X-ray Made Simple: Lay Radiology Report Generation and Robust Evaluation

Anonymous ACL submission

Abstract

While multimodal generative models have advanced radiology report generation (RRG), challenges remain in making reports accessible to patients and ensuring reliable evaluation. The technical language and templated nature of professional reports hinder patient comprehension and enable models to artificially boost lexical metrics such as BLEU by reproducing common report patterns. To address these limitations, we propose the Layman's RRG framework, which leverages layperson-friendly language to enhance patient accessibility and promote more robust evaluation and report generation by encouraging models to focus on semantic accuracy over rigid templates. Our approach also introduces and releases two refined layman-style datasets (at the sentence and report levels), along with a semantics-based evaluation metric that mitigates inflated lexical scores and a layman-guided training strategy. Experiments show that training on layman-style data helps models better capture the meaning of clinical findings. Notably, we observe a positive scaling law: model performance improves with more layman-style data, in contrast to the inverse trend observed with templated professional language.

1 Introduction

With the advancement of generative models, image captioning has made significant progress in producing accurate textual descriptions from visual inputs. This capability has been increasingly applied in the medical domain, particularly in Radiology Report Generation (RRG) (Lin et al., 2022; Wang et al., 2022; Lee et al., 2023; Hou et al., 2023; Yan et al., 2023; Li et al., 2023; Liu et al., 2024). RRG aims to generate descriptive reports from medical images, such as chest X-rays, to reduce radiologists' workload while improving the quality, consistency, and efficiency of clinical documentation. Despite recent progress, two critical challenges remain underexplored. First, the generated reports often lack patient accessibility due to their use of highly technical language and rigid clinical templates, making them difficult for non-experts to understand. Second, current evaluation metrics and training paradigms emphasize surface-level textual similarity rather than true semantic understanding, potentially masking important deficiencies in report quality (Stent et al., 2005; Callison-Burch et al., 2006; Smith et al., 2016; Li et al., 2019; Yan et al., 2021; Dalla Serra et al., 2022; Kale et al., 2023; Yan et al., 2021; Dalla Serra et al., 2022). Although these challenges may appear distinct, they are closely linked: the templated language that hinders patient comprehension also leads models to overfit to surface patterns, inflating evaluation scores and hindering semantic generalization.

A patient-centered approach is becoming increasingly vital in modern healthcare, emphasizing transparency and shared decision-making. With policies like the 21st Century Cures Act requiring immediate access to electronic health records (EHR), patients now often receive radiology reports before any clinical interpretation. However, these reports—designed primarily for clinician communication and billing—are written in highly technical language, with fewer than 4% meeting the eighth-grade reading level typical of U.S. adults (Martin-Carreras et al., 2019). This mismatch presents major barriers to understanding and engagement, frequently resulting in confusion, anxiety, and poor adherence to follow-up or treatment plans (Domingo et al., 2022; Mabotuwana et al., 2018). The challenge is compounded by the fact that only 50% of recommended follow-ups are completed (Mabotuwana et al., 2019), in part due to unclear communication of incidental findings. While prior studies have explored barriers from the patient's perspective, little work has addressed the need to redesign the reports themselves. Improving report accessibility is therefore both a practical necessity and an ethical obligation in advancing patient-centered AI.

Beyond the challenge of patient accessibility, radiology report generation also faces a fundamental lack of robustness in both evaluation and training. On the evaluation side, most RRG models are still assessed using lexical overlap-based metrics like BLEU and ROUGE (Papineni et al., 2002; Lin, 2004), which remain dominant in the field (Liu et al., 2023). However, these metrics operate at the surface level, capturing word-level similarity while ignoring clinical meaning. For example, the phrases "*there is a focal consolidation*" and "*there is no focal consolidation*" receive similarly high BLEU scores due to shared structure, despite expressing opposite clinical conclusions (Stent et al., 2005). This shortcoming is magnified by the highly templated nature of radiology reports (Li et al., 2019; Kale et al., 2023), where rigid formats enable models to achieve high scores by mimicking patterns rather than grasping content. Prior work has shown that template-based substitutions can produce strong lexical scores even when semantic accuracy is lost (Kale et al., 2023). Moreover, such structural rigidity in professional reports could also effect training, as models exposed to these templates often overfit to superficial cues instead of learning generalizable semantic representations.

We hypothesize that adopting layman-style language in radiology report generation can simultaneously address the dual challenges of accessibility and robustness. From the patient's perspective, layman terms enhance the readability and comprehensibility of reports, making them more inclusive and actionable. From the modeling perspective, the linguistic diversity and absence of rigid templates in layman-style reports encourage models to focus on semantic understanding rather than overfitting to superficial patterns. Building on this insight, we propose a new framework for radiology report generation grounded in layman's terms. Our framework includes: (1) creating two high-quality **layman-style datasets**: a *sentence-leve*l dataset and a *report-level* dataset; (2) a **semantics-based evaluation method** based on layman's terms, which provides fairer assessments that mitigates inflated BLEU scores; and (3) a **training strategy based on layman's terms** that improves the model's semantic learning and reduces its reliance on templated language in professional reports.

To validate the effectiveness of the Layman's RRG framework, we conduct extensive experiments using the publicly available MIMIC-CXR dataset (Johnson et al., 2019). Results show that our semantics-based evaluation method, combined with the sentence-level layman dataset, provides significantly more robust assessments. Furthermore, models trained with our layman-guided strategy exhibit stronger semantic generalization compared to those trained on templated professional reports. Notably, we observe a promising scaling trend: as the amount of layman-style training data increases, model performance continues to improve—unlike the diminishing gains seen with professional report training. These findings offer strong empirical support for our hypothesis that layman-style language enhances both accessibility and robustness in radiology report generation. In summary, our contributions are as follows:

- We introduce two high-quality layman-style radiology report generation datasets: a sentence-level dataset and a report-level dataset. To the best of our knowledge, this is the first systematic effort to create patient-friendly datasets for RRG, offering a valuable resource for future research aimed at enhancing the readability and inclusiveness of medical AI systems.

- We propose a layman-guided evaluation method for RRG that leverages LLM-based embedding models to substitute professional report sentences with semantically matched layman equivalents from our dataset. This method enables fairer and more robust assessment using both traditional lexical metrics and our proposed semantics-based metric.

- We demonstrate training on our report-level layman dataset enhances the model's semantic understanding and reveals a promising scaling law: performance improves consistently with more layman-style data—contrasting with the diminishing returns seen when training on professional reports.

2 Related work

2.1 Patient-Centric Reports

Some medical researches show that a direct link between patients' understanding of their medical information with adherence to recommended prevention and treatment processes, better clinical outcomes, better patient safety within hospitals, and less health care utilization (Anhang Price et al.,

2014; López-Úbeda et al., 2024; Martin-Carreras et al., 2019). Radiology reports, although written primarily for healthcare providers, are read increasingly by patients and their family. However, few researches have focused on patient-centric reports.

2.2 Evaluation Metrics for Radiology Report Generation

Evaluation metrics are essential for RRG as they provide measurements of the quality of the produced radiology reports from various approaches and ensure a fair comparison among counterparts. Similar to other AI research domains, prevailing approaches in RRG evaluation adopt automatic metrics by comparing the generated reports with gold standard references (i.e., doctor-written reports). Generally, metrics for this task are categorized into five types: natural language generation (NLG) (Papineni et al., 2002; Lin, 2004; Banerjee and Lavie, 2005; Zhao et al., 2023, 2024; Yang et al., 2024), clinical efficacy (CE) (Peng et al., 2018; Irvin et al., 2019; Smit et al., 2020; Jain et al., 2021), standard image captioning (SIC) (Vedantam et al., 2015), embedding-based metrics, and task-specific features-based metrics. Among these, NLG metrics and CE metrics are the most widely adopted in current approaches. However, most of these metrics primarily focus on word overlap and do not adequately consider the semantic meaning between the ground truth and generated reports.

