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

Anonymous ACL submission

Abstract

Radiology Report Generation (RRG) has advanced considerably with the development of multimodal generative models. Despite the progress, the field still faces significant challenges in evaluation, as existing metrics lack robustness and fairness. We reveal that, RRG with high performance on existing lexical-based metrics (e.g. BLEU) might be more of a mirage - a model can get a high BLEU only by learning the template of reports. This has become a pressing issue for RRG due to the highly patternized nature of these reports. In addition, standard radiology reports are often highly technical. Helping patients understand these reports is crucial from a patient's perspective, yet this has been largely overlooked in previous work. In this work, we un-intuitively approach these problems by proposing the **Layman's RRG** framework that can systematically improve RRG with day-to-day language. Specifically, our framework first contributes a translated *Layman's terms dataset*. Building upon the dataset, we then propose a *semantics-based evaluation* method, which is effective in mitigating the inflated numbers of BLEU and provides more robust evaluation. We show that *training* on the layman's terms dataset encourages models to focus on the semantics of the reports, as opposed to overfitting to learning the report templates. Last, we reveal a promising scaling law between the number of training examples and semantics gain provided by our dataset, compared to the inverse pattern brought by the original formats.

1 Introduction

With the advancement of generative models, image captioning has made significant progress, enabling accurate generation of text descriptions based on images. This capability has been effectively 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 produce textual descriptions of medical images, such as X-rays, for alleviating the burden on radiologists and enhancing the quality and standardization of healthcare. However, standard radiology reports are often highly technical and difficult to comprehend. Helping patients understand these reports is crucial from a patient's perspective, yet this has been largely overlooked in previous work.

Despite the proliferation of methods for RRG and the improvement in the quality of generated reports, the field suffers from a lack of robust evaluation metrics. Traditionally, BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) metrics have been widely used to evaluate generated reports[1]. These metrics focus on calculating the word overlap rate between the ground truth and the generated reports. However, they are limited to surface-level evaluation and are not sensitive to small perturbations of the text that significantly change the semantics (Stent et al., 2005; Callison-Burch et al., 2006; Smith et al., 2016). Taking the example of negation words, the sentences "*there is a focal consolidation*" and "*there is no focal consolidation*" would receive a high BLEU score due to their high word overlap, despite conveying opposite meanings.

Moreover, a distinctive feature of radiology reports is their highly patterned nature (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). Physicians typically follow specific sentence structures related to diseases and organs, and adhere to certain templates when constructing reports. This pattern creates a pressing need for evaluation metrics beyond word overlap, as strictly adhering to these templates can yield high scores on lexical-based evaluations, even when the

[1] In recent surveys, lexical-based methods are still predominantly used to compare performance across methods as the main measures in the field (Liu et al., 2023).

report quality is substandard. For example, Kale et al. (2023) used a standard template and replaced some sentences to generate reports, achieving high scores on BLEU and other metrics based on word-overlapping. Equally importantly, because of its highly patterned nature, we hypothesize that training on such datasets would lead the model to pay more attention to its structure as opposed to its semantics, thus generating reports which have a low semantic similarity with the groundtruths. In addition, most of the radiology reports are highly professional and thus hard for people to understand it.

In this paper, we propose a new framework to approach the problem of Radiology Report Generation, facilitated by the use of layman's terms. Our framework includes 1) two high-quality **Layman's Terms datasets**; 2) **a semantics-based evaluation framework** based on layman's terms, which can provide fairer evaluation and mitigate the inflated numbers of BLEU; 3) **a training framework** based on layman's terms, which is shown to enhance the model's learning on semantics, as opposed to overfitting to the patterned templates of raw reports.

Our Layman's RRG framework first introduces two datasets: a sentence-level dataset and a report-level dataset. Both dataests leverage the advancements in generative models and a rigorous self-refinement system based on semantic similarity and LLM checking. The sentence-level dataset is used to facilitate the design of a fairer evaluation framework, while the report-level dataset replaces raw reports, enabling the training process to better focus on semantics. To validate the efficacy of Layman's RRG framework, we have designed a series of experiments and analyses. The experimental results show that combining our sentence-level Layman's terms dataset and semantics-based method yields a significantly more robust evaluation. Furthermore, we demonstrate that training with our report-level Layman's dataset enhances the model's semantic understanding, demonstrating a promising scaling law for the model's performance as the number of training examples increases.

