectively. The performance metric used is macro F1 score, presented as mean ± standard deviation (e.g., 0.47 ± .04). The difference from Avg. RRIEF is expressed in parentheses using Δ. For each method, the chest X-ray finding with the lowest Δ value is highlighted in bold, while the finding with the highest Δ value is underlined. Detailed performance for each model is shown in **Table S1-S4**. RRIEF = Radiology Report Information Extraction Framework, B = Billion, Enlarged Cardiom. = Enlarged Cardiomediastinum, LLaMA = Large Language Model Meta AI.

Dataset and finding	Avg. RRIEF (Ours)	Avg. CheXpert & CheXbert	Avg. LLaMA 0-shot	Avg. LLaMA 3-shot	Avg. Proprietary LLMs 0-shot	Avg. Proprietary LLMs 3-shot
Support Devices	0.85 ± .10	0.53 ± .01 (Δ -0.32)	0.31 ± .08 (Δ -0.54)	0.49 ± .04 (Δ -0.36)	0.59 ± .09 (Δ -0.26)	0.64 ± .06 (Δ -0.21)
Average	0.85 ± .02	0.71 ± .01 (Δ -0.14)	0.42 ± .08 (Δ -0.43)	0.54 ± .09 (Δ -0.31)	0.61 ± .05 (Δ -0.24)	0.66 ± .04 (Δ -0.19)
<i>Open-i (n = 751)</i>						
Enlarged Cardiom.	0.87 ± .04	0.72 ± .04 (Δ -0.16)	0.31 ± .03 (Δ -0.56)	0.35 ± .11 (Δ -0.53)	0.51 ± .05 (Δ -0.36)	0.56 ± .03 (Δ -0.31)
Cardiomegaly	0.91 ± .01	0.78 ± .05 (Δ -0.13)	0.50 ± .06 (Δ -0.41)	0.57 ± .03 (Δ -0.34)	0.52 ± .03 (Δ -0.39)	0.60 ± .01 (Δ -0.30)
Lung Lesion	0.86 ± .07	0.81 ± .10 (Δ -0.05)	0.28 ± .05 (Δ -0.59)	0.44 ± .01 (Δ -0.43)	0.45 ± .14 (Δ -0.41)	0.60 ± .09 (Δ -0.26)
Lung Opacity	0.69 ± .04	0.71 ± .11 (Δ + 0.02)	0.40 ± .06 (Δ -0.29)	0.41 ± .01 (Δ -0.28)	0.50 ± .06 (Δ -0.18)	0.55 ± .04 (Δ -0.14)
Edema	0.81 ± .09	0.48 ± .00 (Δ -0.33)	0.26 ± .11 (Δ -0.55)	0.43 ± .07 (Δ -0.38)	0.47 ± .05 (Δ -0.34)	0.52 ± .12 (Δ -0.29)
Consolidation	0.82 ± .10	0.73 ± .06 (Δ -0.09)	0.36 ± .04 (Δ -0.46)	0.46 ± .03 (Δ -0.36)	0.48 ± .27 (Δ -0.34)	0.49 ± .16 (Δ -0.32)
Pneumonia	0.89 ± .05	0.96 ± .04 (Δ + 0.07)	0.27 ± .02 (Δ -0.63)	0.56 ± .28 (Δ -0.34)	0.54 ± .12 (Δ -0.35)	0.70 ± .10 (Δ -0.19)
Atelectasis	0.89 ± .04	0.89 ± .01 (Δ -0.01)	0.50 ± .35 (Δ -0.40)	0.52 ± .05 (Δ -0.38)	0.50 ± .02 (Δ -0.39)	0.55 ± .05 (Δ -0.34)
Pneumothorax	0.88 ± .04	0.61 ± .04 (Δ -0.27)	0.69 ± .16 (Δ -0.19)	0.81 ± .17 (Δ -0.07)	0.57 ± .10 (Δ -0.31)	0.63 ± .06 (Δ -0.25)
Pleural Effusion	0.75 ± .05	0.80 ± .01 (Δ + 0.06)	0.53 ± .13 (Δ -0.22)	0.59 ± .14 (Δ -0.16)	0.64 ± .12 (Δ -0.11)	0.72 ± .06 (Δ -0.03)
Pleural Other	0.53	0.46 ± .06	0.21 ± .16	0.43 ± .02	0.45 ± .05	0.45 ± .02

Note: Avg. RRIEF represents the average performance across all RRIEF-LLaMA models (LLaMA1: 7B, 13B, 30B, 65B; LLaMA3: 8B, 70B). Avg. CheXpert & CheXbert is the average performance of CheXpert Labeler and CheXbert. Avg. LLaMA 0-shot and 3-shot represent the average performance of LLaMA1-65B and LLaMA3-70B in 0-shot and 3-shot settings, respectively. Avg. Proprietary LLMs 0-shot and 3-shot represent the average performance of GPT-4o, Gemini-1.5-Flash, and Claude-3.5-Sonnet in 0-shot and 3-shot settings, respectively. The performance metric used is macro F1 score, presented as mean ± standard deviation (e.g., 0.47 ± .04). The difference from Avg. RRIEF is expressed in parentheses using Δ . For each method, the chest X-ray finding with the lowest Δ value is highlighted in bold, while the finding with the highest Δ value is underlined. Detailed performance for each model is shown in **Table S1-S4**. RRIEF = Radiology Report Information Extraction Framework, B = Billion, Enlarged Cardiom. = Enlarged Cardiomeastinum, LLaMA = Large Language Model Meta AI.

Dataset and finding	Avg. RRIEF (Ours)	Avg. CheXpert & CheXbert	Avg. LLaMA 0-shot	Avg. LLaMA 3-shot	Avg. Proprietary LLMs 0-shot	Avg. Proprietary LLMs 3-shot
	$\pm .10$	$(\Delta-0.07)$	$(\Delta-0.32)$	$(\Delta-0.10)$	$(\Delta-0.07)$	$(\Delta-0.08)$
Fracture	$0.82 \pm .06$	$0.57 \pm .00$ $(\Delta-0.25)$	$0.37 \pm .09$ $(\Delta-0.45)$	$0.44 \pm .06$ $(\Delta-0.38)$	$0.45 \pm .02$ $(\Delta-0.37)$	$0.46 \pm .06$ $(\Delta-0.36)$
Support Devices	$0.88 \pm .03$	$0.47 \pm .04$ $(\Delta-0.41)$	$0.32 \pm .13$ $(\Delta-0.56)$	$0.59 \pm .01$ $(\Delta-0.29)$	$0.60 \pm .04$ $(\Delta-0.28)$	$0.65 \pm .19$ $(\Delta-0.22)$
Average	$0.82 \pm .02$	$0.66 \pm .00$ $(\Delta-0.16)$	$0.38 \pm .10$ $(\Delta-0.44)$	$0.51 \pm .08$ $(\Delta-0.31)$	$0.51 \pm .08$ $(\Delta-0.30)$	$0.58 \pm .04$ $(\Delta-0.24)$
Note: Avg. RRIEF represents the average performance across all RRIEF-LLaMA models (LLaMA1: 7B, 13B, 30B, 65B; LLaMA3: 8B, 70B). Avg. CheXpert & CheXbert is the average performance of CheXpert Labeler and CheXbert. Avg. LLaMA 0-shot and 3-shot represent the average performance of LLaMA1-65B and LLaMA3-70B in 0-shot and 3-shot settings, respectively. Avg. Proprietary LLMs 0-shot and 3-shot represent the average performance of GPT-4o, Gemini-1.5-Flash, and Claude-3.5-Sonnet in 0-shot and 3-shot settings, respectively. The performance metric used is macro F1 score, presented as mean $\pm$ standard deviation (e.g., $0.47 \pm .04$). The difference from Avg. RRIEF is expressed in parentheses using Δ. For each method, the chest X-ray finding with the lowest Δ value is highlighted in bold, while the finding with the highest Δ value is underlined. Detailed performance for each model is shown in Table S1-S4. RRIEF = Radiology Report Information Extraction Framework, B = Billion, Enlarged Cardiom. = Enlarged Cardiomedastinum, LLaMA = Large Language Model Meta AI.						

