GREEN: Generative Radiology Report Evaluation and Error Notation

Anonymous ACL submission

Abstract

Evaluating radiology reports is a challenging problem as factual correctness is extremely important due to its medical nature. Existing automatic evaluation metrics either suffer from failing to consider factual correctness (e.g., BLEU and ROUGE) or are limited in their interpretability (e.g., F1CheXpert and F1RadGraph). In this paper, we introduce GREEN (Generative Radiology Report Evaluation and Error Notation), a radiology report generation metric that leverages the natural language understanding of language models to identify and explain clinically significant errors in candidate reports, both quantitatively and qualitatively. Compared to current metrics, GREEN offers: 1) a score aligned with expert preferences, 2) human interpretable explanations of clinically significant errors, enabling feedback loops with end-users, and 3) a lightweight open-source method that reaches the performance of commercial counterparts. We validate our GREEN metric by comparing it to GPT-4, as well as to error counts of 6 experts and preferences of 2 experts. Our method demonstrates not only higher correlation with expert error counts, but simultaneously higher alignment with expert preferences when compared to previous approaches.

1 Introduction

Machine learning has enabled great progress in the automatic interpretation of images, where vision language models (VLMs) translate features of images into text (Radford et al., 2021; Liu et al., 2024). In the medical domain, patient images are interpreted by radiologists, which is referred to as radiology report generation (RRG). Automated and high-quality RRG has the potential to greatly reduce the repetitive work of radiologists, alleviate burdens arising from shortage of radiologists, generally improve clinical communication (Kahn Jr et al., 2009), and increase the accuracy of radiology reports (Rajpurkar and Lungren, 2023).

Commonly used evaluation metrics in RRG literature (Lin, 2004; Zhang et al., 2019; Smit et al., 2020; Delbrouck et al., 2022) seek to evaluate a generated radiology report against a reference report written by a radiologist by leveraging simple n-grams overlap, general language similarity, pathology identification within specific imaging modalities and disease classes, and commercially-available large language models. To achieve performance on par with radiologists, evaluation metrics must be adept with the radiology language in order to accurately assess factual correctness and levels of uncertainties. Additionally, RRG metrics should be interpretable in a scalable fashion to enable a feedback loop between the generated reports and the experts who review them. Moreover, these metrics should be open-source to allow for assessment of private datasets that require the safeguarding of patient information.

Current RRG metrics fall short of capturing the nuanced and multifaceted nature of radiology reports. To mitigate the current gaps in appropriate metrics for RRG, we introduce **GREEN (Generative Radiology Evaluation and Error Notation)**. The GREEN metric introduces five major contributions:

- **Score**: We introduce and validate a score, which ranges from 0 for the weakest assessment, to 1, marking the highest score achievable. We show that GREEN is adept with the radiology language and can accurately assess factual correctness and levels of uncertainties that surpass prior approaches.
- **Interpretable Evaluation Summary**: We provide a method to generate a clear, human-readable evaluation summary independent of the test set size. By providing detailed error categorization with explanations, GREEN enables machine learning practitioners and experts to pinpoint areas for improvement in their trained systems.
- **Practicability**: We open-source the GREEN model that leverages a < 7B parameter language model with similar report evaluation abilities as larger counterparts. This approach decreases GPU requirements and enhances processing speed.
- **Applicability**: Leveraging high-performing commercially-available large language model (LLM) services typically requires a de-identification procedure and institutional review

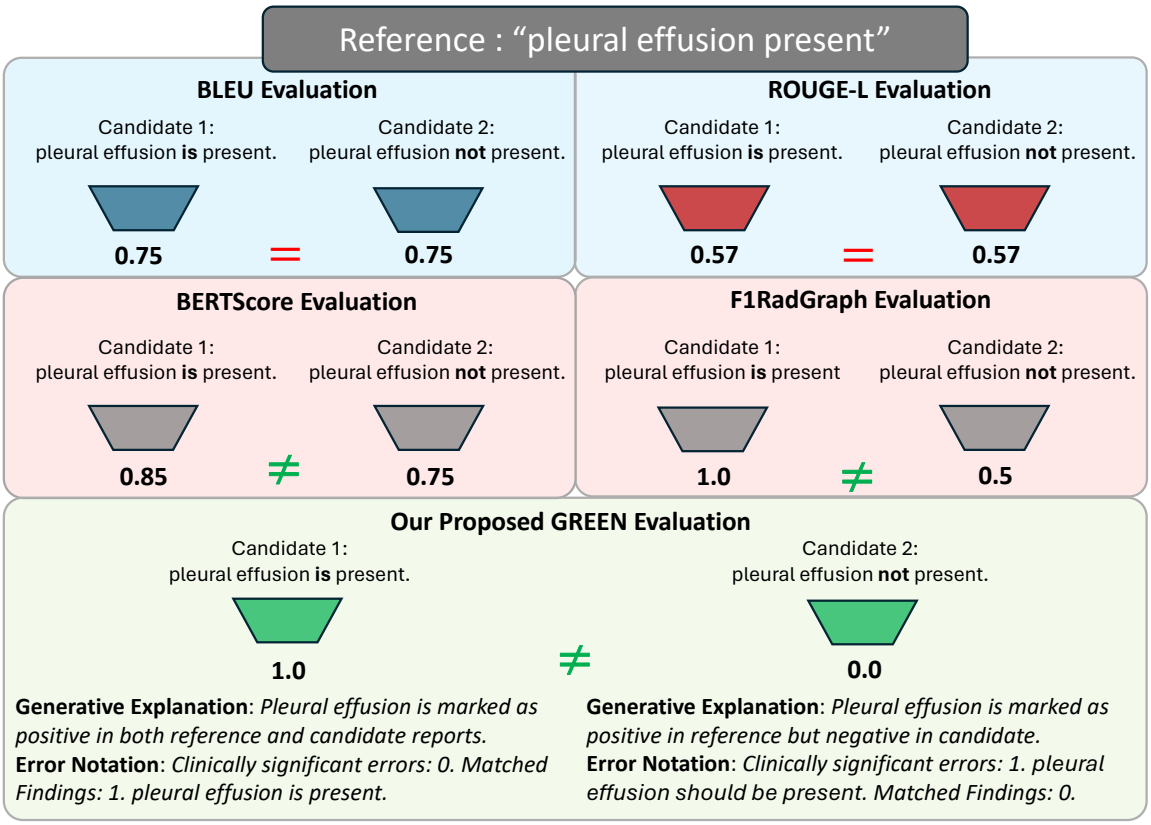


Figure 1: Motivation of GREEN.

board approval for protected health information (PHI). GREEN is free, open-source, and designed for use in confidential datasets without worries of leaking medical information about individuals.