3 Layman's Term RRG

In this section, we present **Layman's term RRG**, a unified framework encompassing {data creation, evaluation, and training}, designed to address the limitations of lexical-based metrics and the rigid, patterned nature of professional radiology reports. The framework (see Figure 1) is supported by two complementary resources: a sentence-level dataset for semantics-based evaluation and a report-level dataset for training models with improved semantic generalization.

3.1 Data Creation

Our data construction pipeline comprises three components: a deduplication preprocessing (applicable only to the sentence-level dataset), a generation–refinement step, and a human verification postprocessing. This pipeline is designed to produce high-quality layman-style sentences and reports.

3.1.1 Deduplication Preprocessing

We first use NLTK to segment each report into individual sentences. Through analyzing large volumes of reports, we found that many repetitive sentences share similar semantics. To simplify the final dataset and reduce the burden of pairwise similarity computation, we apply extensive deduplication to the sentence-level inputs. To this end, we use GritLM (Muennighoff et al., 2024), a decoder-based embedding model that achieves state-of-the-art performance on the Massive Text Embedding Benchmark (MTEB) and the Reasoning as Retrieval Benchmark (RAR-b), to encode sentences and obtain their vector representations. We then iteratively compute pairwise cosine similarities between sentences, retaining those that do not exceed a similarity threshold of 0.8 with previously selected sentences and discarding the rest. Through this deduplication procedure, the number of sentences is reduced from approximately 490,000 to 50,000, substantially lowering computational cost and improving the efficiency of subsequent processing.

3.1.2 Generation–Refinement Step

Generation. After the deduplication on sentences, we use `GPT-4o` to translate professional sentences or reports into layman-style language. The prompt design—detailed in Appendix A.1—specifies the generation objectives, enables batch processing, and instructs the model to return outputs in JSON format. This approach largely reduces cost and improves output consistency through referencing in-batch examples.

Refinement. To enhance translation quality, we introduce a self-refinement method involving a semantic-checking module built upon embedding models, and a correctness self-checking module using the same LLM in the generation step. Details of the self-check prompt are provided in Appendix A.2. For each professional–layman sentence pair, we combine self-check feedback from `GPT-4o` with semantic similarity scores from GritLM to ensure the quality of translated sentence. A translation must pass both checks to be accepted; otherwise, the sentence is resubmitted for regeneration. The full procedure of the generation-refinement step is outlined in Appendix A.4.

3.1.3 Human Verification

Following the refinement process, the dataset quality improved substantially. As shown in Ap-

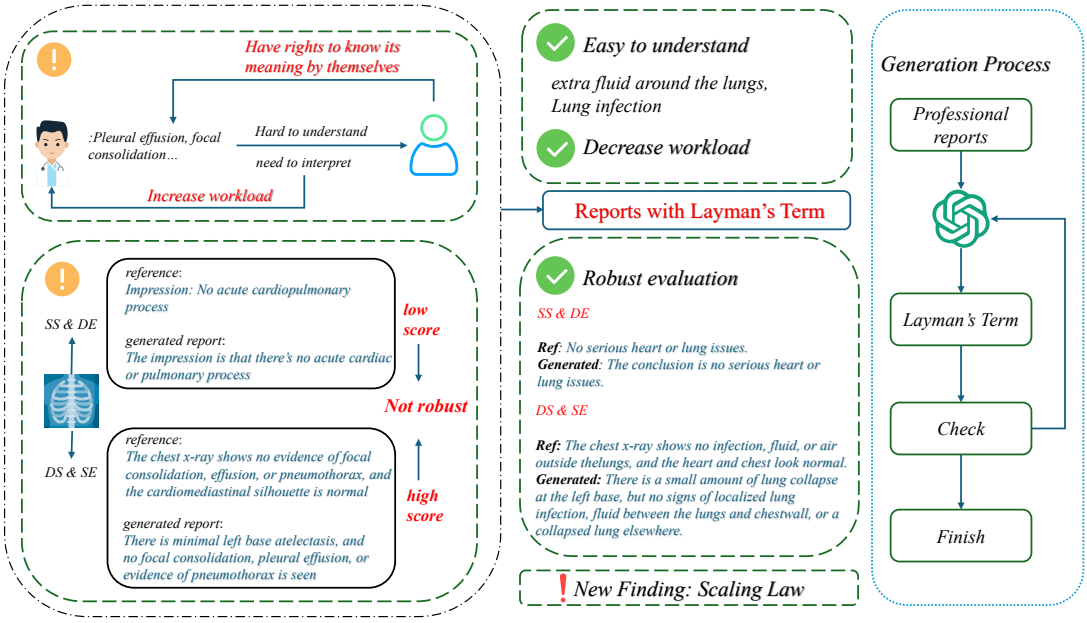


Figure 1: The Layman's RRG Framework. The "DS & SE" denotes different semantics and similar expressions. The "SS & DE" denotes similar semantics and different expressions.

pendix A.6, correction rates increase across self-refinement iterations. Additionally, we randomly sampled 500 sentence pairs for human verification, where over 98% were judged as correct matches.

3.2 Beyond Lexical Overlap: Semantics-Based Evaluation

Through thorough analysis of radiology reports, we observed that word-overlap metrics such as BLEU, ROUGE, and METEOR do not accurately reflect the quality of generated reports. This discrepancy arises due to the presence of semantically similar sentences with different wordings, as well as semantically different sentences with high lexical overlap. For example, the sentences *"There is a definite focal consolidation, no pneumothorax is appreciated"* and *"There is no focal consolidation, effusion, or pneumothorax"* convey distinct clinical meanings but achieve a BLEU-1 score greater than 0.6. This demonstrates that even when the underlying pathology differs, high BLEU scores may still be obtained due to surface-level similarity. Conversely, the sentences *"Impression: No acute cardiopulmonary process"* and *"The impression is that there's no acute cardiac or pulmonary process"* convey the same meaning but receive a low BLEU-1 score due to differences in phrasing. We categorize these inconsistencies into two types: **expression difference** issues and **semantics difference** issues. An expression difference issue occurs when the candidate and reference sentences share similar semantics but exhibit low word overlap. A semantics difference issue arises when the sentences differ in meaning but have high word overlap. Both issues can result in misleading BLEU scores, as illustrated in Table 1.

To address these issues, we propose a novel evaluation method for assessing generated radiology reports. In brief, the method compares a candidate report with a reference report by first splitting both into individual sentences. Each sentence is then replaced with its most semantically similar counterpart from our constructed sentence-level dataset, using GritLM to compute semantic similarity. Sentences exceeding a predefined similarity threshold are considered matched. We then calculate the proportion of matched sentences in both the candidate and reference reports as an additional metric, reported alongside traditional word-overlap metrics such as BLEU, ROUGE, and METEOR. This complementary metric enables our evaluation framework to mitigate the limitations of lexical-based evaluation and provide a more semantically grounded assessment of report quality. The detailed evaluation algorithm is provided in Appendix A.5.

