SICAR at RRG2024: GPU Poor's Guide to Radiology Report Generation

Kiartnarin Udomlapsakul[1,†], **Parinthapat Pengpun**[2,†], **Tossaporn Saengja**[3], **Kanyakorn Veerakanjana**[1,4,5]
Krittamate Tiankanon[1], **Pitikorn Khlaisamniang**[7], **Pasit Supholkhan**[7], **Amrest Chinkamol**[3]
Pubordee Aussavavirojekul[1], **Hirunkul Phimsiri**[1], **Tara Sripo**[7], **Chiraphat Boonnag**[7], **Trongtum Tongdee**[7]
Thanongchai Siriapisith[7], **Pairash Saiviroonporn**[7], **Jiramet Kinchagawat**[1,‡], **Piyalitt Ittichaiwong**[1,4,6,‡]

[1] PreceptorAI team, CARIVA Thailand, [2] Bangkok Christian International School,
[3] Vidyasirimedhi Institute of Science and Technology (VISTEC),
[4] Siriraj Informatics and Data Innovation Center (SIData+), Faculty of Medicine, Siriraj Hospital, Mahidol University,
[5] Social, Genetic and Developmental Psychiatry Centre,
Institute of Psychiatry, Psychology and Neuroscience, King's College London,
[6] School of Biomedical Engineering & Imaging Sciences, King College London,
[7] Radiology Department, Faculty of Medicine, Siriraj Hospital, Mahidol University
[†] Equal first contributions, [‡] Corresponding authors.

Abstract

Radiology report generation (RRG) aims to create free-text radiology reports from clinical imaging. Our solution employs a lightweight multimodal language model (MLLM) enhanced with a two-stage post-processing strategy, utilizing a Large Language Model (LLM) to boost diagnostic accuracy and ensure patient safety. We introduce the **"First, Do No Harm" SafetyNet**, which incorporates X-Raydar, an advanced X-ray classification model, to cross-verify the model outputs and specifically address false negative errors from the MLLM. This comprehensive approach combines the efficiency of lightweight models with the robustness of thorough post-processing techniques, offering a reliable solution for radiology report generation. Our system achieved fourth place on the F1-Radgraph metric for findings generation in the Radiology Report Generation Shared Task (RRG24).[1]

1 Introduction

Radiology is indispensable in healthcare, offering non-invasive methods to diagnose and monitor medical conditions. Central to this practice are radiology reports, which provide detailed interpretations of medical images crucial for clinical decision-making (Mityul et al., 2018). However, writing these reports is a meticulous process that demands significant domain expertise (Hartung et al., 2020). Radiologists must manually review images and formulate descriptive narratives, a task that is not only time-consuming but also susceptible to variability and errors, potentially affecting patient care and outcomes (Alexander et al., 2022).

One of the primary challenges in radiology report writing is the sheer volume of imaging studies that radiologists must interpret (Bruls and Kwee, 2020; Zhan et al., 2020). With the increasing use of imaging modalities such as computed tomography (CT), magnetic resonance imaging (MRI), and X-ray, radiologists are facing a growing workload that exceeds their capacity to provide timely and accurate reports (Winder et al., 2021; Bruls and Kwee, 2020). This challenge is further compounded by the rising demand for imaging services due to an aging population and the increasing prevalence of chronic diseases.

Another imperative issue in radiology report generation is the variability in report quality and consistency (Minn et al., 2015; Pool and Goergen, 2010). Different radiologists may interpret the same set of images differently, leading to inconsistencies in the information provided in the reports. This variability can stem from differences in writing styles, experience levels, and individual biases, all of which can have significant implications for patient care (Plumb et al., 2009; Naik et al., 2001; Brady et al., 2012). Inconsistencies in reports may lead to missed diagnoses or incorrect treatment decisions, underscoring the importance of standardized and automated approaches to report generation.

To address these challenges, Automated systems

[1] https://stanford-aimi.github.io/RRG24/

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 635–644

have the potential to enhance the efficiency and accuracy of radiology report generation (Liao et al., 2023; Pang et al., 2023; Liu et al., 2023). These systems can reduce the time and effort required by radiologists while standardizing reporting practices to ensure consistency and relevance in reports. Moreover, automation can help address the increasing workload and demand for imaging services.