In summary, our contribution are as follows:

- We introduce two high-quality layman's terms radiology report generation datasets, including a sentence-level dataset and a report-level dataset. To the best of our knowledge, this is the first attempt to create patient-friendly report datasets, which can be used for future research on generating patient-friendly reports.
- We design a Layman's terms-based RRG evaluation method, facilitated by leveraging LLM-based embedding models to replace each sentence in the professional RRG reports with our sentence-level dataset. The method is proved to provide fairer and more robust evaluation on lexical-based metrics and our proposed semantics-based metric.
- We prove that training with our report-level Layman's dataset leads to model's enhanced understanding on the semantics, providing a promising scaling law of model's performance with increased training examples, as opposed to the inverse effect brought by the professional reports.

2 Related work

2.1 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), 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.

2.2 Representation Learning

With the built-in vulnerability of lexical-based evaluation, we seek semantics-based methods.

Representation learning concerns the optimal way to numerically represent semantics of texts in the vector space, enabling their relative relationship to be reflected by similarity search. In recent years, contextualized text embedding models (Reimers and Gurevych, 2019) have established the capabilities to represent semantic textual similarity, topical alignment (Cer et al., 2017; Thakur et al., 2021). Recently, the field has grown an emergent consensus on training LLM-based embedding models, and position this as an alignment with an alignment with their generative abilities (Xiao et al., 2024; Muennighoff et al., 2024), showing exceptional abilities on reaching reasoning-level language understanding abilities. In this work, we opt for leveraging dense representational models to both of our semantics-based evaluation methods, and the deduplication procedures.

2.3 Style Transfer

Style Transfer is the task of transforming input content into a different style. Researchers have investigated this task across various modalities, including images and videos (Kim et al., 2022; Jing et al., 2019), as well as music (Cífka et al., 2020). Some of the developed systems have already been applied in industrial solutions. Text Style Transfer (TST) operates on the same principle as other modalities: rewriting textual input with a different attribute while minimizing information loss.

For the Expertise Style Transfer task, Cao et al. (2020) evaluated models from the three macro-categories of TST. Lyu et al. (2023) utilized ChatGPT to translate radiology reports into plain language and found that ChatGPT achieved favorable results, as confirmed by professional radiologists.

3 Radiology Report Generation in Layman's Term

In this section, we formally introduce **Layman's RRG**, a {dataset, evaluation, training} framework that systematically solve the inherent weakness of lexical-based evaluation and patterned nature of professional radiology report, as shown in the Figure 1.

Our framework is facilitated by a sentence-level dataset and a report-level dataset. The sentence-level dataset supports the design of our evaluation framework. The report-level dataset facilitates our training framework.

3.1 Dataset Creation

The process of our generation process is outlined in Algorithm 1, which involves generation and refinement to ensure optimal quality. The refinement system consists of semantics-checking module built upon embedding models, and a correctness self-checking module utilizing the same LLM used for generation. This algorithm is used to generate both the sentence-level dataset and report-level Dataset. We utilize the MIMIC-CXR dataset to create following datasets.

3.1.1 Sentence-level Dataset

The creation of dataset comprises 4 parts: deduplication, translation, refinement and checking.

Deduplication. We first use NLTK to separate each report into sentences. Through analyzing large amounts of reports, we found that there exist many repetitive sentences which have similar semantics. In order to simplify the final dataset and reduce the burden of pairwise similarity computation, we apply extensive deduplication on 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 Massive Text Embedding Benchark (MTEB) and Reasoning as Retrieval Benchmark (RAR-b), to encode sentences and acquire representations. We iteratively calculate pairwise cosine similarity between sentences and maintain sentences that do not have cosine similarity higher than 0.8 with the rest of the sentences, while dropping the ones that do. Through the deduplication procedures, the number of sentences is decreased from 490000 to 50000 approximately, largely reducing the cost and improving the efficiency for the following process.