Table 4 demonstrates RRIEF's performance advantages in mammography report labeling across individual findings, using the same comparative approach. Compared to zero-shot and few-shot performances of LLaMA1-65B and LLaMA3-70B, RRIEF models showed notable improvements in the internal test for 'Nodule', 'Calcification', and 'Architectural Distortion' with macro F1 score differences of 0.21–0.50, while in the external test, 'Nodule' and 'Lymph Node Enlargement' showed improvements with score differences of 0.45–0.71. Against proprietary LLMs in both settings, RRIEF models demonstrated higher F1 scores for 'Architectural Distortion' and 'Lymph Node Enlargement' across both internal and external tests, with improvements ranging from 0.11 to 0.63.

Additionally, we observed substantial performance variations among RRIEF models as detailed in Table 4. 'Skin Thickening', 'Skin Retraction', and 'Trabecular Thickening' showed considerable variability with standard deviations ranging from 0.12 to 0.30 across both datasets. **Tables S5 and S6** indicate this variability primarily stems from the lower performance of smaller models (RRIEF-LLaMA1-7B, RRIEF-LLaMA3-8B). This size-based performance disparity was most evident in the external test, where RRIEF-LLaMA1-7B scored 0.50 (95% CI: 0.49, 0.51), 0.60 (95% CI: 0.45, 0.74), and 0.24 (95% CI: 0.24, 0.24) for these three findings respectively, while RRIEF-LLaMA3-70B achieved perfect scores of 1.00 (95% CI: 1.00, 1.00) for all three findings.

Table 4
Average Performance Comparison of Methods for Mammography Report Labeling by Findings.

Dataset and finding	Avg. RRIEF (Ours)	Avg. LLaMA 0-shot	Avg. LLaMA 3-shot	Avg. Proprietary LLMs 0-shot	Avg. Proprietary LLMs 3-shot
<i>SiteA-MMG (n = 1,000)</i>					
Nodule	0.95 ± .04	0.45 ± .23 (Δ-0.50)	0.69 ± .25 (Δ-0.26)	0.77 ± .02 (Δ-0.18)	0.77 ± .01 (Δ-0.19)
Mass	0.87 ± .05	0.55 ± .17 (Δ-0.32)	0.72 ± .21 (Δ-0.16)	0.77 ± .02 (Δ-0.11)	0.78 ± .00 (Δ-0.09)
Calcification	0.97 ± .02	0.56 ± .23 (Δ-0.41)	0.76 ± .06 (Δ-0.21)	0.77 ± .02 (Δ-0.20)	0.80 ± .01 (Δ-0.17)
Asymmetry	0.90 ± .06	0.51 ± .19 (Δ-0.39)	0.80 ± .06 (Δ-0.10)	0.73 ± .03 (Δ-0.17)	0.86 ± .12 (Δ-0.04)
Architectural Dist.	0.97 ± .05	0.56 ± .17 (Δ-0.41)	0.77 ± .25 (Δ-0.20)	0.72 ± .03 (Δ-0.25)	0.81 ± .14 (Δ-0.16)
Skin Thickening	0.81 ± .13	0.55 ± .10 (Δ-0.26)	0.63 ± .10 (Δ-0.18)	0.76 ± .03 (Δ-0.05)	0.80 ± .02 (Δ-0.01)
Lymph Node Enlarge.	0.79 ± .03	0.49 ± .03 (Δ-0.30)	0.67 ± .00 (Δ-0.12)	0.56 ± .05 (Δ-0.23)	0.63 ± .08 (Δ-0.16)
Intra. Lymph Node	0.96 ± .05	0.46 ± .19 (Δ-0.50)	0.84 ± .10 (Δ-0.12)	0.73 ± .05 (Δ-0.23)	0.75 ± .01 (Δ-0.20)
Nipple Retraction	0.90 ± .05	0.53 ± .28 (Δ-0.38)	0.91 ± .08 (Δ + 0.01)	0.71 ± .02 (Δ-0.19)	0.82 ± .15 (Δ-0.08)
Skin Retraction	0.95 ± .12	0.45 ± .30 (Δ-0.51)	0.88 ± .18 (Δ-0.08)	0.69 ± .05 (Δ-0.26)	0.79 ± .08 (Δ-0.16)
Trabecular Thickening	0.78 ± .19	0.67 ± .08 (Δ-0.12)	0.67 ± .47 (Δ-0.12)	0.62 ± .15 (Δ-0.16)	0.82 ± .16 (Δ + 0.04)
Average	0.90 ± .01	0.53 ± .18 (Δ-0.37)	0.76 ± .16 (Δ-0.14)	0.71 ± .01 (Δ-0.18)	0.79 ± .06 (Δ-0.11)

Note: Avg. RRIEF represents the average performance across all RRIEF-LLaMA models (LLaMA1: 7B, 13B, 30B, 65B; LLaMA3: 8B, 70B). Avg. LLaMA 0-shot and 3-shot represent the average performance of LLaMA1-65B and LLaMA3-70B in 0-shot and 3-shot settings, respectively. Avg. Proprietary LLMs 0-shot and 3-shot represent the average performance of GPT-4o, Gemini-1.5-Flash, and Claude-3.5-Sonnet in 0-shot and 3-shot settings, respectively. The performance metric used is macro F1 score, presented as mean ± standard deviation (e.g., 0.47 ± .04). The difference from Avg. RRIEF is expressed in parentheses using Δ. For each method, the mammography finding with the lowest Δ value is highlighted in bold, while the finding with the highest Δ value is underlined. Detailed performance for each model is shown in **Table S5-S8**. RRIEF = Radiology Report Information Extraction Framework, B = Billion, Architectural Dist. = Architectural distortion, Lymph Node Enlarge. = Lymph node enlargement, Intra. Lymph Node = Intramammary lymph node, LLaMA = Large Language Model Meta AI.

