

# QUANTRAD: ADVANCING QUANTITATIVE RELIABILITY IN RADIOLOGY REPORT GENERATION WITH CASCADED DECODERS

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Radiology report generation using artificial intelligence has shown promise in enhancing clinical workflows. However, due to limitations of language modeling loss, existing approaches struggle with quantitative accuracy (e.g., measuring the size of nodules), and lack the ability to produce confidence scores for medical findings, which is crucial for quantitative metrics required by regulatory approval. This paper introduces QuantRad, a novel approach utilizing cascaded decoders to address these challenges in radiology report generation. QuantRad pairs a vision encoder with three decoders that operate sequentially: the first conducts sentence-level topic planning by generating a series of questions, the second recognizes abnormal targets and their quantitative and categorical attributes, and the third generates the final report by answering each question based on the recognized targets. With the dedicated target recognition step, our method integrates the quantitative strength of a perception model to text generation. Specifically, QuantRad recognizes abnormal targets without being biased by language priors, and produces probability scores along with each finding, allowing adjustments of sensitivity for clinical adoption and producing ROC curves for regulatory compliance. Besides, the disentangled topic planning captures the uncertainties in the omission of medical findings and their presentation order, allowing the report generation decoder to be trained with less ambiguity. Our method advances the accuracy and reliability of radiology report generation, offering a promising path for clinical applications and regulatory validation.

## 1 INTRODUCTION

Medical imaging is crucial in healthcare, but a worldwide shortage of radiologists, exacerbated by high burnout rates, poses significant risks to patient care Ganeshan et al. (2020); Parikh et al. (2020); Cao et al. (2023). AI models that generate radiology reports automatically have been attracting growing interest. While it takes thirteen to fifteen years to train a radiologist, these AI models could alleviate the shortage problem at scale.

Radiology Report Generation is typically framed as an image captioning task, with notable progress in recent years. However, the inherent limitations of language modeling loss hinder clinical accuracy, particularly in quantitative aspects. A radiology report should correctly identify abnormalities with accurate attributes. These attributes include quantitative measurements such as nodule sizes (e.g., “2cm”) and categorical descriptors such as severity (e.g., “mild”). However, the language modeling loss treats all mismatches equally, failing to prioritize critical factual correctness. Moreover, quantitative measurements are often biased by language patterns in the training data rather than being grounded in the input image. These constraints significantly compromise the clinical reliability of existing models. Compared to text generation models, perception models such as classification and detection models demonstrate superiority in quantitative accuracy. Therefore, this paper aims to integrate the strengths of perception models into radiology report generation, addressing the quantitative limitations of current encoder-decoder text generation models.

By using a modified ViT-DET Li et al. (2022) model for recognizing abnormal targets as a set prediction task, we directly supervise the prediction of quantitative and categorical attributes rather

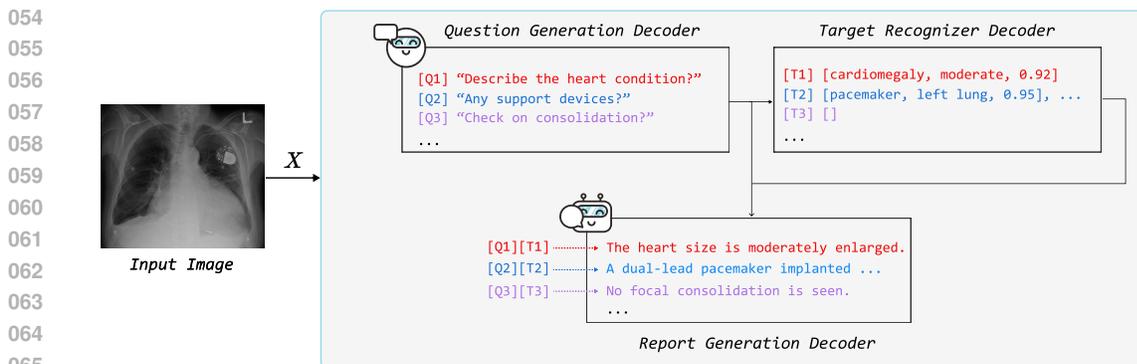


Figure 1: QuantRad pipeline with cascaded decoders. Paired with an image encoder, the first decoder generates a sequence of questions based on the image features. The second decoder recognizes abnormal targets, including their names, attributes, and confidence scores. The third decoder answers each question given the image and recognized targets. The final report is constructed by concatenating all answers. The ground-truth question-answer pairs and abnormal targets are extracted from the report with a private deployment of ChatGPT.

than generating them through the next-word prediction process of the report. This method enables the production of confidence scores for each medical finding, facilitating sensitivity level adjustments in clinical applications. Crucially, it allows for the generation of quantitative metrics such as Receiver Operating Characteristic (ROC) curves, which are essential for regulatory processes. In contrast, existing methods that generate text-only outputs are constrained to binary predictions (positive or negative). The numerical confidence scores produced by our model provide a more nuanced evaluation of the reliability in clinical scenarios, potentially accelerating the path to clinical deployment. This approach bridges the gap between the quantitative requirements of regulatory bodies and the qualitative nature of radiology reports, addressing a significant challenge in the field of medical report generation.

