RadAdapt: Radiology Report Summarization via Lightweight Domain Adaptation of Large Language Models

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, **Akshay S. Chaudhari**, **John Pauly**
Stanford University
{vanveen, cvanuden}@stanford.edu

Abstract

We systematically investigate lightweight strategies to adapt large language models (LLMs) for the task of radiology report summarization (RRS). Specifically, we focus on domain adaptation via pretraining (on natural language, biomedical text, or clinical text) and via discrete prompting or parameter-efficient fine-tuning. Our results consistently achieve best performance by maximally adapting to the task via pretraining on clinical text and fine-tuning on RRS examples. Importantly, this method fine-tunes a mere 0.32% of parameters throughout the model, in contrast to end-to-end fine-tuning (100% of parameters). Additionally, we study the effect of in-context examples and out-of-distribution (OOD) training before concluding with a radiologist reader study and qualitative analysis. Our findings highlight the importance of domain adaptation in RRS and provide valuable insights toward developing effective natural language processing solutions for clinical tasks.

1 Introduction

Radiology reports are comprehensive documents that capture and interpret the results of a radiological imaging examination. Reports are often structured into three main sections: (1) a *background* section that provides general information about the exam and the patient, e.g. medical history (2) a *findings* section that presents detailed exam analysis and results, and (3) an *impression* section that concisely summarizes the most salient findings. In a typical workflow, a radiologist first dictates the detailed findings and then distills them into a concise impression. This impression is the most significant part of a radiology report, as it contains crucial information for clinical decision-making (Kahn Jr et al., 2009). However, performing radiology report summarization (RRS) manually can be labor-intensive and prone to errors (Gershanik et al., 2011), motivating the importance of automating this task.

Large language models (LLMs) have demonstrated remarkable capabilities in natural language understanding and generation, serving as foundation models that can be adapted to various domains and tasks. However, their sheer size, sometimes exceeding 100B parameters, makes training for domain-specific tasks prohibitively expensive in terms of computation and training data. We address this challenge by exploring lightweight strategies for domain adaptation in the context of RRS, culminating in the following contributions:

- We systematically evaluate a variety of LLMs and lightweight adaptation methods, achieving the best performance by pretraining on clinical text and performing parameter-efficient fine-tuning with LoRA (Hu et al., 2021). Generated impressions showcase the effectiveness of lightweight adaptation strategies for RRS.
- We investigate the impact of few-shot prompting by conducting ablation studies on the number of in-context examples provided to each model. Our findings reveal that increased context leads to improved performance across almost all cases, shedding light on the value of prompt engineering when adapting LLMs for RRS.
- We evaluate "out-of-distribution" (OOD) model performance and specifically examine the model's ability to generalize to different imaging modalities and anatomies. Our results indicate that anatomy plays a more crucial role than modality, and best performance is achieved when training on a larger dataset which encompasses all modalities and anatomies.

*Equal contribution

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 449–460

- We conduct a reader study with radiologists who provide qualitative insights and quantitative scores on the model's correctness, coherence, and ability to capture critical information. While our results are promising, we emphasize the need for further improvements and evaluation before clinical deployment.

Overall, our research presents a comprehensive investigation of lightweight strategies for domain adaptation in the context of RRS, offering insights into the effectiveness of different approaches and highlighting the potential of LLMs for this task. Our findings contribute toward the advancement of applied natural language processing (NLP) to radiology with implications for improving radiologists' workflow and patient care.

2 Related Work

In recent years, transformer-based (Vaswani et al., 2017) language models have become ubiquitous in NLP due to their state-of-the-art performance across many tasks including language generation, question answering, and machine translation. Transformer models BERT (Devlin et al., 2018) and GPT-2 (Radford et al., 2019) established a new paradigm of first training on large amounts of general data and then fine-tuning on domain-specific data, as opposed to direct training on domain-specific data. This has led to training transformers with more parameters on increasingly more data, resulting in LLMs such as GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2022), and the "text-to-text transfer transformer," or T5 (Raffel et al., 2020).

However, end-to-end fine-tuning LLMs like GPT-3 (175B parameters) requires substantial computational resources, creating a high barrier to entry. As a result, there has been a growing interest in lightweight methods for domain adaptation. One such method is prompting, in which one provides initial text input to the LLM so it has context for the given task. Performance depends heavily on the provided prompt (Brown et al., 2020), motivating principled prompting methods. Various works have pursued this in the form of natural language instructions (Liu et al., 2023; Wei et al., 2022) or supplying "in-context" examples of desired output (Lampinen et al., 2022). An alternative approach consists of parameter-efficient fine-tuning, where one freezes existing model weights and inserts a small number of tunable parameters (Rebuffi et al., 2017; Houlsby et al., 2019; Lin et al., 2020). We focus on the two highest regarded parameter-efficient fine-tuning methods: prefix tuning (Li and Liang, 2021; Lester et al., 2021) and LoRA (Hu et al., 2021), discussed further in Section 3.