- **Multimodality**: GREEN is designed to understand a wide array of pathologies, linguistic styles, and terminologies. We demonstrate that GREEN exhibits a generalized understanding of medical language that spans various imaging modalities and anatomical structures on out-of-distribution (OOD) data, specifically by examining its application to abdominal computed tomography (CT) reports in a zero-shot fashion.

- **Datasets**: Lastly, we share the dataset used to develop our models. This dataset encompasses 100,000 annotations from GPT-4 related to chest X-rays (spanning various datasets) and 50,000 annotations across a diverse set of imaging modalities. By making these resources available, we hope to facilitate further research and improvement in the accuracy and reliability of automated radiology report generation systems.

2 Related Work

The literature demonstrates various advances in generating radiology reports from medical images (Ramesh et al., 2022; Jeong et al., 2024; Li et al., 2023; Yang et al., 2022; Nguyen et al., 2021; Chen et al., 2024; Chaves et al., 2024). For instance, a set of evaluation metrics are commonly utilized to assess the quality of the generated reports and focus on lexical similarity (e.g., ROUGE-L (Lin, 2004) and BLEU (Papineni et al., 2002)) and factual correctness (e.g., F1CheXbert (Smit et al., 2020) and F1RadGraph (Delbrouck et al., 2022)). F1CheXbert assesses the accuracy of identified disease labels in reports against a narrow reference, covering only 14 CheXbert classes of common-but-specific chest x-ray findings. F1RadGraph enhances factual correctness evaluations by comparing the agreement on anatomical and observational entities between candidate reports and reference reports, using a graph model trained on human annotations. However, the correlation of F1RadGraph with manual evaluations by radiologists is low, leading to the development of more closely-aligned metrics such as RadCliQ (Yu et al., 2023a). RadCliQ consists of an ensemble of ROUGE, BLEU, CheXbert embedding similarities, and RadGraph to form a composite metric which aims to match expert-generated error counts. While RadCliQ is effective in mirroring these error counts, it has low interpretability as the individual metric weights are unknown and the single numerical score is inadequate for clinical integration (Figure 1).

Our approach stands out from previous metrics by emphasizing clinical relevance and interpretability, showing higher alignment with expert error counts and

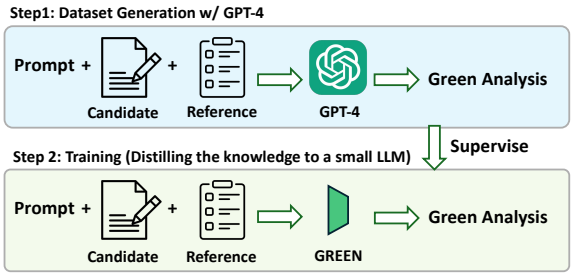


Figure 2: Training procedure of the GREEN model.

expert preferences while still leveraging an open-source <7B parameter model.

3 GREEN

GREEN (Generative Radiology Evaluation and Error Notation) involves **three primary components**.

First, we describe the construction of our **generative language model**, which is designed to identify and classify errors in radiology reports into six categories (Section 3.1). This section is subdivided into the collection of training data (Section 3.1.1) and details of the training process (Section 3.1.2). Second, we elaborate on the **GREEN score** in Section 3.2, including its rationale and significance. Third, we explain the **GREEN summary** and its usefulness.

We then outline the steps we took to validate the effectiveness and relevance of GREEN, both quantitatively and qualitatively, in Section 3.4.

3.1 Generative Large Language Model

3.1.1 Chest X-rays Data Collection

To compile our training dataset, we selected 100,000 reference and generated candidate report pairs from six publicly-available de-identified chest X-ray datasets: MIMIC-CXR (Johnson et al., 2019), MIMIC-PRO (Ramesh et al., 2022), CandidPTX (Feng et al., 2021), PadChest (Bustos et al., 2020), BIMCV-covid19 (Vayá et al., 2020) and OpenI (Demner-Fushman et al., 2012). We employed the prompt shown in Appendix A.1 to task GPT-4 to identify and categorize differences in natural language across six unique clinically-defined categories detailed previously Yu et al. (2023b).

The pairing process used five different heuristics, generating 20,000 unique pairs for each heuristic: i) randomly matching candidates and references, ii) modifying the candidate by removing and shuffling sentences from the original report, iii) using a trained RRG model to create the candidate based on the referenced image (to bridge the gap noted in literature between actual reports and those generated by baseline models), iv) pairing candidates with the closest semantically similar report assessed using BERTScore (Zhang et al., 2019), and v) creating candidates through RadGraph (named-entity recognition dataset) (Jain et al., 2021) permutations of the reference reports, incorporating changes to the presence of findings or by making modifications to size, severity, or location throughout the reports. The number of RadGraph permutations and the BERTScore distribution of the pairs are presented in Figure 3. Additionally, a sample GPT-4 response of a pair with a candidate that includes exactly one RadGraph permutation can be found in Appendix A.2.

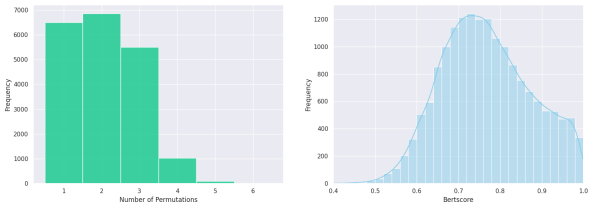


Figure 3: Number of RadGraph permutations among the candidates for 20,000 pairs (left) and BERTScore distribution across 20,000 pairs (right).

To maintain uniqueness across the dataset, once pairs were formed using one heuristic, they were excluded from consideration in others, ensuring each of the 100,000 pairs is distinct. Overall, 174,329 unique reports were utilized either as references or candidates in this study.