3.3 Robust Training with Layman-style Data

To investigate how training data style affects the semantic generalization ability of generative models, we design a scaling-based training protocol

Examples of DS & SE			
Candidate	Reference	Candidate layman term	Reference layman term
The chest x-ray shows a normal cardiomediastinal contour and heart size.	The chest x-ray shows low lung volumes and a mildly enlarged heart size	The chest x-ray shows a normal heart and chest.	The chest x-ray shows lower than normal lung volumes and a slightly enlarged heart.
The chest x-ray shows well-expanded and clear lungs without any focal consolidation, effusion or pneumothorax	The chest x-ray shows left mid lung linear atelectasis/scarring, without any focal consolidation or large pleural effusion	The chest x-ray shows clear lungs without any infection, fluid, or air outside the lungs.	The chest x-ray shows some minor scarring or collapse in the left lung without any signs of localized lung infection or significant fluid.
Examples of SS & DE			
Impression: No acute cardiopulmonary process	The impression is that there's no acute cardiac or pulmonary process	No serious heart or lung issues.	The conclusion is no serious heart or lung issues.
The cardiac and mediastinal silhouettes are grossly stable	The cardiomediastinal silhouette appears stable	The heart and central chest area look stable.	The heart and central chest structures appear stable.
Additionally, there is no sign of pleural effusion or pneumothorax	There are no pleural effusions and pneumothorax	There are no indications of fluid build-up or air leakage in your lungs.	There is no fluid build-up in the chest, and no air leaks from the lungs.

Table 1: Samples can be categorized based on different semantics but similar expressions, as well as similar semantics but different expressions. The upperpart showcases examples of different semantics and similar expressions. Although these sentences yield a high BLEU score, they convey distinct meanings. Conversely, the lower part section presents examples of similar semantics and different expressions. Despite having a high BLEU score, these sentences express different meanings. The blue box and orange box denote the differing expressions in the reference and candidate texts.

using both professional and layman-style radiology reports. Our central hypothesis is that heavily templated professional reports encourage models to focus on surface structure rather than semantic content, while translating these reports into layman's terms removes rigid formatting and introduces linguistic diversity, thereby promoting semantic learning.

We construct a series of training subsets for both datasets (professional and layman-style), with sizes of 5k, 10k, 15k, 20k, 25k, and 50k samples. For each subset, we fine-tune the `MiniGPT-4` model. The training is conducted for 10 epochs with a batch size of 50, using gradient accumulation on NVIDIA A6000 GPUs. After training, we generate 500 radiology reports for each setting.

To evaluate model performance, we adopt our proposed semantics-based evaluation method. Specifically, for each generated report, we compute the semantic similarity between every sentence in the candidate report and each sentence in the reference report using GritLM embeddings. Sentence pairs exceeding a cosine similarity threshold of 0.8 are considered semantically matched. The proportion of matched sentences is used to assess semantic fidelity. In addition, we analyze the distribution of sentence pairs across similarity score ranges to better understand how different training regimes affect the semantic quality and variability of model outputs.

4 Experimental Results

4.1 Readability of Layman-Style Reports

We first evaluated the readability of LLM-generated layman-style radiology reports using two publicly available models, Kimi[1] and DeepSeek[2], on the MIMIC-CXR dataset, denoted as LLM1 and LLM2, respectively. To assess readability, we employed a suite of text-statistics-based metrics[3]. The abbreviations and descriptions of these metrics are listed in Appendix A.10. The `Baseline` approach refers to layman-style reports generated using the prompt provided in Appendix A.1 via `ChatGPT-4o`, while the `Original` approach corresponds to the professional radiology reports without modification. In addition to the baseline prompt (`P1`), we designed an instruction-following prompt (`P2`) that guides the model to generate layman-style reports based on provided examples. An illustration of this prompt is shown in Figure 4. As shown in Table 2, the layman-style reports produced by all three LLM approaches demonstrate substantially higher readability than the original professional reports across all evaluation metrics.

[1] Kimi (`www.moonshot.cn`)

[2] DeepSeek (`www.deepseek.com/`)

[3] We use the open-source Python library available at `pypi.org/project/textstat`

Data	Model	Easy Level↑	Level of Grade Required for Reading↓								
			M1	M2	M3	M4	M5	M6	M7	M8	M9
MIMIC CXR	(Original)	43	9	11	11	11	14	5	11	5	11
	Baseline	76	6	8	8	8	9	7	10	5	**19**
	LLM1+P1	84	**5**	8	8	7	**7**	7	**8**	**4**	21
	LLM1+P2	**85**	**5**	**7**	**7**	**6**	**7**	**6**	**8**	**4**	**19**

Table 2: Readability of Layman-Style Reports. Original represents professional reports. Baseline, LLM1+P1 and LLM1+P2 indicate layman-style reports generated by different LLMs and different prompts.

4.2 Limitations of Lexical-based Evaluation

In this section, we reveal the behavioral differences between lexical-based evaluation metrics and our proposed semantics-based evaluation metric.

To verify the effectiveness of layman-style reports in addressing expression difference and semantic difference issues, we construct two diagnostic subsets: (1) Similar Semantics & Different Expressions (SS & DE) and (2) Different Semantics & Similar Expressions (DS & SE). The way lexical-based and semantics-based metrics respond to these subsets serves as a characterization of their robustness.

For both raw professional reports and their layman-style counterparts, we compute BLEU, ROUGE, and METEOR scores, along with semantic similarity between candidate and reference sentences, within each diagnostic subset. The results are shown in Table 3. In the "DS & SE" subset, sentence pairs in the professional reports are mistakenly assigned high scores by lexical metrics—for example, 0.644 (BLEU-1), 0.505 (BLEU-2), 0.393 (BLEU-3), and 0.312 (BLEU-4). In contrast, their layman-translated counterparts significantly mitigate this mirage effect, reducing the scores to 0.312, 0.116, 0.064, and 0.042, respectively. Furthermore, our semantics-based metric correctly reflects the lack of semantic similarity in these pairs, with the proportion of sentences scoring above 0.8 dropping to only 2% and 1%.

Conversely, in the "SS & DE" subset, an ideal evaluation metric should be robust to surface-level differences and assign high scores to semantically aligned sentence pairs. However, lexical-based metrics fail to capture this relationship, yielding significantly lower scores for professional report pairs. Our translated layman pairs alleviate this weakness, producing higher perceived scores under lexical metrics. More importantly, the combination of our layman-style dataset and semantics-based metric yields the most robust evaluation: it not only achieves a high proportion of semantically similar pairs (over 50% scoring above 0.8), but also maintains a small perceptual gap between professional and layman versions.

Dataset	SS&DE		DS&SE	
Type	raw	layman	raw	layman
B-1	0.192	0.381	0.644	0.314
B-2	0.131	0.251	0.505	0.116
B-3	0.100	0.178	0.393	0.064
B-4	0.066	0.116	0.312	0.042
R-1	0.349	0.407	0.622	0.286
R-2	0.169	0.210	0.399	0.072
R-L	0.341	0.383	0.581	0.250
Meteor	0.386	0.452	0.627	0.310
Semantics	0.5	0.507	0.02	0.01

Table 3: BLEU and ROUGE score in professional report and its layman's term. SS&DE represent similar semantics and different expressions; DS&SE means different semantics and similar expressions. Semantic scores are calculated with the proportion of semantic similarity over 0.8 among all sentences.

In summary, lexical-based metrics suffer from inherent limitations, particularly when applied to the highly patterned structure of professional radiology reports. These metrics often fail to reflect the true semantic relationships between sentence pairs—frequently assigning higher scores to DS pairs than to SS pairs. Our layman-style dataset helps correct this imbalance, reversing the trend and enabling lexical metrics to better align with semantic intent. Most importantly, the combination of semantics-based evaluation and layman-style reports provides the most robust and faithful assessment of generated report quality.

4.3 Improving Model Training with Layman's Terms: Insights from a Scaling Law

To evaluate the impact of training data style on semantic learning, we compare models trained on professional versus layman-style radiology reports using our semantics-based evaluation metric. As shown in Figure 2(b), the model trained on layman-style data demonstrates a clear positive scaling law: semantic performance steadily improves as the training set size increases from 5k to

50k. In contrast, the model trained on professional reports peaks at 10k samples and declines thereafter, suggesting that prolonged exposure to highly templated language leads to overfitting and reduced semantic generalization. Notably, the layman-style dataset starts to outperform professional reports when the training size reaches 50k.

