Exploring the Potential of LLMs as Radiology Report Evaluator

Yuyang Jiang
yuyang2001@uchicago.edu
University of Chicago
Chicago, Illinois, USA

Abstract

The rise of large language models (LLMs) has opened new possibilities in medical AI, particularly in evaluating radiology reports. This study explores the potential of LLMs as radiology report evaluators. Current evaluation frameworks for radiology report generation are often shallow and limited, leading to inconclusive results about AI systems' capabilities. To address this, we propose a comprehensive pipeline for evaluating generated reports using LLMs, grounded in tasks such as label extraction and clarity scoring. Our experiments demonstrate that LLMs consistently outperform state-of-the-art (SOTA) labelers in labeling accuracy and show promise in clarity evaluation, particularly when leveraging agentic collaboration like majority voting. However, challenges remain, including mixed results with prompt engineering strategies, the limitations of multi-step prompting, and failure to fully align with radiologists' preferences for filtering more readable reports. This work represents a step forward in developing a robust, LLM-based evaluation framework, ultimately aimed at enhancing the quality and reliability of radiology report generation pipelines.

Keywords

Natural Language Understanding, Large Language Models, Radiology Report Evaluation

ACM Reference Format:
Yuyang Jiang. 2025. Exploring the Potential of LLMs as Radiology Report Evaluator. In . ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Radiology continues to play a pivotal role in the advancement of AI, supported by the growing approval of AI-driven radiology medical services from the FDA [7]. Among the various applications of AI in radiology, radiology report generation has emerged as one of the most dynamic and widely researched areas [12]. However, current evaluation frameworks for state-of-the-art AI systems remain shallow and incomplete, limiting their ability to test true capabilities or pinpoint specific bottlenecks. This lack of rigor often leads to inconclusive or even contradictory findings regarding the effectiveness of AI in radiology report generation [4]. To address these challenges, there is an urgent need for a coherent and efficient evaluation framework to accurately assess the quality of generated radiology reports.

Current radiology report evaluation metrics can be broadly categorized into three types, as illustrated in Figure 1:

(1) **Lexical Metrics:** These metrics focus on the surface form and exact word matches between generated and reference texts. They can be further divided into two subtypes based on their mechanisms: (i) Word Overlap Metrics, such as BLEU [9] and ROUGE [6] evaluate textual overlap. While intuitive, they fail to account for negation, synonyms, or semantic nuances, often overlooking the factual accuracy of the content; (ii) Embedding Similarity Metrics, like BERTScore [15], incorporate semantic awareness by comparing text embeddings. However, they may underemphasize critical medical terms, potentially missing errors in key conclusions, which are vital in a clinical context.
(2) **Clinical Efficacy Metrics:** These metrics assess the clinical correctness of radiology reports by analyzing extracted features and can be divided into two subtypes based on the feature types: (i) Metrics based on Label Extraction, such as Pos F1, cmeasure accuracy using outputs from state-of-the-art labelers; (ii) Metrics based on Entity Extraction, such as RadGraph F1 [3], MEDCON [13] and RaTEScore [16], evaluate the similarity of entity embeddings extracted from trained medical named entity recognition (NER) models. Although these metrics better align with human radiologists' evaluations, the logical validity and clinical impact of the heuristics determining which conditions or features are included remain underexplored.
(3) **Unified Metrics:** These metrics leverage large language models (LLMs) to perform comparisons at either the sentence level (e.g., GREEN [8]) or the report level (e.g., FineRadScore [1]). They aim to balance the assessment of clinical accuracy and lexical similarity simultaneously. While this approach shows promise, it still faces challenges in optimizing the trade-off between simplicity for adaptation to open-source models and the high deployment cost associated with large-scale evaluations using closed-source models.

Thus, I aim to develop a unified, accurate, and scalable evaluator leveraging LLMs, given their exceptional performance in Natural Language Understanding (NLG) tasks. This marks the initial step toward creating an effective evaluator grounded in a series of NLG tasks, with the long-term goal of ensuring quality control across the entire report generation pipeline, including the integration of image-based information.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

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).