As Large Language Models (LLMs) have become widely available, numerous studies have explored the development of Multimodal LLMs (MLLMs) capable of natively processing additional modalities, such as images (Lu et al., 2023; Yang et al., 2023). Although there have been significant advancements in the development of MLLMs for various tasks (Chen et al., 2023; Wu et al., 2023), none have specifically focused on lightweight models for the medical domain.

Local deployment is critical as many hospitals are concerned that uploading images to the cloud for AI processing may violate privacy laws such as the General Data Protection Regulation (GDPR) in Europe or the Personal Data Protection Act (PDPA) in Thailand. Addressing this issue is essential to ensure that patients can receive enhanced medical services while maintaining their privacy. Additionally, most hospitals in developing countries are GPU-constrained and lack access to high-end GPUs which are typically required for deployment. Therefore, it is imperative to develop lightweight models capable of performing inference on-premise using consumer-grade GPUs.

Motivated by these challenges, we investigate various architectures with a focus on identifying models that offer the optimal cost-to-performance ratio for local deployment. For the purposes of this study, we concentrate on the task of findings generation.

Our contributions are summarized as follows:

- We developed and trained a lightweight Multimodal Large Language Model (MLLM) for the radiology report generation task using a two-stage training strategy, achieving performance metrics comparable to those of larger models.
- We introduced a novel two-stage post-processing strategy. The first stage enhances the readability and clarity of the reports. The second stage, "First, Do No Harm" SafetyNet, employs the X-Raydar classification model to cross-verify the model outputs, significantly improving diagnostic accuracy and ensuring patient safety.

2 Methodology

2.1 Model Architecture

Impressed by its superior performance, which surpasses even some larger models despite its lightweight nature in general domain, we decided to follow model architecture design of Bunny for this study (He et al., 2024). Our model components include the SigLIP-so400m[2] (Zhai et al., 2023) as the visual encoder, a two-layer Multi-layer perceptron (MLP) with a GELU activation as the vision-language connector, and the Phi-2 2.7B as our LLM (Hughes, 2023).

The SigLIP visual encoder extracts meaningful features from chest X-ray images, enabling the model to capture relevant visual information. The MLP integrates these visual features with language representations. Phi-2, a 2.7 billion parameter lightweight language model trained on high-quality data, achieves performance metrics comparable to substantially larger models. It demonstrates exceptional proficiency in benchmarks such as commonsense reasoning, language comprehension, question-answering, and coding tasks, frequently surpassing models with significantly more parameters.

2.2 Training Strategy & Datasets

We employ a two-stage training strategy to optimize our model's performance. In the first stage, we train only the MLP connector using the LLaVa-Med alignment 500k dataset (Li et al., 2023; Zhang et al., 2023), while keeping the rest of the model frozen. LLaVa-Med is a large-scale dataset specifically curated for medical vision-language tasks, containing a diverse collection of medical imaging modalities and tasks. By pretraining on this dataset, the MLP connector learns to effectively map visual features to language representations in the medical domain.

The second stage involves fine-tuning both the vision-language connector and the Language Model (LLM), while keeping the visual encoder frozen. This fine-tuning process utilizes the interpret-cxr dataset (Xu et al., 2024) comprising a mixture of multiple chest X-ray datasets: CheXpert (Chambon et al., 2024), PadChest (Bustos et al., 2020), BIMCV COVID-19 (Vayá et al., 2020), and MIMIC-CXR-JPG (Johnson et al., 2019). This

[2] SigLIP HuggingFace Link

dataset includes chest X-ray images along with their corresponding radiology reports, providing task-specific training data. For our study, we combined both findings and impressions into a single dataset, totaling 710,807 image-text pairs. In our preliminary study, retaining only the first image from each study outperformed using all images, as shown in 3. Therefore, we heuristically kept only the first image to preserve report diversity.

2.3 Two-Stage Post-Processing Strategy

In addition to our model's architecture and training strategies, we implement a crucial post-processing strategy, wherein the model outputs undergo sequential processing to enhance the overall quality of the reports (See Appendix C for detailed prompts).

2.3.1 First stage: Report Refinement

In the first stage, we utilize a Large Language Model (LLM) to enhance the comprehensiveness of the findings reports. Our key objectives are to improve readability and clarity, eliminate nonsensical words, and remove duplicated sentences from the model hallucinations. For normal chest X-ray (CXR) findings, we provide detailed, standardized explanations to clarify the condition. We use a recommended vocabulary list to maintain consistency across reports. Our methodology promotes concise reporting by focusing on critical findings, while still adhering to a professional radiology report format. This includes transforming simple statements like "No significant findings" into comprehensive and detailed descriptions.