To further assess semantic quality, we analyze the distribution of sentence-level similarity scores under the 50k training setting, as shown in Figure 2(a), with full statistics across all training scales provided in Appendix A.11. The layman-style model yields more sentence pairs with high similarity scores (e.g., >0.8), indicating stronger alignment with the reference semantics. In contrast, the professional model produces more outputs in the mid-to-low similarity range, reflecting weaker semantic fidelity.

To understand why the model trained on 10k professional reports achieves the highest semantic performance, we conduct further analysis and identify signs of representation collapse. Specifically, we compute the pairwise cosine similarity of generated reports on the test set. The 10k professional model exhibits an average cosine similarity of 0.893 with a variance of 0.008, suggesting that the model learns to mimic the dominant class (e.g., no findings or normal reports) to minimize loss, rather than capturing diverse semantic content. In contrast, the 10k layman-style model yields a lower average similarity of 0.802 with a higher variance of 0.012, reflecting greater report diversity and semantic richness. These findings, combined with the overfitting trend observed in Figure 2, support the conclusion that the layman-style dataset promotes a more robust and natural progression in semantic learning as the dataset scales—unlike the shortcut behavior observed with professional reports. Furthermore, we evaluate specialized clinical metrics and find that at the 10k scale, the layman-trained model outperforms its professional counterpart (CheXbert: 0.447 vs. 0.398; RadCliQ-v0: 0.413 vs. 0.405).

4.4 Evaluating Semantic Fidelity: Human vs. Automated Metrics

Due to the obscurity of professional radiology reports and the high cost of involving clinicians as annotators, few studies have explored the correlation between human scores and automated metrics such as BLEU in this domain. However, it is well documented in other fields that word-overlap-based metrics often fail to capture semantic accuracy and typically exhibit weak correlation with human evaluations. Therefore, relying solely on such metrics to assess the quality of generated radiology reports is inadequate. To enable a fair comparison between models trained on professional and layman-style reports, and to make professional reports more comprehensible to non-clinician human evaluators, we first translate all professional references into layman terms. We then recruit three human annotators—fluent English speakers with non-clinical backgrounds—to score the generated reports using a unified evaluation protocol: *"Given the generated text and the reference, calculate the proportion of sentences in the generated text that semantically match each sentence in the reference."* This protocol is consistently applied to evaluate both types of model outputs. After collecting scores from all annotators, we compute the final report score by averaging across annotators and across reference-matched sentences. The inter-annotator agreement (IAA), measured by Cohen's Kappa, is 0.63 for professional reports and 0.58 for layman-style reports, indicating fair to good agreement (0.4–0.75 range). Details about the annotators and scoring procedures are provided in Appendix A.12. The correlation results between human evaluations and automated metrics are presented in Table 4. Across the board, reports generated in layman terms show stronger alignment with human judgments. This holds not only for lexical metrics such as BLEU, ROUGE, and METEOR, but also for clinically relevant Clinical Efficacy (CE) metrics, including CheXbert-F1, RadGraph-F1, and RadCliQ. Although CE metrics are designed to assess named entity correctness in medical texts, we find them equally applicable to layman-style reports. Notably, the correlation between CE metrics and human scores is consistently higher for layman-style outputs, reinforcing their semantic fidelity and accessibility.

4.5 Case Study

Table 5 presents several sentence-level examples demonstrating how translating professional radiology terminology into layman's language can substantially improve clarity and patient understanding. For instance, the clinical term *pleural effusion* is rephrased as *extra fluid around the lungs*, offering a more intuitive explanation. Similarly, *bibasilar atelectasis*, which may be obscure or confusing to non-experts, becomes *collapsed lung areas*, conveying the concept in simpler terms. These examples highlight the value of plain language in

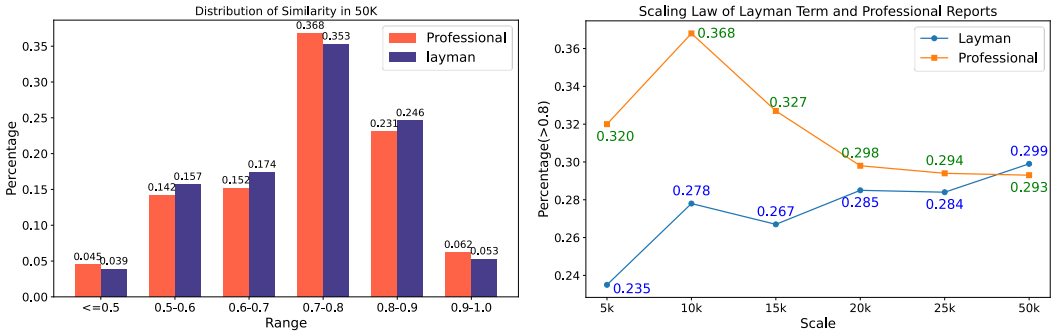


Figure 2: Scaling law of the model's semantic understanding by training on report-level datasets.

Correlation	Pearson		Spearman	
Type	raw	layman	raw	layman
B-1	0.533	0.534↑	0.536	0.524
B-2	0.526	0.573↑	0.532	0.538↑
B-3	0.480	0.557↑	0.502	0.519↑
B-4	0.420	0.519↑	0.450	0.472↑
R-1	0.543	0.586↑	0.550	0.565↑
R-2	0.430	0.524↑	0.441	0.485↑
R-L	0.526	0.561↑	0.532	0.534↑
Meteor	0.527	0.586↑	0.538	0.556↑
Semantics	0.559	0.601↑	0.558	0.576↑
Chexbert	0.570	0.600↑	0.620	0.703↑
Radgraph	0.521	0.652↑	0.536	0.658↑
RadCliQ-v0	0.616	0.710↑	0.633	0.724↑
RadCliQ-v1	0.613	0.719↑	0.630	0.728↑

Table 4: The correlation of automated metrics (BLEU, ROUGE and semantic scores) and human evaluators, for both professional reports and their layman's terms counterpart. Semantic scores are calculated with the proportion of semantic similarity over 0.8 among all sentences.

original	layman
Both lung fields are clear	Both lungs look healthy with no problems
No evidence of pleural effusion	There is no extra fluid around the lungs
The chest x-ray shows subtle patchy lateral left lower lobe opacities, which are most likely vascular structures and deemed stable with no definite new focal consolidation	The x-ray shows faint cloudy spots in the lower part of the left lung, likely blood vessels, and overall stable with no new clear lung infection
Overall impression suggests appropriate positioning of the tubes and bibasilar atelectasis, along with findings consistent with small bowel obstruction	The overall impression suggests proper placement of tubes and some collapsed lung areas, along with signs of small bowel obstruction
However, cephalization of engorged pulmonary vessels has probably improved	The congested blood vessels in the lungs have likely improved

Table 5: Examples from the sentence-level dataset.

enhancing communication and promoting patient comprehension in medical settings.

5 Conclusion

In this paper, we presented the Layman's RRG framework to jointly address the challenges of accessibility and robustness in radiology report generation. At the core of our framework are two high-quality layman-style datasets—at the sentence and report levels—constructed through a rigorous generation and self-refinement pipeline. These datasets serve as the foundation for both evaluation and training. Building on this, we introduced a semantics-based evaluation method that, when paired with our sentence-level dataset, mitigates the overestimated scores produced by traditional word-overlap metrics and more accurately captures the semantic quality of generated reports. Furthermore, we proposed a layman-guided training strategy utilizing the report-level dataset, which enhances the model's semantic understanding and exhibits a positive scaling behavior, where performance continues to improve as the training data grows. Collectively, these contributions provide a foundation for building radiology report generation systems that are not only semantically faithful, but also more accessible to patients and non-experts.