Dataset and finding	Avg. RRIEF (Ours)	Avg. LLaMA 0-shot	Avg. LLaMA 3-shot	Avg. Proprietary LLMs 0-shot	Avg. Proprietary LLMs 3-shot
<i>CDD-CESM (n = 326)</i>					
Nodule	0.96 ± .06	0.46 ± .35 (Δ-0.51)	0.52 ± .29 (Δ-0.45)	0.83 ± .15 (Δ-0.13)	0.97 ± .03 (Δ + 0.01)
Mass	0.95 ± .04	0.56 ± .25 (Δ-0.40)	0.77 ± .24 (Δ-0.18)	0.88 ± .12 (Δ-0.08)	0.98 ± .01 (Δ + 0.03)
Calcification	0.92 ± .06	0.58 ± .26 (Δ-0.34)	0.81 ± .12 (Δ-0.11)	0.88 ± .12 (Δ-0.04)	0.93 ± .06 (Δ + 0.02)
Asymmetry	0.95 ± .12	0.70 ± .41 (Δ-0.25)	0.80 ± .08 (Δ-0.16)	0.72 ± .04 (Δ-0.23)	0.96 ± .08 (Δ + 0.01)
Architectural Dist.	0.96 ± .08	0.70 ± .43 (Δ-0.27)	0.87 ± .19 (Δ-0.10)	0.77 ± .04 (Δ-0.19)	0.86 ± .13 (Δ-0.11)
Skin Thickening	0.71 ± .30	0.62 ± .52 (Δ-0.09)	0.81 ± .27 (Δ + 0.10)	0.77 ± .15 (Δ + 0.07)	0.73 ± .08 (Δ + 0.03)
Lymph Node Enlarge.	0.94 ± .14	0.24 ± .02 (Δ-0.71)	0.25 ± .01 (Δ-0.70)	0.31 ± .06 (Δ-0.63)	0.53 ± .11 (Δ-0.42)
Intra. Lymph Node	0.86 ± .22	0.62 ± .34 (Δ-0.24)	0.76 ± .30 (Δ-0.10)	0.86 ± .12 (Δ-0.00)	0.98 ± .03 (Δ + 0.13)
Nipple Retraction	0.90 ± .15	0.71 ± .42 (Δ-0.19)	0.95 ± .05 (Δ + 0.05)	0.83 ± .15 (Δ-0.07)	1.00 ± .00 (Δ + 0.10)
Skin Retraction	0.85 ± .15	0.60 ± .57 (Δ-0.25)	0.76 ± .34 (Δ-0.09)	0.69 ± .08 (Δ-0.16)	0.85 ± .27 (Δ-0.00)
Trabecular Thickening	0.64 ± .30	0.66 ± .49 (Δ + 0.01)	0.83 ± .25 (Δ + 0.18)	0.86 ± .12 (Δ + 0.22)	1.00 ± .00 (Δ + 0.36)
Average	0.88 ± .10	0.58 ± .37 (Δ-0.30)	0.74 ± .19 (Δ-0.14)	0.76 ± .04 (Δ-0.11)	0.89 ± .03 (Δ + 0.01)
Note: Avg. RRIEF represents the average performance across all RRIEF-LLaMA models (LLaMA1: 7B, 13B, 30B, 65B; LLaMA3: 8B, 70B). Avg. LLaMA 0-shot and 3-shot represent the average performance of LLaMA1-65B and LLaMA3-70B in 0-shot and 3-shot settings, respectively. Avg. Proprietary LLMs 0-shot and 3-shot represent the average performance of GPT-4o, Gemini-1.5-Flash, and Claude-3.5-Sonnet in 0-shot and 3-shot settings, respectively. The performance metric used is macro F1 score, presented as mean ± standard deviation (e.g., 0.47 ± .04). The difference from Avg. RRIEF is expressed in parentheses using Δ. For each method, the mammography finding with the lowest Δ value is highlighted in bold, while the finding with the highest Δ value is underlined. Detailed performance for each model is shown in Table S5-S8. RRIEF = Radiology Report Information Extraction Framework, B = Billion, Architectural Dist. = Architectural distortion, Lymph Node Enlarge. = Lymph node enlargement, Intra. Lymph Node = Intramammary lymph node, LLaMA = Large Language Model Meta AI.					

Report Labeling Performance on Coronary CT Angiography

As shown in Table 5, RRIEF-LLaMA3-8B demonstrated significant performance improvements in both internal and external tests compared to its 3-shot counterpart and LLaMA3-70B across all categories: stenosis severity, plaque burden, and modifiers. When compared to DeepSeek-R1-Distill-Qwen-14B, RRIEF-LLaMA3-8B showed significant performance advantages in all categories with the sole exception of plaque burden assessment in the external test. Against proprietary LLMs, RRIEF-LLaMA3-8B exhibited significantly superior performance showing F1 score of 0.87 in stenosis severity labeling compared to Gemini-1.5-Flash in the internal test (0.83, $P = .02$) and against GPT-4o in the external test (0.68, $P < .001$). Notably, for modifiers, RRIEF-LLaMA3-8B achieved significantly higher performance than all proprietary models in the external test, with an F1 score of 1.00 compared to 0.93 for the best-performing proprietary model ($P = .004$).

Table 5
Performance on Coronary CT Angiography Reports.

Method	Model	SiteB-CCTA (n = 50)			SiteC-CCTA (n = 51)		
		Stenosis Severity	Plaque Burden	Modifiers (averaged)	Stenosis Severity	Plaque Burden	Modifiers (averaged)
3-shot	LLaMA3-8B	0.31*	0.77*	0.84*	0.41*	0.54*	0.91*
		(0.29, 0.34)	(0.72, 0.82)	(0.81, 0.87)	(0.35, 0.47)	(0.48, 0.59)	(0.89, 0.94)
	LLaMA3-70B	0.61*	0.85*	0.99*	0.74*	0.44*	0.94*
		(0.56, 0.65)	(0.81, 0.90)	(0.98, 1.00)	(0.68, 0.81)	(0.35, 0.54)	(0.91, 0.97)
	DeepSeek-R1-Distill-Qwen-14B	0.67*	0.84*	0.99*	0.73*	0.94	0.90*
		(0.63, 0.70)	(0.79, 0.89)	(0.98, 1.00)	(0.67, 0.79)	(0.93, 0.96)	(0.87, 0.93)
	GPT-4o	0.89	1.00	1.00	0.68*	1.00	0.93*
		(0.85, 0.93)	(1.00 , 1.00)	(1.00 , 1.00)	(0.63, 0.73)	(1.00 , 1.00)	(0.91, 0.96)
	Gemini-1.5-Flash	0.83*	1.00	1.00	0.91	1.00	0.91*
		(0.77, 0.88)	(1.00 , 1.00)	(1.00 , 1.00)	(0.88, 0.94)	(1.00 , 1.00)	(0.89, 0.94)
	Claude-3.5-Sonnet	0.99	1.00	0.92*	0.94	0.93	0.93*
		(0.97 , 1.00)	(1.00 , 1.00)	(0.91, 0.92)	(0.91 , 0.98)	(0.91, 0.95)	(0.91, 0.95)
RRIEF (Ours)	LLaMA3-8B	0.87	0.98	1.00	0.83	0.93	1.00
		(0.83, 0.91)	(0.96, 0.99)	(1.00 , 1.00)	(0.77, 0.88)	(0.90, 0.96)	(1.00 , 1.00)

