Coarse-to-Fine Personalized LLM Impressions for Streamlined Radiology Reports

By

Steven Sun

Bimalsen Rajbhandari

Changhao Zhang

Colleen Jung

Supervisor: Utku Pamuksuz

A Capstone Project

Submitted to the University of Chicago in partial fulfillment of the Requirements for the degree of

Master of Science in Applied Data Science

Division of Physical Sciences

April 2024

Abstract

In radiology, creating the "Impression" section of reports is critical for clinical decision-making but is also a significant contributor to radiologist burnout due to its time-intensive nature. This paper introduces a novel approach leveraging open-sourced pre-trained Large Language Models (LLMs) to automate and personalize the crafting of radiological impressions. Utilizing a coarse-to-fine methodology, this strategy begins by generating a preliminary impression based on imaging findings and patient data, which is then refined to align with the individual radiologist's stylistic preferences through advanced machine learning techniques. The refinement process employs reinforcement learning from human feedback (RLHF) to ensure that the impressions are not only stylistically consistent but also factually accurate. By automating the impression generation process with tailored, high-quality outputs, this approach aims to alleviate the administrative burden on radiologists, enabling them to devote more time to patient care and reducing the risk of burnout. The proposed system will be fine-tuned using diverse LLM architectures such as LLaMA 3B, LLaMA 7B, and Mistral 7B, sourced from radiology reports at the University of Chicago Medicine, to balance computational efficiency and accuracy. This initiative promises to enhance the efficiency and personalization of radiology reporting, improving overall clinical practice and patient outcomes.

Keywords: healthcare, Impression generation, GenAI, Large Language Model, Fine-tuning, GPU, deep learning

Table of Contents

Introduction

Healthcare is fundamental to individual and societal well-being. At the front lines are physicians who work tirelessly to provide the best patient care. Their work is pivotal in creating a healthier and more vibrant society. Due to the importance of their work, medical practice can be highly stressful as errors can be costly—a matter of life and death in many cases. To make matters worse, physicians must often work long hours under stressful conditions. Hence, there is no surprise that burnouts are common in the medical profession (Shanafelt et al. 2003; Siu et al. 2012). Burnout leads to an increase in errors and a reduction in care quality (West et al. 2006; Williams et al. 2007). Therefore, creating tools to reduce the workload of physicians will reduce burnout, which leads to an improvement in patient care and a reduction in medical errors.

Research Problem

In a radiology report, the "Findings" section presents the observations made by the radiologist during the examination of the imaging studies. This section provides a detailed description of the abnormalities, if any, detected in the images. It typically includes information such as the size, location, shape, and characteristics of any lesions, masses, fractures, or other abnormalities found in the images. The "Impression" section offers a synthesized interpretation of these findings, aiding in clinical decision-making and patient management, which is why it is considered the most important part of the report.

Recognizing the substantial time investment radiologists allocate to crafting impressions, this paper proposes leveraging open-sourced pre-trained Large Language Models (LLMs) to expedite the generation of personalized impressions. By incorporating both the findings section and pertinent clinical data, such as patient information, into the LLMs' framework, the proposed approach seeks to produce impressions that align with individual radiologists' unique writing styles while ensuring factual correctness.

The adoption of LLMs promises manifold benefits. Firstly, it streamlines the radiology reporting process, enabling radiologists to redirect their attention towards direct patient care, thereby potentially expanding patient outreach. Additionally, by alleviating the burden associated with impression creation, the proposed methodology holds promise in mitigating radiologist burnout with its attendant negative ramifications.

Goals of Analysis

This paper proposes a methodological approach centered on fine-tuning a selection of open-sourced pre-trained LLMs using radiology reports sourced from the University of Chicago (UC) Medicine. The objective is to generate impressions of high quality that are both factually accurate and tailored to individual preferences.

This paper leverages open-source LLMs to ensure the accessibility of our research framework to a broad audience. We explore the utilization of LLMs with varying sizes and architectures, including LLaMA 3B, LLaMA 7B, and Mistral 7B, to strike a balance between computational efficiency and accuracy.

