Structuring Radiology Reports: Challenging LLMs with Lightweight Models

Anonymous ACL submission

Abstract

Radiology reports are critical for clinical decision-making but often lack a standardized format, limiting both human interpretability and machine learning (ML) applications. While large language models (LLMs) like GPT-4 can effectively reformat these reports, their proprietary nature, computational demands, and data privacy concerns limit clinical deployment. To address this challenge, we employed lightweight encoder-decoder models (<300M parameters), specifically T5 and BERT2BERT, to structure radiology reports from the MIMIC-CXR and CheXpert databases. We benchmarked our lightweight models against five open-source LLMs (3-8B parameters), which we adapted using in-context learning (ICL) and low-rank adaptation (LoRA) finetuning. We found that our best-performing lightweight model outperforms all ICL-adapted LLMs on a human-annotated test set across all metrics (BLEU: 212%, ROUGE-L: 63%, BERTScore: 59%, F1-RadGraph: 47%, GREEN: 27%, F1-SRRG-Bert: 43%). While the overall best-performing LLM (Mistral-7B with LoRA) achieved a marginal 0.3% improvement in GREEN Score over the lightweight model, this required 10× more training and inference time, resulting in a significant increase in computational costs and carbon emissions. Our results highlight the advantages of lightweight models for sustainable and efficient deployment in resource-constrained clinical settings.

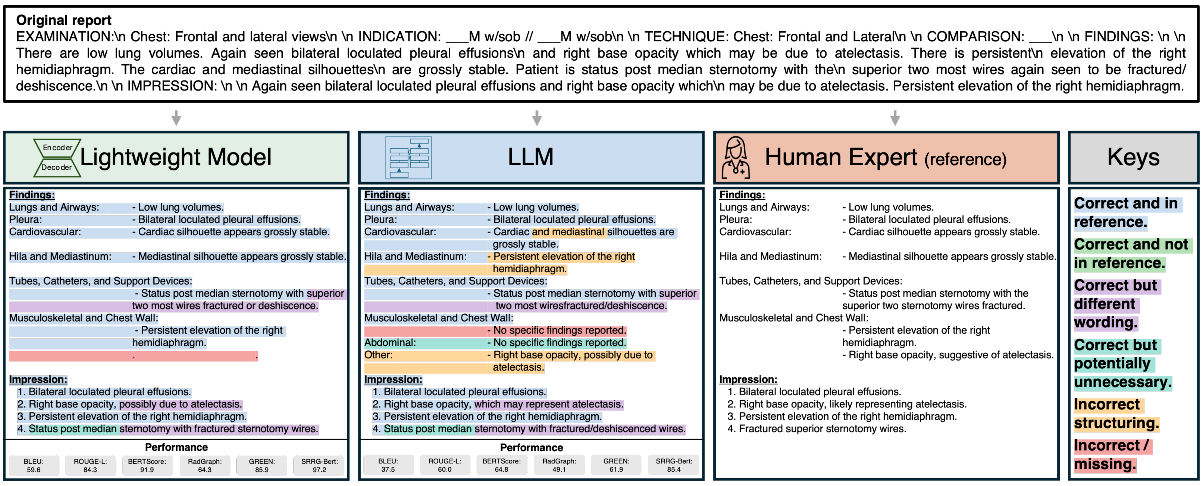
1 Introduction

Radiology reports play a critical role in clinical workflows by summarizing imaging findings that guide medical decisions (Kahn Jr et al., 2009). However, variations in reporting style due to individual and institutional practices as well as regional guidelines create inconsistencies that hinder interpretability for physicians and patients (Hartung et al., 2020). Moreover, the lack of structured formats limits their usefulness as training data for machine learning (ML) applications (dos Santos et al., 2023; Steinkamp et al., 2019).

Large language models (LLMs) offer a promising solution for generating structured reports from free-form text (Adams et al., 2023; Busch et al., 2024; Hasani et al., 2024). However, deploying these models locally remains infeasible for most institutions due to the significant computational resources required (Zhang et al., 2025). Cloud-based solutions provide an alternative but introduce concerns related to data security, confidentiality, and regulatory compliance (Arshad et al., 2023; Thirunavukarasu et al., 2023). While proprietary LLMs can also be accessed via Application Programming Interface (API), this approach entails drawbacks such as dependency on a third-party vendor, potential cost increases and unpredictable changes in usage terms. (Tian et al., 2024). These limitations highlight the need for smaller, open-source models that can be deployed on-device with minimal hardware requirements.

To address these challenges, we propose lightweight (<300M parameters), task-specific models for structuring free-text chest X-ray radiology reports (see Figure 1) efficiently. These models substantially reduce computational demands (Chen et al., 2024a), eliminating the need for cloud-based hosting, and enhancing data security by enabling offline deployment. We train these models on the MIMIC-CXR (Johnson et al., 2019) and CheXpert Plus (Chambon et al., 2024) datasets and structure the originally free-form reports with GPT-4 (Achiam et al., 2023) as a weak annotator, enabling large-scale supervision. We evaluate model performance on an independent test set, annotated by five radiologists (Anonymous, 2025). Our contributions include:

- **Lightweight Model Development and Evaluation**: We train and systematically evaluate lightweight (<300M parameters), task-



Original report

EXAMINATION:\n Chest: Frontal and lateral views\n \n INDICATION: ___M w/sob // ___M w/sob\n \n TECHNIQUE: Chest: Frontal and Lateral\n \n COMPARISON: ___\n \n FINDINGS: \n \n There are low lung volumes. Again seen bilateral loculated pleural effusions\n and right base opacity which may be due to atelectasis. There is persistent\n elevation of the right hemidiaphragm. The cardiac and mediastinal silhouettes\n are grossly stable. Patient is status post median sternotomy with the\n superior two most wires again seen to be fractured/ deshiscence.\n \n IMPRESSION: \n \n Again seen bilateral loculated pleural effusions and right base opacity which\n may be due to atelectasis. Persistent elevation of the right hemidiaphragm.

Lightweight Model

Findings:
Lungs and Airways: - Low lung volumes.
Pleura: - Bilateral loculated pleural effusions.
Cardiovascular: - Cardiac silhouette appears grossly stable.

Hila and Mediastinum: - Mediastinal silhouette appears grossly stable.

Tubes, Catheters, and Support Devices:
- Status post median sternotomy with superior two most wires fractured or deshiscence.
Musculoskeletal and Chest Wall:
- Persistent elevation of the right hemidiaphragm.
. .

Impression:
1. Bilateral loculated pleural effusions.
2. Right base opacity, possibly due to atelectasis.
3. Persistent elevation of the right hemidiaphragm.
4. Status post median sternotomy with fractured sternotomy wires.

Performance

BLEU:	ROUGE-L:	BERTScore:	RadGraph:	GREEN:	SRRG-Bert:
59.6	84.3	91.9	64.3	85.9	97.2

LLM

Findings:
Lungs and Airways: - Low lung volumes.
Pleura: - Bilateral loculated pleural effusions.
Cardiovascular: - Cardiac and mediastinal silhouettes are grossly stable.
Hila and Mediastinum: - Persistent elevation of the right hemidiaphragm.
Tubes, Catheters, and Support Devices:
- Status post median sternotomy with superior two most wiresfractured/deshiscence.
Musculoskeletal and Chest Wall:
- No specific findings reported.
Abdominal: - No specific findings reported.
Other: - Right base opacity, possibly due to atelectasis.

Impression:
1. Bilateral loculated pleural effusions.
2. Right base opacity, which may represent atelectasis.
3. Persistent elevation of the right hemidiaphragm.
4. Status post median sternotomy with fractured/deshiscenced wires.

Performance

BLEU:	ROUGE-L:	BERTScore:	RadGraph:	GREEN:	SRRG-Bert:
37.5	60.0	64.8	49.1	61.9	85.4

Human Expert (reference)

Findings:
Lungs and Airways: - Low lung volumes.
Pleura: - Bilateral loculated pleural effusions.
Cardiovascular: - Cardiac silhouette appears grossly stable.

Hila and Mediastinum: - Mediastinal silhouette appears grossly stable.

Tubes, Catheters, and Support Devices:
- Status post median sternotomy with the superior two sternotomy wires fractured.
Musculoskeletal and Chest Wall:
- Persistent elevation of the right hemidiaphragm.
- Right base opacity, suggestive of atelectasis.