3.1.2 Model Architecture and Training

To enhance performance with medical data, we pre-trained LLaMA-2 and Phi-2 using a comprehensive set of domain-specific datasets to form RadLLaMA-2 and RadPhi-2. These datasets include MIMIC-IV Radiology Reports (Johnson et al., 2023), MIMIC-IV Discharge Summaries, MIMIC-CXR Radiology Reports, and a variety of sources from PubMed (Abstracts and Patient Reports). Specialized datasets such as Wiki Medical Terms[1] and Medical Guidelines[2] (Vashishth et al., 2021) were also used.

To obtain a local RRG evaluator for the GREEN metric model, we opted to train open-source models instead of relying on API-based models. Specifically, we further fine-tuned RadLLaMA-2 and RadPhi-2, as well as other models of different sizes, architectures, and pre-training datasets, such as LLaMA-2 (Touvron et al., 2023), Phi-2 (Javaheripi and Bubeck, 2023), and Mistral-v0.1 (Jiang et al., 2023) (Figure 2). Models were trained on 8x NVIDIA A100 Tensor Core GPUs with 40GB VRAM using the Huggingface framework with Flash Attention 2, DeepSpeed Stage 3, and the AdamW optimizer. An effective batch size of 2,048 was used for 12 epochs, as well as a base learning rate of 1e-4, a warm-up ratio of 0.05, and a weight decay of 0.1. Training for 7B-parameter models averaged 40 GPU hours, while the 2.7B-parameters models averaged 28 GPU hours. For fast and reliable inference, we employed data

[1] www.huggingface.co/datasets/gamino/wiki_medical_terms

[2] www.huggingface.co/datasets/epfl-llm/guidelines

GREEN Summary

```
[Summary]:
Green score: mean 0.23 std 0.04

[Clinically Significant Errors]:
(a) False report of a finding in the candidate: 0.9
[Small right pleural effusion, Small right pleural effusion, Small right pleural effusion]
(b) Missing a finding present in the reference: 0.7
[Residual opacification in the right upper lobe raising concern for pneumonia.,
Underlying chronic upper lobe scarring., Underlying chronic upper lobe scarring.]
(c) Misidentification of a finding's anatomic location/position: 0.4
[The opacity is in the right lower lobe, not the right upper lobe., Right lower lobe
consolidation instead of residual opacification in the right upper lobe., Opacity reported in
the left lower lobe instead of the right upper lobe]
(d) Misassessment of the severity of a finding: 0.8
[Bilateral pleural effusion, Bilateral pleural effusion, Bilateral pleural effusion.]
(e) Mentioning a comparison that isn't in the reference: 0.7
[The candidate report mentions a discussion between doctors, which is not present in the
reference report, The candidate report mentions monitoring and support devices which is not
mentioned in the reference report]
(f) Omitting a comparison detailing a change from a prior study: 0.5
[The candidate report does not mention the absence of disease progression, The candidate
report does not mention the position of the right IJ catheter, The candidate report does not
mention that the blunting of the costophrenic angle is not significantly changed from a prior
study]
```

Figure 4: Sample GREEN summary. For each error subcategory, we provide the most representative error explanations, enabling users to pinpoint areas for improvement for their trained systems.

parallelism and deterministic sampling with a maximum token length of 2,048 to ensure the reproducibility of the GREEN metric.

3.2 GREEN Score

We employed regular expressions (regex) to parse the counts of errors from the model's output. Specifically, we denoted the count of each type of error as # $\text{error}_{s,i}$, where the error's clinical significance $s \in \{\text{sig.}, \text{insig.}\}$ and subcategory $i \in \{(a), (b), \ldots, (f)\}$.

To calculate the GREEN score, we prioritized # $\text{error}_{\text{sig.},i}$ (errors with the potential to alter clinical decision-making processes) alongside the counts of accurate matched findings, # matched findings, for inversion. The formula for the GREEN score is then expressed as:

$$\text{GREEN} = \frac{\#\text{ matched findings}}{\#\text{ matched findings} + \sum_{i=(a)}^{(f)} \#\text{ error}_{\text{sig.},i}} \tag{1}$$

if # matched findings > 0, otherwise 0. Thus, the GREEN score (↑) is bounded between 0 and 1.

3.3 GREEN Summary

We present a method for a detailed analysis of error explanation per error subcategory i. Initially, the process involves clustering the embeddings of the explanation sentence, $e_i \in E_i$, into k clusters using Sentence Transformers (Ni et al., 2021). Subsequent steps include quantifying the size of these k clusters, where k is determined by the silhouette distance (Shahapure and Nicholas, 2020). We then identify the largest cluster and isolate the top-3 members exhibiting the closest distance to the cluster's mean as determined through cosine distance measures. The top-3 members are then named for each error category in the GREEN summary (Appendix A.5). For example, in Figure 4, subcategory (a) has "*(Small right pleural effusion, Small right pleural effusion, Small right pleural effusion)*" as the three closest members to the largest cluster, which indicates significant hallucinations of a "*small right pleural effusion*". This opens up the possibility for targeted detection of data biases or quality issues, as well as specific areas for model improvement.

3.4 Validation

3.4.1 Expert Error Counts Dataset

To validate the GREEN score in a clinical setting, we utilized the publicly-available ReXVal dataset, which includes assessments from six board-certified radiologists on 200 pairs of generated radiology reports from 50 cases of the MIMIC-CXR test set (Yu et al., 2023b). Each radiologist counted the occurrences of six spe-

Table 1: Results on the internal test set (10,000 examples) when compared to GPT-4 error counts. [1]Mean absolute error ± standard deviation, [2]Average error count.