Translation. After attaining the simplified dataset, we utilize `GPT-4o` to translate them. The design of prompt is detailed in Appendix A.1, where we detail the objective, and combine batch processing, and instruct the model to return the results in JSON format, which largely reduces the cost and allows the model to provide more consistent output through referencing in-batch examples. The translated results will be refined rigorously in next step.

Refinement In order to further improve the quality of translated sentences, we use a two-module self-refine method to improve it. Specifically, for each professional sentence and its layman's term counterpart, we combine the results of `GPT-4o` self-checking and leverage semantic similarity from GritLM to ensure the quality of translated sentence. Our method requires the translated sentence to meet both standards in order to pass the quality assessment, otherwise another round of translation will be prompted, until meeting both criteria. The design

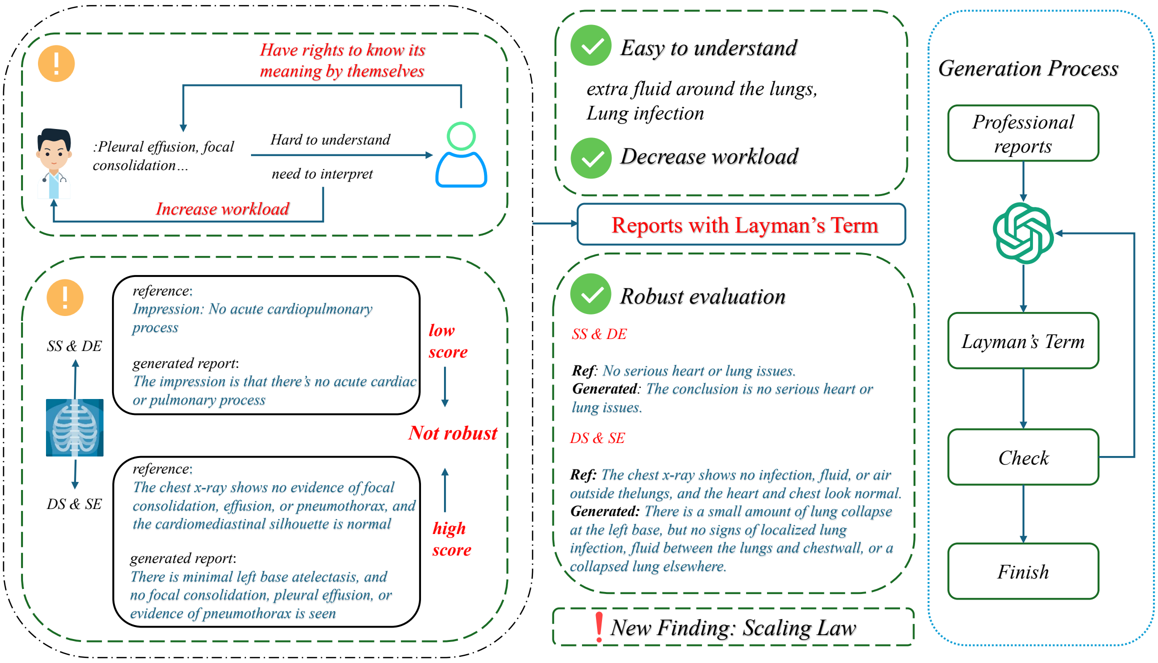


Figure 1: The Layman's RRG Framework. The "DS & SE" denotes different semantics and similar expressions. The "SS & DE" denotes similar semantics and different expressions.

of self-checking prompt is defined in Appendix A.2.

Human Checking After the refinement process, the quality of the datasets has achieved significant improvements. We refer to Appendix A.6 for how correction rates improve across self-refinement iteration. Further, we randomly select 500 sentence pairs and ask human annotators to check whether these pairs are matched. The final accuracy has exceeded 98%. The detailed algorithm of Dataset Generation and Refinement Algorithm is in Appendix A.4.

3.1.2 Report-level Dataset

In parallel, We have also created a report-level dataset, which is used to train the model. Our hypothesis is: because of the patternized nature of the professional reports, models are prone to overfit to the templates of the reports instead of their semantics. On the other hand, the diverse wordings in reports with layman's terms prevent the models to learn any templates, but in turn enforce them to learn to attend the aspects essential to the diagnosis. We also utilized Algorithm 1 to create this dataset. We also did some additional experiments using different LLMs and different datasets, which is shown as Appendix A.12.