Note: Data represent the **Macro F1** Score and 95% Confidence Interval from 10 Bootstrap Tests. For each column, the highest performance is shown in bold and the second-highest is underlined. Values marked with an asterisk (*) indicate statistically significant improvements of RRIEF-LLaMA3-8B compared to each model. 'Modifiers (average)' represents the simple average of F1 scores for E, I, N, G, HRP, and S. Detailed performance for each model is shown in **Table S9** and **S10**. RRIEF = Radiology Report Information Extraction Framework, LLaMA = Large language model Meta AI.

Effect of Training Sample Size on Model Performance

Figure 4 demonstrates a strong positive correlation between training data size and performance of RRIEF-LLaMA3-8B on the MIMIC-CXR dataset (Pearson's correlation coefficient 0.75, $P < .001$). Notably, with as few as 200 training reports, RRIEF-LLaMA3-8B achieved a macro F1 score of 0.79 (95% CI: 0.77, 0.82) and surpassed all baseline models including CheXbert and 3-shot configurations of Claude-3.5-

Sonnet, LLaMA3-70B, and DeepSeek-R1-Distill-Qwen-14B ($P < .001$ for all comparisons). Performance began to plateau around 500 training reports, where RRIEF-LLaMA3-8B achieved an F1 score of 0.84 (95% CI: 0.82, 0.86), with only marginal improvement observed when increasing to 1,000 reports (0.85, 95% CI: 0.85, 0.86).

Discussion

In this study, we developed and validated an efficient framework for information extraction from different types of radiology reports. Our approach achieved three key objectives. First, we established a privacy-preserving framework that significantly outperformed existing methodologies, achieving F1 scores of 0.87 and 0.85 in internal and external tests for chest X-ray reports, surpassing CheXpert Labeler, CheXbert, and all proprietary models ($P < .001$). Second, we demonstrated the framework's effectiveness across different imaging modalities with distinct characteristics, showing high performance in mammography (F1 scores: 0.91 and 0.99 in internal and external tests) and coronary CT angiography reports (F1 scores: 0.87 for stenosis severity, 0.98 for plaque burden, and 1.00 for modifiers in internal testing). Third, to facilitate widespread adoption, we analyzed the relationship between training data size and performance, finding that 200–500 annotated reports were sufficient to achieve superior results compared to specialized methods and proprietary models ($P < .001$), while providing our implementation as an open-source resource.

Recent studies have highlighted the potential of LLMs for extracting information from radiology reports, yet most prior approaches relied on zero-shot or few-shot methods without additional training (13, 29, 31). These attempts yielded performance levels inadequate for clinical implementation (22), as confirmed by our comparative analysis where even leading proprietary models like GPT-4o, Gemini-1.5-Flash, and Claude-3.5-Sonnet underperformed significantly. Additionally, previous methodologies have predominantly focused on reports from single modalities, particularly CXR (5, 6, 30), creating siloed solutions with limited transferability to different imaging types and reporting formats. RRIEF addresses these limitations through parameter-efficient fine-tuning of open-source LLMs, enabling local deployment while preserving patient privacy.

The performance variations observed across different findings in our mammography experiments highlight the impact of model size on generalizability when applying RRIEF. We found that smaller models like RRIEF-LLaMA1-7B showed considerable variability in handling certain findings such as 'Skin Thickening', 'Skin Retraction', and 'Trabecular Thickening', achieving F1 scores as low as 0.24 in the external test set. In contrast, larger models like RRIEF-LLaMA3-70B attained perfect scores of 1.00 for these same findings. This disparity reflects larger models' superior ability to bridge discrepancies between training and test datasets, particularly when faced with terminology variations or uncommon findings. For instance, certain findings had varying expressions across datasets (e.g., 'intramammary lymph node' versus 'IMLN'), while others were absent from the training set but present in the test set. The robust performance of larger models in these challenging scenarios demonstrates their enhanced capacity to generalize and adapt to novel or variant terms. These findings underscore the advantages of

our RRIEF framework when implemented with larger models, particularly for handling diverse medical terminologies and unseen concepts—a crucial capability for real-world medical practices where terminology and reporting styles vary substantially across institutions, radiologists, and over time.

Our analysis also reveals a positive correlation between training sample size and performance on the MIMIC-CXR dataset, with our approach outperforming all baselines using 200 training samples. While performance improves with more training data, we observed diminishing returns beyond 500 reports for CXR labeling. This efficiency extends across modalities, as demonstrated by RRIEF achieving comparable performance to leading proprietary models on CCTA datasets using only 100 training reports.

To implement RRIEF for new datasets, users need only follow a straightforward process: first, define appropriate findings and labels according to clinical requirements or specialty guidelines. This flexibility to freely set findings and labels underscores the method's extensive scalability, as demonstrated in our mammography implementation where we created specialized laterality labels (e.g., 'right breast', 'bilateral breasts') to capture this critical diagnostic information. Second, annotate a reasonable number of reports (200–500 based on our experiments), which can be expedited by using off-the-shelf LLMs for initial annotations followed by expert review. Finally, apply parameter-efficient fine-tuning using techniques like QLoRA (41), which dramatically reduces computational requirements while maintaining performance—enabling even the largest models (e.g., 70B parameters) to be fine-tuned on a single 48GB GPU. Our publicly shared implementation code allows easy adoption of this framework, enabling customized report labeling solutions across diverse clinical settings without specialized expertise.

However, this study faces the limitation of having a relatively small dataset for testing. Each dataset contains 1,000 reports or fewer, with a large proportion—around 20%—comprising normal cases devoid of any abnormalities. This condition leads to the presence of numerous rare findings, each with fewer than ten positive cases, such as 'Pneumothorax' in CXR and 'Skin Retraction' in mammography. Furthermore, the diversity of labels assigned to each finding, which totals four for CXR and five for mammography, may somewhat diminish the statistical rigor of our reported performance for these rare findings.

In conclusion, RRIEF demonstrates high-performance information extraction across chest X-ray, mammography, and coronary CT angiography reports while preserving patient privacy through locally deployed fine-tuned LLMs. Our framework significantly outperformed existing specialized methods and proprietary LLMs with minimal training requirements. We anticipate that our method will serve as an effective automatic report labeling strategy for various imaging modalities and reporting styles, facilitating large-scale retrospective studies and offering a practical tool to advance radiological research and clinical practice.

