Structuring Radiology Reports: Challenging LLMs with Lightweight Models

Johannes Moll, Louisa Fay, Asfandyar Azhar, Sophie Ostmeier,
Tim Lueth, Sergios Gatidis, Curtis Langlotz, Jean-Benoit Delbrouck

jomoll@stanford.edu github.com/jomoll/rad-report-structuring huggingface.co/collections/StanfordAIMI

What is structured reporting?

Generating radiology reports using strict, **template-driven rules** to ensure **clear**, **consistent**, and **standardized** clinical documentation.

EXAMPLE

Exam Type: Chest radiograph (PA and lateral views).
History: Evaluation for interval change following a non-ST elevation myocardial infarction (NSTEMI).
Technique: Standard posteroanterior and lateral chest radiograph.
Comparison: Small bilateral pleural effusions.
Findings:
Lungs & Airways: - No evidence of pneumothorax.
- Indistinctness of the hila with increased interstitial markings suggestive of pulmonary edema.
Cardiovascular: - Cardiomediastinal contours are stable.
Pleura: - Moderate bilateral pleural effusions..
Impression: 1. Worsening bilateral pleural effusions.
2. Suggestive of pulmonary edema.

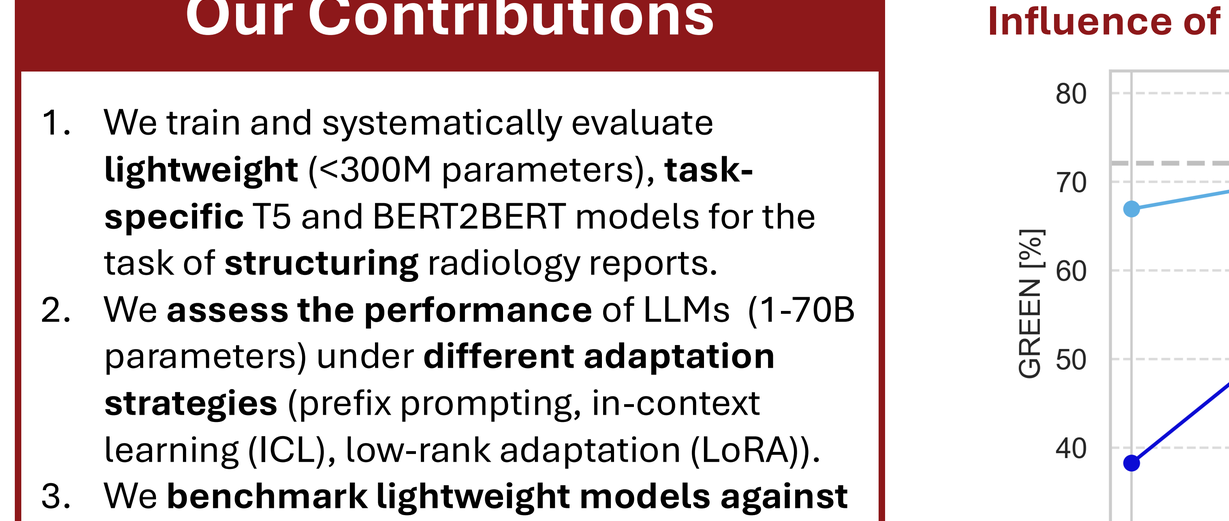
Motivation

- Structured reports are easier to read and understand.
- Structured reporting can benefit downstream machine learning applications.
- High computational requirements, lack of transparency, and data privacy concerns hinder practical deployment of existing approaches.

Our Contributions

1. We train and systematically evaluate **lightweight** (<300M parameters), **task-specific** T5 and BERT2BERT models for the task of **structuring** radiology reports.
2. We **assess the performance** of LLMs (1-70B parameters) under **different adaptation strategies** (prefix prompting, in-context learning (ICL), low-rank adaptation (LoRA)).
3. We **benchmark lightweight models against LLMs of increasing size**, considering model performance as well as training time, inference speed and costs, and environmental impact.