3.2 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 the generated reports. This discrepancy arises due to the presence of semantically similar sentences with different wordings and semantically different sentences with high word 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 different meanings but achieve a BLEU-1 score greater than 0.6. This highlights that despite differing in pathology, sentences can receive a high BLEU score. Conversely, the sentences *"Impression: No acute cardiopulmonary process"* and *"The impression is that there's no acute cardiac or pulmonary process"* have identical meanings but receive a low BLEU-1 score.

We categorize these discrepancies into two types: expression difference issues and semantics difference issues. The expression difference issue arises when the reference sentence and the candidate sentence have similar semantics but low word overlap. The semantics difference issue occurs when the reference sentence and the candidate sentence have different semantics but high word overlap. These issues collectively result in misleading BLEU scores, as illustrated in Table 1.

Candidate	Reference	Candidate layman term	Reference layman term
Examples of DS & SE			
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 no evidence of focal consolidation, effusion, or pneumothorax, and the cardiomediastinal silhouette is normal	There is minimal left base atelectasis, and no focal consolidation, pleural effusion, or evidence of pneumothorax is seen	The chest x-ray shows no infection, fluid, or air outside the lungs, and the heart and chest look normal.	There is a small amount of lung collapse at the left base, but no signs of localized lung infection, fluid between the lungs and chest wall, or a collapsed lung elsewhere.
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.
The heart size is mildly enlarged	Heart size appears at least mild to moderately enlarged	The heart is slightly larger than normal.	The heart is slightly larger than normal.

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.

Moreover, the specialized nature of radiology reports makes it challenging to find human annotators for evaluation. Therefore, calculating the correlation between automatic metrics and human scores is crucial for assessing the reliability of automatic evaluation metrics.

To address these issues, we propose a novel evaluation method for assessing generated reports. The evaluation process is detailed in Appendix A.5. In summary, it takes a candidate report and a reference report, splits them into sentences, and replaces each sentence with the most semantically similar sentence from a predefined sentence-level dataset. The semantic similarity is measured using GritLM, and sentences with a similarity score higher than a specified threshold are counted. The proportion of sentences in the candidate and reference reports that meet this similarity threshold are calculated, alongside traditional metrics such as BLEU, ROUGE, and METEOR scores. Our evaluation framework aims to provide a more accurate evaluation of the generated reports by addressing the limitations of word-overlap metrics and templated RRG reports.

3.3 Training

In parallel to the evaluation framework, we extensively showcase that training on our layman's terms dataset provides significant benefits to models.

We depict the scaling law between the number of training examples and model's semantic gain, which is brought exclusively by our layman's terms dataset, while the opposite effect is brought by the raw dataset format. This discrepancy highlights that training on the original dataset longer brings overfitting, and is detrimental to models, while our dataset shows continual improvements.

We use `MiniGPT4` as the base model. We split each dataset (raw and layman's) into 5k, 10k, 15k, 20k, 25k and 50k to train the model respectively. The models are trained with 10 epochs, a batch size of 50 (achieved with gradient accumulation), on A6000 GPUs.

3.4 On the limitations of lexical-based evaluation

In this section, we reveal the behavioural difference between lexical-based evaluation metric and

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 2: 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.

semantics-based evaluation metric.

In order to verify the effectiveness of layman's term report which could resolve expression difference issues and semantics difference issues, we construct two subsets: Similar Semantics (SS) & Different Expressions (DE), Different Semantics (DS) & Similar Expressions (SE). The perception of lexical-based and semantics-based metric towards the two subsets characterizes their robustness.