Declarations

Data Availability

For the publicly available datasets (MIMIC-CXR, Open-i, CDD-CESM), we have shared the annotated labels for the radiology reports used in our training and testing sets. The SiteA-MMG dataset contains identifiable protected health information and therefore cannot be shared publicly. For the SiteB-CCTA, we have made both the reports and annotated labels publicly available. The SiteC-CCTA dataset is not publicly available but can be shared upon request to the corresponding author. Data are available at https://github.com/reonaledo/report_labeler.

Code Availability

The analytical code for reproducing the results of this study and a user-friendly implementation that can be adapted for specific research purposes are currently in development. The complete implementation code will be made publicly available at https://github.com/reonaledo/report_labeler upon publication of this manuscript.

Acknowledgements

This study was supported by the National Research Foundation of Korea (NRF) grants funded by the Ministry of Science and ICT (MSIT) (Grant No. RS-2024-00354666) and the Seoul National University Hospital Research Fund (Grant No. 03-2023-0410). The funders played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Author Contribution

Conceptualization: D.M, C.M.P, J.M.C Supervision: C.M.P, J.M.C Writing: D.M Data acquisition: K.N.J, S.B, W.G.J, J.M.C, S.K Data analysis: D.M, S.K Critical review: S.H, J.C, J.M.C, C.M.P All authors read and approved the final manuscript and had final responsibility for the decision to submit it for publication.

References

1. Kong HJ. Managing Unstructured Big Data in Healthcare System. *Healthc Inform Res.* 2019;25(1):1–2.
2. Casey A, Davidson E, Poon M, Dong H, Duma D, Grivas A, et al. A systematic review of natural language processing applied to radiology reports. *BMC Med Inform Decis Mak.* 2021;21(1):179.
3. Reichenpfader D, Müller H, Denecke K. A scoping review of large language model based approaches for information extraction from radiology reports. *Npj Digit Med.* 2024;7(1):1–12.
4. Pivovarov R, Coppleson YJ, Gorman SL, Vawdrey DK, Elhadad N. Can Patient Record Summarization Support Quality Metric Abstraction? *AMIA Annu Symp Proc AMIA Symp.* 2016;2016:1020–9.
5. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison [Internet]. *arXiv*; 2019 [cited 2024 Jan 4]. Available from: <http://arxiv.org/abs/1901.07031>

6. Smit A, Jain S, Rajpurkar P, Pareek A, Ng AY, Lungren MP. CheXbert: Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT [Internet]. arXiv; 2020 [cited 2024 Jan 4]. Available from: <http://arxiv.org/abs/2004.09167>
7. Kuling G, Curpen B, Martel AL. BI-RADS BERT and Using Section Segmentation to Understand Radiology Reports. *J Imaging*. 2022;8(5):131.
8. Torres-Lopez VM, Rovenolt GE, Olcese AJ, Garcia GE, Chacko SM, Robinson A, et al. Development and Validation of a Model to Identify Critical Brain Injuries Using Natural Language Processing of Text Computed Tomography Reports. *JAMA Netw Open*. 2022;5(8):e2227109.
9. Wood DA, Lynch J, Kafiabadi S, Guilhem E, Busaidi AA, Montvila A, et al. Automated Labelling using an Attention model for Radiology reports of MRI scans (ALARM) [Internet]. arXiv; 2020 [cited 2024 Nov 20]. Available from: <http://arxiv.org/abs/2002.06588>
10. Wood DA, Kafiabadi S, Al Busaidi A, Guilhem EL, Lynch J, Townend MK, et al. Deep learning to automate the labelling of head MRI datasets for computer vision applications. *Eur Radiol*. 2022;32(1):725–36.
11. Zech J, Pain M, Titano J, Badgeley M, Schefflein J, Su A, et al. Natural Language–based Machine Learning Models for the Annotation of Clinical Radiology Reports. *Radiology*. 2018;287(2):570–80.
12. Adams LC, Truhn D, Busch F, Kader A, Niehues SM, Makowski MR, et al. Leveraging GPT-4 for Post Hoc Transformation of Free-text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study. *Radiology*. 2023;307(4):e230725.
13. Mukherjee P, Hou B, Lanfredi RB, Summers RM. Feasibility of Using the Privacy-preserving Large Language Model Vicuna for Labeling Radiology Reports. *Radiology*. 2023;309(1):e231147.
14. Wiest IC, Ferber D, Zhu J, van Treeck M, Meyer SK, Juglan R, et al. Privacy-preserving large language models for structured medical information retrieval. *Npj Digit Med*. 2024;7(1):1–9.
15. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models [Internet]. arXiv; 2023 [cited 2024 Jan 5]. Available from: <http://arxiv.org/abs/2307.09288>
16. Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, et al. arXiv.org. 2024 [cited 2024 Sep 11]. The Llama 3 Herd of Models. Available from: <https://arxiv.org/abs/2407.21783v2>
17. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. GPT-4 Technical Report [Internet]. arXiv; 2024 [cited 2024 Mar 14]. Available from: <http://arxiv.org/abs/2303.08774>
18. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. LLaMA: Open and Efficient Foundation Language Models [Internet]. arXiv; 2023 [cited 2024 Jan 5]. Available from: <http://arxiv.org/abs/2302.13971>
19. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language Models are Few-Shot Learners. In: *Advances in Neural Information Processing Systems* [Internet]. Curran Associates, Inc.; 2020 [cited 2024 Mar 14]. p. 1877–901. Available from: <https://papers.nips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>

20. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. *Commun Med*. 2023;3(1):1–8.
21. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge [Internet]. *arXiv*; 2023 [cited 2024 Mar 15]. Available from: <http://arxiv.org/abs/2303.14070>
22. Bhayana R. Chatbots and Large Language Models in Radiology: A Practical Primer for Clinical and Research Applications. *Radiology*. 2024;310(1):e232756.
23. Ueda D, Mitsuyama Y, Takita H, Horiuchi D, Walston SL, Tatekawa H, et al. ChatGPT's Diagnostic Performance from Patient History and Imaging Findings on the Diagnosis Please Quizzes. *Radiology*. 2023;308(1):e231040.
24. Amin KS, Davis MA, Doshi R, Haims AH, Khosla P, Forman HP. Accuracy of ChatGPT, Google Bard, and Microsoft Bing for Simplifying Radiology Reports. *Radiology*. 2023;309(2):e232561.
25. Bera K, O'Connor G, Jiang S, Tirumani SH, Ramaiya N. Analysis of ChatGPT publications in radiology: Literature so far. *Curr Probl Diagn Radiol*. 2024;53(2):215–25.
26. Fink MA, Bischoff A, Fink CA, Moll M, Kroschke J, Dulz L, et al. Potential of ChatGPT and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer. *Radiology*. 2023;308(3):e231362.
27. Li D, Gupta K, Chong J. Evaluating Diagnostic Performance of ChatGPT in Radiology: Delving into Methods. *Radiology*. 2023;308(3):e232082.
28. Ramasamy SK. Response to Performance of ChatGPT on a Radiology Board-style Examination. *Radiology*. 2023;307(5):e231330.
29. Lehnen NC, Dorn F, Wiest IC, Zimmermann H, Radbruch A, Kather JN, et al. Data Extraction from Free-Text Reports on Mechanical Thrombectomy in Acute Ischemic Stroke Using ChatGPT: A Retrospective Analysis. *Radiology*. 2024;311(1):e232741.
30. Gu J, Cho HC, Kim J, You K, Hong EK, Roh B. CheX-GPT: Harnessing Large Language Models for Enhanced Chest X-ray Report Labeling [Internet]. *arXiv*; 2024 [cited 2024 Apr 19]. Available from: <http://arxiv.org/abs/2401.11505>
31. Dorfner FJ, Jürgensen L, Donle L, Al Mohamad F, Bodenmann TR, Cleveland MC, et al. Comparing Commercial and Open-Source Large Language Models for Labeling Chest Radiograph Reports. *Radiology*. 2024;313(1):e241139.
32. Minssen T, Vayena E, Cohen IG. The Challenges for Regulating Medical Use of ChatGPT and Other Large Language Models. *JAMA*. 2023;330(4):315–6.
33. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *Npj Digit Med*. 2023;6(1):1–6.
34. Wu C, Lin W, Zhang X, Zhang Y, Wang Y, Xie W. PMC-LLaMA: Towards Building Open-source Language Models for Medicine [Internet]. *arXiv*; 2023 [cited 2024 Mar 15]. Available from: <http://arxiv.org/abs/2304.14454>

35. Raeini M. Privacy-Preserving Large Language Models (PPLLMs) [Internet]. Rochester, NY: Social Science Research Network; 2023 [cited 2024 Nov 21]. Available from: <https://papers.ssrn.com/abstract=4512071>
36. Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng C ying, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data*. 2019;6(1):317.
37. Demner-Fushman D, Kohli MD, Rosenman MB, Shooshan SE, Rodriguez L, Antani S, et al. Preparing a collection of radiology examinations for distribution and retrieval. *J Am Med Inform Assoc JAMIA*. 2016;23(2):304–10.
38. Khaled R, Helal M, Alfarghaly O, Mokhtar O, Elkorany A, El Kassas H, et al. Categorized contrast enhanced mammography dataset for diagnostic and artificial intelligence research. *Sci Data*. 2022;9(1):122.
39. D’Orsi CJ, Sickles EA, Mendelson EB, Morris EA. 2013 ACR BI-RADS Atlas: Breast Imaging Reporting and Data System [Internet]. American College of Radiology; 2014. Available from: <https://books.google.co.kr/books?id=nhWSjwEACAAJ>
40. CAD-RADS™ 2.0–2022 Coronary Artery Disease-Reporting and Data System - *Journal of Cardiovascular Computed Tomography* [Internet]. [cited 2024 Nov 21]. Available from: [https://www.journalofcardiovascularct.com/article/S1934-5925\(22\)00240-4/fulltext](https://www.journalofcardiovascularct.com/article/S1934-5925(22)00240-4/fulltext)
41. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRA: Efficient Finetuning of Quantized LLMs [Internet]. arXiv; 2023 [cited 2024 Jan 5]. Available from: <http://arxiv.org/abs/2305.14314>
42. DeepSeek-AI, Guo D, Yang D, Zhang H, Song J, Zhang R, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning [Internet]. arXiv; 2025 [cited 2025 Feb 14]. Available from: <http://arxiv.org/abs/2501.12948>

Figures

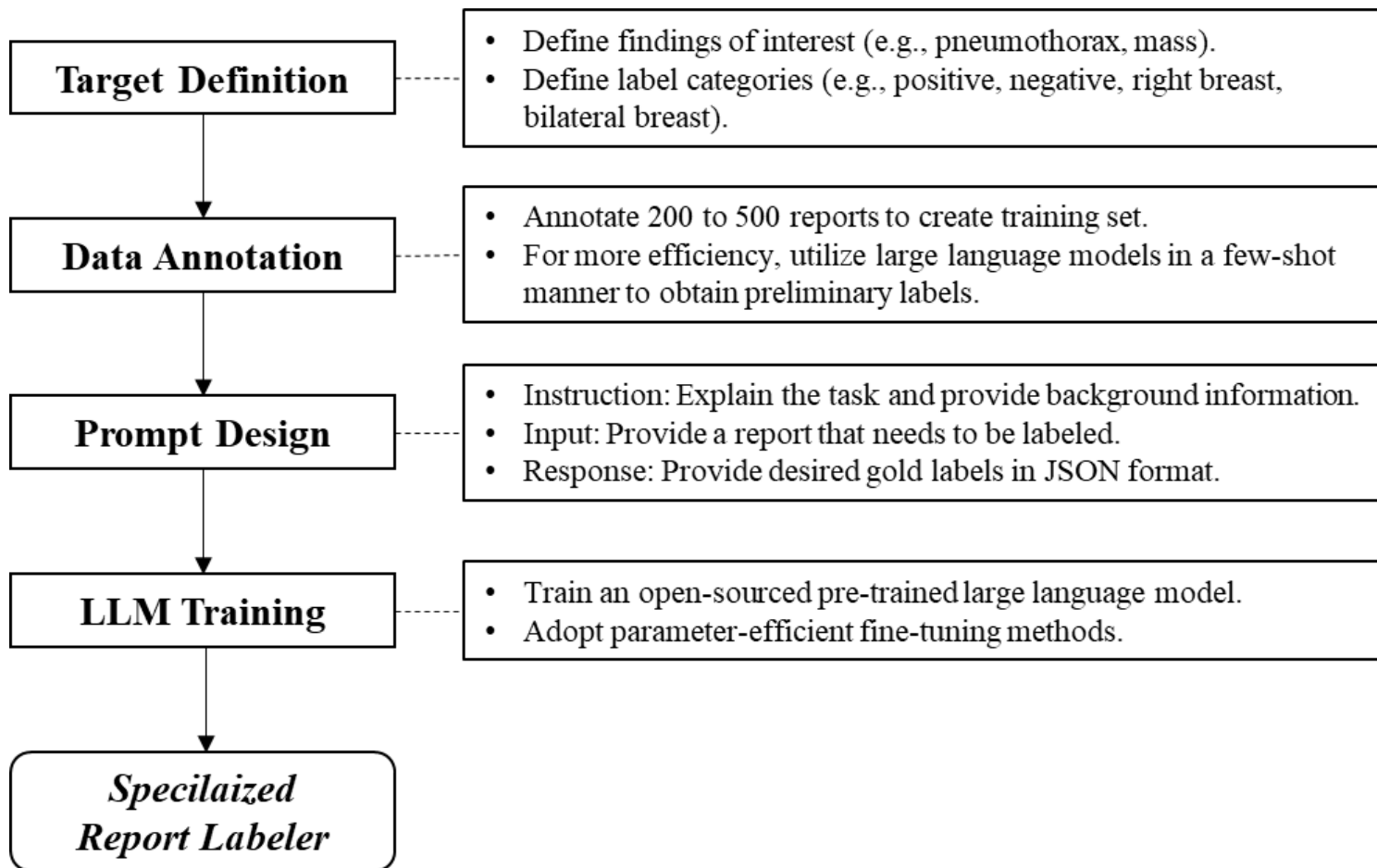


Figure 1

An overview of the proposed Radiology Report Information Extraction Framework (RRIEF). LLM = Large Language Model.