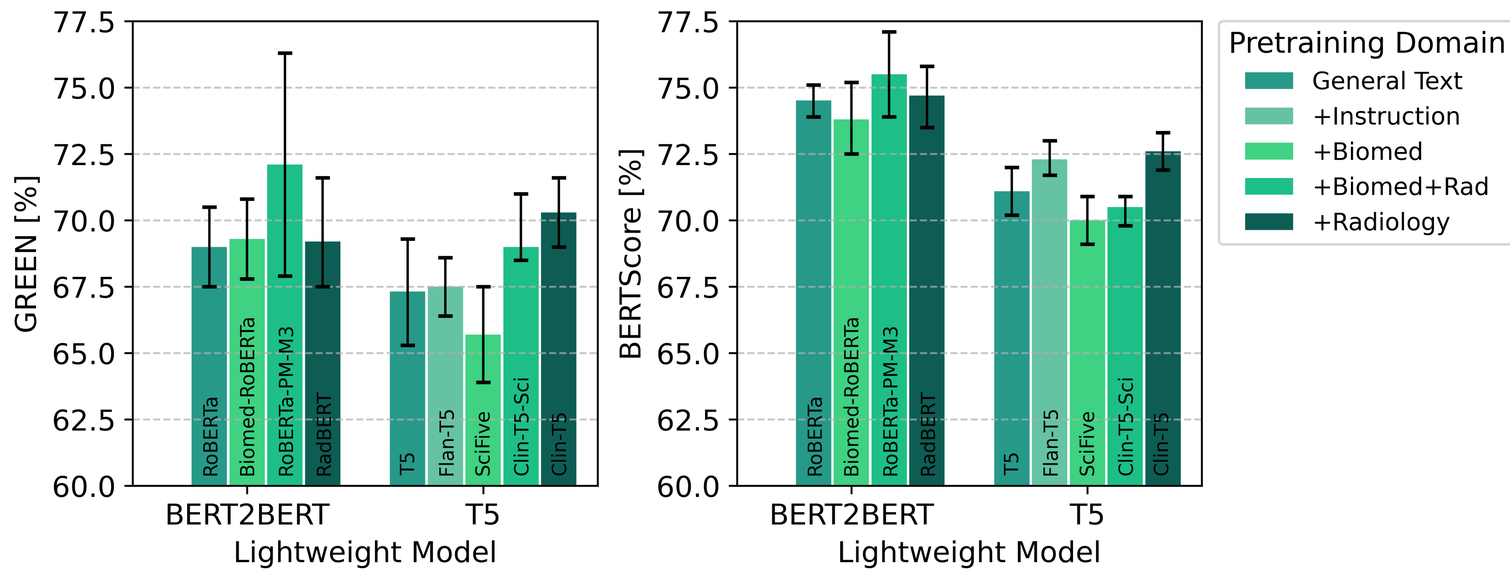
Methods



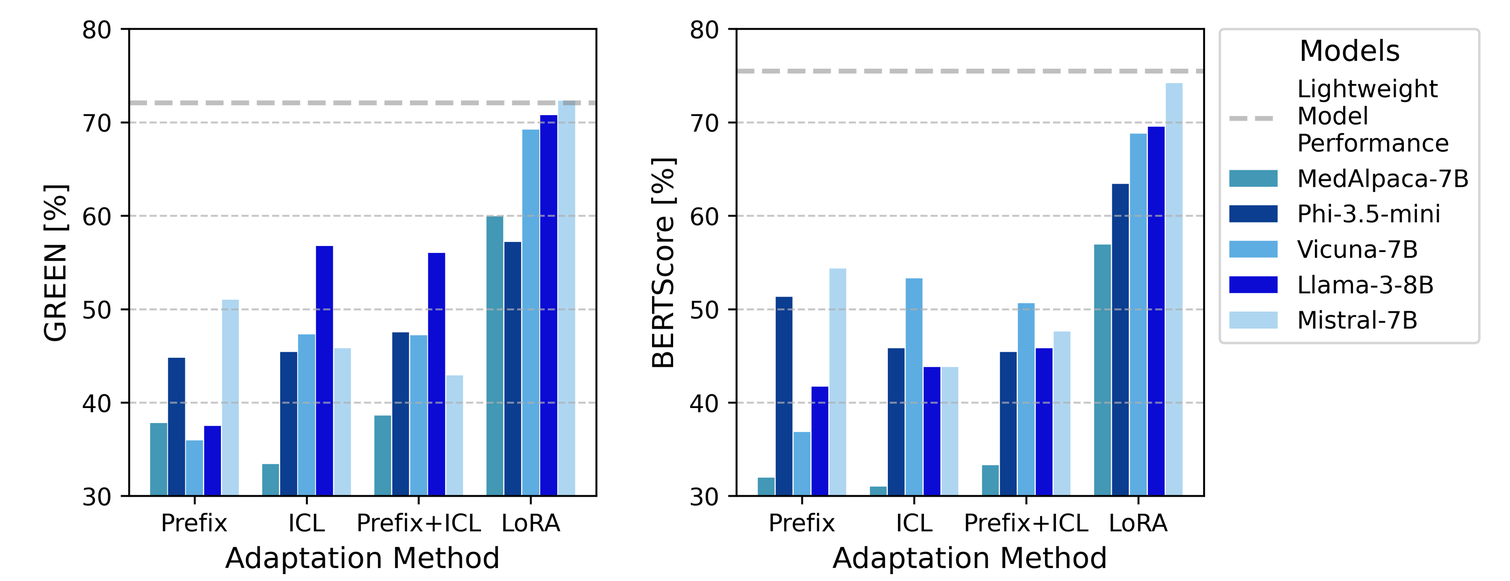
Finetune **lightweight, task-specific** encoder-decoder models and **benchmark** them **against state-of-the-art LLMs** with 1-70B parameters. Evaluate for structural consistency, linguistic similarity, and clinical accuracy.

