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

Outline

Stanford AIMI

- Motivation
- Goals and Hypothesis
- Dataset
- Experiments
  - Models
  - Methods
- Results
  - Quantitative (overall, model size, out-of-distribution)
  - Qualitative (reader study, error analysis)
- Conclusions
- Next Steps

Motivation

We investigate the task of **radiology report summarization (RRS)**.

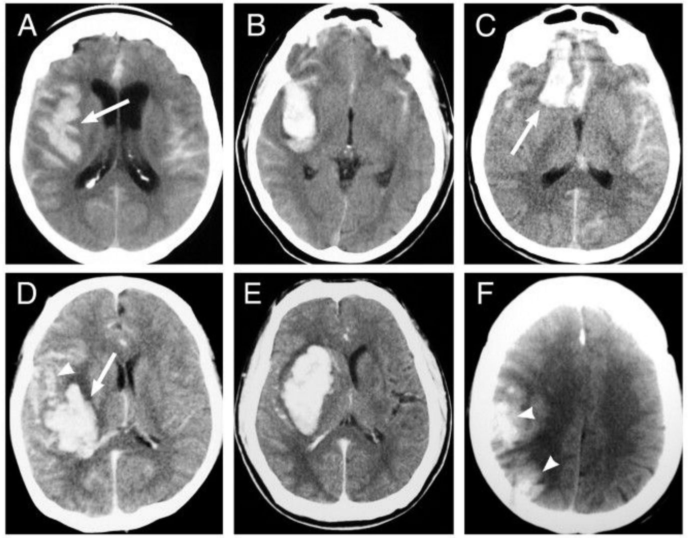
Why?

- Radiology reports communicate crucial information from medical imaging studies.
- RRS could be a useful clinical task in practice.
  - Radiologists write summaries manually – time-consuming, could lead to errors.
  - Downstream clinicians sometimes only look at the summary!
- Technically interesting!
  - Lots of information/jargon specific to the clinical domain.
  - Interpretability, coherence, and factual correctness are crucial.

How?

- Lightweight adaptation methods for large language models (LLMs).

Dataset: MIMIC-III



*not real paired image - just for example of head CT

FINDINGS:
There is no evidence of acute intracranial hemorrhage, mass effect or shift of normally midline structures. There is no cerebral edema or loss of grey/white matter differentiation to suggest an acute ischemic event. The sulci and ventricles are prominent, most likely age-related involutionary changes. Confluent hypodensities in the deep white matter and periventricular distribution most likely represent small vessel ischemic disease. Air-fluid levels are seen in bilateral sphenoid sinuses. Scattered ethmoid air cells are opacified. Mastoid air cells appear well aerated. no acute fracture is seen. Right anterior scalp laceration is noted.