Impression:
1. Bilateral loculated pleural effusions.
2. Right base opacity, likely representing atelectasis.
3. Persistent elevation of the right hemidiaphragm.
4. Fractured superior sternotomy wires.

Keys

Correct and in reference.
Correct and not in reference.
Correct but different wording.
Correct but potentially unnecessary.
Incorrect structuring.
Incorrect / missing.

Figure 1: Overview of our study and qualitative comparison. An unstructured radiology report is structured using lightweight task-specific models and adapted large language models (LLMs) compared to human expert annotations.

specific T5 and BERT2BERT models for the task of structuring radiology reports.

- **Analysis of LLMs**: We assess the performance of five LLMs (3-8B parameters) under different adaptation strategies (prefix prompting, in-context learning (ICL), low-rank adaptation (LoRA)).

- **Benchmarking and Cost Analysis**: We benchmark lightweight models against LLMs, considering model performance on the BLEU, ROUGE-L, BERTScore, F1-RadGraph, GREEN, and F1-SRRG-Bert metrics, as well as training time, FLOPs per forward pass, inference speed and costs, and environmental impact.

2 Related Work

Beyond LLMs: Lightweight Models for Medical Text Processing

Recent studies have explored the use of LLMs, namely GPT-3.5 (OpenAI, 2022) and GPT-4, to transform free-form radiology reports into structured formats (Adams et al., 2023; Bergomi et al., 2024; Hasani et al., 2024). A recent review by Busch et al. highlights that these approaches achieve low error rates and minimal accuracy loss compared to human experts (Busch et al., 2024). However, their reliance on proprietary architectures, lack of transparency, and restrictions on patient data privacy pose significant challenges for clinical deployment (Khullar et al., 2024; Rezaeikhonakdar, 2023). To address these limitations, similar tasks in medical NLP have adopted lightweight, task-specific models that maintain high accuracy while considerably reducing computational costs (Chen et al., 2024a; Griewing et al., 2024; Pecher et al., 2024). Existing task-specific models for radiology NLP fall into two categories: hybrid models and lightweight transformer models. Hybrid models combine rule-based methods with deep learning, enforcing domain-specific constraints but lacking flexibility (Gabud et al., 2023). In contrast, lightweight transformer models have been successfully applied to relation extraction, report coding, and summarization (Jain et al., 2021; Yan et al., 2022; Van Veen et al., 2023). While they require careful tuning to avoid hallucinations and overfitting, recent studies suggest that well-tuned lightweight models can match larger LLMs in accuracy while being far more computationally efficient (Pecher et al., 2024). Our work builds on this foundation by introducing a lightweight, task-specific model explicitly optimized for structured radiology report generation.

Model Adaptation and Finetuning

Prior work has explored a range of adaptation strategies for LLMs, from prompt-based methods to parameter-efficient finetuning (PEFT) and full finetuning, each balancing performance, data requirements, and computational cost. Prompting techniques such as prefix prompting and ICL (Brown et al., 2020; Lampinen et al., 2022) adapt models without modifying their weights. Prefix prompting typically provides instructions to guide model responses, while ICL enhances adaptation by incorporating task-specific examples within the prompt. However, these methods suffer from context length

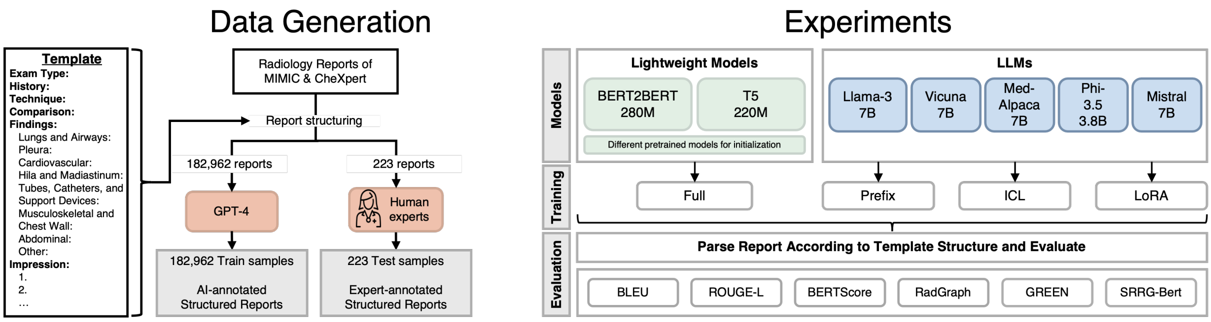


Figure 2: Left: Dataset generation from free-form radiology reports to structured radiology reports using GPT-4 (AI-based) and human experts (manual annotation). Right: Overview about our experiments including selection of lightweight models and LLMs, training/adaptation methods, and evaluation strategy and metrics.

constraints and sensitivity to prompt phrasing (Li et al., 2023). PEFT techniques like LoRA (Hu et al., 2021), prefix-tuning (Li and Liang, 2021), and adapter layers (Houlsby et al., 2019) enable efficient adaptation with minimal computational overhead, making them well-suited for clinical NLP. While effective in low-data settings, PEFT often struggles with complex reasoning and generalization across domains (Lialin et al., 2023). In contrast, full finetuning updates all model parameters, often achieving stronger adaptation when sufficient labeled data and computational resources are available. Building on this, our approach applies full finetuning to lightweight models while leveraging GPT-4-generated structured labels to address data scarcity, enabling large-scale supervised training while preserving domain-specific accuracy.

AI-Based Dataset Generation

A major challenge in developing models for structuring radiology reports is the limited availability of high-quality annotated datasets, i.e., datasets that contain both free-form and corresponding structured reports. Recent work in similar fields has explored leveraging LLMs such as GPT-4 as weak annotators to generate labels, providing a scalable alternative to manual annotation (Liyanage et al., 2024; Savelka et al., 2023). Despite their successes, studies suggest that models trained on GPT-generated data should still be rigorously evaluated against human-annotated ground truth to ensure reliability and validity (Pangakis et al., 2023).

3 Methods

In this study, we transform free-text chest X-ray radiology reports into a standardized format using deep learning. The structured reports follow a predefined template based on ’RPT144’ of RSNA’s RadReport Template Library (Radiological Society of North America (RSNA), 2011). This template comprises the sections: Exam Type, History, Technique, Comparison, Findings, and Impression. The Findings section is further organized into organ systems: ’Lungs and Airways’, ’Pleura’, ’Cardiovascular’, ’Tubes, Catheters, and Support Devices’, ’Musculoskeletal and Chest Wall’, ’Abdominal’, and ’Other’. The Impression section is structured as a numbered list, prioritizing the most clinically relevant findings. As shown in Figure 2, this template is incorporated into the prompt during data annotation, and deviations from it in a structured report are penalized during evaluation. Unlike previous approaches that rely on large, general-purpose models like GPT-4, we explore the effectiveness of lightweight, task-specific models for this task.

3.1 Data

We use unstructured radiology reports from the publicly available MIMIC-CXR (Johnson et al., 2019) and CheXpert Plus (Chambon et al., 2024) datasets, preserving their original training and validation splits. To train our models in a supervised manner, we employed GPT-4 as a weak annotator, using the prompt provided in Appendix A.1 to generate structured reports that conform to our template. We obtained a total of 182,962 reports, 125,447 samples from MIMIC and 57,515 from CheXpert Plus. For evaluation and benchmarking, we conducted a human expert review of 223 reports, comprising 161 from the MIMIC-CXR test set and 72 from the CheXpert Plus validation set. Five board-certified radiologists from our institution reviewed the structured reports alongside their original free-form counterparts, assessing them for errors and adherence to our predefined template (detailed in (Anonymous, 2025)).

3.2 Evaluation Strategies

Even though all models generate full reports, we focus our quantitative analysis on the Findings and Impression sections due to their clinical significance. Before applying our metrics, we parse these sections to assess adherence to the predefined template. In the Findings section, we identify predefined organ system headers (e.g., 'Lungs and Airways', 'Cardiovascular') and extract their corresponding observations. Metrics are computed separately for each organ system and then averaged across all identified systems. In the Impression section, we enforce a sequentially numbered format and flag any inconsistencies in ordering. To assess both linguistic quality and clinical accuracy, we use a combination of lexical and radiology-specific metrics.