2.3.2 Second stage: "First, Do No Harm" SafetyNet

This post-processing strategy, termed "First, Do No Harm" SafetyNet, involves using an advanced X-ray classification model, X-Raydar, to provide a second opinion on chest X-ray images. This methodology mirrors the practice of doctors consulting with colleagues to validate the diagnoses, thereby mitigating the risk of errors that could potentially harm patients.

X-Raydar Integration X-Raydar, a state-of-the-art X-ray classification model, is trained on a substantial dataset of 1.8 million chest X-rays, covering a wide range of pathologies (Cid et al., 2024). By integrating X-Raydar into our post-processing strategy, we leverage its robust performance to cross-verify and refine the outputs generated by our MLLM.

Second Opinion Inference A major challenge in findings generation is the occurrence of false negative errors, such as incorrectly reporting "lungs are clear" or "no cardiomegaly" To mitigate this issue, we use Llama3 70B[3] with a specially designed prompt to detect and correct such critical errors. The prompt incorporates the classification results from X-Raydar to specifically address common false negative errors. For example, if X-Raydar identifies signs of cardiomegaly but the initial report states "no cardiomegaly," our tailored prompt for Llama3 ensures that the final report accurately reflects the patient's condition. This dual-check strategy significantly increases agreement with the ground truth report, thereby improving diagnostic accuracy and enhancing patient safety.

3 Experimental Setup

3.1 Evaluation

We evaluated our approach using metrics for natural language generation (NLG) quality and clinical accuracy, as implemented by the Vilmedic framework (Delbrouck et al., 2022b).

NLG Metrics

- BLEU measures the precision of n-grams in the generated text compared to a reference text (Papineni et al., 2002).
- ROUGE-L focuses on the longest common subsequence between the generated and reference texts (Lin, 2004).
- BERTscore uses contextual embeddings to compare semantic similarity between the generated and reference texts (Zhang et al., 2019).

Clinical Accuracy Metrics

- F1-CheXbert computes the F1 score based on the similarity of indicator vectors for 14 pathologies (Smit et al., 2020).
- F1-RadGraph calculates the overlap in clinical entities and relations extracted from the reports (Delbrouck et al., 2022a).

These metrics provide a comprehensive evaluation of our model's performance in generating accurate and clinically relevant radiology reports.

3.2 Model Architecture Ablations

To investigate the complex relationship between model architecture and overall performance across

[3]LLaMa3 70B Instruct HuggingFace Link

Table 1: Performance of Various LLM and Visual Encoder Combinations on the Public Findings Benchmark (One Epoch ≈ 5000 Steps)

Model	Step	BLEU4	ROUGEL	Bertscore	F1-cheXbert	F1-RadGraph
Phi-2 + SigLIP	4000	5.83	20.98	46.72	49.69	19.21
Phi-2 + SigLIP	8000	6.93	**23.41**	50.81	**55.70**	22.05
Phi-2 + SigLIP	12000	6.96	23.26	51.63	52.91	**22.86**
Phi-2 + SigLIP (S2)	4000	5.08	19.85	45.67	47.96	18.53
Phi-2 + SigLIP (S2)	8000	**7.47**	23.38	50.56	55.30	22.32
Phi-2 + SigLIP (S2)	12000	7.3	22.9	**50.82**	52.84	21.93
Llama3 (OpenBio) + SigLIP (S2)	4000	2.26	16.03	41.42	37.49	12.98
Llama3 (OpenBio) + SigLIP (S2)	8000	5.21	20.32	47.06	45.78	18.11
Llama3 (OpenBio) + SigLIP (S2)	12000	6.01	20.75	48.24	49.72	18.09

various metrics and tasks, we designed and conducted the following series of experiments to isolate specific architectural elements and their effects:

- Language Model:
  - Phi-2 2.7B
  - Llama3-OpenBioLLM 8B[4]
- Visual Encoder:
  - SigLIP
  - SigLIP with S2-Wrapper

In addition to our base model, Phi-2 2.7B with the SigLIP visual encoder, we conducted further ablation studies to understand the impact of different model architectures. For the language model (LLM), we selected Llama3-OpenBioLLM 8B as our larger model to investigate whether initializing from a medical LLM could enhance the performance of a MLLM on findings generation task. The model was fine-tuned using a comprehensive dataset of high-quality biomedical data, allowing it to comprehend and generate text with precise domain-specific accuracy and fluency. The Llama3-OpenBioLLM 8B demonstrated exceptional performance on multiple medical LLM benchmarks[5], surpassing even some larger models.