In addition to methods for prompting and fine-tuning, previous work has demonstrated adaptation to the medical domain via pretraining on biomedical or clinical text. Consider SciFive (Phan et al., 2021), Clinical-T5 (Lehman and Johnson, 2023), and Med-PaLM (Singhal et al., 2022), which leveraged LLMs for text generation on medical tasks. Both SciFive (biomedical) and Clinical-T5 (clinical) achieved state-of-the art results for their respective domains in tasks such as named entity recognition, natural language inference, and question-answering. Additionally, Med-PaLM achieved success aligning text generation models to clinical tasks via methods such as chain-of-thought prompting, prompt tuning (Lester et al., 2021), and instruction tuning (Chung et al., 2022a).

Within the clinical domain, we focus specifically on the task of RRS. Previous work has approached this task with a focus on consistency and factual correctness (Zhang et al., 2020; Dai et al., 2021; Miura et al., 2021; Delbrouck et al., 2022a). To the best of our knowledge, this work is the first to leverage lightweight domain adaptation strategies on LLMs for the RRS task. Additionally, most prior work on RRS uses only chest x-rays (Dai et al., 2021; Abacha et al., 2021) from large datasets like MIMIC-CXR (Johnson et al., 2019a) and CheXpert (Irvin et al., 2019). In contrast, our work employs the MIMIC-III dataset (Johnson et al., 2016), which contains longer radiology reports from a diverse set of two imaging modalities (MR, CT) and seven anatomies (head, chest, etc.), presenting a more difficult summarization task.

3 Methods

As depicted in Figure 1, we investigate adapting LLMs to the task of RRS along two axes: (1) models pretrained on various combinations of natural language, biomedical, and clinical text data, as described in Section 3.1, and (2) various methods for discrete prompting and parameter-efficient fine-tuning, as described in Section 3.2.

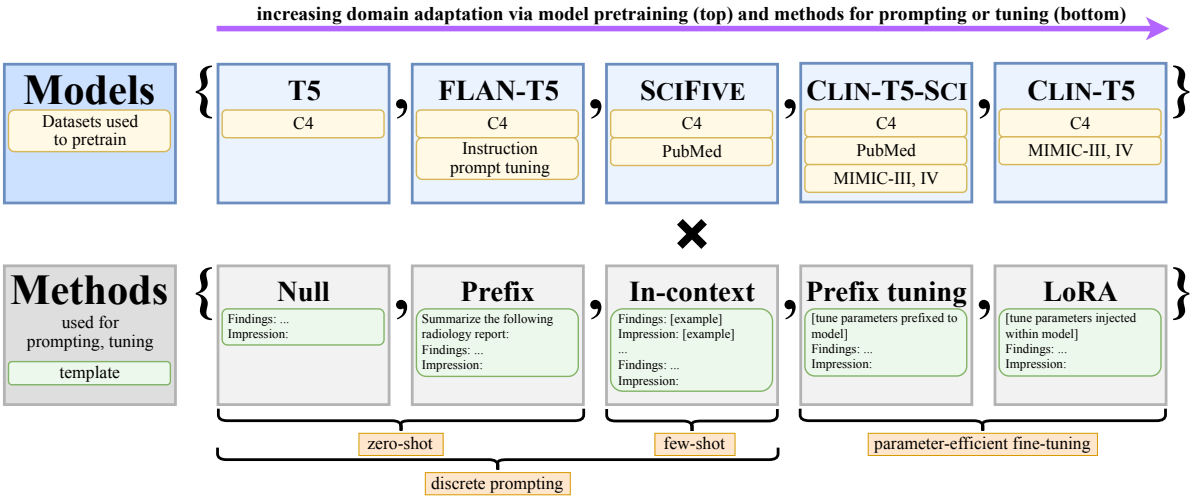


Figure 1: Diagram of experiments. We evaluate every combination of pretrained LLM (top) and lightweight adaptation method (bottom). Moving from left to right, the models and methods become increasingly adapted to the downstream clinical task of RRS.

3.1 Pretrained Models

To mitigate variance introduced by different model architectures, we focus this study on T5, i.e. "text-to-text transfer transformer," a highly regarded encoder-decoder architecture available for public use (Raffel et al., 2020). T5's text-to-text framework enables the model to be used on any NLP task, and its pretraining over the C4 (Colossal Clean Crawled Corpus) dataset (Raffel et al., 2020) enables excellent performance in transfer learning. We include results for two architecture sizes: base (223M parameters), and large (738M parameters). From hereon we refer to T5 as the original model pretrained on C4 alone. The remaining four models are simply a version of T5 that was subsequently end-to-end fine-tuned on datasets of various relevance to the RRS task:

- FLAN-T5 (Chung et al., 2022b) tuned via instruction prompt tuning.
- SCIFIVE (Phan et al., 2021) tuned on the biomedical text dataset of PubMed, i.e. Pubmed Abstract (NCBI, 1996) and PubMed Central (NCBI, 2000).
- CLIN-T5-SCI or Clinical-T5-Sci (Lehman and Johnson, 2023) tuned on PubMed and two clinical text datasets (MIMIC-III (Johnson et al., 2016) and MIMIC-IV (Johnson et al., 2020a)).
- CLIN-T5 or Clinical-T5 (Lehman and Johnson, 2023) tuned on MIMIC-III and IV alone.