Lexical Metrics To ensure comprehensive evaluation of text quality, we apply the following metrics: *BLEU* (Papineni et al., 2002) measures n-gram overlap, serving as a proxy for fluency and syntactic similarity. *ROUGE-L* (Lin, 2004) evaluates the longest common subsequence, capturing sentence-level similarity. *BERTScore* (Zhang et al., 2019) computes semantic similarity by comparing contextual embeddings from a pretrained transformer model.

Radiology-Specific Metrics To capture clinical accuracy, we apply the following metrics: *F1-RadGraph* (Yu et al., 2023; Delbrouck et al., 2022) evaluates the precision and recall of key clinical terms and relationships extracted from generated reports. *GREEN* (Ostmeier et al., 2024) assesses the factual correctness of generated radiology reports using a finetuned LLM. *F1-SRGG-Bert* (Anonymous, 2025) uses a fine-tuned BERT model to classify extracted findings into 55 disease labels, assigning each as Present, Absent, or Uncertain. It then computes the F1-score by comparing predictions from the generated report to the ground truth. Throughout this paper, our visualizations primarily focus on GREEN and BERTScore, as GREEN correlates most strongly with expert evaluations of clinical accuracy (Ostmeier et al., 2024), while BERTScore captures semantic similarity, making their combination effective for assessing structured radiology reports.

3.3 Lightweight Models

We introduce lightweight models, which are specifically trained to structure radiology reports according to a predefined template. Our lightweight models are based on encoder-decoder architectures given their recent success in similar tasks such as radiology report generation (Aksoy et al., 2023; Chen et al., 2024b) and radiology report summarization (de Padua and Qureshi, 2024; Van Veen et al., 2023; Zhang et al., 2018). Specifically, we focused on two architectures, *T5-Base* (Raffel et al., 2020), which has 223M parameters, and *BERT2BERT* (Rothe et al., 2020), where two identical BERT models are used as the encoder and decoder, resulting in a total of 278M parameters. To investigate the influence of pretraining domains, we initialize our models with the parameters from five open-source T5 variants (Table 2) - *T5-Base* (Raffel et al., 2020)(general text), *Flan-T5-Base* (Chung et al., 2024)(instruction-tuning), *SciFive* (Phan et al., 2021)(biomedical text), *Clin-T5-Sci* (Lehman and Johnson, 2023)(biomedical text and radiology reports), and *Clin-T5-Base* (Lehman and Johnson, 2023)(radiology reports) - and four BERT variants (Table 3) - *RoBERTa-base* (Liu, 2019)(general text), *BioMed-RoBERTa* (Gururangan et al., 2020)(biomedical text), *RoBERTa-base-PM-M3-Voc-distill-align* (Lewis et al., 2020)(for simplicity named RoBERTa-PM-M3 here, biomedical text and radiology reports), and *RadBERT-RoBERTa* (Yan et al., 2022)(radiology reports). We train our lightweight models end-to-end, updating all parameters, for a maximum of ten epochs using a cosine learning rate scheduler with initial leaning rate of $1e^{-4}$, an effective batch size of 128, and the Adam optimizer. A detailed description of hyperparameters can be found in Appendix A.3. To account for variability, each configuration is trained three times with different random seeds. Following prior work (Van Veen et al., 2023), we rank pretraining datasets by relevance, assuming radiology reports to be the most relevant, followed by biomedical text (e.g., PubMed abstracts) and general-domain text (e.g., Wikipedia). However, we acknowledge that this ranking is inherently subjective and may vary depending on the specific task.

3.4 Comparison LLMs

To benchmark our lightweight models (<300M parameters), we conduct a comprehensive comparison with instruction-tuned LLMs ranging from 3 to 8 billion parameters: Llama-3.1-8B-Instruct (Grattafiori et al., 2024); its derivatives Vicuna-7B-v1.5 (Chiang et al., 2023), optimized for conversational tasks, and Med-Alpaca-7B (Han et al.,

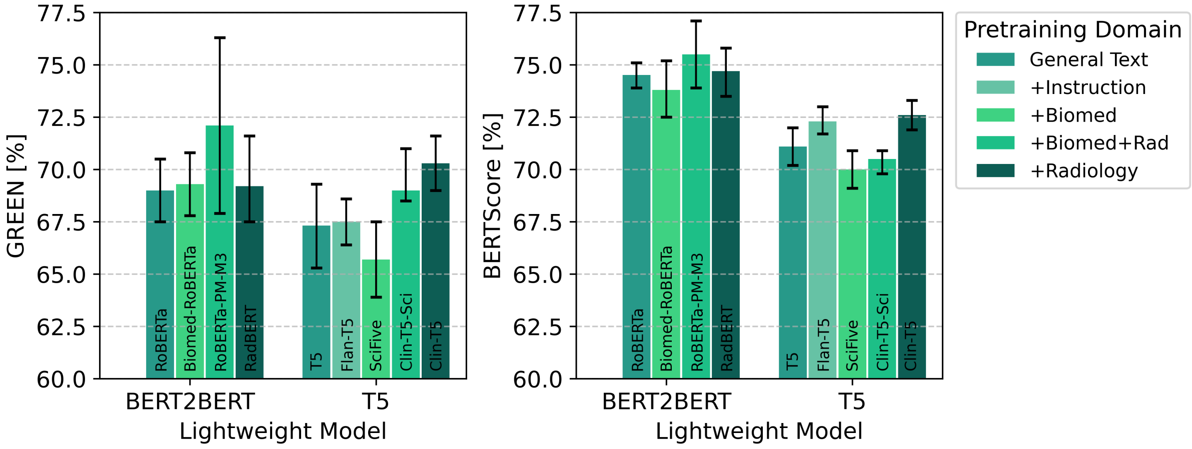


Figure 3: Performance comparison of lightweight models, initialized from pretrained models of increasing domain relevance. The plot shows the finetuned BERT2BERT and T5 models evaluated using GREEN (left) and BERTScore (right), initialized from various pretrained models, with pretraining datasets ranging from general text (least domain-specific) to radiology (most domain-specific). Error bars denote 95% confidence intervals over the three training runs.

2023), finetuned for medical question-answering; as well as Phi-3.5-Mini-Instruct (Abdin et al., 2024) and Mistral-7B (Jiang et al., 2023). We assess three adaptation techniques: **1. Prefix Prompting.** The model is prompted using the same instructions employed during training data generation (Appendix A.1). **2. ICL.** The model is given either one (1-shot) or two (2-shot) free-form reports along with their structured counterparts. These examples are manually selected from the training set to optimally represent the data distribution. **3. LoRA Finetuning.** The LLMs are finetuned for five epochs on the complete training set using LoRA with a rank of eight, modifying approximately 0.1% of the model's parameters. We use a cosine learning rate scheduler with an initial learning rate of $1e^{-4}$, an effective batch size of 256 and the Adam optimizer. Detailed finetuning configurations are provided in Appendix A.4 and the impact of different LoRA ranks is analyzed in Section 4.3.

3.5 Benchmarking Lightweight Models Against LLMs

We benchmark our best lightweight model against the top-performing LLM from previous experiments (Mistral-7B with LoRA), adding GPT-4 as a reference. Given the lightweight model's significantly smaller size, we quantify the finetuning required for the LLM to match its performance by varying LoRA rank (r) and tracking the number of training epochs. Our evaluation focuses on two key aspects: We systematically finetune the selected LLM and measure the point at which its performance matches or surpasses that of the lightweight model. We then compare the computational costs associated with training and deploying the lightweight model, Mistral-7B, and GPT-4. This comparison includes the average GREEN score, training time per epoch, floating-point operations (FLOPs) for a single forward-pass, inference time per sample, inference costs per sample, and CO2 emissions per sample. All models are trained and evaluated on a single Nvidia A100 (80GB) GPU. Inference costs are estimated using the Google Cloud pricing calculator[1], and CO2 emissions are calculated with CodeCarbon (Lacoste et al., 2019). These comparisons provide insights into the trade-offs between large-scale LLMs and compact lightweight models in terms of both performance and resource efficiency.