Figure 1: Current Evaluation Metrics Review.

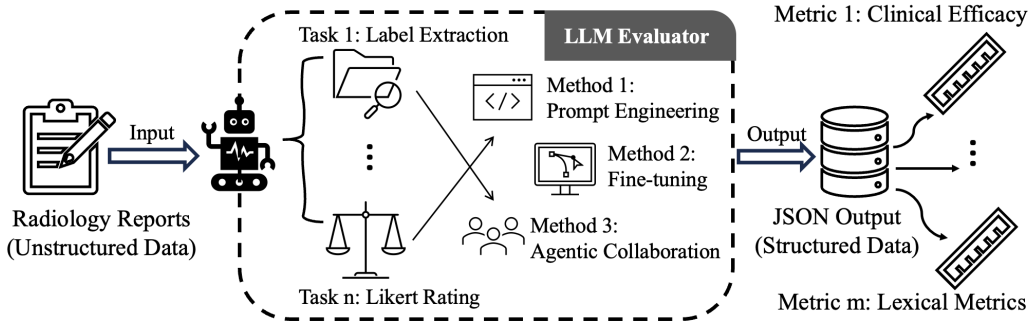

Figure 2: General Pipeline for an LLM-based Evaluator.

2 Methodology

2.1 General Pipeline for an Evaluator

For this general pipeline design, I aim to follow three key words **"unified, accurate, scalable"**:

(1) **Accurate:** This requires the evaluator to not only outperform on standard evaluation benchmarks in the lab setting but also remain robust when translating to clinical settings.
(2) **Unified:** The evaluator should go beyond a single evaluation metric and provide comprehensive indices, akin to a "physical examination" of generated radiology reports. These detailed indicators help not only in assessing a report's quality but also in diagnosing specific strengths and weaknesses.
(3) **Scalable:** The pipeline must balance input costs (e.g., time, computational resources, and financial investment) against output utility (e.g., model performance). This consideration impacts model selection and determines the methods for deploying and utilizing the evaluator efficiently.

Building on these principles, I propose a pipeline as illustrated in Figure 2. This pipeline is divided into two main stages:

(1) **Processing Unstructured Data:** An LLM-based evaluator processes unstructured data, such as radiology reports, and extracts structured features—such as entities and labels—formatted as JSON. This step enables downstream analysis and interpretation.
(2) **Evaluating with Custom Metrics:** A set of evaluation metrics is applied to the structured features to assess various aspects of radiology reports, such as clarity, accuracy, and completeness, while pinpointing specific areas of failure.

To further enhance scalability, I decompose the LLM evaluation task in stage (1) into smaller, focused sub-tasks. Each sub-task is matched to the most suitable methods—such as prompt engineering or an agentic framework—depending on budgetary constraints. This flexible approach ensures that the pipeline remains efficient and adaptable to different resource settings, making it practical for real-world clinical applications while maintaining high standards of performance and reliability.

2.2 Representative Sub-Tasks

In this report, I select two representative tasks as a pilot study to explore the potential of LLMs within this evaluator pipeline:

Task 1: Label Extraction Building on prior work [2], I aim to obtain an LLM-based automatic labeler to extract labels for 14 CheXpert medical conditions in this task. To evaluate different LLMs' performance as a labeler, I compute both positive F1 and negative

F1 scores against a human-annotated benchmark. Each condition is labeled as `present, absent, uncertain, unmentioned`, with positive F1 focusing exclusively on positive labels compared to all other categories while negative F1 considers negative labels as 1 and all other labels as 0. I report macro-averaged F1 scores across all 14 conditions and the top 5 conditions (the five most common: Pneumothorax, Pneumonia, Edema, Pleural Effusion, and Consolidation). This label extraction task is applied downstream to establish pathology-based accuracy metrics for measuring the clinical efficacy of generated radiology reports.