Ethics Statement

In this paper, we introduce a Layman RRG framework for radiology report generation and evaluation. The advantage of our framework is that it is better for models to enhance the understanding on the semantics, as well as provide a more robust evaluation framework. However, a potential downside is that some layman's terms may express inappropriate or offensive meanings because of the hallucination issues of LLMs. Therefore, it is crucial to carefully review the content of training datasets prior to training the layman models to mitigate this issue.

Limitations

Although our Layman RRG framework could provide a promising training process and provide a robust evaluation process, it has certain limitations. Primarily, as we utilized `GPT-4o` to translate the professional reports to layman's terms and proceed a strict modification process to improve the quality of translated layman's term, it may also include a few of professional reports that do not translate perfectly. In future work, we will focus more on continuing to improve the quality of translated reports.

References

Rebecca Anhang Price, Marc N Elliott, Alan M Zaslavsky, Ron D Hays, William G Lehrman, Lise Rybowski, Susan Edgman-Levitan, and Paul D Cleary. 2014. Examining the role of patient experience surveys in measuring health care quality. *Medical Care Research and Review*, 71(5):522–554.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *11th conference of the european chapter of the association for computational linguistics*, pages 249–256.

Francesco Dalla Serra, William Clackett, Hamish MacKinnon, Chaoyang Wang, Fani Deligianni, Jeff Dalton, and Alison Q O'Neil. 2022. Multimodal generation of radiology reports using knowledge-grounded extraction of entities and relations. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 615–624.

Jane Domingo, Galal Galal, Jonathan Huang, Priyanka Soni, Vladislav Mukhin, Camila Altman, Tom Bayer, Thomas Byrd, Stacey Caron, Patrick Creamer, et al. 2022. Preventing delayed and missed care by applying artificial intelligence to trigger radiology imaging follow-up. *NEJM Catalyst Innovations in Care Delivery*, 3(4):CAT–21.

Wenjun Hou, Kaishuai Xu, Yi Cheng, Wenjie Li, and Jiang Liu. 2023. Organ: observation-guided radiology report generation via tree reasoning. *arXiv preprint arXiv:2306.06466*.

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 QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *Preprint*, arXiv:2106.14463.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Kaveri Kale, Pushpak Bhattacharyya, and Kshitij Sharad Jadhav. 2023. Replace and report: Nlp assisted radiology report generation. *ArXiv*, abs/2306.17180.

Suhyeon Lee, Won Jun Kim, Jinho Chang, and Jong Chul Ye. 2023. Llm-cxr: Instruction-finetuned llm for cxr image understanding and generation. In *The Twelfth International Conference on Learning Representations*.

Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. 2019. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6666–6673.

Yaowei Li, Bang Yang, Xuxin Cheng, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. 2023. Unify, align and refine: Multi-level semantic alignment for radiology report generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2863–2874.

Chen Lin, Shuai Zheng, Zhizhe Liu, Youru Li, Zhenfeng Zhu, and Yao Zhao. 2022. Sgt: Scene graph-guided

transformer for surgical report generation. In *International conference on medical image computing and computer-assisted intervention*, pages 507–518. Springer.

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

Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. 2024. Bootstrapping large language models for radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18635–18643.

Chang Liu, Yuanhe Tian, and Yan Song. 2023. A systematic review of deep learning-based research on radiology report generation. *arXiv preprint arXiv:2311.14199*.

Pilar López-Úbeda, Teodoro Martín-Noguerol, Jorge Escartín, Alberto Cabrera-Zubizarreta, and Antonio Luna. 2024. Automated mri pituitary structured reporting from free-text using a fine-tuned llama model: a feasibility study. *Japanese Journal of Radiology*, pages 1–9.

Thusitha Mabotuwana, Christopher S Hall, Vadiraj Hombal, Prashanth Pai, Usha Nandini Raghavan, Shawn Regis, Brady McKee, Sandeep Dalal, Christoph Wald, and Martin L Gunn. 2019. Automated tracking of follow-up imaging recommendations. *American Journal of Roentgenology*, 212(6):1287–1294.

Thusitha Mabotuwana, Christopher S Hall, Joel Tieder, and Martin L Gunn. 2018. Improving quality of follow-up imaging recommendations in radiology. In *AMIA annual symposium proceedings*, volume 2017, page 1196.

Teresa Martin-Carreras, Tessa S Cook, and Charles E Kahn Jr. 2019. Readability of radiology reports: implications for patient-centered care. *Clinical imaging*, 54:116–120.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. *arXiv preprint arXiv:2402.09906*.

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.

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*, 2018:188.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew 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, Online. Association for Computational Linguistics.

Aaron Smith, Christian Hardmeier, and Jörg Tiedemann. 2016. Climbing mont bleu: the strange world of reachable high-bleu translations. In *Proceedings of the 19th annual conference of the European association for machine translation*, pages 269–281.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *International conference on intelligent text processing and computational linguistics*, pages 341–351. Springer.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*.

An Yan, Zexue He, Xing Lu, Jiang Du, Eric Chang, Amilcare Gentili, Julian McAuley, and Chun-Nan Hsu. 2021. Weakly supervised contrastive learning for chest x-ray report generation. *arXiv preprint arXiv:2109.12242*.

Benjamin Yan, Ruochen Liu, David E Kuo, Subathra Adithan, Eduardo Pontes Reis, Stephen Kwak, Vasantha Kumar Venugopal, Chloe P O'Connell, Agustina Saenz, Pranav Rajpurkar, et al. 2023. Style-aware radiology report generation with radgraph and few-shot prompting. *arXiv preprint arXiv:2310.17811*.

Bohao Yang, Kun Zhao, Chen Tang, Liang Zhan, and Chenghua Lin. 2024. Structured information matters: Incorporating abstract meaning representation into llms for improved open-domain dialogue evaluation. *arXiv preprint arXiv:2404.01129*.

Kun Zhao, Bohao Yang, Chenghua Lin, Wenge Rong, Aline Villavicencio, and Xiaohui Cui. 2023. Evaluating open-domain dialogues in latent space with next sentence prediction and mutual information. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 562–574.

Kun Zhao, Bohao Yang, Chen Tang, Chenghua Lin, and Liang Zhan. 2024. SLIDE: A framework integrating small and large language models for open-domain dialogues evaluation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15421–15435, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

A Appendix

A.1 Prompt for Translation

Given a series of sentences that are split from radiology reports.

Sentences:
{placeholder for 50 sentences}

Please finish the following tasks.
Tasks:
1. Translation: Please translate each sentence into plain language that is easy to understand. You must translate all the sentences.

For each task, return a dict. Here are some examples:
Task 1:
“‘*json*
{
"0": "No signs of infection, fluid, or air outside of the lung—everything looks normal.",
"1": "The unclear spots seen in both lungs are most likely just shadows from nipples.",
...
}
“‘

A.2 Prompt for Refinement

Given a series of Original sentences that are split from radiology reports and their translated layman's terms sentence.

Original Sentences:
{placeholder for 50 sentences}

Translated Layman's Term:
{placeholder for 50 sentences}

Please finish the following tasks.
Tasks:
1. Check and Modification: Please check if the translated sentence is semantically consistent and has the same detailed description as the given original sentence. If it is, make no changes; otherwise, make modifications.