Language Model	MAE ± STD [1]			Δ GREEN ↓
	ΔSig. Error Count ↓ 3.1± 2.6[2]	ΔInsig. Error Count ↓ 0.15 ± 0.52[2]	ΔMatched Findings ↓ 2.07 ± 1.84[2]	
Mistral-v0.1 (7B)	0.97 ± 1.18	0.22 ± 0.58	0.44 ± 0.70	0.11 ± 0.17
LLaMA-2 (7B)	1.35 ± 1.40	**0.15 ± 0.52**	1.62 ± 1.67	0.29 ± 0.26
Phi-2 (2.7B)	0.84 ± 1.14	0.20 ± 0.58	0.34 ± 0.59	0.09 ± 0.14
RadLLaMA-2 (7B)	0.70 ± 0.99	0.20 ± 0.57	0.29 ± 0.54	0.08 ± 0.13
RadPhi-2 (2.7B)	**0.63 ± 0.99**	0.18 ± 0.57	**0.26 ± 0.53**	**0.06 ± 0.12**

cific error types denoted earlier, distinguishing between errors of clinical significance and those considered insignificant.

3.4.2 Expert Preference Dataset

Since we lacked a ground-truth score for evaluating radiology reports, we turned to a radiologist preference dataset. This helped to address the shortcomings of comparing errors counts in the previous section. In particular, when radiologists compare two reports, they do so with an intuitive weighting of matched findings and significant and insignificant errors. As such, the preference dataset enabled us to determine which metric most effectively replicates expert evaluations and allowed us to assess GREEN as a preference generator (Rafailov et al., 2024; Ethayarajh et al., 2024).

We collected 100 pairwise preferences by two board-certified radiologists (with over 5 and 25 years of experience). The dataset comprised of 50 cases of the ReXVal dataset (Yu et al., 2023b), supplemented by an additional 50 cases randomly selected from the MIMIC-CXR test set. The two radiologists were presented with a chest X-ray alongside two corresponding candidate reports generated by an image-captioning model fine-tuned on the MIMIC CXR training set[3]. The primary task for the radiologist was to select the candidate report they preferred and to quantify their confidence in this selection on a scale ranging from 1 to 10. The intention behind this was to categorize the complexity of the task. In essence, when radiologists have a high degree of confidence in their chosen report for a given case, it is anticipated that the automated preference generator will show the highest concordance, as the task is deemed less challenging. The quality of the two candidates is considered to be equal when experts had differing preferences, hence implying no preference. We then excluded such cases from consideration.

4 Experiments

4.1 Inter-Expert Analysis

The baseline for model performance was established by comparing the correlation between experts to each other from the ReXVal dataset using Kendall's Tau coefficient (Yu et al., 2023b). The correlation between the 6 experts was less than the average correlation across experts, which spans from 0.41 to 0.60 (Figure 5). Additionally, we assessed the discrepancy between experts by computing the mean absolute error of significant error counts, resulting in a mean difference of 0.83 ± 0.13 (Table 2). These inter-expert measures serve as upper bound for the GREEN model performance outlined below.

Table 2: Difference between ReXVal experts and the GREEN model measured using mean absolute error of significant error counts.

	0	**1**	**2**	**3**	**4**	**5**	GREEN
0	–	0.505	0.835	0.675	0.495	1.130	1.160
1	0.505	–	1.100	0.870	0.660	1.365	1.485
2	0.835	1.100	–	0.730	0.770	0.725	0.715
3	0.675	0.870	0.730	–	0.570	0.965	0.895
4	0.495	0.660	0.770	0.570	–	1.025	1.005
5	1.130	1.365	0.725	0.965	1.025	–	0.930
GREEN	1.160	1.485	0.715	0.895	1.005	0.930	–

4.2 Performance on Training Data Distribution

We measured performance of the GREEN metric models by sampling deterministically and comparing the mean absolute errors and classical lexical metrics against reference labels from GPT-4. We found that RadPhi-2 and RadLLaMA-2 exhibit the lowest mean absolute difference for clinically significant errors of 0.63 ± 0.99 (Table 1) and the highest classical lexical metrics with a mean BERTScore of 0.84 ± 0.10 (Table 3). We measured clustering and summary consistency with language similarity metrics like BERTScore. Consistent with the quantitative results, we found that RadPhi-2 and RadLLaMA-2 yielded the best natural language agreement with GPT-4 (Table 1).

4.3 Performance on Validation Data Distribution

To analyze the performance upper bound of GREEN, we inferred GREEN scores from GPT-4 responses on the validation set, and referred to it as GREEN-GPT-4.

[3]https://huggingface.co/nlpconnect/vit-gpt2-image-captioning

Table 3: Results on the Internal test set (10,000 examples) when compared to GPT-4 responses.

Language Model	Lexical		
	BERTScore↑	ROUGE-L ↑	BLEU ↑
Mistral-v0.1 (7B)	0.80 ± 0.11	0.68 ± 0.18	0.54 ± 0.22
LLaMA-2 (7B)	0.78 ± 0.12	0.65 ± 0.19	0.53 ± 0.21
Phi-2 (2.7B)	0.80 ± 0.11	0.70 ± 0.18	0.54 ± 0.23
RadLLaMA-2 (7B)	0.83 ± 0.24	0.73 ± 0.17	0.59 ± 0.23
RadPhi-2 (2.7B)	**0.84 ± 0.10**	**0.76 ± 0.17**	**0.64 ± 0.23**

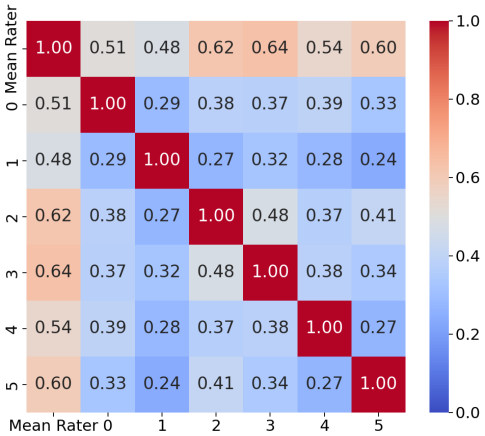


Figure 5: Mean-expert and inter-expert correlation matrix (Kendall's Tau) for fine-grained error counts on the external validation set (RexVal (Yu et al., 2023b)).