For the visual encoder, we employed the S2-Wrapper, an extension designed to extract multi-scale features from images (Shi et al., 2024). This approach was chosen to evaluate the impact of multi-scale feature extraction on the findings generation task. The integration of the S2-Wrapper aims to enhance the model's ability to handle complex visual features and improve the overall accuracy of the generated reports.

4 Results & Discussion

4.1 Model Architecture Ablations

Our best architecture, Phi-2 combined with SigLIP visual encoder, demonstrates superior performance as indicated by the F1-Radgraph metric as presented in Table 1. Notably, this configuration outperforms the S2-wrapper extension. We hypothesize that the general domain SigLIP visual encoder encounters difficulties in effectively extracting useful information from X-ray images at multiple scales. Additionally, this architecture surpasses the performance of the larger medical domain Llama3-OpenBioLLM 8B, suggesting that the success in this specific findings generation task may be more dependent on the quality of image information extracted by the visual encoder rather than the pretrained knowledge of LLMs.

4.2 Post-processing

Table 2: Performance improvement of each post-processing stage on Hidden Findings Benchmark.

Model	F1-RadGraph
Phi-2 + SigLIP	22.61
Phi-2 + SigLIP (Stage 1)	23.11 (+0.5)
Phi-2 + SigLIP (Stage 1&2)	24.62 (+1.51)

Our two-stage post-processing strategy markedly improves the performance metrics for our findings generation task, as demonstrated by the hidden-findings test results in Table 2. In the first stage, report refinement increased the F1-Radgraph metric from 22.61 to 23.11 (+0.5). The incorporation of the "First, Do No Harm" SafetyNet in the second stage further elevated the F1-Radgraph metric from 23.11 to 24.62 (+1.51), resulting in a total improvement of 2.01 points over the default model. This comprehensive approach not only enhances report readability but also significantly boosts diagnostic accuracy and patient safety, leading to higher quality radiology reports.

5 Conclusion

We present our approach to the Radiology Report Generation task in the BioNLP 2024 shared task. This study investigates various training configurations and data mixtures to develop lightweight models for generating radiology reports from chest X-ray images. Our findings demonstrate that even a smaller model, such as the Phi-2 language model, can perform comparably to larger models in the

[4]LLaMa3 OpenBioLLM 8B HuggingFace Link
[5]OpenLLM Leaderboard

report generation task. Additionally, incorporating post-processing techniques significantly enhances the quality of the reports and ensures patient safety. This is particularly crucial for hospitals in resource-constrained settings. By focusing on models that can be fine-tuned on a single A100 GPU and operated on-premises with a consumer-grade GPU, we address privacy concerns and improve the accessibility of this technology.

Limitations

In this work, we utilized the Llama3-70b-instruct model on HuggingChat for post-processing in both stages, demonstrating that it improves the metric (F1-RadGraph) of the generated reports. However, we did not explicitly analyze the quality of post-processing with smaller LLMs to determine if they can achieve similar results. Future research could explore post-processing with multiple LLM sizes to understand the impact of model size on performance. Additionally, our current approach involves sequential two-stage post-processing, which may not fully leverage the LLM's capabilities and could introduce unnecessary complexity and latency. Combining these stages into a single step could reduce latency and streamline the overall process.

Acknowledgments

This research is supported by PreceptorAI team and CARIVA Thailand. The experiments were conducted with the resources and services provided by the Siriraj Radiology Department (SIRAD). Their resources and domain expertise were instrumental in the successful completion of this study.

References

Robert Alexander, Stephen Waite, Michael A Bruno, Elizabeth A Krupinski, Leonard Berlin, Stephen Macknik, and Susana Martinez-Conde. 2022. Mandating limits on workload, duty, and speed in radiology. *Radiology*, 304(2):274–282.

Adrian P. Brady, Risteárd Ó Laoide, Peter A McCarthy, and Ronan McDermott. 2012. Discrepancy and error in radiology: Concepts, causes and consequences. *The Ulster Medical Journal*, 81:3 – 9.