4 Results

The models are evaluated using all metrics introduced in Section 3.2. We primarily report results using GREEN and BERTScore, as they provide complementary assessments of clinical accuracy and semantic similarity. However, unless stated otherwise, the observed trends hold across all metrics. A detailed comparison across all metrics is provided in Appendix A.5.

[1] https://cloud.google.com/products/calculator (Assessed January 2025)

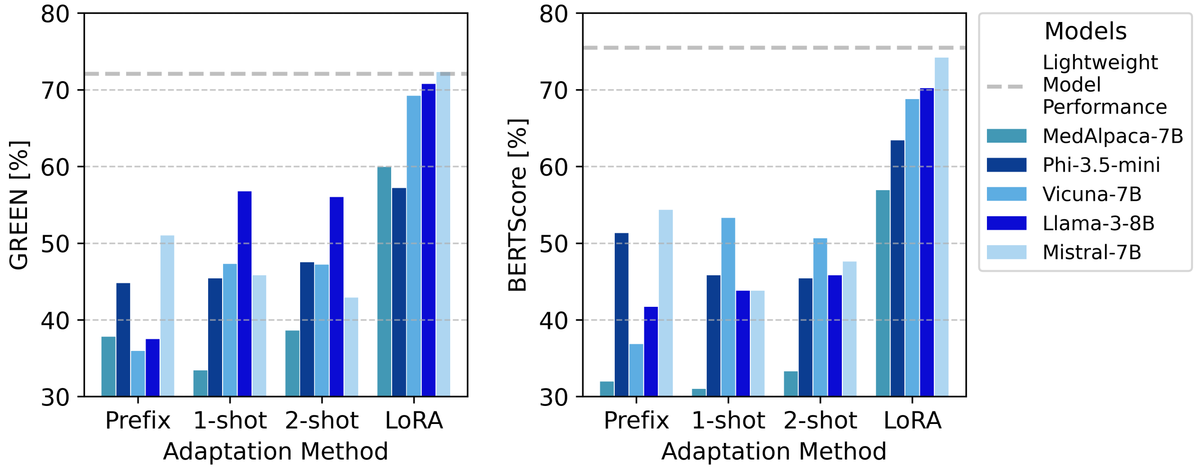


Figure 4: Comparison of LLM Adaptation Methods and the best performing lightweight model (BERT2BERT initialized from RoBERTa-PM-M3). (Left)/(Right)The figure depicts the GREEN Score/BERTScore for five different LLMs across various adaptation methods, including prefix prompting, in-context learning (ICL) with 1-shot and 2-shot settings, and LoRA finetuning for five epochs.

4.1 Comparison of Lightweight Models and Domain Adaptation

As introduced in Section 3.3, we initialized our lightweight models with the weights from different pretrained models. Specifically, we evaluate four different pretrained models as initializations for the BERT2BERT model and five for the T5 model (Tables 2 and 3). Each pretraining configuration was trained three times with different random seeds. Figure 3 presents the model performance for the GREEN and BERTScore metrics, while a more comprehensive overview can be found in Table 4. For the BERT2BERT model, domain adaptation shows a clear but non-linear impact on performance. Pretraining on biomedical text improves GREEN by 0.4% over the general-text baseline, while adding radiology reports yields a more substantial 4.5% improvement. However, pretraining exclusively on radiology reports (RadBERT) provides only a marginal 0.3% increase. For the T5 model, instruction-tuning alone leads to 0.3% improvement over the general-text baseline. Pretraining on biomedical text and radiology reports achieves a 2.5% gain, while using exclusively radiology reports leads to 4.4% increase. However, the biomedical text initialization (SciFive) underperforms the general baseline by 2.4%. Table 4 confirms that these trends persist across both datasets and sections, with scores for the Impression section being on average by $\approx 20\%$ higher. Overall, BERT2BERT models outperform T5 variants, with the best BERT2BERT model (RoBERTa-PM-M3) beating the best T5 (Clin-T5-Base) by 2.6% on GREEN and 2.9% on BERTScore.

4.2 Adaptation of LLMs

We present the results of adapting LLMs to the structuring task as outlined in Section 3.4. Figure 4 visualizes the average test set performance on the GREEN and BERTScore metrics across the proposed adaptation methods: prefix prompting, 1-shot and 2-shot in-context learning (ICL), and LoRA finetuning. LoRA finetuning consistently achieves the highest performance across all models. The detailed breakdown of results across the structured Findings and Impression sections is provided in Tables 5 and 6 of the Appendix. Averaged across all five LLMs, 1-shot ICL improves performance compared to prefix prompting by $26.1\%/41.4\%$ in GREEN/BERTScore on Findings and $6.5\%/-6.8\%$ on Impression. Similarly, 2-shot ICL shows a $22.2\%/33.0\%$ improvement on Findings and $9.6\%/-9.6\%$ on Impression. LoRA finetuning achieves the highest scores overall, outperforming prefix prompting by $263\%/277\%$ on Findings and $8.7\%/2.2\%$ on Impression. Among the evaluated metrics, BLEU exhibits the largest improvements (up to 562% for LoRA on Findings), whereas F1-SRRG-Bert and GREEN show the smallest gains. Across LLMs, Llama-3 and Vicuna perform best in ICL, while Mistral-7B achieves the highest performance in both prefix prompting and LoRA finetuning. The overall best-performing configuration is Mistral-7B with LoRA finetuning.

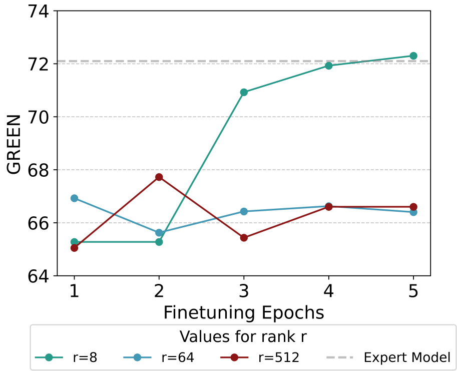


Figure 5: Performance of Mistral-7B finetuned with LoRA across varying ranks and an increasing number of fine-tuning epochs in comparison to the best lightweight model's performance.

4.3 Benchmarking

Building on these results, we benchmark our best lightweight model against Mistral-7B (the best-performing LLM) and GPT-4. Figure 5 illustrates the performance of Mistral-7B over training epochs when finetuned using LoRA with different ranks r. After five epochs, $r = 8$ marginally surpasses the lightweight model's GREEN score by 0.3%, while other configurations show no significant improvement over time. Table 1 details the trade-offs in training and inference, comparing training time per epoch, FLOPs per forward pass, inference latency, cost per sample, and CO2 emissions. Training the lightweight model requires only 10% of the time needed for Mistral-7B (with LoRA, $r = 8$, for 5 epochs). A single forward pass during inference consumes just 1.4% of the floating-point operations, leading to 7% of the inference time and 12.5% of the costs on our infrastructure. This translates to 14% of the CO2 footprint. GPT-4, estimated to be over 100× larger than Mistral-7B, outperforms both models, exceeding their GREEN scores by more than 27%, but at a substantially higher computational cost and C02 footprint.

4.4 Qualitative Analysis

To complement the quantitative analysis, Figure 1 presents a qualitative comparison of BERT2BERT, Mistral-7B, and expert-reviewed reports. Both models successfully adhere to our predefined template (see Figure 2 for reference), particularly in the Findings section, where content is well-aligned with organ system categories. A full test set analysis shows that the lightweight model correctly applies the Findings and Impression section headers in all cases, while the LLM deviates in 5% of instances, occasionally using all capital letters or omitting section names in less than 1% of reports. Both models, as well as expert annotations, generally include only relevant organ systems, but occasionally report less relevant negative findings (e.g., *"Pleura: - No specific findings reported"*). Complete omission of relevant findings occurs in less than 1% of cases, indicating high completeness in capturing clinical details. Differences in prioritization in the Impression section are observed in fewer than 5% of reports for both models, demonstrating occasional variation but overall consistency with expert-reviewed reports.