Task 2: Clarity Scoring In this task, I aim to leverage LLMs as judges to evaluate the linguistic quality of radiology reports, focusing on clarity, conciseness, and overall readability. Each report is rated on a Likert scale, where a score of 5 represents exceptional linguistic quality, and a score of 1 indicates poor performance. To assess the effectiveness of LLMs in measuring lexical performance, I calculate the alignment between the LLM-generated ratings and human annotations. This analysis provides insight into the LLMs' ability to evaluate and quantify the linguistic quality of generated radiology reports.

2.3 Evaluate an Evaluator

To evaluate an evaluator, the process involves selecting radiology reports for testing and comparing metric scores with radiologist evaluations [14]. The alignment between metric scores and radiologist evaluations is assessed using statistical measures such as the Kendall rank correlation coefficient. This provides a quantitative measure of how effectively the metrics reflect clinical relevance based on radiologist judgments.

3 Experiments

3.1 Task 1: Zero-shot Label Extraction

In this experiment, I aim to explore how accurately LLMs can label radiology reports in a zero-shot setting.

Datasets I utilize 557 studies from the human-annotated MIMIC-CXR dataset [5], the most widely used benchmark for chest X-ray analysis. One radiologist reviewed and curated this dataset to ensure its quality and relevance, creating a robust evaluation benchmark for this task.

Models Three state-of-the-art open-source models were tested in a zero-shot prompting setting: Llama-3.1-70B-Instruct, Mixtral-8x22B-Instruct-v0.1, and Qwen2.5-72B-Instruct. For comparison, I selected three widely used labelers from the literature as baselines: (1) Rule-based labelers: NegBio [10] and CheXpert [2]; (2) Bert-based labelers: CheXbert [11].

Results As shown in Figure 3(A), all three LLMs outperform the SOTA labelers on average, particularly excelling in identifying positive conditions. Among the LLMs, Llama-3.1-70B-Instruct demonstrates the consistently best performance across four macro-averaged metrics. Notably, Qwen2.5-72B-Instruct excels in accurately identifying the five most common conditions. Figure 3(B) further validates these findings, showing that the three LLMs outperform baseline models across nearly all medical conditions, with the exception of "Pleural Other." These results underscore the potential of LLMs to improve automated labeling in radiology, even in challenging zero-shot scenarios.

3.2 Task 1 (Cont'd): Prompt Engineering or Agentic Framework?

In this extended experiment, I investigate how far LLM performance can be pushed using advanced techniques such as prompt engineering and agentic collaboration before turning to compute-intensive fine-tuning.

Prompt Engineering I tested two distinct prompting strategies: (1) 5-shot prompting: Five cases were carefully selected to cover a diverse range of conditions. One radiologist revised the labels and annotated explanations for these cases. These refined examples, along with their explanations, were used as 5-shot examples to prompt the model; (2) Multi-step prompting: The single-task zero-shot setting was restructured into a 14-step independent labeling process, where each condition was labeled separately.

Agentic Framework For agent collaboration, I implemented majority voting across the three LLMs. The results from the best-performing model, Llama-3.1-70B-Instruct, were set as the default labels, serving as a baseline for comparison.

Results As shown in Figure 4, majority voting consistently improved performance across three LLMs compared to their zero-shot settings. This demonstrates the effectiveness of aggregating predictions from multiple models to enhance robustness and accuracy. The impact of 5-shot prompting, however, was mixed. It significantly boosted the performance of Mixtral-8x22B-Instruct-v0.1 and Qwen2.5-72B-Instruct, but had only a minimal effect on Llama-3.1-70B-Instruct, the best-performing model in the zero-shot setting. Surprisingly, multi-step prompting performed poorly in the Llama experiments, likely due to the loss of in-context learning when breaking the task into independent steps. These results suggest that while prompt engineering and agentic frameworks can improve LLM performance, their efficacy depends on the model and task design.

3.3 Task 2: Zero-shot Clarity Scoring

Table 1: Spearman Rank Correlation: GPT-4o Alignment with [Dr. Ben]

Dataset	Spearman Rank Correlation
Overall (50×3)	0.29
MIMIC-CXR reports (50)	0.38
GPT-4-generated reports (50)	0.12
Llama-generated reports (50)	0.69

In this experiment, I investigate how well GPT-4o rates the clarity of radiology reports compared to radiologist annotations.