RJM Bruls and RM Kwee. 2020. Workload for radiologists during on-call hours: dramatic increase in the past 15 years. *Insights into imaging*, 11:1–7.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P. Langlotz. 2024. Chexpert plus: Hundreds of thousands of aligned radiology texts, images and patients.

Ke Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *ArXiv*, abs/2306.15195.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, Emily B. Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S. Chaudhari, and Curtis Langlotz. 2024. Chexagent: Towards a foundation model for chest x-ray interpretation. *Preprint*, arXiv:2401.12208.

Yashin Dicente Cid, Matthew Macpherson, Louise Gervais-Andre, Yuanyi Zhu, Giuseppe Franco, Ruggiero Santeramo, Chee Lim, Ian Selby, Keerthini Muthuswamy, Ashik Amlani, et al. 2024. Development and validation of open-source deep neural networks for comprehensive chest x-ray reading: a retrospective, multicentre study. *The Lancet Digital Health*, 6(1):e44–e57.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022a. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360.

Jean-benoit Delbrouck, Khaled Saab, Maya Varma, Sabri Eyuboglu, Pierre Chambon, Jared Dunnmon, Juan Zambrano, Akshay Chaudhari, and Curtis Langlotz. 2022b. Vilmedic: a framework for research at the intersection of vision and language in medical ai. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 23–34.

Michael P Hartung, Ian C Bickle, Frank Gaillard, and Jeffrey P Kanne. 2020. How to create a great radiology report. *Radiographics*, 40(6):1658–1670.

Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. 2024. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*.

Alyssa Hughes. 2023. Phi-2: The surprising power of small language models — microsoft.com. [Accessed 18-05-2024].

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng.

2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*.

Yuxiang Liao, Hantao Liu, and Irena Spasic. 2023. Deep learning approaches to automatic radiology report generation: A systematic review. *Informatics in Medicine Unlocked*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chang Liu, Yuanhe Tian, and Yan Song. 2023. A systematic review of deep learning-based research on radiology report generation. *ArXiv*, abs/2311.14199.

Yuzhe Lu, Sungmin Hong, Yash Shah, and Panpan Xu. 2023. Effectively fine-tune to improve large multimodal models for radiology report generation. *ArXiv*, abs/2312.01504.

Matthew J Minn, Arash R Zandieh, and Ross W Filice. 2015. Improving radiology report quality by rapidly notifying radiologist of report errors. *Journal of digital imaging*, 28:492–498.

Marina I Mityul, Brian Gilcrease-Garcia, Mark D Mangano, Jennifer L Demertzis, and Andrew J Gunn. 2018. Radiology reporting: current practices and an introduction to patient-centered opportunities for improvement. *American Journal of Roentgenology*, 210(2):376–385.

Sandeep S Naik, Anthony Hanbidge, and Stephanie R Wilson. 2001. Radiology reports: examining radiologist and clinician preferences regarding style and content. *American Journal of Roentgenology*, 176(3):591–598.

Ting Pang, Peigao Li, and Lijie Zhao. 2023. A survey on automatic generation of medical imaging reports based on deep learning. *BioMedical Engineering OnLine*, 22.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

AAO Plumb, FM Grieve, and SH Khan. 2009. Survey of hospital clinicians' preferences regarding the format of radiology reports. *Clinical radiology*, 64(4):386–394.

Felicity Pool and Stacy Goergen. 2010. Quality of the written radiology report: a review of the literature. *Journal of the American College of Radiology*, 7(8):634–643.

Baifeng Shi, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. 2024. When do we not need larger vision models? *arXiv preprint arXiv:2403.13043*.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. 2020. Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519.

Maria De La Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, et al. 2020. Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients. *arXiv preprint arXiv:2006.01174*.

Mateusz Winder, Aleksander Jerzy Owczarek, Jerzy Chudek, Joanna Pilch-Kowalczyk, and Jan Baron. 2021. Are we overdoing it? changes in diagnostic imaging workload during the years 2010–2020 including the impact of the sars-cov-2 pandemic. In *Healthcare*, volume 9, page 1557. MDPI.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. Next-gpt: Any-to-any multimodal llm. *ArXiv*, abs/2309.05519.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Curtis Langlotz, and Jean-Benoit Delbrouck. 2024. Overview of the first shared task on clinical text generation: Rrg24 and "discharge me!". In *The 23rd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Bangkok, Thailand. Association for Computational Linguistics.

Ling Yang, Zhanyu Wang, Zhenghao Chen, Xinyu Liang, and Luping Zhou. 2023. Medxchat: A unified multimodal large language model framework towards cxrs understanding and generation.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *Preprint*, arXiv:2303.15343.