Recognizing the diversity in radiologists' writing styles—ranging from succinct to elaborate and encompassing structural preferences such as bullet points—the study endeavors to employ reinforcement learning from human feedback (RLHF) to achieve personalization. This approach empowers radiologists to select the reporting style that best suits their preferences and enhances their workflow efficiency.

Ensuring the factual correctness of generated impressions poses a critical challenge, particularly given the limitations of traditional text evaluation methods such as string-matching and n-gram-based metrics like ROUGE and BLEU. These metrics may yield inflated scores for generated text that diverges significantly from the intended meaning. To address this issue, a reinforcement learning framework is proposed to train the model in generating factually accurate impressions while mitigating the occurrence of hallucinations—a phenomenon where the model produces outputs that lack grounding in reality.

By integrating these methodological components, the study aims to advance the development of LLM-based frameworks for radiology reporting, fostering improved efficiency, personalization, and accuracy in clinical practice.

Background

Literature review

GPT and Impression Generation

Ziegelmayer et al. 2023 and Sun et al. 2023 Ziegelmayer et al. (2023) and Sun et al. (2023) explored the potential of GPT-4, a large language model, to generate radiology impression reports using zero-shot learning. In their studies, the researchers compared impressions crafted by GPT-4 based on prompts containing findings sections, images, or both, with those generated by radiologists. For evaluation, radiologist, physician's assessments, and common automatic evaluation metrics like Bleu, RadCliQ, and RadGraph were employed.

The findings highlight both promise and areas for improvement. Radiologists consistently rated their own impressions higher in terms of coherence, comprehensiveness, factual consistency, and potential for causing medical harm compared to GPT-4 impressions generated using only findings sections. However, post-hoc analysis revealed that these differences were not statistically significant. Interestingly, radiologists displayed a relatively low accuracy of just 61% in identifying AI-generated impressions derived from findings sections. This suggests that GPT-4, even without specific training on radiology reports, can produce outputs that mimic human-written impressions to a certain extent.

Another intriguing finding is the potential presence of bias. Even when radiologists themselves generated the impressions, those perceived as AI-generated received lower radiological scores. This suggests that radiologists might hold preconceived notions about the quality of AI-generated text, potentially influencing their assessment even of human-written reports. Furthermore, there were significant discrepancies between the radiological assessments and the scores generated by conventional automatic evaluation methods and physicians. This underscores the need for further development of evaluation metrics specifically tailored to the nuances of radiology impression generation.

A limitation of both studies is the use of small, non-diverse datasets. Despite this limitation, the research offers promising insights. The fact that even a zero-shot learning model with restricted data achieved acceptable performance suggests significant potential for LLMs in this domain. A fine-tuned Large Language Model (LLM) leveraging a more expansive and varied dataset has the potential to significantly improve the quality and accuracy of AI-generated radiology impressions.

Open-Sourced LLMs and Impression

Hu et al. 2022 proposes an advanced method for generating automatic impressions from radiology reports. It uses 2 datasets: OPENI: Contains 3,268 radiology reports collected by Indiana University, used in a random split of 2400 train, 292 validation, and 576 test reports. And MIMIC-CXR: A larger dataset containing 124,577 reports used with both official and random splits.

To deal with these 2 datasets, the researchers use text Encoding: Utilizes BioBERT pre-trained on biomedical texts to encode input findings into vector representations. Graph Construction: Forms a graph from the findings using extracted entities and their dependencies to model relationships among key terms. Graph Encoder: Employs Graph Neural Networks (GNNs) to encode relational information from the graph. Contrastive Learning Module: Implements contrastive learning to enhance the model's focus on critical information by comparing "positive" examples (where non-key words are masked) against "negative" examples (where key words are masked) and Sequence Generation: The decoder, a Transformer-based model, generates the final summary based on the enhanced embeddings provided by the graph and contrastive encoders.

Researchers utilize ROUGE scores to measure the quality of generated summaries. The performance of the model is better than the other public models: For OPENI, the model reaches ROUGE-1, ROUGE-2, and ROUGE-L scores of 64.97, 55.59, and 64.45, respectively. For MIMIC-CXR, it scores 49.13 in ROUGE-1, 33.76 in ROUGE-2, and 47.12 in ROUGE-L. Also, the Human Evaluation shows that the model generally matches or exceeds the quality of human-written references in terms of key information and accuracy but sometimes lacks in readability.