Please see the top row in Figure 1 for a visual illustration of these five models. We acknowledge the difficulty of ranking pretraining datasets' relevance for a particular downstream task. In the case of RRS, it may seem reasonable to assert that clinical text is more relevant than biomedical text, and that biomedical text is more relevant than general natural language text. However, it becomes more difficult to compare FLAN-T5's instruction tuning, which drastically improves performance on prompting benchmarks but has not yet been explored for medical tasks (Longpre et al., 2023). We also note the complication introduced by CLIN-T5-SCI and CLIN-T5 being trained on a subset of MIMIC-III, the same dataset used for evaluation; please see Section 5.4 for further discussion.

3.2 Lightweight task adaptation methods

For each pretrained model discussed in Section 3.1, we evaluate five lightweight domain adaptation methods for prompting and tuning (Fig. 1, bottom row). Prompting provides adaptation by simply supplying particular tokens to the frozen pretrained model. In contrast, parameter-efficient fine-tuning provides adaptation by adding a small number of parameters to the model and optimizing them for the task, while the original model parameters remain frozen. Compared to updating all model parameters via end-to-end fine-tuning, parameter-efficient methods require much less computation and training data. We describe these five methods below in order of increasing adaptation to the downstream RRS task:

Table 1: We employ parameter-efficient fine-tuning methods for domain adaptation that modify <0.4% of model parameters while keeping other parameters frozen.

Model size	Method	Tunable parameters #	Tunable parameters % of total	Training time (hr) per epoch	Training time (hr) total	# epochs
Base (223M)	prefix tuning	0.37M	0.17%	0.98	9.83	10
	LoRA	0.88M	0.39%	1.32	6.60	5
Large (738M)	prefix tuning	0.98M	0.13%	2.93	29.3	10
	LoRA	2.4M	0.32%	3.85	19.3	5

1. *Null prompting* (Zhao and Schütze, 2021) is a simple discrete (sequence of real natural-language tokens), zero-shot prompt. We supply the radiology report findings section and the basic prompt, "impression:".

2. *Prefixed prompting* (Zhao and Schütze, 2021) is a discrete, zero-shot prompt with a brief instruction prepended to the original null prompt above. For our instruction we use "summarize the following radiology report:". This provides the model some context for the RRS task. We note that a slight modification to the prepended instruction may significantly change the generated output for an individual sample. However, that same modification does not meaningfully alter quantitative metrics when applied over the entire dataset.

3. *In-context learning* (Lampinen et al., 2022) is a type of discrete, few-shot prompt. We begin with the null prompt and prepend one, two, or four task examples using the same template. Particular examples are chosen by computing the k-nearest neighbors (Johnson et al., 2019b) of the findings section for each training example in the embedding space of a PubMedBERT model (Deka et al., 2022). This provides the most relevant examples for the RRS task.

4. *Prefix tuning* (Li and Liang, 2021) is a parameter-efficient fine-tuning method which prepends and optimizes an additional task-specific vector called the *prefix* as input to the model. For our base and large architectures, this requires tuning a mere 0.17% and 0.13% of total parameters, respectively (see Table 1). This approach provides the model a task-specific prompt that is very well aligned to the downstream task.

5. *LoRA* (Hu et al., 2021), or low-rank adaptation, approximates the fine-tuning process by injecting trainable rank decomposition matrices into each architecture layer. Compared to prefix tuning, which requires some portion of the input sequence for adaptation, LoRA beneficially preserves the entire sequence length for the downstream task. For our base and large architectures, this requires tuning a mere 0.39% and 0.32% of total parameters, respectively (see Table 1). Because LoRA modifies slightly more parameters than prefix tuning, we characterize this method as having greater domain adaptation.

4 Experiments

4.1 Data

Our primary dataset is MIMIC-III (Johnson et al., 2016), which contains 79,790 radiology reports across two imaging modalities and seven anatomies. Table 2 contains a dataset overview. Figure 2 contains an example of a CT head report; recall the task of RRS is to predict the report impressions

Table 2: Number of reports in MIMIC-III by modality, anatomy, and dataset split.