Henry Zhan, Kevin M. Schartz, Matthew E. Zygmont, Jamlik-Omari Johnson, and Elizabeth A. Krupinski. 2020. The impact of fatigue on complex ct case interpretation by radiology residents. *Academic radiology*.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. 2023. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Preliminary Study

To investigate the impact of data composition and transfer learning on model performance, we conducted the following experiments:

A.1 Initialization and Training Strategies:

We explored two initialization and training strategies:

- Initialization from MLLM Pretrained on General Domain: This approach involves continually fine-tuning the pretrained multimodal language model (MLLM) on the interpret-cxr dataset, focusing solely on the final stage of training (stage 2).
- Initialization from LLM and Randomly Initialized Adapter: In this method, the language model (LLM) is pretrained on the LLaVa-MED dataset (stage 1) with a randomly initialized adapter, followed by fine-tuning on the interpret-cxr dataset (stage 2). This two-stage process aims to leverage domain-specific pretraining to enhance performance.

Our results indicated that this two-stage approach, which leverages domain-specific knowledge from LLaVa-MED, is beneficial and enhances performance.

A.2 Image Selection

We examined the effect of different image selection techniques:

- Reusing the Same Report When Multiple Images Are Provided: This technique involves using all available images for a given report, resulting in 1 million image-text pairs. This approach aims to maximize the amount of visual information provided to the model.
- Using Only the First X-ray Image When Multiple Images Are Provided: Here, only the first image from each study is used, leading to a dataset of 700,000 image-text pairs. This method is intended to reduce redundancy and potential bias in the reports by focusing on the most relevant image.

Our data mixture study revealed that using only the first image from each study yielded slightly better performance than using all images, ensuring the diversity of the radiology reports.

B Dataset Cleaning

In the preliminary inspection of the dataset, we observed that numerous reports within the interpret-cxr dataset contained sentences with information that could not be derived solely from the X-ray images. These sentences included details such as dates, doctor information, references to other imaging modalities, and comparisons with previous findings. Such extraneous information introduces noise that may lead the model to hallucinate incorrect dates, numbers, and comparisons with non-existent prior studies (Chen et al., 2024).

To mitigate this issue, we attempted to utilize GPT-3.5 Turbo to remove this irrelevant information from the dataset. The dataset cleaning prompts and examples are detailed in Appendix B.1 and B.2. However, during the evaluation, we observed a slight decline in performance metrics, as illustrated in Table 4, following the removal of these sentences. We suspected that the public-test and hidden-test datasets did not undergo similar cleaning procedures, resulting in uncleaned test sets. Therefore, to maximize of our performance metrics, we decided to use the original dataset without data cleaning for the remaining of our study.

B.1 Cleaning Prompt

We provide the prompt used for preprocessing and cleaning the training dataset to remove information that cannot be obtained solely from X-ray images.
Findings: "Remove non-x-rays discernible information from chest x-ray findings i.e. date, previous report mentions and comparison, and information from other imaging modality. Keep all remaining sentences unchanged:"
Impression: "Remove non-x-rays discernible information from chest x-ray impression i.e. date, doctor information, previous report mentions and comparison, and information from other imaging modality. Keep all remaining sentences unchanged. But if there is nothing left, return |Nonel and stop generating:"

B.2 Examples

B.2.1 Comparison with previous report

Original: Compared with the previous one, the x-ray is slightly inspired. no lung consolidations or pleural effusion are observed.
Clean: No lung consolidations or pleural effusion are observed.

B.2.2 Date Mentions

Original: AP chest radiograph on 12/11/08 at 2315 demonstrates a dual lead AICD. Stable cardiomegaly and stable left basilar opacities, likely

Table 3: Preliminary Evaluation of Initialization Strategies and Image Selection for Radiology Report Generation. This table compares the performance metrics of models initialized from a pretrained MLLM versus those initialized from an LLM with a randomly initialized adapter, as well as the impact of using only the first image from each study versus using all provided images. The results indicate that initializing from an LLM with a randomly initialized adapter yields better performance, and selecting the first image from each study slightly improves the metrics. Consequently, we heuristically retained only the first image to reduce redundancy and maintain report diversity.