Karn et al. 2023, uses instruction-tuned LLMs for generating radiology report impressions. First, they pre-train an instruction-tuned LLM (Bloomz LLM is used) on a massive dataset of medical text data. This pre-training step aims to improve the LLM's understanding of medical terminology and concepts. Subsequently, they instruction-tune the pre-trained LLM on a dataset of radiology reports specifically for the task of generating impressions from findings sections. This instruction-tune step tailors the LLM to the specific structure and language used in radiology reports. Their system achieved the top ranking in the RadSum23 Task 1B challenge. This contribution is significant because it demonstrates the potential of instruction-tuned LLMs to produce impression from findings. We propose a distinct approach that leverages patient clinical information alongside findings sections during the fine/instruction-tuning stage. This inclusion aims to enrich the model's understanding of the patient's context, potentially leading to more accurate impression generation. Furthermore, we forgo the pre-training step, focusing solely on fine/instruction-tuning with radiology reports to reduce computational cost.

Tie et al. 2023, investigate the use of fine-tuned LLMs for generating personalized impressions in PET reports. They argue that current methods for impression generation lack personalization and may not fully capture all relevant findings. The authors train various LLMs on a dataset of PET reports, with the report findings and original impressions used as input and reference, respectively. They introduce an additional input token encoding the reading of physician's identity to personalize the generated impressions. The performance of the LLM-generated impressions is evaluated using metrics commonly employed in text summarization tasks, along with physician assessments for clinical utility and overall impression quality. The paper argues that fine-tuning LLMs on PET reports with physician identity information leads to the generation of clinically useful and personalized impressions. The PEGASUS model achieved the best results, with physician evaluations indicating high agreement on the acceptability and overall utility of the generated impressions compared to those written by other physicians. This paper demonstrates that several open-sourced LLMs can be utilized to generate personalized findings aligning closely with the primary focus of our research endeavor.

Liu et al. 2023, leveraged the Alpaca instruction-tuning framework to train Radiology-GPT on a radiology-focused dataset. The paper used the publicly available MIMIC-CXR dataset to generate an impression text given a findings text as an instruction. The paper argues that String-matching and n-gram based methods such as ROGUE and BLEU were deemed inappropriate due to the variability in radiologists' writing styles and interpretations. Instead, radiologists evaluated the mode based on understandability, coherence, relevance, conciseness and clinical utility. Radiology-GPT showed comparable understandability and slightly better coherence compared to ChatGPT but lagged slightly in relevance due to producing shorter responses. However, Radiology-GPT scored higher in conciseness and clinical utility, reflecting its design focus on delivering succinct and focused outputs. Other models tested, such as StableLM-7B, Dolly-12B, and LLaMA-7B, were outperformed by both Radiology-GPT and ChatGPT, indicating the value of domain-specific tuning and instruction comprehension in healthcare applications of LLMs.

Zhang et al. 2020, proposes a novel framework for training neural summarization models with a focus on factual correctness. They argue that existing models prioritize fluency and often generate summaries with factual errors. The paper implements a reinforcement learning framework where the summarization model is trained with a reward function that incorporates both fluency and factual correctness. They show significant improvements in factual correctness compared to existing models, with the generated summaries approaching human-authored quality in radiology reports.

Van Veen et al. 2023 presents a study on optimizing the adaptation of large language models (LLMs) for radiology report summarization (RRS) using lightweight domain adaptation techniques. Researchers use MIMIC-III, which is a dataset consisting of radiology reports across various imaging modalities and anatomical regions, used to train and evaluate the models.

Domain Adaptation used for this research that focuses on adapting pre-trained language models to the radiology domain by using a combination of pre-training on clinical text and fine-tuning on RRS examples. Parameter-Efficient Fine-Tuning such as LoRA (Low-Rank Adaptation) were employed to fine-tune only a small fraction (0.32%) of the model's parameters, which improves efficiency and reduces computational cost. And Prompt Engineering that experiments with few-shot prompting and the use of in-context examples to improve model performance without extensive fine-tuning.