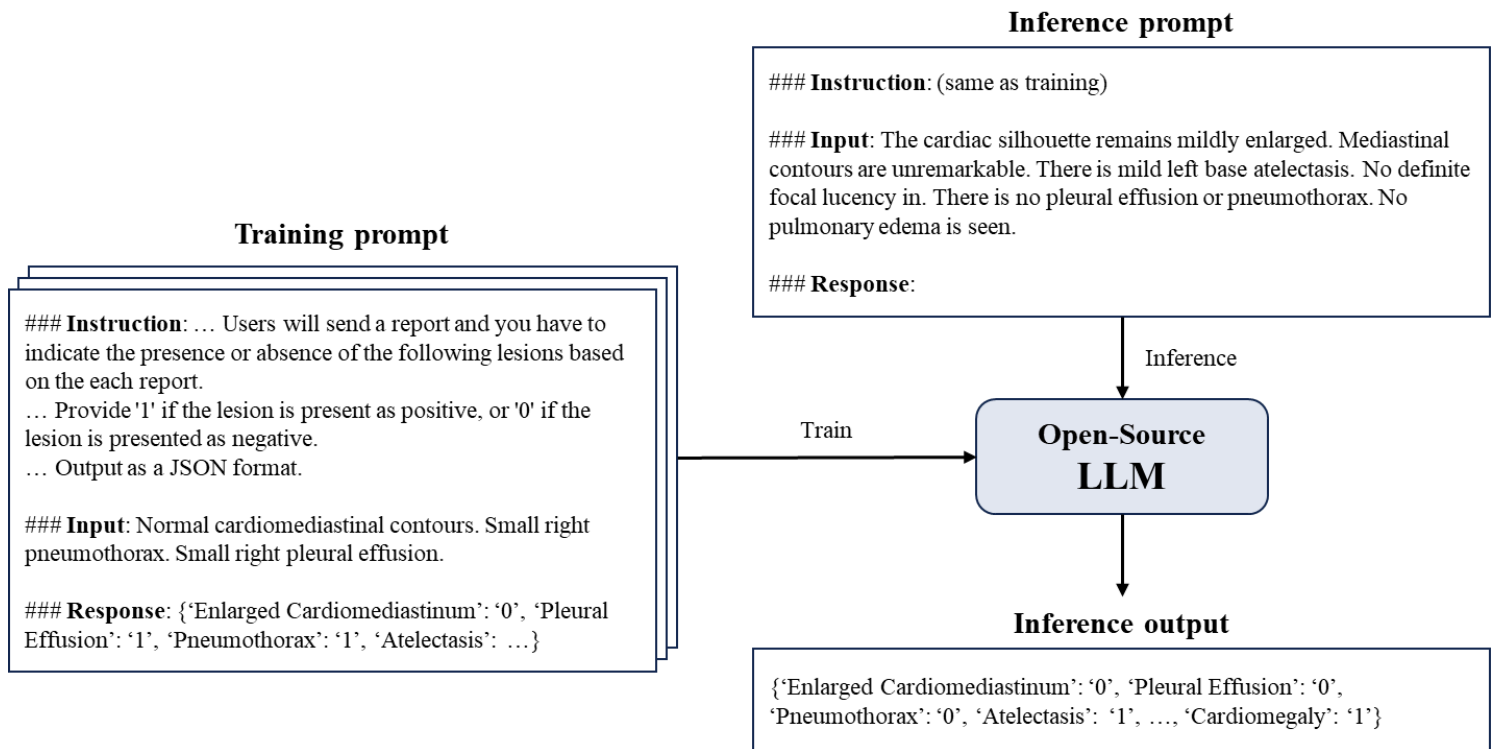


Figure 2

Example of a prompt designed for training and inference phases within the Radiology Report Information Extraction Framework (RRIEF). The instructions provided describe the task which requires generating outputs in JSON format upon receiving reports. The input consists of a target report. When the trained language model receives identical instructions and a report to be labeled, it generates labels in JSON format. In the figure, '0' represents a label 'negative', while '1' represents 'positive'. B = billion, LLM = Large Language Model.