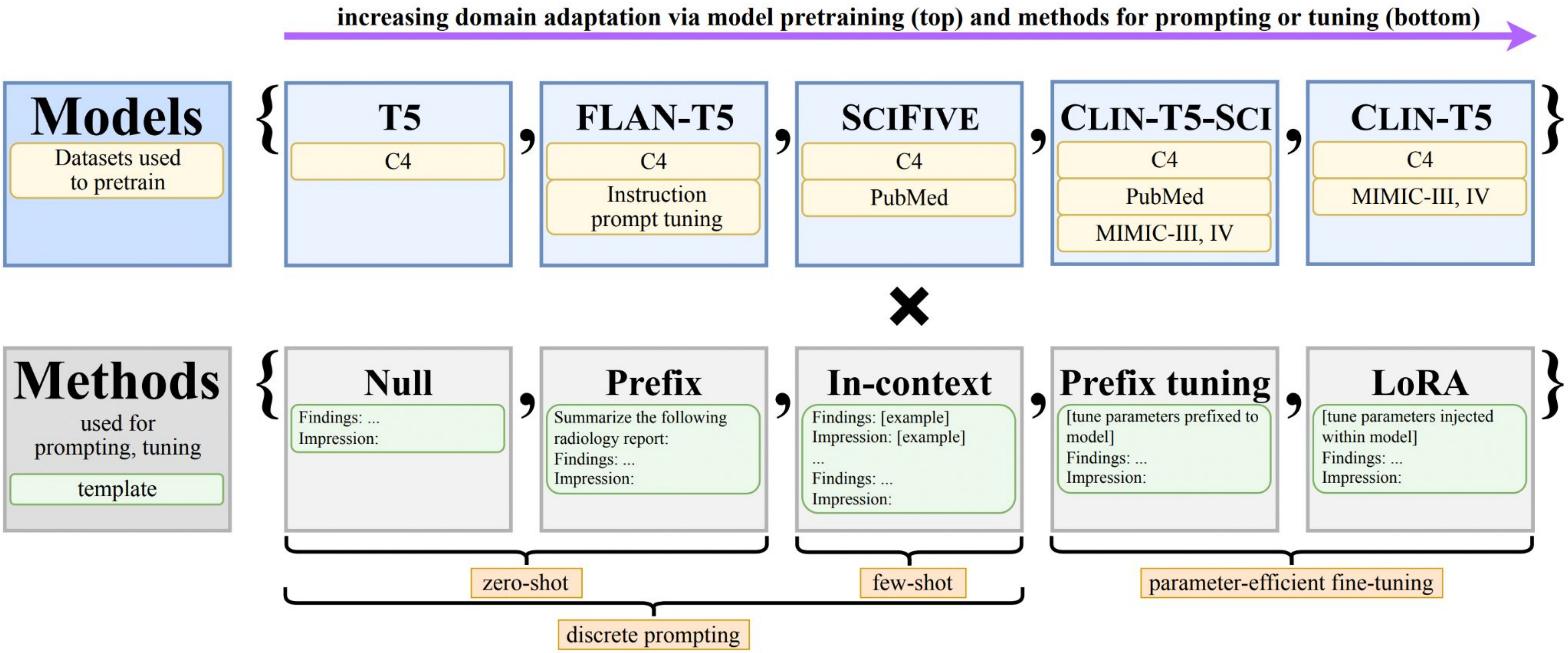
IMPRESSION:
1. No acute intracranial process.
2. Small vessel ischemic disease. Prominent sulci and ventricles, likely age-related involutionary changes.
3. Sinus disease, as above.

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

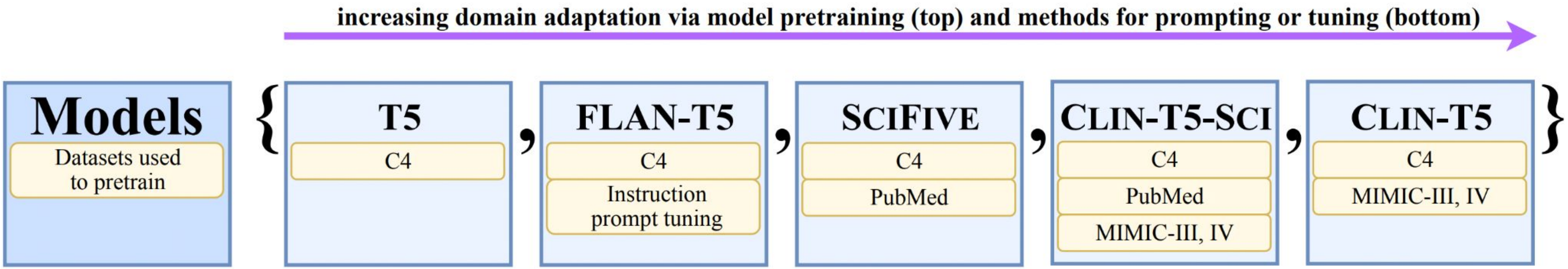
Modality/ Anatomy	**Number of reports** Train	Val	Test
CT head	25,122	3,140	3,141
CT abdomen	12,792	1,599	1,599
CT chest	10,229	1,278	1,280
MR head	5,851	731	732
CT spine	4,414	551	553
CT neck	912	114	115
MR spine	-	-	2,822
CT sinus	-	-	1,268
MR abdomen	-	-	1,062
MR pelvis	-	-	254
MR neck	-	-	231

Johnson et al, 2016.