The adapted models showed superior performance in generating concise and coherent summaries of radiology reports. The best results were achieved by models that were pre-trained on clinical text and fine-tuned on specific RRS tasks. For the evaluation, both quantitative metrics (like ROUGE scores) and qualitative assessments (via radiologist reader studies) were used to evaluate the models. The models adapted with clinical pre-training and lightweight fine-tuning techniques outperformed other models.

In conclusion, the research highlights the effectiveness of lightweight domain adaptation strategies in applying LLMs to specialized tasks like radiology report summarization. It demonstrates that targeted pre-training and efficient fine-tuning can significantly enhance the performance of LLMs in domain-specific applications while maintaining computational efficiency.

In their groundbreaking study, Radhakrishnan et al. 2023 unveiled "R2GenGPT: Radiology Report Generation with Frozen LLMs," a model that innovatively combines visual and textual data to facilitate the automated generation of radiology reports. This model is particularly notable for its utilization of a visual encoder that extracts significant features from radiological images. These features are then seamlessly integrated into the textual data stream via a sophisticated visual mapper. This mapper plays a crucial role, as it bridges the gap between the high-dimensional image data and the language model's textual feature space, allowing for coherent and accurate report generation.

A unique aspect of R2GenGPT is its strategy of employing "frozen" language model parameters. This approach ensures that the pre-trained LLM is utilized without undergoing further training during the model's application, thus maintaining the integrity and robust understanding capabilities of the LLM while focusing computational resources on optimizing the integration of visual inputs. This enhances the model's efficiency and significantly reduces the computational overhead associated with training expansive neural network parameters.

The study by Radhakrishnan et al. delves into various alignment strategies to optimize the interaction between visual inputs and textual outputs. These strategies are critical for ensuring that the visual data accurately informs the generated text, maintaining the clinical accuracy and utility of the automatically generated reports.

In essence, the R2GenGPT model shows a significant advancement in applying AI in healthcare, specifically in radiology. By automating the generation of detailed radiological reports, the model can ease the workload of radiologists, thereby allowing them to concentrate more on patient care and less on the administrative aspects of report generation. This could lead to more timely diagnostics, improved patient outcomes, and potentially, a reduction in radiologist burnout.

Data

This study capitalizes on a comprehensive dataset comprising 789,278 anonymized radiology reports procured from UChicago Medicine. The dataset encompasses a rich array of clinical patient information alongside unstructured textual findings and structured impression sections, as exemplified in Table 1. Noteworthy is the marked disparity in textual freedom between the findings and impression sections, with the former exhibiting a notably higher level of textual variability. Moreover, the dataset spans reports derived from eight distinct modalities, with CT (Computed Tomography) scans constituting approximately 90% of the dataset, as delineated in Table 2. Furthermore, the findings section exhibits a substantially higher word and sentence count relative to the impression and clinical information sections, as elucidated in Table 3. The exceptional quality and extensive scope of this dataset confer a distinct advantage over prior studies constrained by inferior data quality.

An additional persona column will be constructed based on the writing style of the impression, encompassing a spectrum ranging from succinct to elaborate, and accommodating structural preferences such as bullet points.

Ensuring diversity within the dataset, encompassing various sexes and races, assumes paramount importance. Although a comprehensive analysis of sex and race diversity within the dataset remains pending, efforts will be directed towards delineating the distribution of these demographic variables to foster a diverse and representative dataset.

Table 1: Example of clinical information, Findings section, and the Impression section

clinical_information	findings	impression
Male, 71 years old.Chronic renal disease. Assess lung fields.	Moderate to severe bilateral interstitial opacities in both mid and lower lungs as well as pleural effusions and cardiomegaly are noted although somewhat improved from previous. These findings may reflect pulmonary edema with some improvement. Right central line tip over the SVC. Dobbhoff tube below the diaphragms tracheostomy tube again noted. LVAD and ICD again present, unchanged. No pneumothorax.	Moderate pulmonary edema pattern, slightly improved from previous.