Modality/ Anatomy	Number of reports Train	Val	Test
CT head	25,122	3,140	3,141
CT abdomen	12,792	1,599	1,599
CT chest	10,229	1,278	1,280
MR head	5,851	731	732
CT spine	4,414	551	553
CT neck	912	114	115
MR spine	-	-	2,822
CT sinus	-	-	1,268
MR abdomen	-	-	1,062
MR pelvis	-	-	254
MR neck	-	-	231

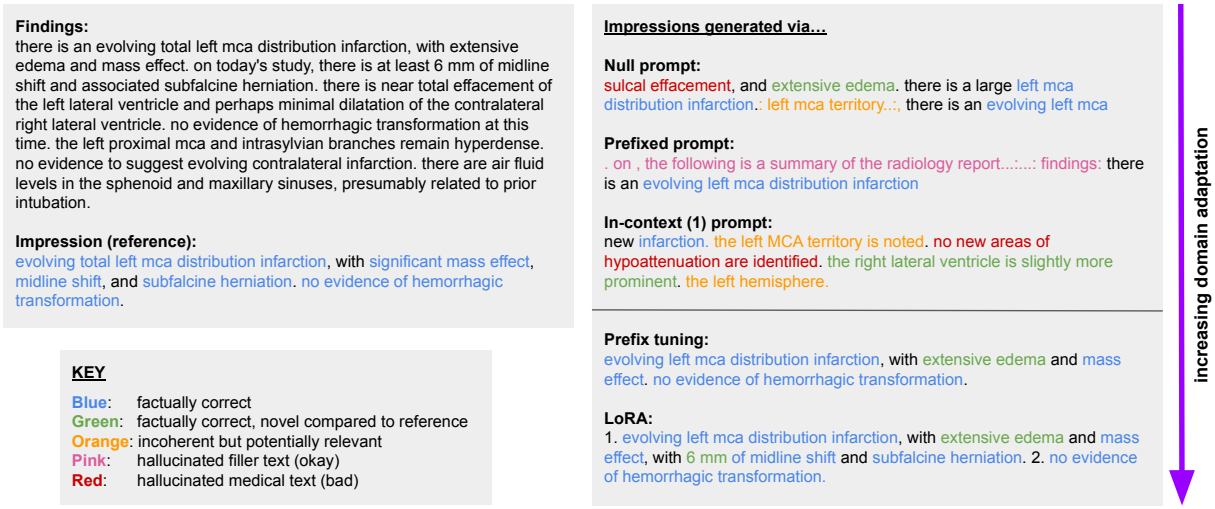


Figure 2: Example radiology report. Left: Findings and reference impression. Right: Generated impressions with various methods for discrete prompting (top) and parameter-efficient fine-tuning (bottom), all using the CLIN-T5-LARGE model. Color annotations were provided by a radiologist who specializes in the relevant anatomy (head).

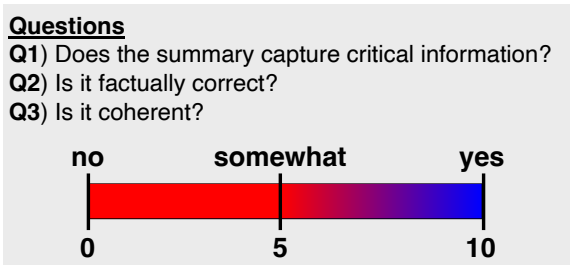
section (label) given the findings section (input). We also provide some evaluations on a secondary radiology report dataset, MIMIC-CXR (Johnson et al., 2019a). This is an easier summarization task as it contains one modality and anatomy (chest x-rays) with generally shorter impression sections than MIMIC-III. PhysioNet (Johnson et al., 2020b) and ViLMedic (Delbrouck et al., 2022b) provided access to pre-processed versions of these datasets, removing confidential patient information. We conduct additional processing according to domain adaptation strategies discussed in Section 3.2.

4.2 Evaluation

For quantitative evaluation of our generated impressions, we employ common summarization metrics such as BLEU and ROUGE-L. Between a given pair of reference and generated text, BLEU evaluates overlap using a weighted average of 1- to 4-gram precision, and ROUGE-L evaluates the longest common subsequence overlap. Beyond these token-level syntactic similarity metrics, we employ by also using metrics like BERTScore (via HuggingFace), which computes the semantic similarity between the reference and generated texts using BERT embeddings (Zhang* et al., 2020). Lastly, following previous work (Delbrouck et al., 2022a), we evaluate our model with F1-RadGraph, a F-score style metric that measures the factual correctness, consistency and completeness of generated radiology reports compared to the reference. F1-RadGraph uses RadGraph (Jain et al., 2021), a graph dataset of entities and relations present in radiology reports.