Experiments



Experiments: Pretraining Datasets and Models



Experiments: Pretraining Datasets and Models

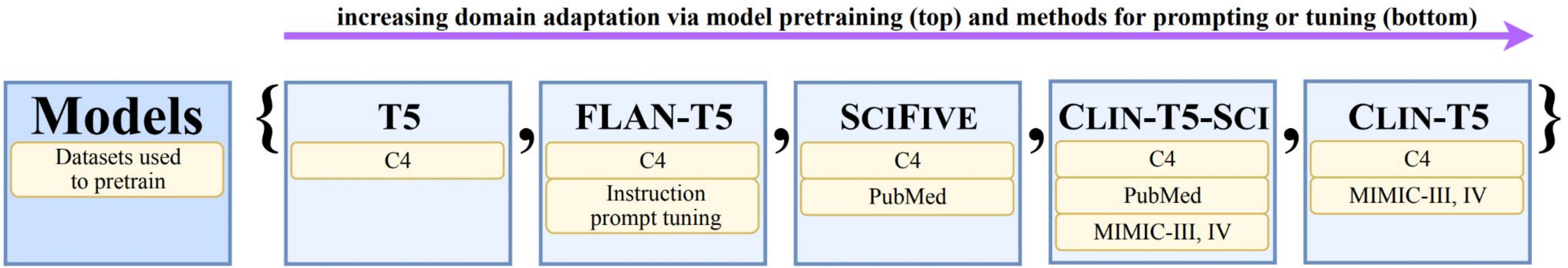
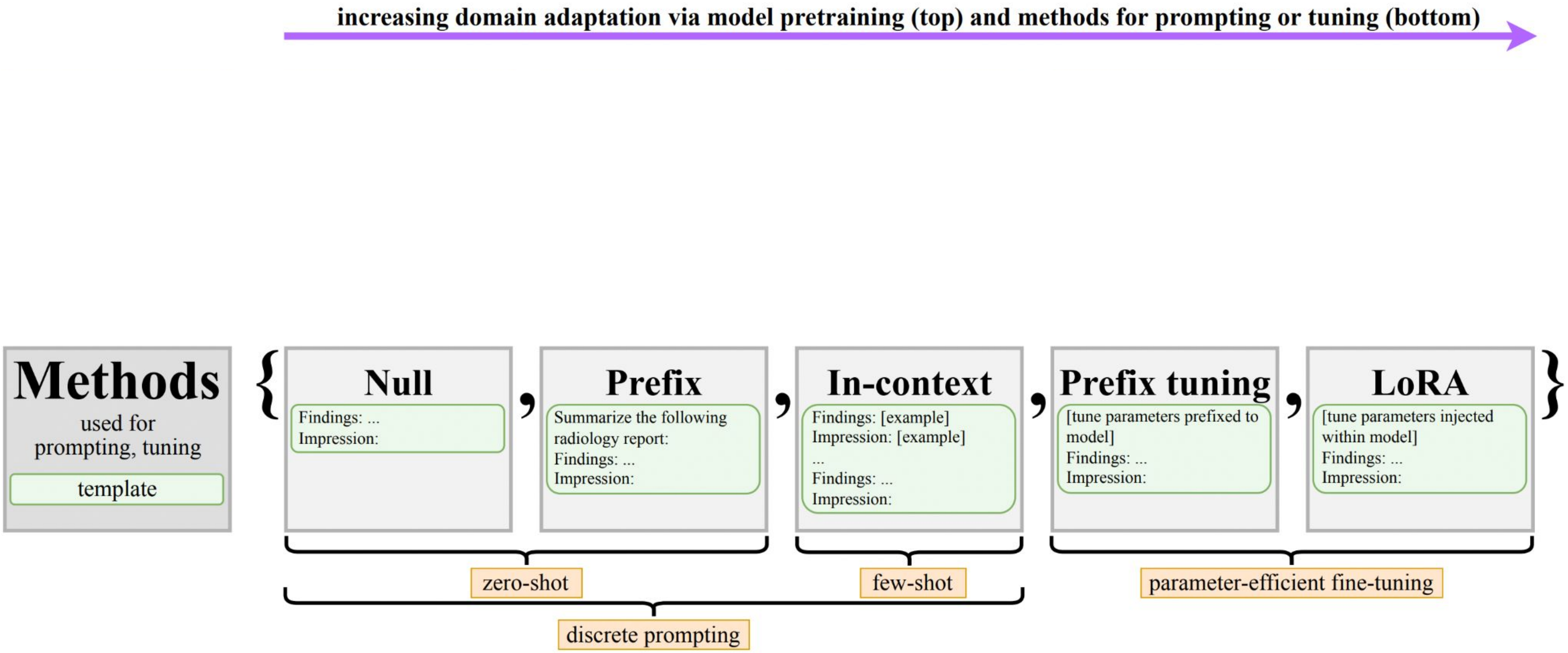


Table 1: We employ parameter-efficient fine-tuning methods for domain adaptation that modify <0.4% of model parameters while keeping other parameters frozen.