Table 2: Dataset distribution based on Modality

Modality	Approx. Percentage
CT	90.97%
X-Ray	0.01%
MRI	6.56%
Ultrasound	1.26%
Fluid	0.06%
MRA	0.36%
PET	0.77%
Sonogram	0.01%
Mammogram	0.00%
Total Reports	789,278.00

Table 3: Average words and sentences for Findings, Impression and Clinical Information

Data	Average Words	Average Sentences
Findings	130	6.3
Impression	29	2.4
Clinical Information	16	1.6

Methodology

This capstone project leverages a fine-tuning approach to adapt pre-trained Large Language Models (LLMs) for the generation of personalized radiology report impressions. Fine-tuning is selected due to its ability to efficiently utilize the extensive knowledge embedded in pre-existing LLMs. This approach conserves computational resources and reduces training time, which is crucial given the complex and data-intensive nature of medical texts. The robust dataset provided by UChicago Medicine enables the precise customization of these models to the specific linguistic and clinical nuances encountered in radiology reports.

The project explores a selection of model architectures, including LLaMA (3B, 7B) and Mistral (7B), to balance computational efficiency with the capability to generate high-quality outputs. The choice of these models is motivated by their open-source availability, which enhances the reproducibility of the research and allows for broader academic engagement.

Personalization of the generated impressions is achieved through the incorporation of Reinforcement Learning from Human Feedback (RLHF). This technique refines the model outputs to align closely with the individual stylistic preferences of radiologists. By adapting to their specific feedback, the model learns to produce text that is not only stylistically consistent but also factually accurate, addressing a critical requirement in medical reporting.

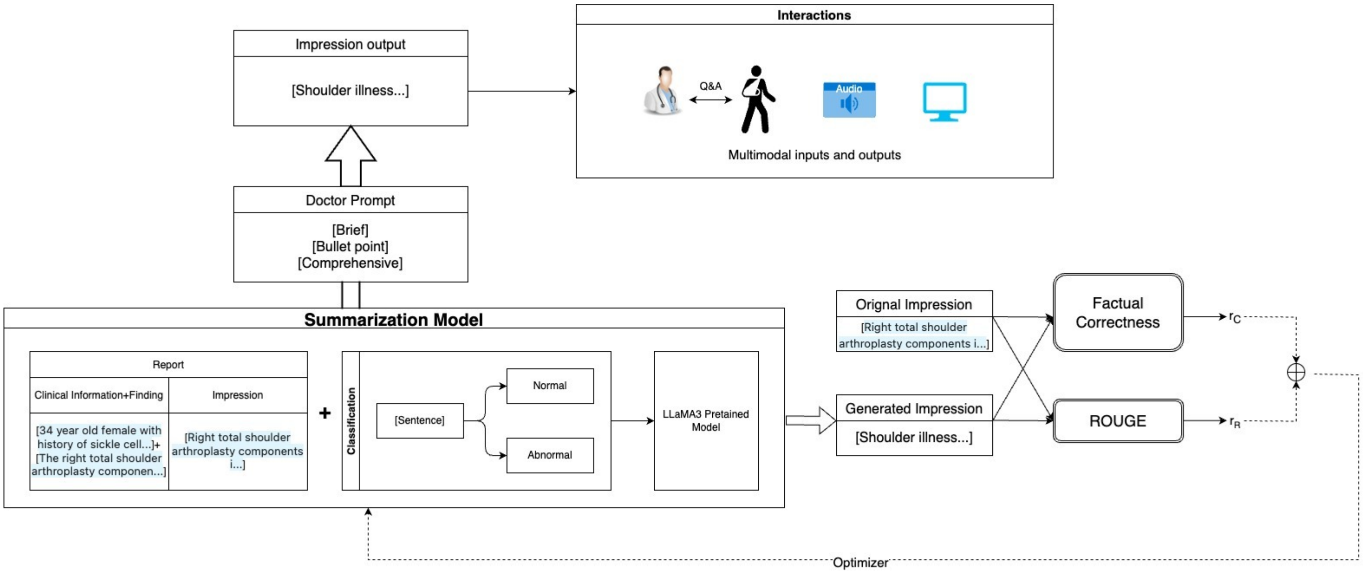
Additionally, the methodology employs a sophisticated instructional design in the tuning process. It integrates clinical data alongside imaging findings to enrich the context within which the model operates. This integration improves the relevance and factual accuracy of the impressions, making them more useful in clinical decision-making. Training prompts are carefully crafted to emulate real-world application scenarios, ensuring that the fine-tuning process optimally prepares the model for its intended operational environment.