Datasets I randomly selected 50 studies from the MIMIC-CXR dataset, where each study includes associated chest X-rays and one human-written report. For each study, I also prepared one generated report using GPT-4V based on given chest X-rays and another using Llama. This resulted in a total of 150 reports (50 human-written, 50 GPT-4V-generated, and 50 Llama-generated), which were subsequently evaluated by one radiologist for Clarity/Readability using a Likert scale.

(A)

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**

(B)

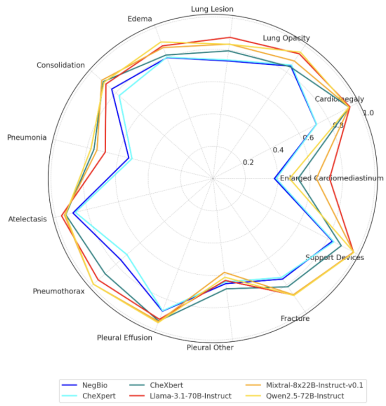

Figure 3: (A) Averaged LLM Performance on Identifying Positive and Negative Conidtions. (B) LLM Performance across 13 Medical Conditions.

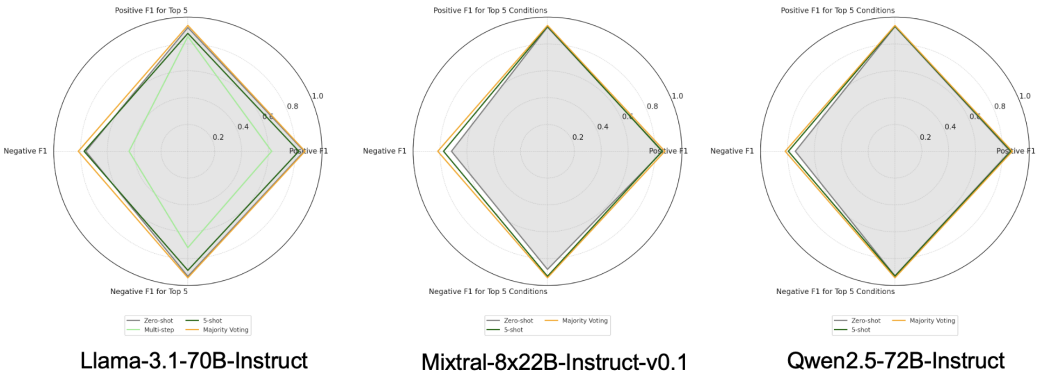

Figure 4: Three LLMs' performance under different prompting strategies and majority voting.

Models In this task, I tested only GPT-4o, as it represents the current strongest LLM and provides a robust benchmark for identifying the upper bound of current LLMs in clarity scoring.

Table 2: Paired Sample t-test Results: Grouped by GPT-4o Scores

Score (by GPT-4o)	p-value
3	0.0001
4	0.0028
5	0.3253

Table 3: Paired Sample t-test Results: Grouped by Dr. Ben's Scores

Score (by Dr. Ben)	p-value
4	1
5	0.0006

Results As shown in Table 1, GPT-4o aligns well with radiologist annotations when scoring Llama-generated reports but performs poorly when evaluating its own generated reports. The alignment results for human-written reports, however, are less clear. To investigate further, I conducted two paired-sample t-tests with the null hypothesis H_0: GPT-4o scores are equal to radiologist scores.

(1) **Grouping by GPT-4o Scores (Table 2**: When grouping studies based on GPT-4o's Likert scores, the results indicate that high-scoring (5) reports identified by GPT-4o align well with radiologists' preferences, confirming its ability to recognize high-quality Llama-generated reports.
(2) **Grouping by Radiologist Scores (Table 3**: When grouping studies by radiologists' Likert scores, the analysis reveals that GPT-4o tends to underrate high-quality (5) human-written reports, often assigning them lower scores (3–4). This discrepancy suggests potential limitations in GPT-4o's criteria for clarity and readability, particularly when evaluating reports it did not generate.