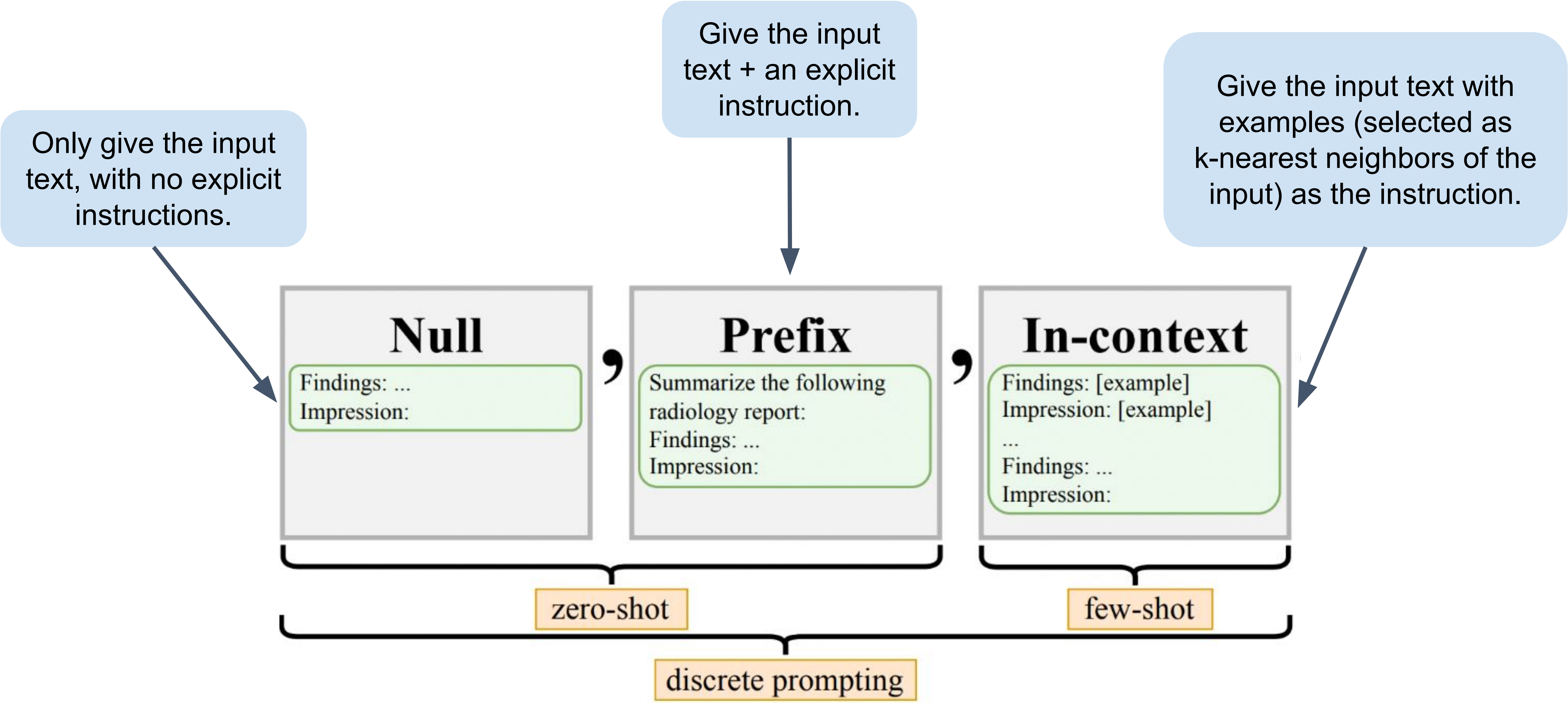
		Tunable parameters		Training time (hr)		
Model size	**Method**	**#**	**% of total**	**per epoch**	**total**	**# epochs**
Base (223M)	prefix tuning	0.37M	0.17%	0.98	9.83	10
	LoRA	0.88M	0.39%	1.32	6.60	5
Large (738M)	prefix tuning	0.98M	0.13%	2.93	29.3	10
	LoRA	2.4M	0.32%	3.85	19.3	5