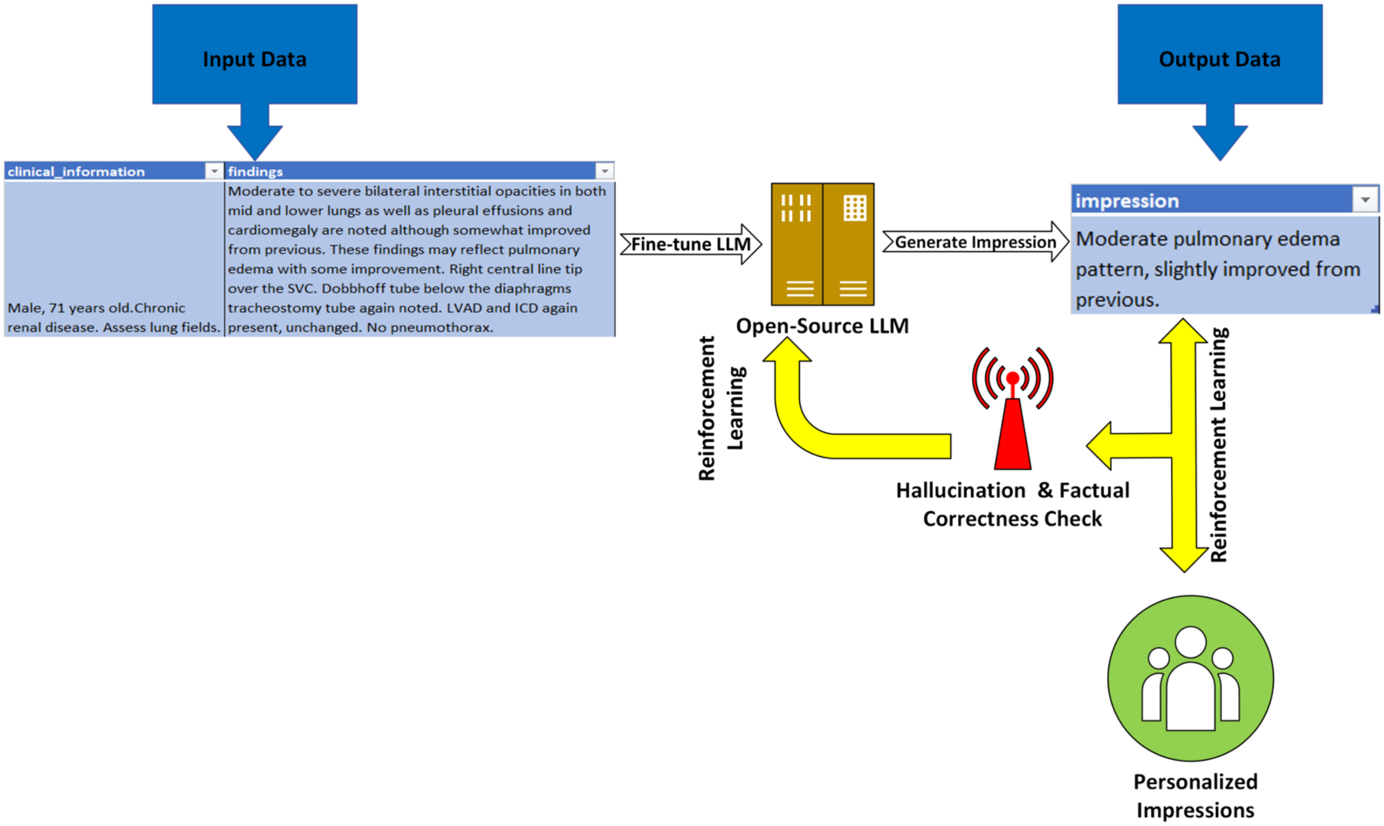
An encoder-decoder framework is utilized, combining traditional transformer models with graph neural networks. This setup enhances the model's capability to encode complex interdependencies among medical entities extracted from imaging findings, thus maintaining the factual integrity crucial for medical applications.