5 Discussion

In this paper, we propose the use of lightweight, task-specific models for structuring radiology reports into a predefined template. Despite being 10–30 times smaller than finetuned LLMs, our models achieve comparable performance while offering significant advantages in speed, cost-efficiency, and sustainability. To enable large-scale supervised training, we leveraged GPT-4 as a weak annotator to generate a training dataset, aligning chest radiology reports from MIMIC-CXR and CheXpert Plus with their corresponding structured versions as ground truth. Since GPT-generated data can introduce inconsistencies and biases, we evaluated all models on a human-annotated test set. Our study focused on two types of lightweight models, BERT2BERT and T5. Overall, our BERT2BERT model performed best when initialized from RoBERTa-PM-M3, surpassing the best

Table 1: Trade-off between model performance and computational costs for training and inference using total training time [h], GREEN Score [%], floating point operations required for a single forward pass [TFLOPs/sample], inference time [s/sample], inference cost [$/sample], and CO2 emissions [mg/sample] across the best-performing BERT2BERT, Mistral-7B, and GPT-4 using a single Nvidia A100 -80GB GPU.

	Model	BERT2BERT	Mistral-7B	GPT-4
	# Parameters	0.28B	7.25B	>1T
	Training time	2.1	21.2	-
Inference	GREEN	72.1	72.3	92.2
	TFLOPs	0.0480	3.51	>1,000
	Time	0.16	2.30	-
	Cost	2e-4*†	0.0032*†	0.03°
	CO_2 eq.	10.2	75.0	500

† Google Cloud Pricing Calculator.
° Accessed via OpenAI's API.

T5 variant, Clin-T5-Base, by 2.6% on GREEN. Our results further indicate that pretraining on biomedical texts - particularly radiology reports - generally improved model performance. However, despite being pretrained exclusively on radiology reports, the RadBERT model did not outperform general-text variants. This suggests that pretraining factors beyond the training corpus, such as architectural choices and optimization techniques, may also influence model performance. For example, RoBERTa-PM-M3 benefited from a distillation process, in which it was derived from RoBERTa-large-PM-M3-Voc, a larger model.

To balance performance with computational feasibility, we restricted our comparison to LLMs within the 3-8B parameter tier, evaluating different adaptation techniques within this range. We showed that finetuning with LoRA consistently yields higher performance compared to prefix prompting and ICL methods. As shown in Table 6, this trend is primarily driven by performance differences on the Findings section. Since our evaluation strategy assesses each organ system separately and missing or inconsistently phrased headers (e.g., *'Lungs and Airways'* vs. *'Lungs'*) receive a score of zero, these results suggest that LoRA fine-tuning more effectively adapts LLMs to follow our predefined template. We believe that although organ system names are provided in both the prefix prompt (see Appendix A.1) and the ICL examples, the absence of iterative feedback mechanisms in these methods makes it challenging for models to internalize and consistently enforce correct structured formatting. In contrast, LoRA allows for parameter updates, enabling the model to refine its representations over multiple training iterations. This reinforcement strengthens structural consistency and improves adherence to the predefined template.

Among the five evaluated LLMs and four adaptation techniques, Mistral-7B with LoRA finetuning achieved the best results. We selected it for benchmarking against our lightweight model in Section 4.3. When using prefix prompting, 1-shot, and 2-shot adaptation, Mistral-7B did not outperform our lightweight model. Comparable performance was only achieved after fine-tuning the LLM with LoRA ($r = 8$) for five epochs. While the LLM ultimately structured radiology reports according to our template and even surpassed the performance of our lightweight model by 0.3%, this came at the cost of significantly longer training times and higher inference costs. With less than 5% the size of the LLM, our lightweight model operated at 12.4% of its inference costs while running 14× faster, resulting in a significantly smaller carbon footprint. We then added GPT-4 to the comparison, which significantly outperformed the smaller models when adapted to this task. However, this superior performance may be partly attributed to GPT-4's prior involvement in data annotation and its use as a reference for radiologist annotations, potentially biasing evaluation in its favor. Despite GPT-4's strong performance, its proprietary nature, and regulatory constraints on patient data may limit its practicality in clinical settings.

Our qualitative analysis in Section 4.4 showed that both models (the lighweight model and Mistral-7B LLM finetuned with LoRA) followed the predefined template when tested on expert-annotated reports, omitting relevant findings in less than 1% of cases. This suggests that lightweight models (<300M parameters) can effectively learn structured formatting while maintaining clinical accuracy. Furthermore, the results indicate that our GPT-generated annotations provided a sufficient training signal, though expert review remains crucial for ensuring data reliability.

6 Conclusion

We demonstrate that lightweight, task-specific models with less than 300M parameters can effectively structure radiology reports according to a predefined template, providing a practical and scalable alternative to LLMs, while addressing concerns around computational efficiency, data privacy, and deployment feasibility. Our best-performing lightweight model, a BERT2BERT architecture initialized from two pretrained RoBERTa-PM-M3 models, achieves competitive performance while maintaining a significantly lower computational footprint. While Mistral-7B (LoRA, $r = 8$) achieves slightly better performance after five epochs of finetuning (0.3% on the GREEN Score), the lightweight model operates at less than 14% of its inference cost and CO2 emissions, making it a far more resource-efficient solution. GPT-4 surpasses both by over 27%, but its reliance on cloud-based APIs and privacy concerns make it impractical for direct clinical deployment. These findings reinforce the lightweight model's viability for real-world clinical applications, where infrastructure limitations, privacy regulations, and sustainability concerns play a critical role.

Limitations

First, as discussed in Section 3.1, the labels used for training our specialized models and adapting the LLMs were generated from MIMIC-CXR and CheXpert reports using GPT-4 as a weak annotator. While our prompt builds on previous work, we refined it to better align with our task's requirements (e.g., explicitly specifying organ systems for the Findings section). However, GPT-4 may introduce biases, and to mitigate this, we evaluate model performance on an independent test set annotated by five radiologists.

Second, both MIMIC-CXR and CheXpert originate from hospitals in the United States - Beth Israel Deaconess Medical Center (Boston, MA) and Stanford Hospital (Stanford, CA) - and contain only chest X-rays from adult patients. As a result, these datasets may lack demographic diversity, potentially limiting generalizability to other populations.

Third, as described in Section 3, all models take full free-form reports as input and generate structured reports comprising the following sections: Exam Type, History, Technique, Comparison, Findings, and Impression. However, for quantitative evaluation, we focus exclusively on Findings and Impression, as these sections are clinically critical and exhibit the highest variability. Other sections, such as Exam Type and History, often remain unchanged and can be directly copied from the original report, making them less relevant for assessing model performance.

Fourth, 1-shot and 2-shot ICL examples were manually selected from the training set to best represent the data distribution. While we initially applied algorithmic methods to optimize alignment, manual selection proved to further improve performance. This introduces a potential selection bias, which may affect the generalizability of our ICL results.

Fifth, due to computational and time constraints, we did not perform full-parameter fine-tuning on LLMs in the 3–8B parameter range. Instead, adaptation methods such as LoRA were used to efficiently finetune models within our resource limits.

Sixth, since GPT-4's exact architecture and parameter count remain undisclosed, Table 1 provides estimated values for its parameter size, floating point operations per forward pass, and CO_2 emissions. These estimates introduce uncertainty in direct comparisons with open-source models.

Seventh, as discussed in Section 3.5 and shown in Tables 5 and 7, GPT-4 was evaluated using prefix prompting and ICL. However, since it was also used for data annotation and provided as a reference for radiologist, its results may be biased in its favor. To account for this, we excluded GPT-4 from large parts of the discussion to avoid misleading comparisons.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Lisa C Adams, Daniel Truhn, Felix Busch, Avan Kader, Stefan M Niehues, Marcus R Makowski, and Keno K Bressem. 2023. Leveraging gpt-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. *Radiology*, 307(4):e230725.

Nurbanu Aksoy, Nishant Ravikumar, and Alejandro F Frangi. 2023. Radiology report generation using transformers conditioned with non-imaging data. In *Medical Imaging 2023: Imaging Informatics for Healthcare, Research, and Applications*, volume 12469, pages 146–153. SPIE.

Anonymous. 2025. Automatic structured radiology report generation. Under review.

Hassaan B Arshad, Sara A Butt, Safi U Khan, Zulqarnain Javed, and Khurram Nasir. 2023. Chatgpt and artificial intelligence in hospital level research: potential, precautions, and prospects. *Methodist DeBakey cardiovascular journal*, 19(5):77.