Experiments: Domain Adaptation Methods



Experiments: Prompting Methods

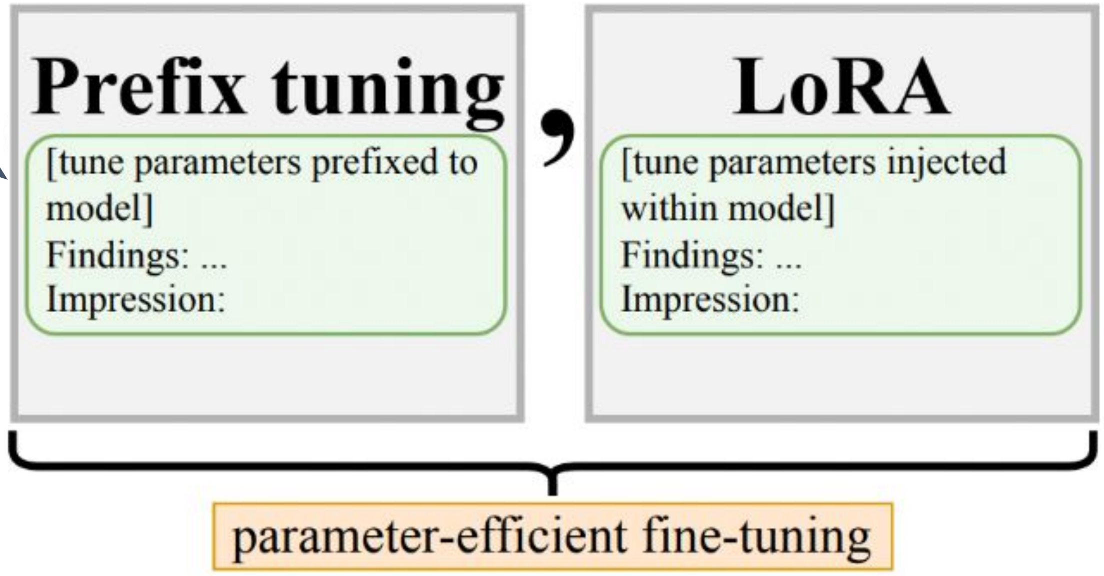
•



Experiments: PEFT Methods

Create tunable "prompt", hold pretrained model constant.

Tune injected model parameters, hold prompt constant.



Results

We achieve best performance by **maximally adapting** to the clinical RRS task via both task-agnostic pretraining (on clinical text) and lightweight task adaptation (LoRA for RRS).

Results

Stanford AIMI

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**

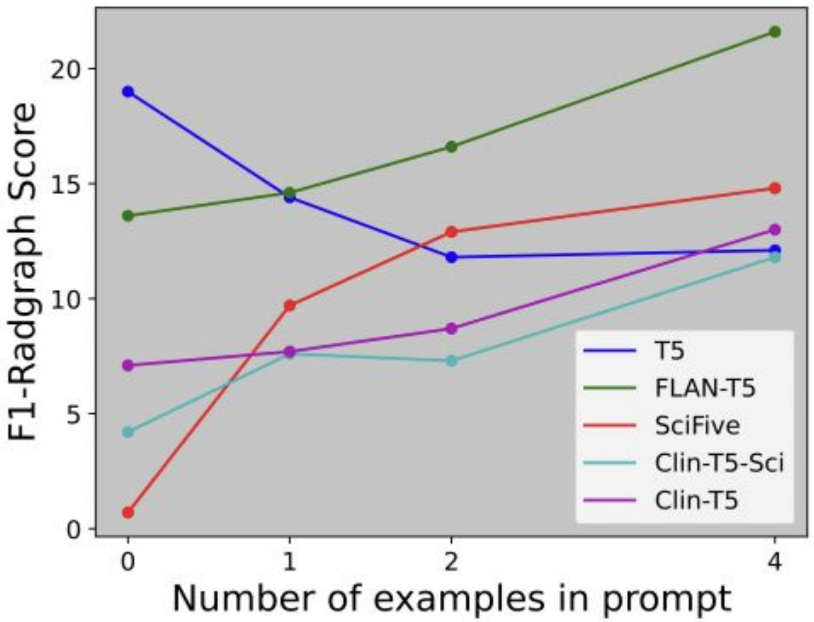


Figure 4: Domain adaptation. Left: Adaptation via pretraining on increasingly relevant data (T5, SciFive, Clin-T5-Sci, Clin-T5) generally leads to improved performance for both fine-tuning methods. Note we exclude FLAN-T5, whose degree of domain adaptation is difficult to rank. See Table 5 in the appendix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

Results: Model Size

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).