4.3.1 Expert Error Counts

To quantitatively validate GREEN, we measured the mean absolute difference and accuracy relative to the average radiologist, as detailed in Section 4.1. We found that, overall, RadLLaMA-2 exhibits the lowest differences to the mean radiologist's error counts (1.54 ± 1.36 sig. error difference), which approaches the performance of GPT-4 (1.51 ± 1.29 sig. error difference). Compared to all experts individually, RadLLaMA-2 exhibits an average difference of 1.02 ± 0.27, which is within the boundaries of the average inter-expert difference of 0.83 ± 0.13. Drawing from these quantitative results, along with the findings presented in Section 3.4, we selected RadLLaMA-2 as the GREEN model for all future experiments.

To validate GREEN against existing metrics, we assessed the correlation between the total error count by radiologists and the classical metrics, alongside GPT-4 and GREEN, using an external dataset (Table 5). Both GREEN-GPT-4 and GREEN demonstrated similar and stronger correlations compared to classical metrics. We noted that the GREEN correlation significantly outperforms that of RadGraph, despite RadGraph being trained on human annotations (RadGraph: 0.47 (95% CI, -0.55 0.39) vs. GREEN: 0.63 (95% CI, 0.69 0.56)).

Compared to the inter-expert correlation, GREEN exhibits a competitive degree of correlation at 0.63 compared to the range from 0.48 to 0.64 on the same examples (Figure 5).

Figure 6: Radiologist confidence vs. accuracy of preference labeling. As the confidence of the experts in their preferences increases, the GREEN score demonstrates the highest alignment with expert preferences as compared to the approach of using just the summed error counts. This difference was quantified using accuracy (green lines). Of note, if GPT-4 is asked directly about a preference, it aligns poorly with the expert preference. However, when the GREEN score formula is applied, a higher accuracy is shown even at lower expert confidence levels. Detailed results can be found in Table 6.

Additionally, we observed that correlating the unweighted summed error counts of GREEN (clinically sig. errors + insig. errors) yields a correlation coefficient of 0.79 (95% CI, 0.74 0.83), which may exceed the performance of GPT-4-based G-Rad (Chaves et al., 2024).

We further used expert preferences to determine whether the summed error counts or the GREEN score best mimics expert evaluation. This approach is based on the assumption that clinically significant errors, insignificant errors, and matched findings carry different weights in determining the quality of a candidate report.

4.3.2 Expert Preferences

The accuracy of the generated preferences was measured by how often the expert-preferred report matched the report that had a higher score from GREEN, a lower summed error count, or the preference of GPT-4. The prompt that was used for GPT-4 preferences is shown in Appendix A.3).

We observed the highest mean accuracy for GREEN and GREEN-GPT-4, which both outperforms the summed error count approach. The preferences of GPT-4 exhibited an accuracy of 0.23 (95% CI, 0.13 0.36) (Table 6).

Upon examining the impact of varying confidence levels (Figure 6), we observed that GREEN's preference alignment improves in conjunction with increased radiologist confidence, distinguishing it from the approach of using just the sum of error counts or the direct GPT-4 preference with low accuracy.

Table 4: Results on the external validation set (200 examples) compared to ReXVal human experts. [1]MAE: Mean Absolute Error of the sum of Sig. Errors and Insig. Errors, [2]STD: Standard Deviation, [3]Average error count. (a)-(f) Accuracy for each significant error category: (a) False report of a finding in the candidate, (b) Missing a finding present in the reference, (c) Misidentification of a finding's anatomic location/position, (d) Misassessment of the severity of a finding, (e) Mentioning a comparison that isn't in the reference, and (f) Omitting a comparison detailing a change from a prior study.

	Language Model	MAE[1] ± STD [2]		Accuracy ↑					
		ΔSig. Error↓	ΔInsig. Error↓	(a)	(b)	(c)	(d)	(e)	(f)
		7.03 ± 1.16 [3]	0.47 ± 0.52[3]						
Compared to Average Expert	Mistral-v0.1 (7B)	2.60 ± 1.91	0.87 ± 0.94	0.13	0.31	0.62	0.59	0.48	0.67
	LLaMA-2 (7B)	2.62 ± 1.25	**0.47 ± 0.52**	0.10	0.23	0.65	0.59	0.68	0.70
	Phi-2 (2.7B)	2.10 ± 1.39	0.65 ± 0.70	0.34	0.08	0.65	0.57	0.66	0.53
	RadLLaMA-2 (7B)	**1.54 ± 1.36**	0.51 ± 0.54	0.34	0.38	0.60	0.54	0.65	0.68
	RadPhi-2 (7B)	2.08 ± 1.15	0.55 ± 0.61	0.19	0.18	0.62	0.57	0.62	0.61
	GPT-4	1.51 ± 1.29	0.52 ± 0.55	0.32	0.40	0.65	0.59	0.68	0.70

Table 5: Correlation analysis between metrics, GREEN score, and GREEN score inferred from GPT-4 (GREEN-GPT-4) to average total error count of 6 radiologists in the ReXVal dataset (200 examples).

Metrics	Kendall's Tau ↑
BLEU	0.33 (95% CI, 0.42 0.23)
BERTScore	0.44 (95% CI, 0.52 0.35)
F1RadGraph	0.47 (95% CI, 0.55 0.39)
ROUGE-L	0.53 (95% CI, 0.61 0.45)
RadCliQ-v1	0.58 (95% CI, 0.51 0.64)
GREEN (ours)	**0.63 (95% CI, 0.69 0.56)**
GREEN GPT-4 (ours)	0.64 (95% CI, 0.70 0.57)
Error counts GREEN	0.79 (95% CI, 0.74 0.83)
Error counts GPT-4	0.79 (95% CI, 0.75 0.83)

Table 6: Accuracies with 95% CI of various preferences when compared to expert preferences.

	Accuracy
Preference GPT-4	0.23 (95% CI, 0.13 0.36)
Error Count GREEN	0.57 (95% CI, 0.43 0.70)
Error Count GPT-4	0.60 (95% CI, 0.47 0.74)
GREEN (ours)	0.62 (95% CI, 0.49 0.75)
GREEN GPT-4 (ours)	0.68 (95% CI, 0.55 0.79)

5 Multimodality Generalizability

We now demonstrate how this method can be applied to various other imaging modalities.

5.1 Out-of-Chest X-ray Dataset

Recent works extended RRG capabilities of VLMs to other imaging modalities (Hamamci et al., 2024; Bai et al., 2024). To extend the GREEN model to new imaging modalities beyond chest X-rays, we created a dataset analogous to the training dataset used for the GREEN chest X-ray (Section 3.1), but without access to RRG models to generate candidate reports for every modality. We did this to validate our method on a range of imaging modalities for which RRG models may not yet exist.

