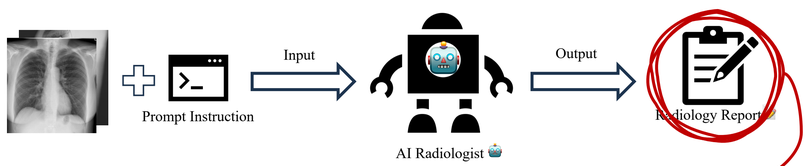
Exploring the Potential of LLMs as Radiology Report Evaluator

THE UNIVERSITY OF CHICAGO
DEPARTMENT OF COMPUTER SCIENCE

Yuyang Jiang
Department of Statistics, University of Chicago

Background: Radiology Report Evaluation

Upstream Task: Automated Radiology Report Generation



Core Task: How to Evaluate Generated Radiology Reports?🤔

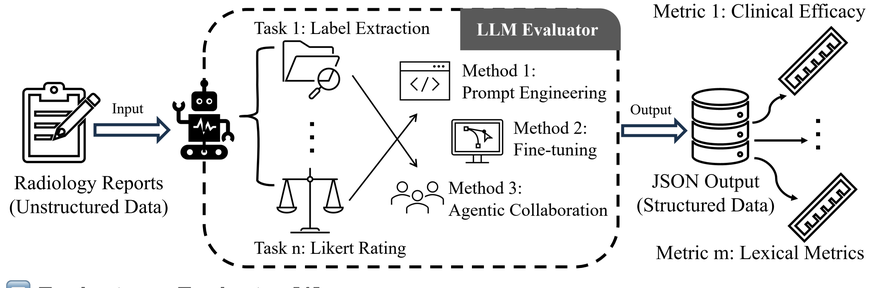
Review for Current Radiology Report Evaluation Metrics

Category	Descripton	Metrics	Concerns
Lexical Performance	Word-overlap Metrics	BLEU, METEOR, ROUGE	Fail to capture negation or synonyms in sentences, thus neglecting the semantic factuality.
	Embedding Similarity Metrics	BERTScore	Fail to emphasize key medical terms, leading to overlooking errors in critical conclusions.
Clinical Efficacy	Metrics based on Label Extraction	Positive/Negative F1 (CheXbert, CheXpert, NegBio)	Whether extracted labels/features are logically consistent and clinically efficient remains unchecked.
	Metrics based on Feature Extraction	RadGraph F1, RaTEScore	
LLM-based Metrics	Report-level Evaluation	GREEN	Rough design to solely identify six clincal significant errors.
	Sentence-level Evaluation to Match Findings	FineRadScore, RadFact	Designed based on closed-source LLMs (costly for large-scale evaluation).

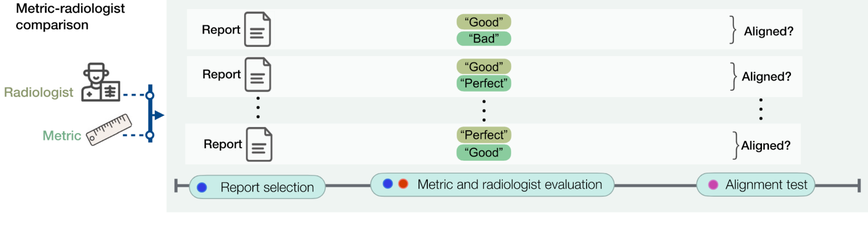
Can we develop a unified, accurate and scalable radiology report evaluator based on LLMs?

Methodology

Design an Evaluator Pipeline



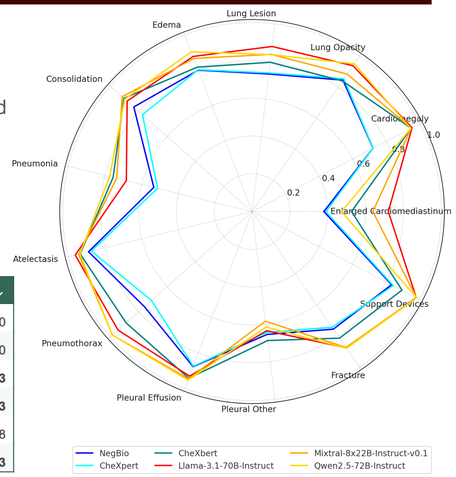
Evaluate an Evaluator [1]



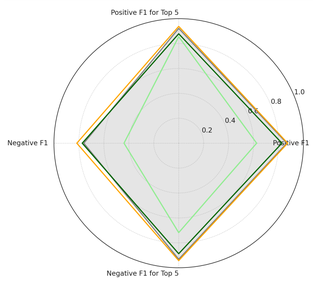
Task 1: How accurate can LLMs label radiology reports? (zero-shot)

Research Question: Can current LLMs outperform existing labelers on *pathology label extraction from chest radiograph reports*?
Downstream Application: Establish accuracy metrics based on pathology labels to *measure clinical efficacy* of generated radiology reports.
Benchmark: Human Annotation from MIMIC-CXR (Curated, 557 samples) [2]
Baselines: (1) Rule-based Labeler (CheXpert [3], NegBio [4]); (2) Bert-based Labeler (CheXbert [5]).