For qualitative evaluation, we include a reader study (Figure 3) with three board-certified radiologists. Each evaluated a randomly selected group of twenty generated impressions in comparison to reference impressions. They responded with either 0 ("no"), 5 ("somewhat"), or 10 ("yes") to the following three questions:

1. Does the generated impression capture critical information? Consider if the model missed any important findings.



Reader	Q1	Q2	Q3
1	8.8 ± 2.2	8.8 ± 2.2	10. ± 0.0
2	8.0 ± 2.5	8.8 ± 2.2	9.0 ± 2.6
3	9.0 ± 2.1	8.9 ± 2.1	10. ± 0.0
Pooled	8.6 ± 2.3	8.8 ± 2.1	9.7 ± 1.6

Figure 3: Radiology reader study. Top: Study design. Bottom: Results via CLIN-T5-LARGE + LoRA on random samples from the CT head dataset. The model scores highest in coherence (Q3) and generally performs well capturing critical information (Q1) in a factually correct way (Q2). Each entry's highlight color corresponds to its location on the above color spectrum.

Method	Model	BLEU	ROUGE-L	BERT	F1-Radgraph
Prefix tuning	T5	12.9	29.1	88.4	30.7
	SciFive	10.3	28.9	88.4	30.2
	Clin-T5-Sci	11.7	33.3	89.3	35.0
	Clin-T5	**11.9**	**33.8**	**89.4**	**35.4**
LoRA	T5	13.7	33.9	89.5	35.2
	SciFive	13.5	34.6	89.6	36.1
	Clin-T5-Sci	13.4	36.4	89.9	37.6
	Clin-T5	**14.8**	**36.8**	**89.9**	**38.2**

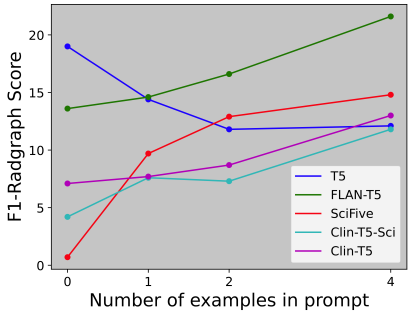


Figure 4: Domain adaptation. Left: Adaptation via pretraining on increasingly relevant data (T5, SciFive, Clin-T5-Sci, Clin-T5) generally leads to improved performance for both fine-tuning methods. Note we exclude FLAN-T5, whose degree of domain adaptation is difficult to rank. See Table 5 in the appendix for comprehensive results. Right: Adaptation via increasing number of in-context examples leads to improved performance in most models. See Section 5.1 for discussion.

2. Is the generated impression factually correct? Given the text that the model did generate, consider its correctness.

3. Is the generated impression coherent, i.e. do you find the syntax comprehensible?

4.3 Experimental details

No further model tuning was needed for the null, prefix, and in-context discrete prompting experiments; we simply run inference with the pre-existing LLM. When tuning model hyperparameters for prefix tuning and LoRA, we discovered that the same set of hyperparameters achieves best performance across all five models. This seems reasonable given that each model employs the same architecture. As such we tuned each model with the same set of hyperparameters. Please see Section A.1 in the appendix for details.

5 Results and Discussion

5.1 Domain Adaptation

A recurring theme throughout our results is that increased domain adaptation leads to improved performance. One axis for domain adaptation is the relevance of pretraining data to the evaluation task, in our case RRS on the MIMIC-III dataset. Figure 4 (left) demonstrates that pretraining the same architecture on increasingly relevant data (T5, SciFive, Clin-T5-Sci, Clin-T5) typically improves performance for prefix tuning and LoRA. We do not include FLAN-T5 in this portion of the analysis, as its relative degree of domain adaptation is difficult to assess.

The second axis for domain adaptation is the amount of context provided either via longer discrete prompts (null, prefix, in-context prompting), or parameter-efficient fine-tuning (prefix tuning, LoRA). For all models, fine-tuning significantly outperforms any discrete prompting technique. Discrete prompting is still useful, however, especially if one only has black-box access to the LLM. Therefore, we compare techniques for discrete prompting in Figure 4 (right), demonstrating that prompting the model with more in-context examples improves performance for almost every model with the lone exception of T5. As T5 is the least domain-adapted model, perhaps a small number of in-context examples may actually hurt performance when examples are sufficiently out-of-domain. We leave this question to future work. Another component of Figure 4 (right) is that the remaining four models improve at different rates: SciFive seems to level off near four in-context examples, while the others continue to steadily increase. This supports our hypothesis that instruction prompt tuning and maximal domain adaptation (in this case, to clinical text) improve results seen with in-context learning. This also motivates future work in instruction prompt tuning a model using domain-specific biomedical or clinical instructions.