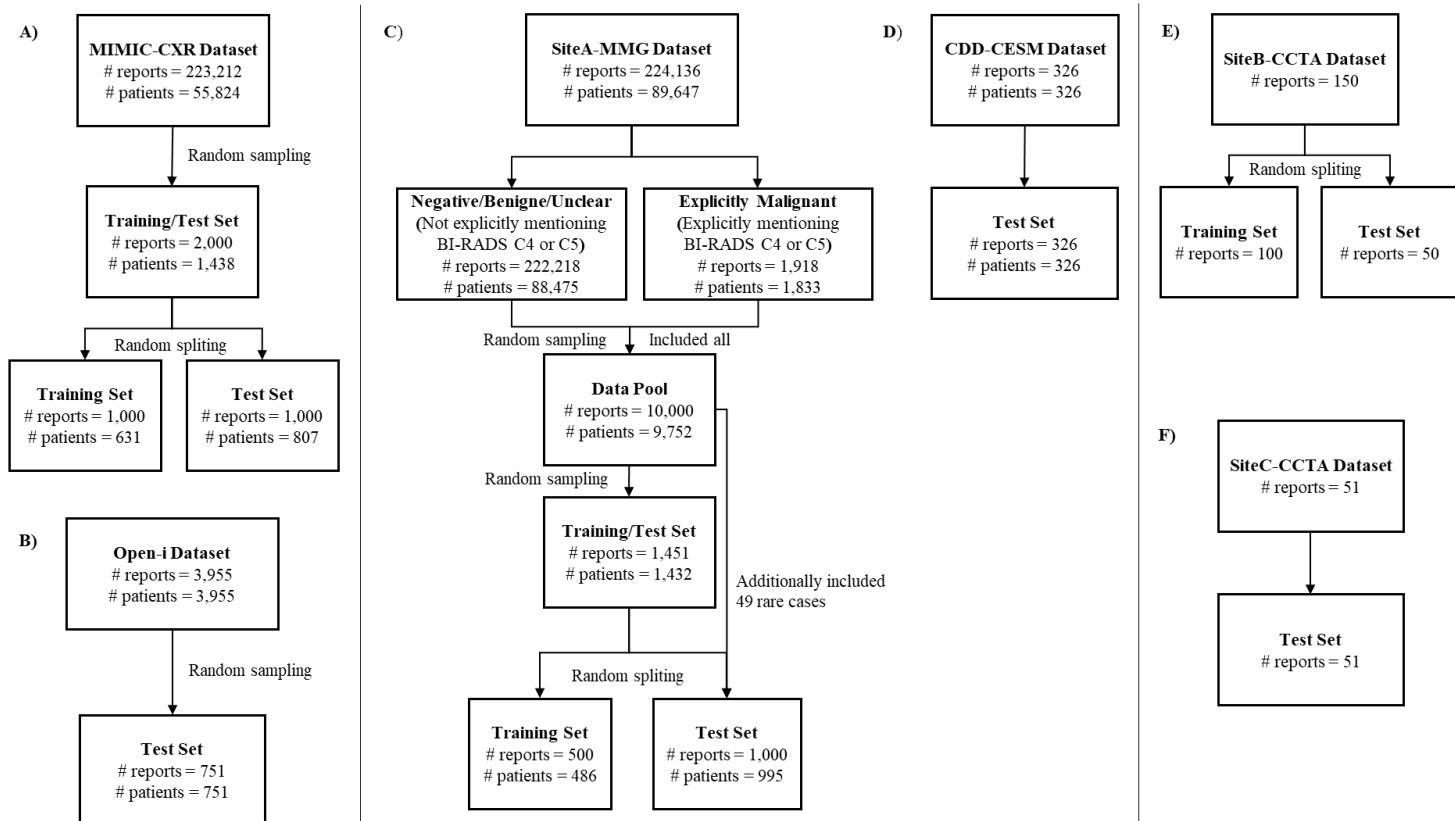


Figure 3

Flowchart shows the data selection process from the (A) Chest X-rays from medical information mart for intensive care (MIMIC-CXR), (B) Open-i, (C) SiteA-MMG, (D) Categorized digital database for low energy and subtracted contrast enhanced spectral mammography images (CDD-CESM), (E) SiteB-CCTA and (F) SiteC-CCTA. For CDD-CESM, conventional mammography report was used. BI-RADS = Breast imaging-reporting and data system.

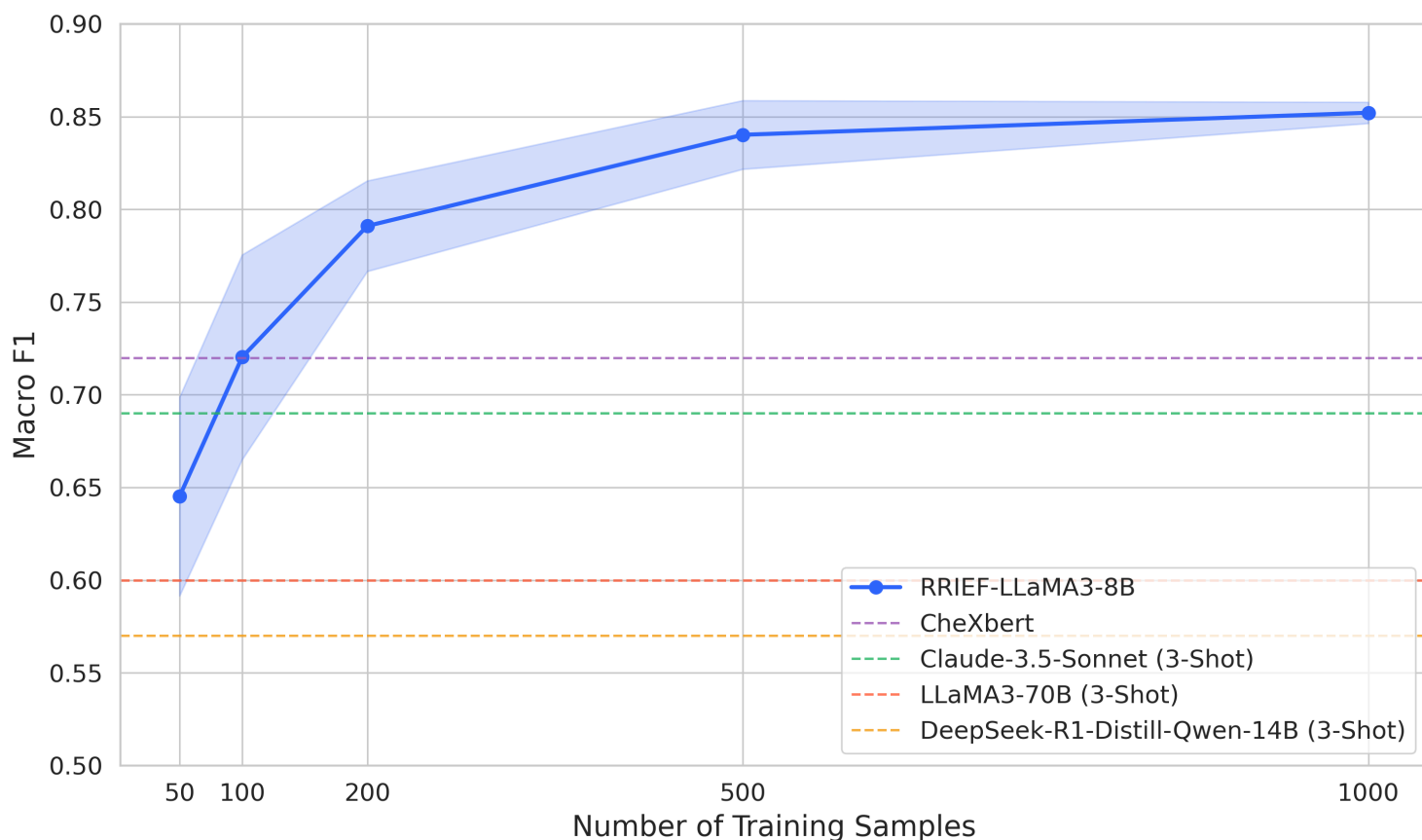


Figure 4

Impact of training data size on model performance. The proposed approach was applied to LLaMA3-8B and evaluated on the MIMIC-CXR dataset with varying training data sizes (50, 100, 200, 500, and 1000 reports). The solid line represents the mean performance across five repetitions with different random samples, while the shaded region indicates the 95% confidence interval. For comparison, results from baseline models are shown: CheXbert and three models in 3-shot setting (Claude-3.5-Sonnet, LLaMA3-70B, and DeepSeek-R1-Distill-Qwen-14B). RRIEF = Radiology Report Information Extraction Framework, B = billion, LLaMA = Large language model Meta AI.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementary.docx](#)