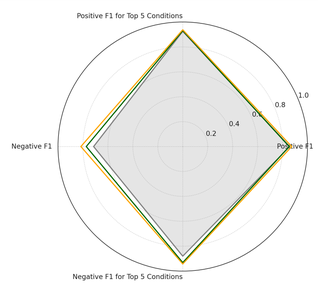
Labeler	# Pos F1	# Pos F1_5	# Neg F1	# Neg F1_5
NegBio	0.74	0.82	0.52	0.70
CheXpert	0.73	0.81	0.47	0.70
CheXbert	0.83	0.91	0.67	**0.93**
Llama-3.1-70B	**0.87**	0.92	**0.76**	**0.93**
Mixtral-8x22B	0.86	0.92	0.71	0.88
Qwen2.5-72B	0.86	**0.93**	0.74	**0.93**



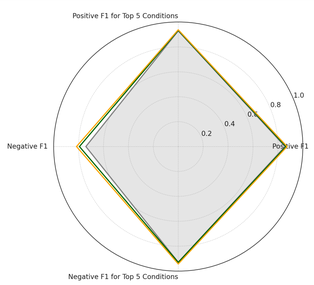
Task 1 (Cont'd): How far can we push LLMs' performance forward before turning to compute-intensive fine-tuning?



Llama-3.1-70B-Instruct



Mixtral-8x22B-Instruct-v0.1



Qwen2.5-72B-Instruct

Task 2: Can current LLMs rate Clarity/Readability on a likert scale (1-5) aligned with radiologist annotations? (zero-shot)

Downstream Application: *Measure lexical performance* of generated radiology reports.
Benchmark: Collected Annotations from Radiologists (150 samples)

Spearman Rank Corr	GPT-4o alignment w [Dr. Ben]
Overall (50*3)	0.29
MIMIC-CXR reports (50)	0.38
GPT-4-generated reports (50)	0.12
Llama-generated reports (50)	0.69

Paired sample t-test	p-value
3 (scored by GPT-4o)	0.0001
4 (scored by GPT-4o)	0.0028
5 (scored by GPT-4o)	0.3253 [verify point 1]

⚠️ H0: GPT-4o score - Human score = 0
(the lower p-value the more significantly to reject H0)

Paired sample t-test	p-value
4 (scored by Dr.Ben)	1 [verify point 2]
5 (scored by Dr.Ben)	0.0006 [verify point 2]

Conclusions & Discussions

Clinical Efficacy

Takeaways

- 🟢 LLMs explicitly outperform current SOTA labelers in labeling radiology reports.
- 🟢 Majority voting across different LLMs works well to improve model performance.
- 🟠 Few-shot prompting improves LLMs in identifying negative mentions but has limited effect on positive mentions.
- 🔴 Multi-step prompting fails terribly.

Open Questions

- Few-shot prompting: How does different samples selected to construct few-shot examples affect LLMs' performance?
- Multi-step prompting: Does independent condition labeling that might lose in-context learning cause the failure?

Lexical Performance

Takeaways

- 🟢 It's reliable to use GPT-4o to filter llama-generated reports in lexical quality.
- 🟠 Looking into human-written reports:
  1. High-score (5) reports filtered by GPT-4o are well aligned with radiologist preference;
  2. GPT-4o tends to annotate high-quality (5) reports confirmed by radiologists as lower scores (3-4).
- 🔴 GPT-4o cannot filter its own generated reports.

Open Questions

- Future To-Do: Does the result generalize across other LLMs or different settings (e.g. prompt engineering)?

References

[1] Yu, Feiyang, et al. "Evaluating progress in automatic chest x-ray radiology report generation." *Patterns* 4.9 (2023).
[2] Johnson, Alistair EW, et al. "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports." Scientific data 6.1 (2019): 317.
[3] Irvin, Jeremy, et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison." Proceedings of the AAAI conference on artificial intelligence. Vol. 33. No. 01. 2019.
[4] Peng, Yifan, et al. "NegBio: a high-performance tool for negation and uncertainty detection in radiology reports." AMIA Summits on Translational Science Proceedings 2018 (2018): 188.
[5] Smit, Akshay, et al. "CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT." arXiv preprint arXiv:2004.09167 (2020).

Acknowledgements

We thank Prof. Samuel Armato and his lab members from Department of Radiology at UChicago for their support for data curation. We also thank Dr. Benjamin M. Mervak from Michigan Medicine for his support for human annotations. We also thank Prof. Chenhao Tan and Chacha Chen from Chicago Human+AI Lab for their thoughtful feedback for experiment design.