Evaluation extends beyond conventional NLP metrics like ROUGE and BLEU to include assessments of clinical utility and physician reviews. These evaluations are essential to ascertain the practical applicability of the impressions in medical settings. Rigorous bias and error analyses are conducted to ensure the model's outputs are reliable and unbiased, safeguarding against potential ethical issues in clinical implementations.

The project documentation covers the computational resources and software frameworks used, adhering to ethical standards and ensuring data privacy in compliance with healthcare regulations. This detailed accounting supports the transparency and replicability of the research, contributing to its value as a reference for future endeavors in applying AI to healthcare.

7 Architecture





clinical_information	findings
Male, 71 years old.Chronic renal disease. Assess lung fields.	Moderate to severe bilateral interstitial opacities in both mid and lower lungs as well as pleural effusions and cardiomegaly are noted although somewhat improved from previous. These findings may reflect pulmonary edema with some improvement. Right central line tip over the SVC. Dobbhoff tube below the diaphragms tracheostomy tube again noted. LVAD and ICD again present, unchanged. No pneumothorax.

impression
Moderate pulmonary edema pattern, slightly improved from previous.

References

1. Shanafelt, T. D., Sloan, J. A., & Habermann, T. M. (2003). The well-being of physicians. *The American Journal of Medicine, 114*(6), 513-519.

2. 蕭鳳兒, 袁仕傑, & 張潔影. (2012). Burnout among public doctors in Hong Kong: Cross-sectional survey. *Hong Kong Medical Journal, 18*(3), 186-192.

3. West, C. P., Huschka, M. M., Novotny, P. J., Sloan, J. A., Kolars, J. C., Habermann, T. M., & Shanafelt, T. D. (2006). Association of perceived medical errors with resident distress and empathy: A prospective longitudinal study. *JAMA, 296*(9), 1071-1078.

4. Williams, E. S., Manwell, L. B., Konrad, T. R., & Linzer, M. (2007). The relationship of organizational culture, stress, satisfaction, and burnout with physician-reported error and suboptimal patient care: Results from the MEMO study. *Health Care Management Review, 32*(3), 203-212.

5. Ziegelmayer, S., Marka, A. W., Lenhart, N., Nehls, N., Reischl, S., Harder, F., Sauter, A., Makowski, M., Graf, M., & Gawlitza, J. (2023). Evaluation of GPT-4's chest X-ray impression generation: A reader study on performance and perception. *Journal of Medical Internet Research, 25*, e50865.

6. Sun, Z., Ong, H., Kennedy, P., Tang, L., Chen, S., Elias, J., Lucas, E., Shih, G., & Peng, Y. (2023). Evaluating GPT-4 on impressions generation in radiology reports. *Radiology, 307*(5), e231259.

7. Karn, S. K., Ghosh, R., & Farri, O. (2023). shs-nlp at radsum23: Domain-adaptive pre-training of instruction-tuned llms for radiology report impression generation. *arXiv preprint arXiv:2306.03264*.

8. Tie, X., Shin, M., Pirasteh, A., Ibrahim, N., Huemann, Z., Castellino, S. M., Kelly, K. M., & et al. (2023). Automatic personalized impression generation for pet reports using large language models. *ArXiv*.

9. Liu, Z., Zhong, A., Li, Y., Yang, L., Ju, C., & Wu, Z. (2023). Radiology-GPT: A large language model for radiology. *arXiv [Preprint]*.

10. Zhang, Y., Merck, D., Tsai, E. B., Manning, C. D., & Langlotz, C. P. (2019). Optimizing the factual correctness of a summary: A study of summarizing radiology reports. *arXiv preprint arXiv:1911.02541*.

11. Van Veen, D., Van Uden, C., Attias, M., Pareek, A., Bluethgen, C., Polacin, M., Chiu, W., & et al. (2023). RadAdapt: Radiology report summarization via lightweight domain adaptation of large language models. *arXiv preprint arXiv:2305.01146*.

12. Hu, J., Li, Z., Chen, Z., Li, Z., Wan, X., & Chang, T.-H. (2022). Graph enhanced contrastive learning for radiology findings summarization. *arXiv preprint arXiv:2204.00203*.

13. Wang, Z., Liu, L., Wang, L., & Zhou, L. (2023). R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology, 1*(3), 100033.