For each task, return a dict. Here are some examples:
Task 1:
“‘*json*
{
"0": "No signs of infection, fluid, or air outside of the lung—everything looks normal.",
"1": "The unclear spots seen in both lungs are most likely just shadows from nipples.",
...
}
“‘

Algorithm 1 Dataset Generation and Refinement

Require: A set of n data items $D = \{d_1, d_2, \ldots, d_n\}$, a threshold θ for semantic similarity
Ensure: Translated set $T = \{t_1, t_2, \ldots, t_n\}$ where each t_i is a valid translation of d_i

```
for i = 1 to n do
    repeat
        t_i ← LLM-Translate(d_i)
        sim ← Semantic-Similarity(d_i, t_i)
        correct ← LLM-Check-Translation(d_i, t_i)
    until sim ≥ θ and correct
end for
return T
```

A.3 Dataset

In this part, we outline the statistics of our datasets as follows in the Table 6.

Datasets	Sentence-level	Report-Level
# Numbers	50000	50000
Avg. # Words per sample	28.68	101.45
Avg. # Sentences per sample	1	5.05

Table 6: Data statistics of the sentence-level and report-level dataset.

A.4 Dataset Generation and Refinement Algorithm

The Dataset Generation and Refinement Algorithm is shown as Algorithm 1.

A.5 Candidate Report Evaluation using GRITLM and Layman Term Replacement

The Candidate Report Evaluation using GRITLM and Layman Term Replacement is shown as Algorithm 2.

A.6 Refinement Rate

In this section, we examine a subset of 100 samples to analyze the refinement process, observing both the accuracy proportion at each stage and the sentence modification rate per step. As illustrated in Figure 3, the refinement process concludes after three iterations.

Algorithm 2 Candidate Report Evaluation using GRITLM and Layman Term Replacement

Require: Candidate report C, Reference report R, Sentence-level dataset S, Semantic similarity threshold $\theta = 0.8$
Ensure: Proportion of sentences in C and R with semantic similarity $\geq \theta$ after replacement, BLEU, ROUGE, and Meteor scores

$C_s \leftarrow$ Split-Sentences(C)
$R_s \leftarrow$ Split-Sentences(R)
for each sentence $c_i \in C_s$ **do**
  $max\_sim \leftarrow 0$
  for each sentence $s_j \in S$ **do**
    $sim \leftarrow$ GRITLM-Similarity(c_i, s_j)
    if $sim > max\_sim$ **then**
      $max\_sim \leftarrow$ sim
      $replacement \leftarrow$ Layman-Term(s_j)
    end if
  end for
  $c_i \leftarrow replacement$
end for
for each sentence $r_i \in R_s$ **do**
  $max\_sim \leftarrow 0$
  for each sentence $s_j \in S$ **do**
    $sim \leftarrow$ GRITLM-Similarity(r_i, s_j)
    if $sim > max\_sim$ **then**
      $max\_sim \leftarrow$ sim
      $replacement \leftarrow$ Layman-Term(s_j)
    end if
  end for
  $r_i \leftarrow replacement$
end for
$similar\_count \leftarrow 0$
for each sentence $c_i \in C_s$ **do**
  for each sentence $r_i \in R_s$ **do**
    $sim \leftarrow$ GRITLM-Similarity(c_i, r_i)
    if $sim \geq \theta$ **then**
      $similar\_count \leftarrow similar\_count + 1$
      break
    end if
  end for
end for
$proportion \leftarrow \frac{similar\_count}{|C_s|}$
$BLEU \leftarrow$ Compute-BLEU(C_s, R_s)
$ROUGE \leftarrow$ Compute-ROUGE(C_s, R_s)
$Meteor \leftarrow$ Compute-Meteor(C_s, R_s)
return $proportion, BLEU, ROUGE, Meteor$

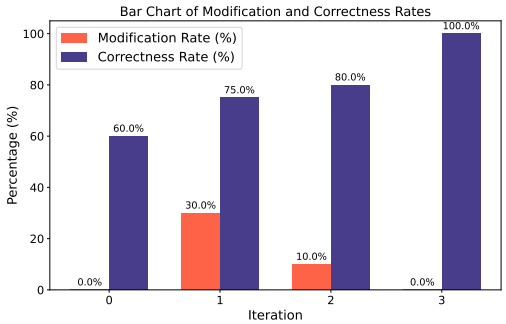


Figure 3: Rate of Refinement as Iterations Increase

A.7 Analysis of Refinement Step

As mentioned in the early parts, our data generation pipeline leverages a rigorous refinement process. This includes a LLM self-refinement module and an embedding model to assess semantic similarity.

Here, we present an example going through 4 steps in the refinement process. As detailed in Table 7, the example includes the translated report at each step and the calculation of semantic similarity between each sentence in the original professional report and the corresponding sentence in layman's terms. Step 0 is the raw professional report that requires translation, and Steps 1-3 present the reports translated to layman's terms. The red numbers display the semantic similarity. It is evident that the semantic similarity increases in each step and remains unchanged at the third step, signifying the conclusion of the refinement process. This analysis demonstrates that the refinement process effectively enhances the quality of the translated layman's reports.

Step	Report
0	Subtle rounded nodular opacity projecting over both lung bases which could represent nipple shadows. Recommend repeat with nipple markers to confirm and exclude underlying pulmonary nodule. Subtle bibasilar opacities likely represent atelectasis or aspiration. No evidence of pneumonia.
1	There are some unclear spots in the lower parts of both lungs which might just be shadows caused by nipples (0.776). We recommend doing another x-ray using nipple markers to be sure (0.731). There are also subtle changes in the lower lungs likely due to collapsed lung areas or inhaled food/liquid (0.704). No signs of pneumonia (0.971).
2	The unclear spots seen in both lung bases are most likely just shadows from nipples (0.778↑). We recommend a repeat x-ray with nipple markers to confirm and exclude any underlying lung nodules (0.911↑). There are also subtle changes in the lower lungs likely due to collapsed lung areas or inhalation of food/liquid (0.712↑). No evidence of pneumonia (0.999↑).
3	The unclear spots seen in both lung bases are most likely just shadows from nipples. We recommend a repeat x-ray with nipple markers to confirm and exclude any underlying lung nodules. There are also subtle changes in the lower lungs likely due to collapsed lung areas or inhalation of food/liquid. No evidence of pneumonia. (Refinement ends)

Table 7: The expression of an example going through the refinement process.

A.8 Instruction Tuning

We further ran an initial experiment for the new application, by concatenating the 50k professional dataset and the 50k layman's dataset, yielding a 100k two-class instruction tuning training set. We hypothesize that seeing both versions with different wordings would encourage the model to pick up the semantic overlaps between the two datasets.

Training set	Similarity >0.8
professional 50k	0.293
layman 50k	0.299
professional + layman 100k	0.323

Table 8: Comparison of Similarity Scores Between Mixed and Single Datasets

For the two datasets, we prepend their corresponding instruction to the example: "Given this X-ray image, generate a professional radiology report.", "Given this X-ray image, generate a radiology report in layman's terms." and in inference, we prepend the same instructions based on our need. The experiments took 5 days on 4 A6000 GPUs.

In Table 8, we reported the model performance on three settings: 1) trained professional & inference professional 2) trained layman & inference layman 3) trained both & inference professional. We show the percentage of generated reports that have over 0.8 cosine similarity with the groundtruth reports for each setting, aligning with the setting in Figure 3 (right) in the paper.

As shown in the results, the instruction-tuned model, when exposed to both professional and layman reports in the training, can generate a higher percentage of professional reports that are more semantically aligned with the groundtruth. This has indicated that the model is able to pick up semantic hints from the layman's dataset in the training to enhance its professional report generation. More importantly, this new unified model can generate both professional and layman's reports when provided with the instructions.