Laura Bergomi, Tommaso M Buonocore, Paolo Antonazzo, Lorenzo Alberghi, Riccardo Bellazzi, Lorenzo Preda, Chandra Bortolotto, and Enea Parimbelli. 2024. Reshaping free-text radiology notes into structured reports with generative question answering transformers. *Artificial Intelligence in Medicine*, 154:102924.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Felix Busch, Lena Hoffmann, Daniel Pinto Dos Santos, Marcus R Makowski, Luca Saba, Philipp Prucker, Martin Hadamitzky, Nassir Navab, Jakob Nikolas Kather, Daniel Truhn, et al. 2024. Large language models for structured reporting in radiology: past, present, and future. *European Radiology*, pages 1–14.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Curtis P Langlotz, et al. 2024. Chexpert plus: Hundreds of thousands of aligned radiology texts, images and patients. *arXiv e-prints*, pages arXiv–2405.

Dong Chen, Shuo Zhang, Yueting Zhuang, Siliang Tang, Qidong Liu, Hua Wang, and Mingliang Xu. 2024a. Improving large models with small models: Lower costs and better performance. *arXiv preprint arXiv:2406.15471*.

Qi Chen, Yutong Xie, Biao Wu, Xiaomin Chen, James Ang, Minh-Son To, Xiaojun Chang, and Qi Wu. 2024b. Act like a radiologist: Radiology report generation across anatomical regions. In *Proceedings of the Asian Conference on Computer Vision*, pages 1–17.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Raul Salles de Padua and Imran Qureshi. 2024. Leveraging summary of radiology reports with transformers. *Artificial Intelligence in Health*, 1(4):85–96.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis P Langlotz. 2022. Improving the factual correctness of radiology report generation with semantic rewards. *arXiv preprint arXiv:2210.12186*.

Daniel Pinto dos Santos, Elmar Kotter, Peter Mildenberger, and Luis Martí-Bonmatí. 2023. Esr paper on structured reporting in radiology—update 2023. *Insights into Imaging*, 14(1):199.

Roselyn Gabud, Portia Lapitan, Vladimir Mariano, Eduardo Mendoza, Nelson Pampolina, Maria Art Antonette Clariño, and Riza Theresa Batista-Navarro. 2023. A hybrid of rule-based and transformer-based approaches for relation extraction in biodiversity literature. In *Proceedings of the 2nd Workshop on Pattern-based Approaches to NLP in the Age of Deep Learning*, pages 103–113.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Sebastian Griewing, Fabian Lechner, Niklas Gremke, Stefan Lukac, Wolfgang Janni, Markus Wallwiener, Uwe Wagner, Martin Hirsch, and Sebastian Kuhn. 2024. Proof-of-concept study of a small language model chatbot for breast cancer decision support–a transparent, source-controlled, explainable and data-secure approach. *Journal of Cancer Research and Clinical Oncology*, 150(10):1–12.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

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.

Amir M Hasani, Shiva Singh, Aryan Zahergivar, Beth Ryan, Daniel Nethala, Gabriela Bravomontenegro, Neil Mendhiratta, Mark Ball, Faraz Farhadi, and Ashkan Malayeri. 2024. Evaluating the performance of generative pre-trained transformer-4 (gpt-4) in standardizing radiology reports. *European Radiology*, 34(6):3566–3574.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2020. Mimic-iv. *PhysioNet. Available online at:*

https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021), pages 49–55.

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.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Charles E Kahn Jr, Curtis P Langlotz, Elizabeth S Burnside, John A Carrino, David S Channin, David M Hovsepian, and Daniel L Rubin. 2009. Toward best practices in radiology reporting. *Radiology*, 252(3):852–856.

Dhruv Khullar, Xingbo Wang, and Fei Wang. 2024. Large language models in health care: Charting a path toward accurate, explainable, and secure ai. *Journal of General Internal Medicine*, pages 1–3.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. 2022. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*.

Eric Lehman and Alistair Johnson. 2023. Clinical-t5: Large language models built using mimic clinical text. *PhysioNet*.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: understanding and extending the state-of-the-art. In *Proceedings of the 3rd clinical natural language processing workshop*, pages 146–157.

Dacheng Li, Rulin Shao, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. 2023. How long can context length of open-source llms truly promise? In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Vladislav Lialin, Vijeta Deshpande, and Anna Rumshisky. 2023. Scaling down to scale up: A guide to parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.15647*.

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

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Chandreen R Liyanage, Ravi Gokani, and Vijay Mago. 2024. Gpt-4 as an x data annotator: Unraveling its performance on a stance classification task. *PloS one*, 19(8):e0307741.

NCBI. 1996. PubMed.

NCBI. 2000. PubMed Central (pmc).

OpenAI. 2022. Gpt-3.5. https://openai.com/.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, et al. 2024. Green: Generative radiology report evaluation and error notation. *arXiv preprint arXiv:2405.03595*.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. Automated annotation with generative ai requires validation. *arXiv preprint arXiv:2306.00176*.

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.

Branislav Pecher, Ivan Srba, and Maria Bielikova. 2024. Comparing specialised small and general large language models on text classification: 100 labelled samples to achieve break-even performance. *arXiv preprint arXiv:2402.12819*.

Long N Phan, James T Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature. *arXiv preprint arXiv:2106.03598*.

Radiological Society of North America (RSNA). 2011. Radreport: Radiology reporting templates. template rpt144. Accessed: 2024-02-07.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Delaram Rezaeikhonakdar. 2023. Ai chatbots and challenges of hipaa compliance for ai developers and vendors. *Journal of Law, Medicine & Ethics*, 51(4):988–995.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Jaromir Savelka, Kevin D Ashley, Morgan A Gray, Hannes Westermann, and Huihui Xu. 2023. Can gpt-4 support analysis of textual data in tasks requiring highly specialized domain expertise? *arXiv preprint arXiv:2306.13906*.

Jackson M Steinkamp, Charles Chambers, Darco Lalevic, Hanna M Zafar, and Tessa S Cook. 2019. Toward complete structured information extraction from radiology reports using machine learning. *Journal of digital imaging*, 32:554–564.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, et al. 2024. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493.

Dave Van Veen, Cara Van Uden, Maayane Attias, Anuj Pareek, Christian Bluethgen, Malgorzata Polacin, Wah Chiu, Jean-Benoit Delbrouck, Juan Manuel Zambrano Chaves, Curtis P Langlotz, et al. 2023. Radadapt: Radiology report summarization via lightweight domain adaptation of large language models. *arXiv preprint arXiv:2305.01146*.

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. 2022. Radbert: adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, et al. 2023. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9).

Kuo Zhang, Xiangbin Meng, Xiangyu Yan, Jiaming Ji, Jingqian Liu, Hua Xu, Heng Zhang, Da Liu, Jingjia Wang, Xuliang Wang, et al. 2025. Revolutionizing health care: The transformative impact of large language models in medicine. *Journal of Medical Internet Research*, 27:e59069.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D Manning, and Curtis P Langlotz. 2018. Learning to summarize radiology findings. *arXiv preprint arXiv:1809.04698*.