This new dataset is also based on MIMIC-IV Radiology Reports, which includes 2,321,355 de-identified radiology reports from 237,427 patients. It covers a variety of imaging modalities such as X-ray, computed tomography, magnetic resonance imaging, and ultrasound, as referenced in (Johnson et al., 2023).

We first uniformly sampled reports to maintain a distribution of cases similar to that described in (Johnson et al., 2023). Secondly, we used 4 methods to modify the radiology reports to generate 50,000 candidate reports: i) re-arranging the order of sentences, ii) removing sentences, iii) randomly pairing sentences, and iv) modifying the report by sampling random combinations of error categories and asking GPT-4 to incorporate errors into the reports to generate a candidate report (if no error categories are sampled, GPT-4 is asked to rephrase the report with the same meaning by changing a small number of words) (Appendix A.4).

We then prompted GPT-4 to evaluate the differences with the same prompt design as with the chest X-ray data. We further split these 50,000 reports into training, validation, and test sets according to the same 80/10/10 ratio and combined them with the initial chest X-ray dataset.

5.2 OOD External Abdominal CT dataset

We randomly chose 15 pairs of reference and candidate reports from an abdominal CT dataset. The dataset originated from the [BLIND] dataset, which includes examinations from December 2012 to October 2018.

Table 7: Adapting GREEN to any imaging modality: Performances on Out-of-Chest X-ray and OOD Data in Zero-Shot and Trained Conditions. [1]Mean absolute error ± standard variation, [2]Average Error, [3]Modalities include X-ray, computed tomography, magnetic resonance imaging, and ultrasound.

Evaluation data	Training data	MAE ± STD [1]			Δ GREEN ↓
		ΔSig. Error Count ↓	ΔInsig. Error Count ↓	ΔMatched Findings ↓	
MIMIC-IV-Notes[3]	CXR dataset	1.05 ± 1.51	0.20 ± 0.51	0.49 ± 1.00	0.10 ± 0.17
	CXR + Out-of CXR dataset	0.61 ± 0.99	0.19 ± 0.48	0.34 ± 1.04	0.07 ± 0.15
Abdominal CT, OOD	CXR dataset	5.31 ± 2.82	0.06 ± 0.24	4.09 ± 2.74	0.21 ± 0.17
	CXR + Out-of CXR dataset	3.12 ± 2.03	0.19 ± 0.53	3.56 ± 3.28	0.17 ± 0.23

Table 8: Adapting GREEN to any imaging modality: Performance on Out-of-Chest X-ray and OOD data distribution in Zero-Shot and Trained Conditions based on Lexical Metrics. [1]Modalities include X-ray, computed tomography, magnetic resonance imaging, and ultrasound.

Evaluation data	Training data	Lexical		
		BERTScore ↑	ROUGE-L ↑	BLEU ↑
MIMIC-IV-Notes[1]	CXR dataset	0.74 ± 0.12	0.62 ± 0.18	0.45 ± 0.21
	CXR + Out-of CXR dataset	0.81 ± 0.09	0.73 ± 0.15	0.60 ± 0.18
Abdominal CT, OOD	CXR dataset	0.68 ± 0.12	0.58 ± 0.15	0.41 ± 0.13
	CXR + Out-of CXR dataset	0.71 ± 0.06	0.58 ± 0.07	0.45 ± 0.06

5.3 Out-of-Chest X-ray Experiments

We first evaluated zero-shot performance of the GREEN model on the Out-of-Chest X-ray dataset (1.05 ± 1.51 sig. error count difference) and on the external OOD (5.31 ± 2.82 sig. error count difference). We fine-tuned the best checkpoint of the GREEN model on the Out-of-Chest X-ray dataset with a batch size of 80 for 8 epochs and the same hyperparameters as mentioned in Section 3.1.2. We used the same evaluation experiments as in the previous section. We found that further fine-tuning on multimodality data improves the sig. error count difference and the text similarity metrics for both the in-distribution and out-of-distribution data (0.61 ± 0.99 and 3.12 ± 2.03 sig. error count difference).

6 Conclusion

In this study, we introduced GREEN (Generative Radiology Report Evaluation and Error Notation), a novel metric aimed at enhancing the evaluation of radiology reports. GREEN outperforms existing metrics by aligning closely with the nuanced requirements of medical diagnostics through its precise assessment of factual correctness and uncertainties. The score's high correlation with expert evaluations underscores its effectiveness.

The open-source nature of GREEN supports widespread use and collaborative improvements without compromising data privacy. Its lightweight design ensures practicality across diverse settings, reducing computational demands. Additionally, GREEN's adaptability across different imaging modalities and extensive datasets encourage broader applicability and research in medical artificial intelligence.

The GREEN metric's ability to maintain robust performance on OOD data further signifies its versatility and potential as a standard for future developments in automated radiology reporting.

7 Limitations

Analyzing each sample takes roughly 3.75 seconds on one A100 GPU. However, using batching can accelerate the processing to four samples in 4.22 seconds (equivalent to about 1.06 seconds per sample). Due to its complexity, it is slower compared to ROUGE, at approximately 0.015 seconds per sample, but faster than GPT-4, at up to 22.0 seconds per sample).

We introduce OOD metrics and suggest a strategy to adjust the GREEN model to different imaging techniques, even in the absence of an initial RRG model for each technique. Nonetheless, fine-tuning GREEN for new imaging modalities might be required in subsequent studies to ensure satisfactory performance.

Although the model operates deterministically to ensure reproducible outputs, the error quantification remains, to some extent, uncontrollable, which introduces a degree of randomness to the counting of errors. This randomness may stem from inherent uncertainties in the task, as evidenced by the disagreement among experts on fine-grained error counts (Section 4.1). This is a characteristic that has been previously observed and noted in inter-expert agreement analyses (Irvin et al., 2019).