A.9 Case Study

In this section, we provide more examples from sentence-level dataset and report-level dataset. The Table 9 include some examples in the sentence-level dataset and Table 10 present samples selected from the report-level dataset.

A.10 Additional Experiments

We also tested the LLM-based approach using two different open-access ChatGPTs[4] in both MIMIC CXR and PadChest (English translated) datasets, denoted as LLM1 and LLM2, respectively.

[4]Kimi (www.moonshot.cn) and DeepSeek (www.deepseek.com/)

raw	layman
Both lung fields are clear	Both lungs look healthy with no problems
No evidence of pleural effusion	There is no extra fluid around the lungs
The chest x-ray shows subtle patchy lateral left lower lobe opacities, which are most likely vascular structures and deemed stable with no definite new focal consolidation	The x-ray shows faint cloudy spots in the lower part of the left lung, likely blood vessels, and overall stable with no new clear lung infection
The impression states that the opacities are bilateral and indicative of an infection that requires follow up attention to ensure resolution	The impression notes the cloudy spots are in both lungs, likely indicating an infection that needs follow-up to ensure it's resolved
Overall impression suggests appropriate positioning of the tubes and bibasilar atelectasis, along with findings consistent with small bowel obstruction	The overall impression suggests proper placement of tubes and some collapsed lung areas, along with signs of small bowel obstruction
A mildly displaced fracture of the right anterior sixth rib and possible additional right anterior seventh rib fracture are noted	There is a slightly displaced fracture of the right front sixth rib and possibly another right front seventh rib fracture
There is increased soft tissue density at the left hilum and a fiducial seed is seen in an unchanged position	Increased tissue density is seen at the left lung root and a tracking marker is in the same place as before
However, cephalization of engorged pulmonary vessels has probably improved	The congested blood vessels in the lungs have likely improved
Moderate bilateral layering pleural effusions are also present along with a notable compression deformity of a lower thoracic vertebral body, without information about the age of the patient	Moderate fluid in both pleura is seen along with a compression deformity in a lower chest spine bone, without age information on the patient
The chest x-ray image reveals worsening diffuse alveolar consolidations with air bronchograms, particularly in the right apex and entire left lung	The x-ray shows worsening of diffuse lung cloudiness with air-filled bronchial tubes, especially in the right lung apex and the entire left lung

Table 9: Some examples of sentence-level dataset.

`Baseline` approach in MIMIC CXR dataset indicates the layman reports which using prompts provided in A.1. `(Original)` approach in MIMIC CXR and PadChest indicate the original radiology reports. We also reported their readability scores. Apart from the baseline prompt (denoted as `P1`), a instruction-following prompt (denoted as `P2`) is designed for GPT to generate layman report by examples provided. An example is shown in Fig. 4.

The evaluation metrics are in three types: i) Clinical accuracy, ii) Relevance, and iii) Readability. For Readability, a set of text statistics metrics[5] to be used. Their abbreviation and the corresponding metrics are listed below:

- `Easy`: The Flesch Reading Ease formula

[5]The open-source Python library is provided on pypi.org/project/textstat

raw	layman
Bilateral nodular opacities, which most likely represent nipple shadows, are observed. There is no focal consolidation, pleural effusion, or pneumothorax. Cardiomediastinal silhouette is normal, and there is no acute cardiopulmonary process. Clips project over the left lung, potentially within the breast, and the imaged upper abdomen is unremarkable. Chronic deformity of the posterior left sixth and seventh ribs is noted.	There are spots seen in both lungs that are likely just nipple shadows. There is no evidence of a specific infection, fluid in the lungs, or air outside the lungs. The shape of the heart and area around it looks normal. There are no immediate heart or lung issues. There are surgical clips in the area of the left lung, likely in the breast, and the upper abdomen appears normal. There is a long-term deformity of the sixth and seventh ribs on the left side.
The chest x-ray shows normal cardiac, mediastinal, and hilar contours with clear lungs and normal pulmonary vasculature. No pleural effusion or pneumothorax is present. However, multiple clips are seen projecting over the left breast, and remote left-sided rib fractures are also demonstrated. The impression is that there is no acute cardiopulmonary abnormality detected.	The chest x-ray shows a normal heart shape and clear lungs with no fluid or air outside the lungs. There are multiple surgical clips seen in the left breast area, and old rib fractures on the left side. There are no immediate heart or lung problems detected.
The chest x-ray shows no evidence of focal consolidation, effusion, or pneumothorax, and the cardiomediastinal silhouette is normal. Multiple clips projecting over the left breast and remote left-sided rib fractures are noted. No free air below the right hemidiaphragm is seen. The impression is that there is no acute intrathoracic process.	The chest x-ray does not show any specific lung infection, fluid, or air outside the lungs. The heart and surrounding area appear normal. Multiple surgical clips are seen in the left breast area, and old rib fractures on the left side are noted. There is no free air under the right side of the diaphragm. There are no immediate issues inside the chest.

Table 10: Some examples of report-level dataset.

- `M1`: The Flesch-Kincaid Grade Level
- `M2`: The Fog Scale (Gunning FOG Formula)
- `M3`: The SMOG Index
- `M4`: Automated Readability Index
- `M5`: The Coleman-Liau Index
- `M6`: Linsear Write Formula
- `M7`: Dale-Chall Readability Score
- `M8`: Spache Readability Formula
- `M9`: McAlpine EFLAW Readability Score

The experimental results are provided in Table 11 and Table 12.

A.11 Scaling Law

As illustrated in Figure 5, the training dataset scales are 5k, 10k, 15k, and 20k from top to bottom, respectively. We use the trained models to generate reports and calculate the semantic similarity between the generated reports and reference reports. The figures on the left represent models trained by layman's terms, while the plots on the right represent those trained using raw professional reports.

A.12 Details of Human Annotators

Institutional Review Board (IRB). Our work does not require IRB approval as it only involves semantic assessment. Our evaluation compares the semantic consistency between paragraph pairs, where the ground truth is sourced from a public dataset available on GitHub. As our task focuses solely on semantic consistency without involving any X-ray images in the evaluation process, it can be considered a common text generation task.

Human Annotators We would like to highlight the nature of the human evaluation of this work as the assessment of semantic alignment, which makes the task fall back to the evaluation of a regular text generation task. This process is without involvement of any medical images. So we recruit human annotators from linguistic students and medical PhD students, who are professional in English reading and understanding. In addition, all of them have the right to access the MIMIC-CXR dataset.

Data	Model	Clinical Accuracy						Relevance			
		Chexbert-F1			RadGraph-F1			B.	M.	R.	Sem.
		Acc	Micro	Macro	R1	R2	R3				
MIMIC CXR	Baseline	0.737	0.576	0.076	0.026	0.023	0.016	0.073	0.299	0.337	0.577
	LLM1+P1	0.771	0.602	0.086	0.012	0.010	0.007	0.085	0.366	**0.348**	0.587
	LLM1+P2	**0.846**	**0.776**	**0.138**	**0.028**	**0.024**	**0.017**	**0.087**	**0.384**	0.347	**0.758**
PadChest	LLM1+P1	0.918	0.655	0.060	0.058	0.039	0.030	**0.068**	**0.436**	**0.251**	0.685
	LLM1+P2	**0.940**	**0.748**	**0.075**	**0.065**	**0.041**	**0.029**	0.065	0.421	0.244	**0.778**
	LLM2+P1	**0.945**	**0.746**	**0.074**	0.095	0.073	0.061	0.084	0.389	0.267	0.778
	LLM2+P2	0.937	0.736	0.073	**0.153**	**0.134**	**0.122**	**0.188**	**0.497**	**0.373**	**0.792**

Table 11: Clinical Accuracy and Relevance of Layman-style reports on MIMIC-CXR and PadChest Dataset. Baseline, LLM1+P1 and LLM1+P2 indicate layman-style reports generated by different LLMs and different prompts.