Table 5 in the appendix includes an ablation of all configurations for domain adaptation using a base architecture (223M parameters). We subsequently take the best model (Clin-T5) for the best methods (prefix tuning, LoRA) and scale to a large architecture (738M parameters) in Table 3. This scaling provides a significant performance boost. Similar to the base architecture, LoRA outperforms prefix tuning in all cases. Interestingly, Clin-T5-

Table 3: Best results overall. Top: Given that the base architecture (223M parameters) performs best via pretraining on clinical text (CLIN-T5) and subsequent fine-tuning, we improve performance on MIMIC-III by scaling to the large architecture (738M). Bottom: LoRA also outperforms prefix tuning on the MIMIC-CXR dataset using CLIN-T5.

Dataset	Method	Size	BLEU	ROUGE-L	BERT	F1-Radgraph	F1-CheXbert
MIMIC-III	prefix tuning	base	11.9	33.8	89.4	35.4	-
		large	14.6	36.7	89.9	38.4	-
	LoRA	base	14.5	36.4	89.9	38.0	-
		large	**16.2**	**38.7**	**90.2**	**40.8**	-
MIMIC-CXR	prefix tuning	large	16.1	43.4	89.7	41.0	70.2
	LoRA	large	**18.9**	**44.5**	**90.0**	**41.8**	**70.9**

BASE + LoRA achieves nearly equivalent performance to CLIN-T5-LARGE + prefix tuning, exemplifying the benefits of LoRA over prefix tuning. Finally, after performing all prior analysis using the MIMIC-III dataset, we apply the best combination (CLIN-T5-LARGE + LoRA) on MIMIC-CXR and include results in Table 3.

5.2 Out-of-distribution performance

Table 4 demonstrates out-of-distribution performance using a CLIN-T5-BASE model prefix tuned on CT head data. When evaluating on a different test set, the model better summarizes reports on a different modality (MR head) than a different anatomy (CT other). This suggests that report finding tokens are more anatomy- than modality-specific, which seems reasonable. Counterintuitively, the model performs worse when shifting just the anatomy (CT other) compared to shifting the modality and anatomy (MR other). This could be due to a myriad of reasons, such as MR findings being 15% longer than CT findings and MR other containing fewer anatomies (four) than CT other (five). Lastly, we find that training CLIN-T5-BASE on all data leads to higher performance than training on CT head alone.

5.3 Qualitative Evaluation

5.3.1 Error Analysis

We perform a qualitative error analysis of 20 randomly selected reference findings, reference impressions, and generated impressions via CLIN-T5-LARGE on the CT head dataset. Please see Figure 2 for an example. We describe four types of deviations from the reference impressions and their corresponding colors used in Figure 2:

1. Factually correct text that is novel compared to the reference impression (green). This text is present in the reference findings but not in the reference impression.

2. Incoherent but potentially relevant text (orange). This contains medically relevant information that is also included in the reference, but is presented with incoherent grammar.

3. Hallucinated filler text (pink). This includes extra punctuation or common words such as "findings." These are filler text because they are undesirable but do not detract from the correctness of the generated impression.

4. Hallucinated medical text (red). This includes text that is either (1) not explicitly included

Table 4: Out-of-distribution (OOD) performance of CLIN-T5 prefix tuned on CT head. Compared to in-distribution (first row), performance suffers increasingly with OOD modalities (second row) and anatomies (third row). Additionally, when evaluating CT head, tuning on a larger dataset comprising all modalities/anatomies (bottom row) improves performance compared to tuning on CT head alone (top row).

Dataset		OOD					
Train	Test	Modality	Anatomy	BLEU	ROUGE-L	BERT	F1-Radgraph
CT head	CT head			11.4	35.0	**89.8**	35.1
CT head	MR head	✓		9.0	27.5	87.8	27.4
CT head	CT other		✓	2.9	19.5	86.7	16.3
CT head	MR other	✓	✓	7.9	24.2	87.2	25.9
All	CT head	N/A	N/A	**12.6**	**35.3**	89.7	**36.4**

in the findings or reference impression, or (2) relevant to the findings but communicated in a factually incorrect manner. This is the worst of the four deviations, as we want the model to avoid inferring information which didn't originate directly from the radiologist.

5.3.2 Reader study

Three radiologists evaluated impressions generated via CLIN-T5-LARGE + LoRA on the CT head dataset according to the procedure described in Section 4.2. Results in Figure 3 are encouraging, as our generated impressions scored well for all three questions. Additionally, the radiologists shared the following observations:

- Occasionally there is extra information included in the "reference" impression that is not available in the findings, i.e. which the model has no chance of summarizing.
- The model may include duplicate or re-summarized when referring to prior studies. For example, the reference will state "area of subarachnoid hemorrhage ... which is unchanged since the patient's prior scan," while the generated impression merely says "no significant change since the prior study." This difference is typically an institutional or personal preference.
- The model made an incorrect reference to the patient's prior medical history. The reference was "subtle hypodensity in the left frontal lobe. given lack of prior studies available for comparison," but the generated impression was "subtle hypodensity ... in the left frontal lobe, which is consistent with prior studies."