LLM + Visual Encoder	Train Data	Epoch	BLEU4	ROUGEL	Bertscore	F1-cheXbert	F1-RadGraph
Phi-2 + SigLIP (init from Bunny)	LLaVa-Med + CXR	1	4.84	20.05	45.81	48.39	18.13
Phi-2 + SigLIP	LLaVa-Med + CXR	1	5.74	20.72	47.32	49.46	19.13
Phi-2 + SigLIP	LLaVa-Med + CXR (First)	1	5.45	21.17	47.43	51.10	19.52

Table 4: Results of the dataset cleaning experiment on findings and impressions. We performed stage 2 finetuning on the SigLIP and Phi-2 2.7B model architecture with different data mixtures for this experiment. "Raw" refers to the original interpret-cxr dataset, while "Clean" denotes the dataset cleaned by GPT-3.5 Turbo using the specified cleaning prompt.

Report Type	Train Data	Epoch	BLEU4	ROUGEL	Bertscore	F1-cheXbert	F1-RadGraph
findings	Raw	1	6.19	24.49	47.61	43.91	18.08
findings	Clean	1	5.84	24.17	47.14	44.19	17.83
impression	Raw	1	9.87	27.65	50.57	51.80	23.96
impression	Clean	1	6.62	23.66	50.74	49	24.62

atelectasis. Persistent right-sided pleural effusion. Diffuse reticular opacities, mild interstitial edema. Elevation of the left hemidiaphragm. Hiatal hernia. Partially visualized abdominal aortic stent graft. AP chest radiograph on 12-11-2008 at 3:11 a.m. demonstrates no significant interval change in cardiopulmonary status.

Clean: AP chest radiograph demonstrates a dual lead AICD. Stable cardiomegaly and stable left basilar opacities, likely atelectasis. Persistent right-sided pleural effusion. Diffuse reticular opacities, mild interstitial edema. Elevation of the left hemidiaphragm. Hiatal hernia. Demonstrates no significant interval change in cardiopulmonary status.

B.2.3 Other Modalities Mentions

Original: Chest x-ray. bilateral bronchiectasis with a predominance on the right side, noting an increase in density around these right basal bronchiectasis in relation to consolidations described in previous ct. there is no pleural effusion. cardiomedastinal silhouette and hila are within normal limits. biapical caps. bone and soft parts without notable findings.

Clean: Chest x-ray. Bilateral bronchiectasis with a predominance on the right side. There is no pleural effusion. Cardiomedastinal silhouette and hila are within normal limits. Biapical caps. Bone and soft parts without notable findings.

B.2.4 Doctor information

Original: 1.Interval development and resolution of a right upper lobe opacification, possibly representing interval resolution of right upper lobe aspiration or asymmetric pulmonary edema. 2. Persistent small bilateral pleural effusions. ""Physician to Physician Radiology Consult Line: (753) 619-1110"" I have personally reviewed the images for this examination and agreed with the report transcribed above.

Clean: 1.Interval development and resolution of a right upper lobe opacification, possibly representing interval resolution of right upper lobe aspiration or asymmetric pulmonary edema. 2. Persistent small bilateral pleural effusions."

C LLM Prompts

We provide a template of our post-processing prompt for the LLM to enhance diverse aspects generated report. The Report Refinement prompt enhance the readability and clarify the report while the "First, Do No Harm" SafetyNet prompt of Llama3 combines the results of our MLLM model and the classification results from X-Raydar.

Prompt 1: First Stage: Report Refinement

Radiology Reporting Instructions

You are an experienced radiologist tasked with interpreting CXR images and generating reports from free-text descriptions. Your primary objectives are to:

- Enhance readability and clarity of the text.
- Conduct Radgraph sterilization to ensure data integrity and accuracy.

When processing normal CXR findings, provide detailed explanations to clarify the condition. For instance:

Input Examples

- No significant findings.
- No acute cardiopulmonary findings.
- No acute cardiopulmonary abnormality.
- The heart is normal in size. The mediastinum is unremarkable. The lungs are clear.
- The heart size and pulmonary vascularity appear within normal limits. The lungs are free of focal airspace disease. No pleural effusion or pneumothorax is seen.
- No acute cardiopulmonary findings.

Expected Output

- The lungs are clear. No cardiomegaly. The cardiomediastinal and hilar contours are normal. There is no focal consolidation, pleural effusion, or pneumothorax. The pulmonary vascular markings are normal. No free air beneath the diaphragm.