These findings highlight the strengths and weaknesses of GPT-4o in clarity scoring. While it performs well with Llama-generated reports, its misalignment on human-written and self-generated reports suggests that further refinement is needed to enhance its consistency and alignment with human judgment.

4 Conclusions and Discussion

4.1 Clinical Efficacy

Overall, the experiment results highlight that LLMs outperform state-of-the-art (SOTA) labelers in labeling radiology reports, with majority voting across different LLMs further improving performance. This underscores their potential as reliable tools for clinical labeling tasks. Few-shot prompting enhances the identification of negative mentions but has limited impact on positive mentions, suggesting room for optimization. In contrast, multi-step prompting performs poorly, likely due to the loss of in-context learning when conditions are labeled independently.

These results raise important questions, such as how sample selection affects few-shot prompting performance and whether integrating context can address the shortcomings of multi-step prompting. Continued exploration of these areas is vital for fully leveraging LLMs in clinical applications.

4.2 Lexical Performance

GPT-4o, as the representative LLM, proves to be a reliable tool for filtering LLaMA-generated radiology reports with regrad to lexical quality, effectively identifying reports that align with human-written standards. Among human-written reports, high-scoring reports (rated 5 by GPT-4o) tend to closely align with radiologists' preferences, indicating that GPT-4o can accurately identify reports that exhibit clarity, conciseness, and readability. However, GPT-4o sometimes underrates high-quality reports confirmed by radiologists, assigning them lower scores (3–4). This discrepancy suggests that while GPT-4o performs well overall, its evaluation criteria may not fully capture all nuances of human judgment.

Interestingly, GPT-4o struggles to filter its own generated reports effectively, raising questions about its ability to self-assess lexical quality. This limitation highlights the need for further exploration into how different LLMs perform in self-evaluation tasks and whether prompt engineering or other settings can improve their reliability.

5 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.

References

[1] Alyssa Huang, Oishi Banerjee, Kay Wu, Eduardo Pontes Reis, and Pranav Rajpurkar. 2024. FineRadScore: A Radiology Report Line-by-Line Evaluation Technique Generating Corrections with Severity Scores. *arXiv preprint arXiv:2405.20613* (2024).

[2] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 590–597.

[3] 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* (2021).

[4] Yuyang Jiang, Chacha Chen, Dang Nguyen, Benjamin M Mervak, and Chenhao Tan. 2024. Gpt-4v cannot generate radiology reports yet. *arXiv preprint arXiv:2407.12176* (2024).

[5] A Johnson, T Pollard, R Mark, S Berkowitz, and Steven Horng. 2019. MIMIC-CXR database (version 2.0. 0). *physionet* 2 (2019), 5.

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

[7] Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, et al. 2023. Artificial intelligence index report 2023. *arXiv preprint arXiv:2310.03715* (2023).

[8] 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* (2024).

[9] 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*. 311–318.

[10] Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. 2018. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings* 2018 (2018), 188.

[11] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. 2020. CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. *arXiv preprint arXiv:2004.09167* (2020).

[12] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421* 9, 1 (2023), 1.

[13] Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data* 10, 1 (2023), 586.

[14] 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 (2023).

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

[16] Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Ratescore: A metric for radiology report generation. *arXiv preprint arXiv:2406.16845* (2024).

A Training Details

Considering potential future publication, I cannot make my codes and prompts fully public in this Appendix, but I will include my computing resources and data sources for the conducted experiments.

GPU Usage: My hardware for inference includes four A100 GPUs, each equipped with 80GiB of memory, and operates on CUDA version 12.4.

Data licenses: MIMIC-CXR license can be found at https://physionet.org/content/mimic-cxr/view-license/2.0.0/. In particular, we accessed the data by following the required steps on https://physionet.org/content/mimic-cxr/2.0.0/. We first registered and applied to be a credentialed user, and then completed the required training of CITI Data or Specimens Only Research. We also signed the data use agreement for the project before we get access to the dataset.