This study provided valuable insights which could not be obtained via quantitative metrics. Fundamentally we advocate for the use of reader studies when evaluating report summarization to facilitate more clinically relevant research.

5.4 Pitfalls

We now discuss weaknesses in our analysis which motivate future study. One potential concern is that CLIN-T5 and CLIN-T5-SCI were pretrained over the MIMIC-III dataset. This serves our purpose for evaluating models with high levels of domain adaptation, but it could result in data leakage if Lehman and Johnson (2023) used the test set for pretraining. We also evaluate each model on a second clinical dataset, MIMIC-CXR, and CLIN-T5 similarly performs the best (Table 6, appendix).

Ultimately though, while we chose CLIN-T5 in Table 3 because it achieves the highest scores, we note the comparable performance of models which did not pretrain with MIMIC-III, such as T5, FLAN-T5, and SCIFIVE (see Appendix Table 5).

Another weakness is that "domain" and "distribution" are not rigorously defined. Our intuitive characterizations could be improved by quantifying or visualizing the distance between different distributions using various embedding- or graph-based methods (Johnson et al., 2019b; Jain et al., 2021).

Lastly, we made assumptions determining the best configuration of model and prompting method. We first performed a comprehensive evaluation of all models and methods using the base architecture (223M parameters) on MIMIC-III (Table 5, appendix) and all models with LoRA on MIMIC-CXR (Table 6, appendix). From these results we chose only the best model (CLIN-T5) for scaling to the large architecture (738M parameters) due to compute constraints. Future work should include a comprehensive evaluation of all configurations on both architecture sizes and datasets.

6 Conclusion

Our research employs innovative lightweight strategies to adapt LLMs for the task of RRS. We investigate how domain adaptation—both via model pretraining on relevant data and via methods for discrete prompting and parameter-efficient fine-tuning—affects downstream RRS task performance. We achieve best performance using a model pretrained on clinical text (CLIN-T5) and subsequently fine-tuned with RRS samples using LoRA. These compelling results require tuning a mere 0.32% of model parameters. While further validation is required before clinical deployment, we believe our findings contribute to the literature and advance the potential for improved radiologist workflows and patient care.

7 Acknowledgements

Our research is a direct continuation of the radiology report summarization track conducted at ACL BioNLP 2023 (Delbrouck et al., 2023).

References

Asma Ben Abacha, Yassine M'rabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. 2021. Overview of the mediqa 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 74–85.

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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

H.W. Chung, L. Hou, S. Longpre, et al. 2022a. Scaling instruction-finetuned language models. *https://doi.org/10.48550/arXiv.2210.11416*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022b. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Songtai Dai, Quan Wang, Yajuan Lyu, and Yong Zhu. 2021. Bdkg at mediqa 2021: system report for the radiology report summarization task. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 103–111.

Pritam Deka, Anna Jurek-Loughrey, et al. 2022. Evidence extraction to validate medical claims in fake news detection. In *International Conference on Health Information Science*, pages 3–15. Springer.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022a. Improving the factual correctness of radiology report generation with semantic rewards. *https://aclanthology.org/2022.findings-emnlp.319*.

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

Jean-Benoit Delbrouck, Maya Varma, Pierre Chambon, and Curtis Langlotz. 2023. Overview of the radsum23 shared task on multi-modal and multi-anatomical radiology report summarization. In *Proceedings of the 22st Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Esteban F Gershanik, Ronilda Lacson, and Ramin Khorasani. 2011. Critical finding capture in the impression section of radiology reports. In *AMIA Annual Symposium Proceedings*, volume 2011, page 465. American Medical Informatics Association.

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 Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Jeremy Irvin et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *https://doi.org/10.48550/arXiv.1901.07031*.

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*.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2020a. Mimic-iv. *PhysioNet. Available online at: https://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)*.

Alistair Johnson, Tom Pollard, and Roger Mark. 2020b. MIMIC-III clinical database.

Alistair Johnson et al. 2019a. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *https://www.nature.com/articles/s41597-019-0322-0*.

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.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019b. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

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.

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*.

E. Lehman and A. Johnson. 2023. Clinical-t5: Large language models built using mimic clinical text. *https://doi.org/10.13026/rj8x-v335*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

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

Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2020. Exploring versatile generative language model via parameter-efficient transfer learning. *arXiv preprint arXiv:2004.03829*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning.

Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P. Langlotz, and Dan Jurafsky. 2021. Improving factual completeness and consistency of image-to-text radiology report generation. In *NAACL-HLT 2021*.

NCBI. 1996. Pubmed.

NCBI. 2000. Pubmed central (pmc).

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*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yuhao Zhang, Derek Merck, Emily Bao Tsai, Christopher D. Manning, and Curtis P. Langlotz. 2020. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *ACL2020*.