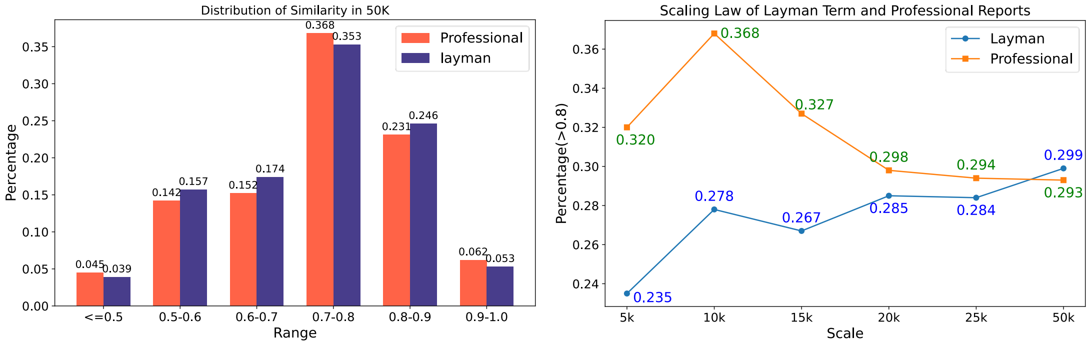
For both raw professional reports and their layman's term's counterpart, we calculate the BLEU score, ROUGE score and Meteor score as well as semantic similarity between candidates and references in the "SS & DE" and "DS & SE" subsets. The result is shown in the Table 2. From this table, for "DS & SE", pairs in the professional subset are wrongly perceived to have a high score by lexical metrics, e.g., 0.644 (BLEU-1), 0.505 (BLEU-2), 0.393 (BLEU-3) and 0.312 (BLEU-4); while its translated layman's terms significantly mitigate this mirage (0.312, 0.116, 0.064 and 0.042 respectively). Moreover, our semantics-based metric is able to reveal their non-semantically similar nature, characterized by the fact that > 0.8 pairs are respectively 2% and 1%. By contrast, for the "SS & DE" subset, the most ideal scenario would be having a metric that is robust to expression differences, and perceive the pairs to have a high performance. We could see that lexical-based metrics fail to reflect this pattern, giving raw professional report pairs significantly low scores. Our translated layman's pairs counteract the weakness of lexical metrics, providing a higher perceived score. However, the combination of our layman's dataset and semantics-based metric provide the strongest robustness, showing both a high semantic score (over 50% >0.8 pairs) and a small perceptual difference between raw pairs and layman's pairs.

In summary, despite the inherent weakness of lexical-based metrics, they are even more vulnerable to the heavily patterned raw professional reports, largely fail to reflect the actual relational semantics of pairs (higher lexical scores for DS pairs than SS pairs). Our layman's terms datasets counteract this weakness to a great extent, leading to a more correct relation between pairs (higher lexical scores for SS pairs than DS pairs). Most importantly, the combination of semantics-based method and our layman's terms dataset provides the most robust evaluation.

3.5 Enhancing training with Layman's Terms: A Scaling Law Perspective

When using heavily patternized professional reports to train generative models, we hypothesize that generative models may pay more attention to the structure of the reports, instead of their semantics. In contrary, translating professional reports into layman's terms largely removes their highly-templated nature. Intuitively, the diverse expressions of layman's reports lead to the models focusing on report semantics, as there is no templates to overfit to anymore. We envision that training with layman's terms reports would largely unleash model's potential, leading to stronger semantics understanding on the reports, compared to training on the original professional formats.

In order to verify our hypothesis, we construct a range of training sets for both datasets with increasing training samples, going from 5k to 50k. We use these samples to fine-tune `MiniGPT4` and generate 500 reports in each setting for evaluation. Given the behavioral advantage of semantics-based metrics shown in the last section, we calculate the semantic similarity between each sentence in the generated report and each sentence in the reference report. We then count the number of occurrences in different score ranges. The result of the largest setting is outlined in Figure 2 (a) (Refer to Appendix A.10 for statistics of all settings). Next, we compare the score ranges for the same training scale with different types of data. We consider a



(a) (b)

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

semantic similarity greater than 0.8 as an indication that the generated sentence correctly reflect the semantics of the groundtruth sentence.