References

Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. 2024. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, et al. 2024. Training small multimodal models to bridge biomedical competency gap: A case study in radiology imaging. *arXiv preprint arXiv:2403.08002*.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. 2024. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dina Demner-Fushman, Sameer Antani, Matthew Simpson, and George R Thoma. 2012. Design and development of a multimodal biomedical information retrieval system. *Journal of Computing Science and Engineering*, 6(2):168–177.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Sijing Feng, Damian Azzollini, Ji Soo Kim, Cheng-Kai Jin, Simon P Gordon, Jason Yeoh, Eve Kim, Mina Han, Andrew Lee, Aakash Patel, et al. 2021. Curation of the candid-ptx dataset with free-text reports. *Radiology: Artificial Intelligence*, 3(6):e210136.

Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Bastian Wittmann, Enis Simsar, Mehmet Simsar, et al. 2024. A foundation model utilizing chest ct volumes and radiology reports for supervised-level zero-shot detection of abnormalities. *arXiv preprint arXiv:2403.17834*.

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*, volume 33, pages 590–597.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew Lungren, Andrew Ng, Curtis Langlotz, Pranav Rajpurkar, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Mojan Javaheripi and Sébastien Bubeck. 2023. Phi-2: The surprising power of small language models. Microsoft Research Blog.

Jaehwan Jeong, Katherine Tian, Andrew Li, Sina Hartung, Subathra Adithan, Fardad Behzadi, Juan Calle, David Osayande, Michael Pohlen, and Pranav Rajpurkar. 2024. Multimodal image-text matching improves retrieval-based chest x-ray report generation. In *Medical Imaging with Deep Learning*, pages 978–990. PMLR.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

A Johnson, T Pollard, S Horng, LA Celi, and R Mark. 2023. Mimic-iv-note: Deidentified free-text clinical notes (version 2.2). physionet.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

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.

Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. 2023. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3334–3343.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Hoang TN Nguyen, Dong Nie, Taivanbat Badamdorj, Yujie Liu, Yingying Zhu, Jason Truong, and Li Cheng. 2021. Automated generation of accurate\& fluent medical x-ray reports. *arXiv preprint arXiv:2108.12126*.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y Zhao, Yi Luan, Keith B Hall, Ming-Wei Chang, et al. 2021. Large dual encoders are generalizable retrievers. *arXiv preprint arXiv:2112.07899*.

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*, pages 311–318.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Pranav Rajpurkar and Matthew P Lungren. 2023. The current and future state of ai interpretation of medical images. *New England Journal of Medicine*, 388(21):1981–1990.

Vignav Ramesh, Nathan A Chi, and Pranav Rajpurkar. 2022. Improving radiology report generation systems by removing hallucinated references to nonexistent priors. In *Machine Learning for Health*, pages 456–473. PMLR.

Ketan Rajshekhar Shahapure and Charles Nicholas. 2020. Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pages 747–748. IEEE.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. 2020. Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Shikhar Vashishth, Denis Newman-Griffis, Rishabh Joshi, Ritam Dutt, and Carolyn P Rosé. 2021. Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets. *Journal of biomedical informatics*, 121:103880.

Maria De La Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, et al. 2020. Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients. *arXiv preprint arXiv:2006.01174*.

Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. 2022. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis*, 80:102510.

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. 2023a. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9).

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, EKU Fonseca, Henrique Lee, Zahra Shakeri, Andrew Ng, et al. 2023b. Radiology report expert evaluation (rexval) dataset.

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

A Appendix

A.1 GPT-4 Prompt Template for Generation of Training Data

The following prompt was used in GPT-4 to generate the GREEN model training data. **Reference Report** and **Candidate Report** fields are replaced with their respective actual reports.

GPT-4 Prompt

```
Objective:
Evaluate the accuracy of a candidate radiology report in comparison to a reference
radiology report composed by expert radiologists.

Process Overview:
You will be presented with:
1. The criteria for making a judgment.
2. The reference radiology report.
3. The candidate radiology report.
4. The desired format for your assessment.

1. Criteria for Judgment:
For each candidate report, determine:
    - The count of clinically significant errors.
    - The count of clinically insignificant errors.

Errors can fall into one of these categories:
    a) False report of a finding in the candidate.
    b) Missing a finding present in the reference.
    c) Misidentification of a finding's anatomic location/position.
    d) Misassessment of the severity of a finding.
    e) Mentioning a comparison that isn't in the reference.
    f) Omitting a comparison detailing a change from a prior study.

Note: Concentrate on the clinical findings rather than the report's writing style.
Evaluate only the findings that appear in both reports.

2. Reference Report:
**Reference Report**

3. Candidate Report:
**Candidate Report**

4. Reporting Your Assessment:
Follow this specific format for your output, even if no errors are found:
```
[Explanation]:
<Explanation>

[Clinically Significant Errors]:
(a) <Error Type>: <The number of errors>. <Error 1>; <Error 2>; ...; <Error n>
....
(f) <Error Type>: <The number of errors>. <Error 1>; <Error 2>; ...; <Error n>

[Clinically Insignificant Errors]:
(a) <Error Type>: <The number of errors>. <Error 1>; <Error 2>; ...; <Error n>
....
(f) <Error Type>: <The number of errors>. <Error 1>; <Error 2>; ...; <Error n>

[Matched Findings]:
<The number of matched findings>. <Finding 1>; <Finding 2>; ...; <Finding n>
```
```

A.2 GPT-4 Response used as Training Data

The following depicts a sample GPT-4 response to the prompt in Appendix A.1 for a pair of reports selected from a 'RadGraph Permutation'. Specifically, the candidate report contains exactly one RadGraph modifier change.

Reference: "Examination. Faint infiltrates in the upper middle right field and doubtful retrocardiac suggestive of respiratory infection. Costophrenic sinuses are clear. No other notable findings."

Candidate: "Examination . Subpleural infiltrates in the upper dorsal right field and doubtful retrocardiac suggestive of respiratory infection . Costophrenic sinuses are clear . No other notable findings ."