Use a recommended vocabulary list to standardize report language and maintain consistency across reports. This list includes ['AC', 'Bony', 'Borderline', 'CHF', 'Calcified', 'Cardiac', 'Cardiomediastinal', 'Cardiomegaly', 'Clips', 'Dense', 'Dobbhoff', 'Esophageal', 'Extensive', 'Heart', 'Healing', 'Hilar', 'Hyperinflated', 'IJ', 'Increase', 'Increased', 'Interval', 'Interposition', 'Lung', 'Lungs', 'Lucency', 'Minimal', 'Moderate', 'Mild', 'Mildly', 'Monitoring', 'Multiple', 'Nasogastric', 'Nearly', 'New', 'Normal', 'Orphaned', 'PICC', 'Pneumomediastinum', 'Pneumothorax', 'Port - A - Cath', 'Pulmonary', 'Right-sided', 'Small', 'Slight', 'Slightly', 'Stable', 'Subcutaneous', 'Tip', 'Venous', 'Widespread', 'Worsening', 'Zone', 'accessory', 'acute', 'adenocarcinoma', 'air', 'air-filled', 'airspace', 'along', 'angles', 'anterior', 'anteriorly', 'apparent', 'appearance', 'appropriately', 'area', 'areation', 'artifact', 'atelectasis', 'axilla', 'benign', 'bibasal', 'bilaterally', 'blunting', 'borderline', 'bowel', 'bronchovascular', 'caliber', 'calcification', 'calcified', 'cancer', 'cardiac', 'cardiomegaly', 'central', 'change', 'chest', 'chf', 'clavicle', 'clavicular', 'clear', 'clips']
Reports should be styled succinctly, focusing on critical findings and summarizing significant observations without omitting essential details. Each report should follow the professional radiology report format:

Example of Good Reports

- The lungs are clear without focal consolidation. No pleural effusion or pneumothorax is seen. Heart size is top-normal. The mediastinal silhouette is unremarkable.
- Portable frontal radiograph of the chest demonstrates a right chest tube in unchanged position ending at the right apex. The right basilar pneumothorax continues to decrease in size. The pneumomediastinum is also decreasing. Extensive subcutaneous emphysema persists. Stable heart size and mediastinal contours. Small left pleural effusion is unchanged.
- The cardiac, mediastinal and hilar contours appear stable. Streaky left basilar opacity suggests minor atelectasis. The lateral view depicts a greater degree of right middle lobe atelectasis than before, more coalescent. There is no definite pleural effusion or pneumothorax.
- Persistent hila with a congestive appearance possibly due to pulmonary edema, but without evidence of significant consolidations or pleural effusion. to be correlated clinically.
- Cardiac silhouette is unchanged. Aortic arch calcification seen. Pulmonary vascularity is within normal limits. There is trace right pleural effusion noted. Bibasilar atelectasis is seen. There is no pneumothorax. Multilevel degenerative changes seen in the thoracic spine.

  (answer only summarize report to text paragraph)

Prompt 2: "Second Stage: First, Do No Harm" SafetyNet

Refine 'Input' with 'Refine information' indicating that the patient has conditions as refine information with the following conditions:

1. If the input and refined information have mismatched information, such as Refine information indicating an additional pathology not mentioned in the input, prioritize the refined information.

2. However, if the 'Refine information' suggests a pathology already included in the 'Input', we will not refine the input.

3. We will remove the sentence "lungs are clear" if there is any abnormality in the lung, pulmonary, or pleura.

Example 1:
Input = The lungs are clear. No cardiomegaly. The cardiomediastinal and hilar contours are normal. There is no focal consolidation, pleural effusion, or pneumothorax. The pulmonary vascular markings are normal.
Refine information = This patient has cardiomegaly and pleural effusion.
Output should be = There is cardiomegaly. The cardiomediastinal and hilar contours are normal. There is pleural effusion. There is no focal consolidation or pneumothorax. The pulmonary vascular markings are normal.
Example 2:
Input = The lungs are clear. No cardiomegaly. The cardiomediastinal and hilar contours are normal. There is no focal consolidation, pleural effusion, or pneumothorax. The pulmonary vascular markings are normal.
Refine information = This patient has cardiomegaly.
Output should be = There is cardiomegaly. The lungs are clear. The cardiomediastinal and hilar contours are normal. There is no focal consolidation, pleural effusion, or pneumothorax. The pulmonary vascular markings are normal.
Your answer should provide only the 'Output' format and not include any other comments.