Here, we reveal a positive **scaling law** brought by layman's terms dataset. As shown in Figure 2 (b), the model trained on layman's terms dataset presents a monotonically increased performance along the increase of training samples. By contrast, the model trained professional reports reaches peak performance with 10k training samples, and starts monotonically degrading with more training samples. This clearly indicates that training on more professional reports is detrimental to model's semantics learning, which we attribute to the overfitting to templates throughout the paper. Notably, the layman's terms dataset surpasses raw reports when the training set size reaches 50k. In order to reveal why training on 10k professional reports reaches the highest performance, we conduct a series of experiments. We find that, the model trained on the 10k professional dataset suffers from a severe representation collapse. We measure the pairwise similarity of the generated reports on the test set and find that the reports generated by the 10k professional model results in an average cosine similarity of 0.893 with a variance of 0.008. This implies that the high cosine similarity in Figure 2 is achieved by quickly learning to generate reports that look alike the majority class in the train set (no findings/normal reports) to decrease the loss. In contrast, its 10k layman terms counterpart results in a pairwise cosine similarity of 0.802 and a variance of 0.012, indicating the diversity of the layman reports makes the reports harder to learn at first. Combining this finding with professional reports' vulnerability to overfitting that we illustrate in Figure 2, we can conclude that our layman's dataset provides a more natural and robust progression in model's semantics understanding by scaling up the dataset, compared to shortcut behavior found above for the professional report model. Additionally, we also measured a few more specialized metrics and show that at the scale of 10k, the layman's model outperforms the professional model (Chexbert 0.447 v.s., 0.398; RadCliQ-v0 0.413 v.s., 0.405).

3.6 Evaluation

Due to the obscurity of professional reports and high cost for clinicians to serve as human evaluators, there is a scarcity of research focusing on the correlation between human scores and BLEU scores in this domain. However, it is well-studied problem that metrics based on word overlaps often fail to capture the semantics of the reports and typically exhibit weak correlations with human annotations in other fields. Thus, relying solely on word overlap metrics to evaluate the quality of generated radiology reports is inadequate.

To align the setting of evaluating the two models trained on raw reports and layman's terms reports and make raw reports more comprehensive to non-clinician human evaluators, we translate the professional reports into layman's terms. We then recruit three human annotators, all proficient in English, to score these generated reports following this 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.

The same protocol is applied to evaluate both generation of models trained by layman's term and raw reports, respectively. After obtaining scores from all human evaluators, an aggregate score is calculated by averaging across all criteria to pro-

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 3: 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.

vide a comprehensive measure of each response's quality. The Inner-Annotator Agreement (IAA) among the three evaluators is assessed using Cohen's Kappa, yielding an agreement of 0.63 for raw reports and 0.58 for layman's terms reports, indicating a consistent fair to good agreement level (0.4-0.75). For more details about human annotators, please see Appendix A.11.

The detailed correlation results are outlined in the Table 3. From the results, correlation between automated metrics and humane evaluators is consistenly higher for layman's terms. We also measure Clinical Efficacy (CE) metrics (e.g., CheXbert-F1, RadGraph-F1 and RadCliQ) for both professional reports and layman's reports. As the CE metrics mainly evaluate name entities of the generated reports, these metrics are as well applicable to layman's reports. As can be seen in Table 3, the correlations between CE metrics and human evaluation are generally higher for reports in layman's terms, which verifies the effectiveness of layman's reports.

3.7 Case Study

We present several sentence-level examples in Table 4, demonstrating how translating professional reports into layman's terms improves clarity and comprehension. For instance, the term *pleural effusion* is translated as *extra fluid around the lungs*, making it easier to understand. Similarly, *bibasilar atelectasis*, which is difficult for patients to comprehend, becomes *collapsed lung areas*, a much clearer and more accessible description. These simplifications make medical information much more accessible and easier for patients to understand.

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 4: Examples from the sentence-level dataset.

4 Conclusion

In this paper, we proposed a novel Layman's RRG framework, which includes two Layman's terms datasets, a semantics-based evaluation framework and a training framework. We introduced two high-quality datasets with rigorous refinement process. We show that, the reports in layman's terms combining with our semantics-based method significantly mitigates the inflated results brought by word-overlap metrics and highly-templated radiology reports. Last, we demonstrate that training with our report-level Layman's dataset enhances the model's semantic understanding, displaying a promising scaling law for the model's performance as the number of training examples increases.

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

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.

Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020. Expertise style transfer: A new task towards better communication between experts and laymen. *arXiv preprint arXiv:2005.00701*.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Ondřej Cífka, Umut Şimşekli, and Gaël Richard. 2020. Groove2groove: One-shot music style transfer with supervision from synthetic data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2638–2650.