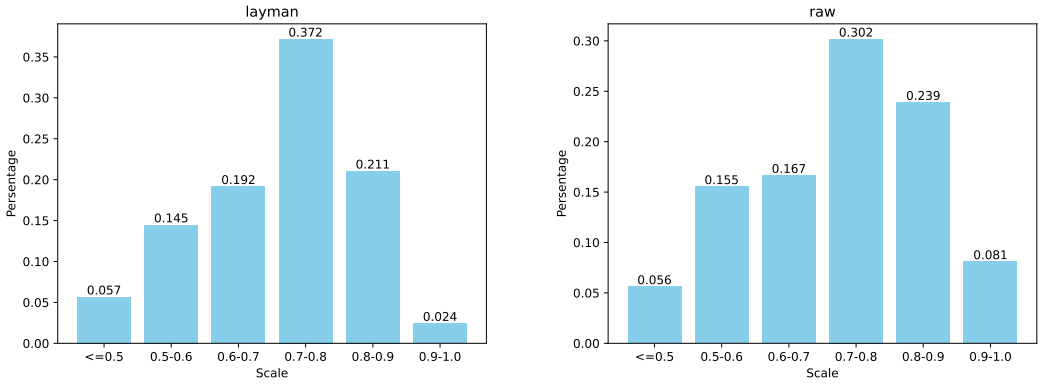
Prompting GPT to Generate Layman Report of Radiology Image Reports

```
message = [ ]

introduction = """You are a writer of science journalism.

Given a radiology reports, please finish the following tasks.
Tasks: 1. Translation: Please translate each report into plain language that is easy to understand (layman's terms). The layman-translated report requires writing factual descriptions, while also paraphrasing complex scientific concepts using a language that is accessible to the general public. Meanwhile, it preserve the details as much as possible. Each translated sentence must correspond to the original sentence. For example, a 4-sentence report should be translated into a 4-sentence layman's termed report. You must translate all the reports.

Here are some examples of layman-version reports:
"""

query = """Report to be translated:\n"""

for example in example_of_layman_reports:
   introduction.append(example)

messages.append({"role":"system", "content": introduction})

for report in radiology_reports:
   messages.append({"role":"user", "content": query})
```

Figure 4: Example of prompting GPT to generate the layman report of the radiology image reports.

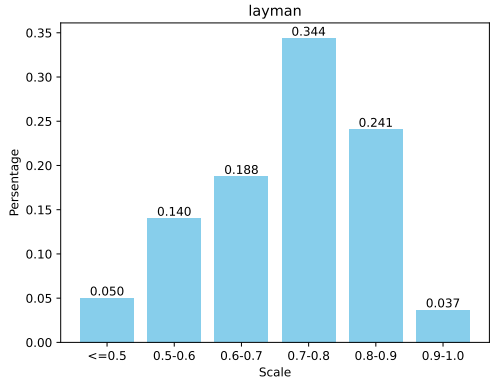
Data	Model	Easy Level↑	Level of Grade Required for Reading↓								
			M1	M2	M3	M4	M5	M6	M7	M8	M9
MIMIC CXR	(Original)	43	9	11	11	11	14	5	11	5	11
	Baseline	76	6	8	8	8	9	7	10	5	**19**
	LLM1+P1	84	**5**	8	8	7	**7**	7	**8**	**4**	21
	LLM1+P2	**85**	**5**	**7**	**7**	**6**	**7**	**6**	**8**	**4**	**19**
PadChest	(Original)	26	12	14	4	14	16	5	14	6	10
	LLM1+P1	69	7	9	4	**8**	9	**7**	**9**	5	19
	LLM1+P2	**73**	**6**	**8**	**3**	**8**	**8**	**7**	**9**	**4**	**18**
	LLM2+P1	**68**	**8**	**9**	4	**9**	**10**	8	**10**	**5**	21
	LLM2+P2	64	**8**	10	**3**	**9**	**10**	**7**	11	**5**	**18**

Table 12: Readability of Layman-Style Reports. Original represents professional reports. Baseline, LLM1+P1 and LLM1+P2 indicate layman-style reports generated by different LLMs and different prompts.

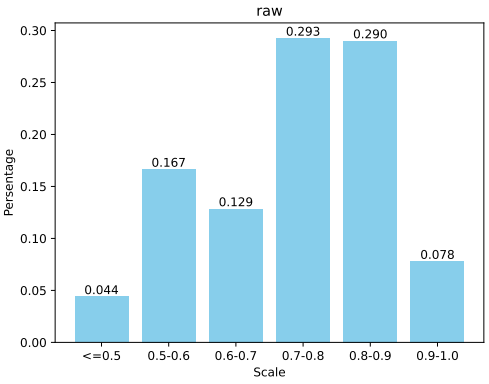


(a)

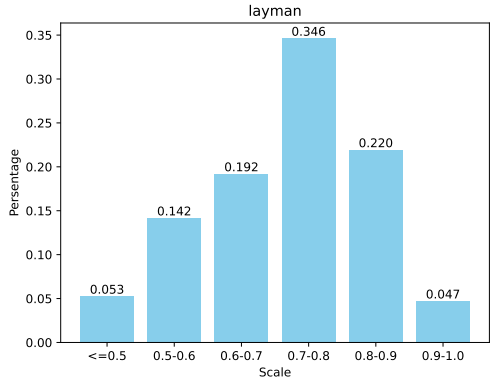
(b)



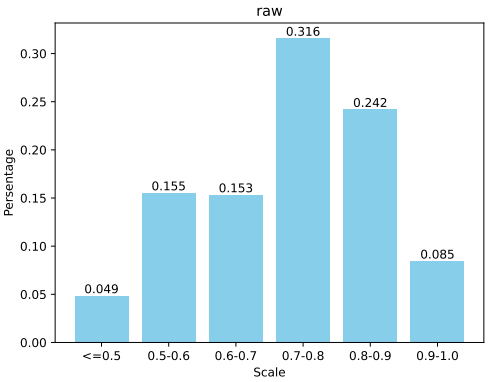
(c)



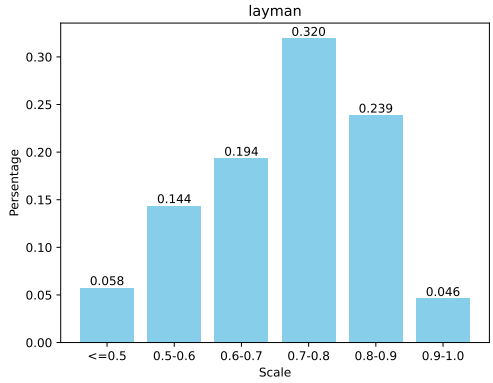
(d)



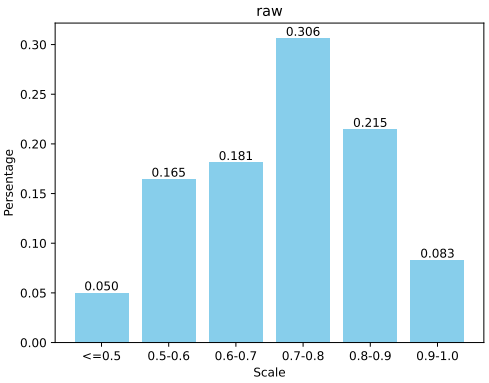
(e)



(f)



(g)



(h)

Figure 5: Scaling law of model's semantic understanding training using report-level datasets. From up to down shows the trend for models trained by 5k, 10k, 15k and 20k respectively.