Mengjie Zhao and Hinrich Schütze. 2021. Discrete and soft prompting for multilingual models. *arXiv preprint arXiv:2109.03630*.

A Appendix

A.1 Hyperparameters for parameter-efficient fine-tuning

For the prefix tuning experiments, we tune each LLM with the following hyperparameters:

- Initial learning rate of $1e^{-2}$ that linearly decays to $1e^{-3}$ after a 100-step warm-up.
- Ten epochs maximum with an early stopping criterion if validation loss has not decayed for five consecutive epochs.
- Batch size of eight (large architecture) or 16 (base architecture) with four gradient accumulation steps, rendering an effective batch size of 32 or 64, respectively.

For the LoRA experiments, we tune each LLM with the following hyperparameters:

- Initial learning rate of $1e^{-3}$ that decays linearly to $1e^{-4}$ after a 100-step warm-up.
- Five epochs with no early stopping criterion.
- Batch size of six with four gradient accumulation steps, rendering an effective batch size of 24.

LoRA requires slightly more memory than prefix tuning, hence we adjusted the effective batch size to comfortably fit on our NVIDIA Quadro RTX 8000 GPU. Despite the larger memory footprint and greater training time per epoch (Table 1), LoRA requires fewer epochs to reach convergence than prefix tuning, resulting in 30% less tuning time overall. We note the importance of a learning rate warm-up over the first 100 gradient steps, which has been shown beneficial for low-data settings (Li and Liang, 2021). We experimented with various learning rate schedulers (step, exponential decay) but found linear decay to give slightly better performance in terms of validation loss. As discussed in Section 4.3, the best set of hyperparameters is constant across each of the five models.

Table 5: Quantitative evaluation across each model and adaptation method using the base architecture size. Parameter-efficient (updating <0.4% of parameters) fine-tuning methods LoRA and prefix tuning drastically outperform discrete prompting strategies. Among these fine-tuning methods, the best performing models are those which have been pretrained on clinical text (CLIN-T5-SCI, CLIN-T5).

Model	Method	BLEU	ROUGE-L	BERT	F1-Radgraph
T5	null	3.4	14.3	84.1	13.8
	prefix	4.7	19.0	86.1	19.0
	in-context (1)	3.4	15.8	85.4	14.4
	in-context (2)	3.3	15.8	85.4	11.8
	in-context (4)	4.4	16.2	85.5	12.1
	prefix tuning	12.9	29.1	88.4	30.7
	LoRA	13.7	33.9	89.5	35.2
FLAN-T5	null	0.5	11.3	83.0	9.7
	prefix	1.1	14.7	84.7	13.8
	in-context (1)	2.9	17.8	85.6	14.6
	in-context (2)	5.3	19.6	86.2	16.6
	in-context (4)	8.6	25.0	87.0	21.6
	prefix tuning	12.1	27.1	87.8	28.0
	LoRA	13.8	34.4	89.5	36.2
SCIFIVE	null	1.0	6.4	80.0	4.2
	prefix	0.3	4.2	78.0	0.7
	in-context (1)	1.8	11.3	82.0	9.7
	in-context (2)	2.8	12.4	82.9	12.9
	in-context (4)	3.4	12.7	83.6	14.8
	prefix tuning	10.3	28.9	88.4	30.2
	LoRA	13.5	34.6	89.6	36.1
CLIN-T5-SCI	null	1.5	7.0	78.7	6.1
	prefix	1.1	5.0	77.9	4.2
	in-context (1)	0.4	9.9	73.3	7.6
	in-context (2)	0.9	11.1	76.1	7.3
	in-context (4)	2.4	14.2	76.7	11.8
	prefix tuning	11.7	33.3	89.3	35.0
	LoRA	13.4	36.4	89.9	37.6
CLIN-T5	null	0.8	12.2	69.4	10.7
	prefix	1.0	9.5	78.6	7.1
	in-context (1)	0.3	8.7	66.1	7.7
	in-context (2)	0.6	9.6	66.6	8.7
	in-context (4)	2.2	11.5	70.9	13.0
	prefix tuning	11.9	33.8	89.4	35.4
	LoRA	**14.8**	**36.8**	**89.9**	**38.2**

Table 6: Quantitative evaluation on MIMIC-CXR with the best adaptation method (LoRA) across each model using the base architecture size. This supports our hypothesis that pretraining with clinical text is beneficial for datasets beyond MIMIC-III.

Model	BLEU	ROUGE-L	BERT	F1-Radgraph	F1-CheXbert
T5	16.9	40.5	89.6	37.6	66.7
FLAN-T5	16.6	41.0	89.3	38.2	68.7
SCIFIVE	17.1	42.4	89.5	39.8	69.0
CLIN-T5-SCI	16.9	42.7	89.5	39.3	69.3
CLIN-T5	**18.1**	**43.6**	**89.7**	**40.1**	**69.5**