A Appendix

A.1 GPT-4 prompt template for structuring of radiology reports

The following prompt was executed with GPT-4 "Turbo 1106 preview" via Azure services to structure free-text radiology reports according to our template. The account was explicitly opted out of human review.

```
"Your task is to improve the formatting of a radiology report to a clear and concise
radiology report with section headings.
Guidelines:
   1. Section Headers: Each section should start with the section header followed by
    a colon. Provide the relevant information as specified for each section.
    2. Identifiers: Remove sentences where identifiers have been replaced with
    consecutive underscores ('\_\_\_').
    3. Findings and Impression Sections: Focus solely on the current examination
    results. Do not reference previous studies or historical data.
    4. Content Restrictions: Strictly include only the content that is relevant to
    the structured sections provided. Do not add or extrapolate information beyond
    what is found in the original report. If the original report doesn't contain
    the information necessary to generate a section, write the section header and
    then leave the section empty. Do not make up any findings.!
    Sections to include (if applicable):
    1. Exam Type: Provide the specific type of examination conducted.
    2. History: Provide a brief clinical history and state the clinical question or
    suspicion that prompted the imaging.
    3. Technique: Describe the examination technique and any specific protocols
    used.
    4. Comparison: Note any prior imaging studies reviewed for comparison with the
    current exam.
    5. Findings:
       Describe all positive observations and any relevant negative observations for
        each organ or organ system under distinct headers.
       Start with the organ system name followed by a colon, then list observations.
        Here is the corresponding template:
            Organ 1:
                - Observation 1
            Organ 2:
                - Observation 1
                - Observation 2
    Use only the following headers for organ systems:
    - Lungs and Airways
    - Pleura
    - Cardiovascular
    - Hila and Mediastinum
    - Tubes, Catheters, and Support Devices
    - Musculoskeletal and Chest Wall
    - Abdominal
    - Other
   6. Impression: Summarize the key findings with a numbered list from the most to the
    least clinically relevant. Ensure all findings are numbered.
The radiology report to improve is the following: \{report\}"
```

A.2 Overview of model checkpoints and pre-training data

Model	Description
T5-BASE (Raffel et al., 2020)	Original model, pre-trained on C4.
FLAN-T5-BASE (Chung et al., 2024)	Additional instruction-prompt tuning.
SCIFIVE (Phan et al., 2021)	Fine-tuned on PubMed Abstract (NCBI, 1996), and PubMed Central (NCBI, 2000).
CLIN-T5-SCI (Lehman and Johnson, 2023)	Fine-tuned on PubMed, MIMIC-III (Johnson et al., 2016), and MIMIC-IV (Johnson et al., 2020).
CLIN-T5-BASE (Lehman and Johnson, 2023)	Fine-tuned on MIMIC-III and MIMIC-IV.

Table 2: Pretrained T5 models used for initialization along with details of their pretraining corpus.

Model	Description
RoBERTa-base (Liu, 2019)	Baseline version, pretrained on Books and Wikipedia.
BioMed-RoBERTa (Gururangan et al., 2020)	Pretrained on PubMed abstracts and PubMed Central.
RoBERTa-base-PM-M3-Voc-distill-align (Lewis et al., 2020)	Pretrained on PubMed abstracts, PubMed Central full-text articles, and MIMIC-III.
RadBERT-RoBERTa (Yan et al., 2022)	Fine-tuned on radiology reports from the Veterans Affairs health care system.

Table 3: Pretrained RoBERTa models used for initialization of the BERT2BERT model along with details of their pretraining corpus.

A.3 Considerations and hyperparameters for end-to-end training

We train all expert models (BERT2BERT and T5 instances) with the following set of hyperparameters:

- Cosine learning rate scheduler, starting at $1e^{-4}$, with 5% warm-up ratio before decay.
- Maximum of 10 epochs, with early stopping enabled by loading the best model at the end based on validation performance.
- Batch size of 32 per device for training and 16 for evaluation, with four gradient accumulation steps, resulting in an effective batch size of 128 for training.
- Adam optimizer with $\beta_2 = 0.95$ and weight decay of 0.1.
- Sequence lengths: Model processes a maximum input length of 370 tokens, with generated outputs constrained between 120 and 286 tokens.

We experimented with different learning rate schedulers and initial learning rates but found the here presented set to give better performance in the validation loss.

A.4 Considerations and hyperparameters for parameter-efficient fine-tuning

As discussed in Section 3.4, we initially finetune all LLMs using the same hyperparameters. We apply LoRA and adjust the target modules to align with each LLM's architecture. We find that, due to their comparable size, using the same LoRA rank and scaling factor leads to a similar proportion of updated parameters across all models ($\sim 0.1\%$). We use the following set of hyperparameters:

- Cosine learning rate scheduler, starting at $1e^{-4}$, with 5% warm-up ratio before decay.
- Maximum of 5 epochs, with early stopping enabled by loading the best model at the end based on validation performance.
- LoRA adaptation with rank $r = 8$ and scaling factor $\alpha = 8$ to enable parameter-efficient fine-tuning.
- Batch size of 16 per device for training and 1 for evaluation, with 16 gradient accumulation steps, resulting in an effective training batch size of 256.
- Adam optimizer with $\beta_2 = 0.95$ and weight decay of 0.1.

We use similar settings as in expert model fine-tuning but reduce the maximum number of epochs due to computational constraints. The results in Section 4.3 later confirm our initial estimate for the optimal LoRA rank.

A.5 Detailed Evaluations of Model Performance

Table 4: Detailed comparison of expert models. This table presents test set evaluations of our fine-tuned expert models initialized from different pretrained checkpoints. Each model was trained three times with different random seeds and evaluated on the Findings sections of the MIMIC (F_M) and CheXpert (F_C) test sets, as well as their corresponding Impression sections (I_M and I_C).

Model	**Section**	**BLEU**	**ROUGE-L**	**BERTScore**	**RadGraph**	**GREEN**	**F1-Score**
BERT2BERT							
roberta-base	F_M	31.3	62.2	**67.4**	54.8	66.1	**73.0**
	F_C	30.6	59.0	64.7	50.1	63.0	69.4
	I_M	41.1	65.4	79.7	57.5	65.6	**81.8**
	I_C	51.1	74.9	86.3	66.1	82.0	94.5
roberta-biomed	F_M	31.6	60.4	65.4	53.1	62.8	70.4
	F_C	29.4	57.8	63.8	48.2	62.1	70.0
	I_M	34.0	65.5	79.9	58.0	69.1	**81.8**
	I_C	48.3	74.1	86.1	65.3	82.0	91.3
roberta-PM	F_M	**33.3**	**62.6**	**67.4**	54.3	**67.0**	71.9
	F_C	**32.8**	**62.5**	**67.3**	**53.8**	**64.2**	**72.8**
	I_M	42.0	66.1	79.8	56.5	**71.8**	81.4
	I_C	**53.4**	**77.6**	**87.5**	67.7	86.4	90.1
roberta-rad	F_M	32.6	62.1	66.8	**54.9**	64.8	71.8
	F_C	29.4	59.2	64.2	50.7	61.0	69.1
	I_M	**42.3**	**67.5**	**80.6**	**58.9**	69.7	81.7
	I_C	52.4	76.6	87.2	65.7	86.7	94.3
T5							
T5-Base	F_M	26.4	52.8	58.8	64.9	58.6	63.6
	F_C	26.0	57.2	61.9	49.1	59.7	66.5
	I_M	35.8	61.7	77.7	56.2	69.8	80.1
	I_C	48.5	73.2	85.8	**67.9**	81.2	87.1
Flan-T5-Base	F_M	27.9	55.9	61.0	48.0	59.3	65.4
	F_C	30.3	59.2	63.5	51.1	62.2	66.2
	I_M	37.3	62.0	77.6	55.5	66.2	77.8
	I_C	51.6	76.1	87.1	68.6	82.3	91.7
SciFive	F_M	24.1	49.3	55.6	43.4	56.4	62.0
	F_C	24.6	54.1	60.5	47.2	56.7	65.7
	I_M	38.6	63.2	78.8	59.5	**71.8**	82.9
	I_C	46.8	71.4	85.1	68.1	77.8	89.4
Clin-T5-Sci	F_M	28.7	59.0	64.4	50.7	62.4	68.9
	F_C	23.4	52.5	57.1	44.0	56.1	62.0
	I_M	33.6	59.4	76.2	51.4	63.8	76.3
	I_C	46.7	71.8	84.6	62.8	84.0	93.0
Clin-T5-Base	F_M	29.8	58.3	64.0	50.9	62.7	68.6
	F_C	27.1	57.3	62.0	49.0	60.9	68.1
	I_M	37.6	63.3	78.9	55.7	68.7	80.2
	I_C	48.4	74.8	85.5	**67.9**	**88.8**	**94.6**

Table 5: Comparison of LLM performance across different adaptation and fine-tuning methods. Results are averaged over all samples in the expert-reviewed MIMIC and CheXpert test sets and reported separately for the Findings and Impression sections. The highest score for each model across adaptation techniques is highlighted.