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**	-

Results: Out-of-Distribution Performance

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**

Results: Out-of-Distribution Performance (2)

Table 6: Quantitative evaluation on Stanford Hospital's dataset of ultrasound radiology reports 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 RRS datasets beyond the MIMIC suite.

Model	BLEU	ROUGE-L	BERT	F1-Radgraph
T5	12.6	31.2	88.2	26.2
FLAN-T5	12.0	30.6	88.3	26.8
SciFive	13.7	30.9	88.2	26.6
Clin-T5-Sci	14.0	32.7	88.6	28.5
Clin-T5	**15.1**	**32.8**	**88.8**	**29.7**

Results: Error Analysis

Stanford AIMI

Findings:
there is an evolving total left mca distribution infarction, with extensive edema and mass effect. on today's study, there is at least 6 mm of midline shift and associated subfalcine herniation. there is near total effacement of the left lateral ventricle and perhaps minimal dilatation of the contralateral right lateral ventricle. no evidence of hemorrhagic transformation at this time. the left proximal mca and intrasylvian branches remain hyperdense. no evidence to suggest evolving contralateral infarction. there are air fluid levels in the sphenoid and maxillary sinuses, presumably related to prior intubation.

Impression (reference):
evolving total left mca distribution infarction, with significant mass effect, midline shift, and subfalcine herniation. no evidence of hemorrhagic transformation.

KEY

Blue: factually correct
Green: factually correct, novel compared to reference
Orange: incoherent but potentially relevant
Pink: hallucinated filler text (okay)
Red: hallucinated medical text (bad)

Impressions generated via...

Null prompt:
sulcal effacement, and extensive edema. there is a large left mca distribution infarction.: left mca territory..:, there is an evolving left mca

Prefixed prompt:
. on , the following is a summary of the radiology report...:...: findings: there is an evolving left mca distribution infarction

In-context (1) prompt:
new infarction. the left MCA territory is noted. no new areas of hypoattenuation are identified. the right lateral ventricle is slightly more prominent. the left hemisphere.

Prefix tuning:
evolving left mca distribution infarction, with extensive edema and mass effect. no evidence of hemorrhagic transformation.

LoRA:
1. evolving left mca distribution infarction, with extensive edema and mass effect, with 6 mm of midline shift and subfalcine herniation. 2. no evidence of hemorrhagic transformation.

increasing domain adaptation

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).

Results: Reader Study

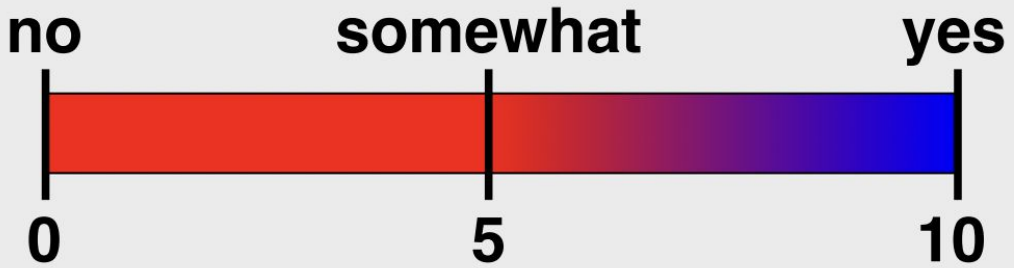
Stanford AIMI

<u>Questions</u>

Q1) Does the summary capture critical information?

Q2) Is it factually correct?

Q3) Is it coherent?



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

Figure 3: Radiology reader study. <u>Top</u>: Study design. <u>Bottom</u>: 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.

Results: Reader Study (2)

Stanford AIMI

"Reference" impression has information that isn't present in the "reference" findings.

The model has no chance of summarizing this information.

Generates repeated information when referring to prior studies.

This difference is typically an institutional or personal preference.

Generates an incorrect conclusion or reference, like nonexistent prior medical history.

This is a model "hallucination".

Conclusions

Employed recent lightweight strategies to adapt LLMs for RRS.

Investigated how domain/task adaptation affects RRS task performance.

Achieved best performance using a larger model maximally adapted to the clinical RRS task.

Evaluated best model quantitatively and qualitatively.

Thank you! Any questions?

github repo

Appendix

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**