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.

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.

Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. 2019. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*, 26(11):3365–3385.

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

Yu Hwan Kim, Se Hyun Nam, Seung Baek Hong, and Kang Ryoung Park. 2022. Gra-gan: Generative adversarial network for image style transfer of gender, race, and age. *Expert Systems with Applications*, 198:116792.

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*.

Qing Lyu, Josh Tan, Michael E Zapadka, Janardhana Ponnatapura, Chuang Niu, Kyle J Myers, Ge Wang, and Christopher T Whitlow. 2023. Translating radiology reports into plain language using chatgpt and gpt-4 with prompt learning: results, limitations, and potential. *Visual Computing for Industry, Biomedicine, and Art*, 6(1):9.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

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.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

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*.

Chenghao Xiao, G Thomas Hudson, and Noura Al Moubayed. 2024. Rar-b: Reasoning as retrieval benchmark. *arXiv preprint arXiv:2404.06347*.

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

```
1: for i = 1 to n do
2:     repeat
3:         t_i ← LLM-Translate(d_i)
4:         sim ← Semantic-Similarity(d_i, t_i)
5:         correct ← LLM-Check-Translation(d_i, t_i)
6:     until sim ≥ θ and correct
7: end for
8: return T
```

A.3 Dataset

In this part, we outline the statistics of our datasets as follows in the Table 5.

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

Table 5: 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.

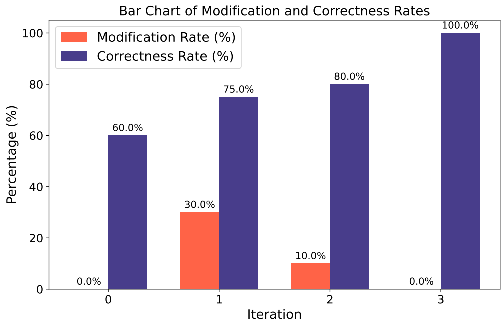


Figure 3: Refinement

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 6, 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.

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.

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 7, 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.

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

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

Table 7: Instruction Tuning

A.9 Case study

In this section, we provide more examples from sentence-level dataset and report-level dataset. The Table 8 include some examples in the sentence-level dataset and Table 9 present samples selected from the report-level dataset.

A.10 Scaling Law

As illustrated in Figure 4, 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.11 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.

A.12 Additional Experiments

We also tested the LLM-based approach using two different open-access ChatGPTs[2] in both MIMIC CXR and PadChest (English translated) datasets, denoted as LLM1 and LLM2, respectively. `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. 5.

The evaluation metrics are in three types: i) Clinical accuracy, ii) Relevance, and iii) Readability. For Readability, a set of text statistics metrics[3] to be used. Their abbreviation and the corresponding metrics are listed below:

- `Easy`: The Flesch Reading Ease formula
- `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 10 and Table 11.

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

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

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 $proportion, BLEU, ROUGE, Meteor$

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 8: Some examples of sentence-level dataset.

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.

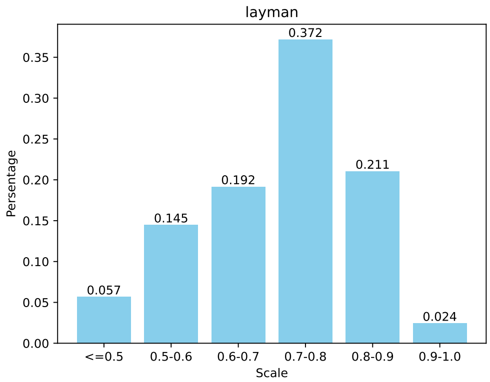
Table 9: Some examples of report-level dataset.