Results

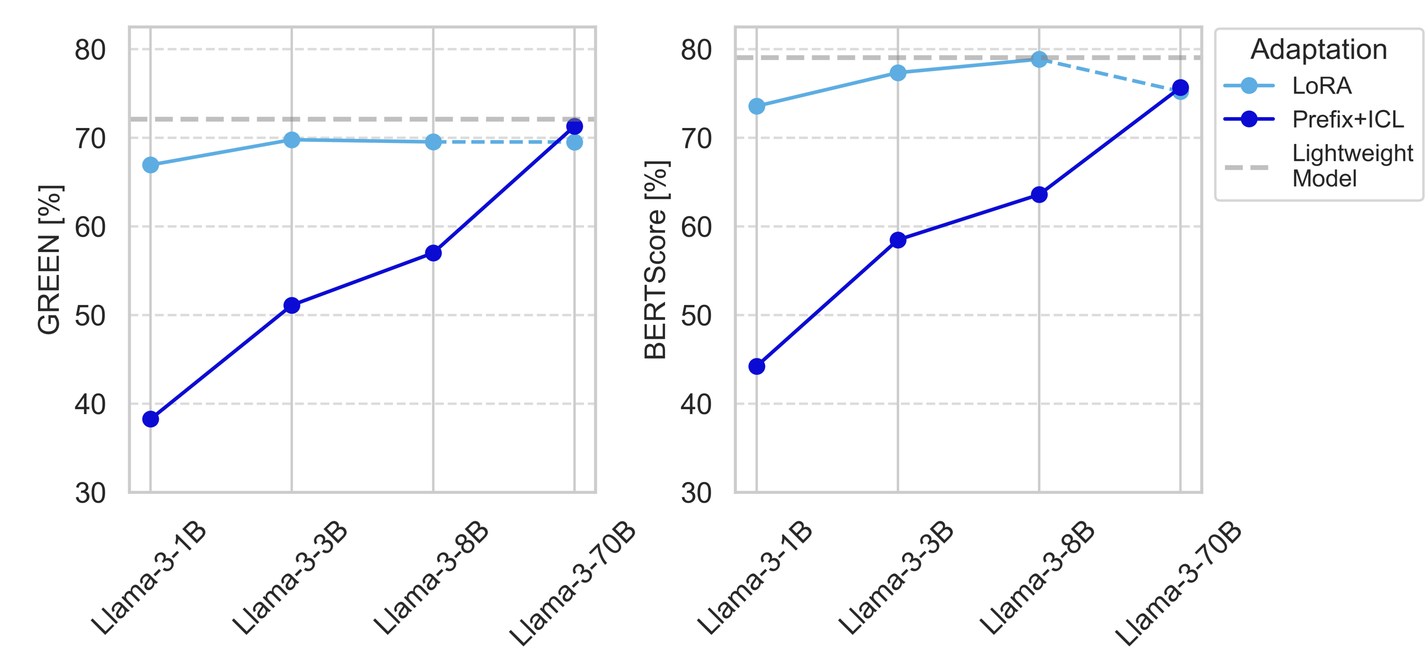
Pretraining Domains:



LLM Adaptation Methods:



Influence of Model Size



Takeaway

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.

Acknowledgments

Paper Code Models