Model	Method	BLEU	ROUGE-L	BERTScore	Radgraph	GREEN	F1-Score
Findings Section							
Medalpaca-7B	Prefix	0.0	0.0	0.0	0.0	0.0	0.0
	1-shot	0.0	0.2	1.4	0.1	0.1	0.9
	2-shot	0.0	0.0	0.0	0.0	0.0	0.0
	LoRA	**19.7**	**45.4**	**50.5**	**41.3**	**51.0**	**57.1**
Phi-3.5-mini	Prefix	11.0	34.6	38.9	26.7	38.1	46.5
	1-shot	8.6	21.5	24.8	20.1	25.6	26.4
	2-shot	6.8	20.1	24.1	18.5	23.2	25.8
	LoRA	**17.8**	**43.8**	**49.5**	**39.0**	**46.7**	**52.9**
Vicuna-7B	Prefix	0.0	0.0	0.0	0.0	0.0	0.0
	1-shot	5.9	21.5	29.2	17.5	22.8	32.4
	2-shot	7.1	19.8	24.6	17.0	22.6	28.2
	LoRA	**32.7**	**62.1**	**66.8**	**54.2**	**66.1**	**70.6**
Llama-3-8B	Prefix	2.4	10.9	12.8	8.6	13.1	12.7
	1-shot	13.1	35.6	42.1	30.6	40.1	46.4
	2-shot	13.7	36.4	42.1	31.1	38.0	46.4
	LoRA	**35.0**	**62.9**	**68.4**	**54.4**	**67.4**	**74.0**
Mistral-7B	Prefix	8.2	26.8	30.3	6.9	32.5	35.8
	1-shot	6.5	15.2	18.4	14.7	16.9	19.4
	2-shot	5.9	14.9	18.1	12.5	18.5	18.4
	LoRA	**37.5**	**69.3**	**73.6**	**61.2**	**72.4**	**77.7**
GPT-4	Prefix	**79.9**	**94.3**	**94.8**	**91.5**	**86.9**	**95.4**
	1-shot	17.1	64.2	77.2	54.7	92.1	92.5
	2-shot	23.1	70.9	81.2	59.9	**93.4**	93.9
Impression Section							
Medalpaca-7B	Prefix	23.6	55.1	63.9	52.0	75.6	80.8
	1-shot	23.3	54.0	60.7	50.3	66.8	74.1
	2-shot	**25.8**	**56.5**	**66.7**	**57.4**	**77.2**	**76.5**
	LoRA	17.4	53.5	63.4	38.4	68.9	86.2
Phi-3.5-mini	Prefix	19.2	45.7	63.7	43.7	51.5	76.0
	1-shot	24.4	48.6	66.8	47.7	65.3	77.8
	2-shot	32.6	48.5	66.8	51.9	71.8	79.2
	LoRA	**39.3**	**64.4**	**77.3**	**56.2**	**67.5**	**78.1**
Vicuna-7B	Prefix	34.0	**64.8**	73.7	57.8	71.9	79.6
	1-shot	**38.8**	64.7	**77.5**	**61.5**	71.9	**84.3**
	2-shot	36.8	62.9	76.8	59.5	71.8	82.3
	LoRA	38.0	63.7	70.9	54.3	**72.4**	81.5
Llama-8B	Prefix	25.5	55.4	70.7	51.3	61.9	77.5
	1-shot	9.7	27.5	45.6	33.1	73.5	63.9
	2-shot	10.6	30.1	49.3	32.6	74.0	68.2
	LoRA	**35.3**	**65.3**	**72.0**	**54.7**	**74.2**	**87.0**
Mistral-7B	Prefix	33.6	63.4	**78.4**	56.0	69.5	78.0
	1-shot	38.3	65.6	76.2	**62.9**	67.4	82.0
	2-shot	39.2	66.0	77.2	**62.9**	67.4	82.0
	LoRA	**42.3**	**67.6**	74.8	57.0	**76.1**	**84.8**
GPT-4	Prefix	**84.1**	**91.2**	**94.6**	**89.0**	**97.5**	**92.8**
	1-shot	26.2	60.0	77.9	54.6	76.7	88.0
	2-shot	36.1	67.4	80.9	64.9	81.6	88.5

Table 6: Detailed comparison of LLM adaptation methods for the Findings and Impression sections. The table shows average values across all five LLMs, along with percentage changes relative to performance under prefix prompting.

Method	**BLEU**	**ROUGE-L**	**BERTScore**	**Radgraph**	**GREEN**	**F1-Score**
Findings Section						
Prefix	4.31	14.4	16.4	8.43	16.7	19.7
1-shot	6.79 +57.5%	18.8 +30.2%	23.2 +41.4%	16.6 +96.8%	21.1 + 26.1%	25.1 + 27.3%
2-shot	6.67 +54.8%	18.2 + 26.3%	21.8 + 33.0%	15.8 +87.4%	20.4 +22.2%	23.8 +20.6%
LoRA	28.5 +562%	56.7 +293%	61.7 +277%	50.0 +493%	60.7 +263%	66.5 + 237%
Impression Section						
Prefix	27.2	56.9	70.1	52.1	66.1	78.4
1-shot	26.9 -1.1%	52.0 -8.5%	65.3 - 6.8%	50.7 - 2.7%	70.4 +6.65%	77.0 -11.8%
2-shot	26.8 -1.5%	52.8 -7.2%	67.3 -3.9%	52.8 +1.3%	72.4 +9.6%	77.6 -1.0%
LoRA	34.4 +26.8%	62.9 +10.6%	71.6 +2.2%	52.1 +0.0%	71.8 +8.7%	83.5 +6.5%

Table 7: Detailed comparison of expert models and LLMs across various NLP metrics. The table presents the best-performing BERT2BERT and T5 models, along with several LLMs fine-tuned for five epochs using LoRA. GPT-4's performance is obtained via prefix prompting, using the same prompt as in dataset generation, with a temperature of 0.3. Note that GPT-4 is excluded when identifying the best scores.

Model	**Section**	**BLEU**	**ROUGE-L**	**BERTScore**	**Radgraph**	**GREEN**	**F1-Score**
BERT2BERT	F_M	33.0	62.6	67.4	54.3	**66.7**	71.9
	F_C	32.8	62.5	67.3	53.8	66.1	72.8
	I_M	**42.0**	66.1	79.8	56.5	69.3	81.4
	I_C	**53.4**	**77.6**	**87.5**	67.7	86.3	90.1
T5	F_M	29.8	58.3	64.0	50.9	62.7	68.6
	F_C	27.1	57.3	62.0	49.0	60.9	68.1
	I_M	37.6	63.3	78.9	55.7	68.7	80.2
	I_C	48.4	74.9	85.5	**67.9**	**88.8**	94.6
Medalpaca-7B	F_M	27.8	51.6	56.7	45.7	55.7	60.0
	F_C	11.6	39.1	44.3	36.8	46.3	54.2
	I_M	19.6	57.5	61.8	52.7	**73.2**	77.4
	I_C	15.2	49.4	64.9	24.1	64.6	95.0
Phi-3.5-mini	F_M	18.3	48.0	54.6	39.9	50.4	57.9
	F_C	17.3	39.6	44.4	38.1	43.0	47.9
	I_M	39.3	**67.4**	**79.9**	**59.8**	67.5	**81.5**
	I_C	39.1	61.3	74.6	52.5	67.5	74.7
Vicuna-7B	F_M	32.2	58.8	64.3	51.2	62.6	68.5
	F_C	33.2	65.3	69.2	57.2	69.5	72.7
	I_M	32.2	57.3	65.3	48.0	63.4	76.0
	I_C	43.8	70.0	76.4	60.6	81.4	87.0
Llama-8B	F_M	32.3	61.2	67.0	53.4	65.7	72.2
	F_C	37.7	64.6	69.8	55.4	69.0	75.8
	I_M	33.4	59.7	67.7	51.8	71.2	77.3
	I_C	37.2	70.8	76.2	57.5	77.1	**96.7**
Mistral-7B	F_M	**33.3**	**63.0**	**68.1**	**56.2**	66.1	**73.7**
	F_C	**41.7**	**75.6**	**79.0**	**66.2**	**78.6**	**81.6**
	I_M	38.4	64.7	71.7	56.6	69.3	79.8
	I_C	46.1	70.4	77.8	57.4	82.9	89.8
GPT-4	F_M	79.0	93.2	93.7	91.1	92.4	94.8
	F_C	80.7	95.3	95.8	91.8	81.3	95.9
	I_M	70.5	83.3	89.5	79.7	95.0	85.5
	I_C	97.7	99.1	99.7	98.3	100.0	100.0