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**

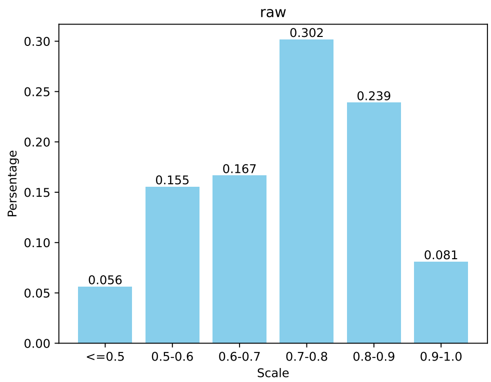
Table 10: Performance

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**

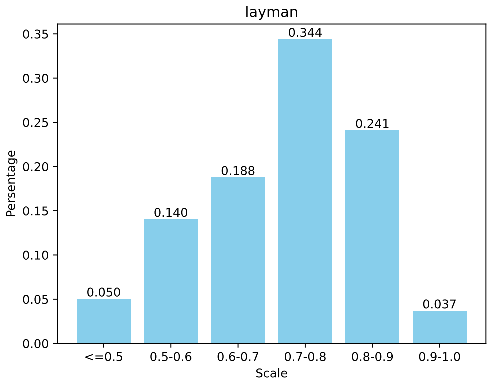
Table 11: Performance



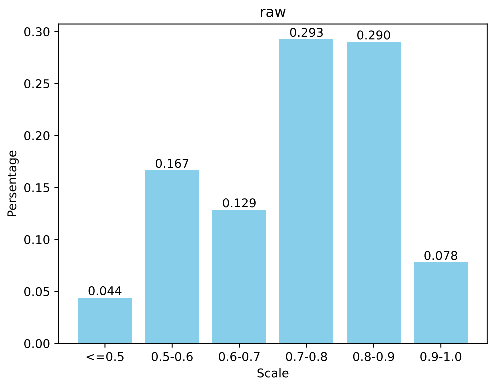
(a)



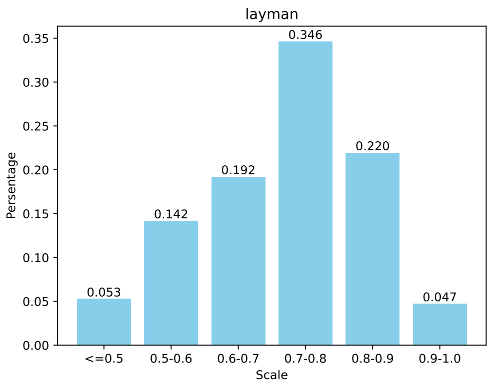
(b)



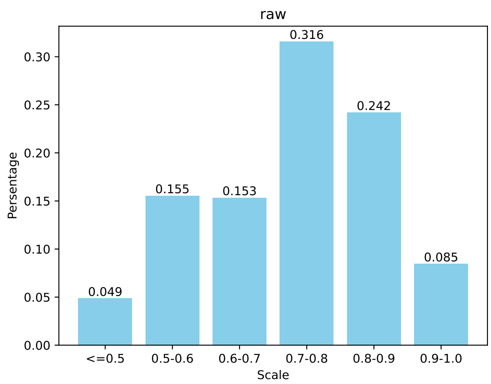
(c)



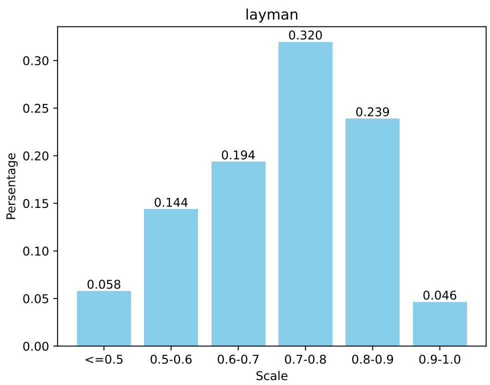
(d)



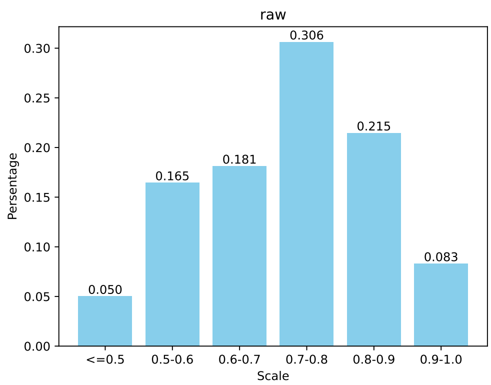
(e)



(f)



(g)



(h)

Figure 4: 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.

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