GPT-4 Response

```
[Explanation]:
The candidate report misidentifies the anatomic location of the infiltrates.
The reference report mentions "upper middle right field" while the candidate report
mentions "upper dorsal right field".

[Clinically Significant Errors]:
(c) Misidentification of a finding's anatomic location/position: 1. The infiltrates
are in the upper middle right field, not the upper dorsal right field.

[Clinically Insignificant Errors]:
(a) False report of a finding in the candidate: 0.
(b) Missing a finding present in the reference: 0.
(c) Misidentification of a finding's anatomic location/position: 0.
(d) Misassessment of the severity of a finding: 0.
(e) Mentioning a comparison that isn't in the reference: 0.
(f) Omitting a comparison detailing a change from a prior study: 0.

[Matched Findings]:
3. Doubtful retrocardiac suggestive of respiratory infection;
Costophrenic sinuses are clear; No other notable findings.
```

A.3 GPT-4 Prompt Template for Preference Evaluation

The following prompt was used in GPT-4 for the evaluation of preferences. **Reference Report**, **Candidate Report 1**, and **Candidate Report 2** fields are replaced with their respective actual reports.

GPT-4 Prompt

```
We would like to request your feedback on the radiology reports generated by two AI
assistants by comparing them to the reference report written by radiologists.

[Reference Report]
**Reference Report**

[Assistant 1]
**Candidate Report 1**

[Assistant 2]
**Candidate Report 2**

[Requirements]
1. The length of the reports is not important.
2. The style of the reports is not important.
3. The clinical accuracy is important especially for positive findings (i.e., diseases).
Therefore, please focus on clinical accuracy instead of the length and style.

Please compare the accuracy of their generated reports. You should tell me whether Assistant 1
is "better than", "worse than", or "equal to" Assistant 2.

Please first compare the generated reports with the reference report to analyze which one is
more in line with the given requirements.

In the last line, please output a single line containing only a single label selecting from
"Assistant 1 is better than Assistant 2", "Assistant 1 is worse than Assistant 2", and
"Assistant 1 is equal to Assistant 2".
```

A.4 Algorithm for Modifying Radiology Reports

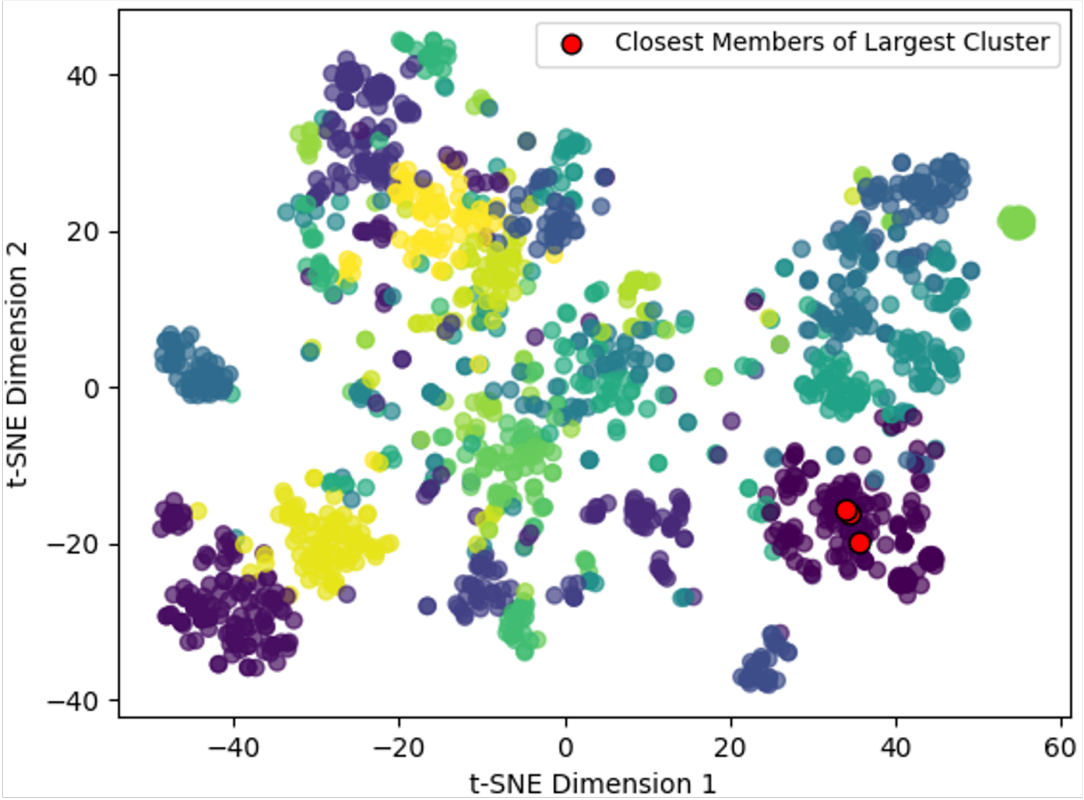
We employed this algorithm to produce prompts for GPT-4 to modify candidate reports for new imaging modalities that lack RRG models.

```
def get_prompt(self, report):
        error_types = self.get_error_combination(report)
        # randomly choose if subtle or not, if subtle add "sentence"
        subtle_change = ""
        if random.random() > 0.5:
            subtle_change = "Aim for subtlety, adjusting only one word where feasible. "
        if not error_types == "no errors":
            return f"[Objective]: Create a candidate radiology report that subtly integrates specific errors based on the provided reference report.
            Process Overview: You will be presented with:
            1. Style of errors.
            2. A reference radiology report to base your candidate report on.
            3. The desired format for your candidate report. Note: Be short in your response!

            Style of errors:
            Introduce errors related to {error_types}. The errors should be woven into the report as if they were genuine observations from a medical image, without any meta-commentary on their accuracy. {subtle_change}

            Reference Report: \n{report}\n Desired format for your candidate report:  \n\n [Candidate]: <Candidate Report>"

        return f"[Objective]: Create a candidate radiology report that has the same
        clinical meaning but is slightly rephrased.
        Process Overview: You will be presented with:  \n 1.A reference radiology report to base your candidate report on. \n 2. The desired format for your candidate report. Note: Be short in your response! \n\n Reference radiology report: \n{report}\n\n Desired format for your candidate report:  \n\n [Candidate]: <Candidate Report>"
```

A.5 Visualization of the GREEN Summary Clustering Technique

Visualization (t-SNE) of the clustering technique used in the GREEN summary. Sentences were clustered for each error subcategory.