Furthermore, we identify that the ambiguities in the textual training data is a key reason affecting the model’s reliability. A radiology report comprises multiple sentences, each focusing on a medical topic like heart conditions, lung effusion, or opacities. The order of presenting these topics varies, reflecting the diverse writing styles of radiologists. For instance, some prefer starting with salient findings, whereas others favor checking visual details upfront to avoid overlooking them. Furthermore, negative findings (i.e. absence of a disease) are sometimes omitted in the report. Such variabilities do not affect the correctness of a report, but they introduce a degree of randomness that is ambiguous for a model to fit. As the image captioning loss is based on categorical cross entropy, it requires a word-by-word exact match with the ground truth. Consequently, a model could be unfairly penalized for accurate, albeit differently paraphrased, predictions. Image captioning datasets like COCO Lin et al. (2014) address such ambiguity by providing multiple ground-truth captions per image. However, this feature is not available in medical report datasets, which hinders the robustness of both training and evaluation.

To overcome the ambiguity in sentence topics, we propose a novel approach by converting report generation into a multi-round visual question answering (VQA) task. In this approach, each sentence in the report is generated with a given topic defined by a question. We utilize ChatGPT Wang et al. (2023)<sup>1</sup> to convert each radiology report into a multi-round VQA format. Then, we train two decoders with the first one generates questions given the image, and another answers each question based on the image and recognized abnormal targets. The question-generation decoder captures the uncertainties of topic omission and ordering, enabling us to train the report generation decoder with less ambiguity for improved test-time reliability. While the question generation decoder may not fit the training data perfectly due to inherent uncertainties, its under-fitting has limited impact on perceivable performance during testing. This pipeline also offers the flexibility to answer new questions that may not be included in the default output.

<sup>1</sup>We utilize a private, in-house deployment of ChatGPT on Azure to satisfy data usage agreements. The converted datasets will be made available upon acceptance.

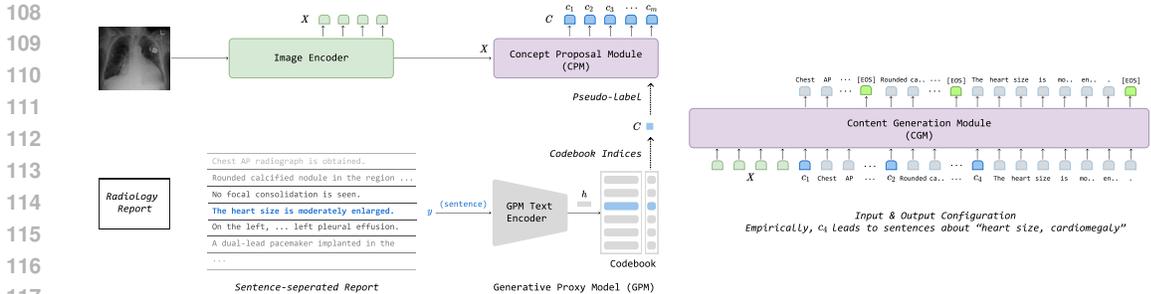


Figure 2: Overview of C2C pipeline. Given a radiology report, the Generative Proxy Model (GPM, ??) gives each sentence a discrete label  $c$  of its concept class. For example, in this figure, the empirical meaning of  $c_2$  for the second sentence is “aortic nodule”. The GPM is used to generate pseudo-labels  $C$  to train the Concept Proposal Module (CPM, Section 3.3), which generates a sequence of concepts  $[c_0, \dots, c_m]$  based on an image. Finally, the Content Generation Module (CGM, Section 3.4) generates the report based on the image feature  $X$  and concept classes  $C$ .  $C$  behave like special begin-of-sentence ( $[BOS]$ ) tokens. The output of an end-of-sentence ( $[EOS]$ ) token signals the completion of a sentence. At this time, we append the next concept token  $c_{i+1}$  to the input to trigger the next sentence, and repeat this autoregressive process until the final  $[EOS]$  token is produced.

Concretely, we propose QuantRad, a novel approach that generates medical report with three decoders that operate in a cascaded manner. As illustrated in Figure 1, the first step, Question Generation, decides the topic per sentence in terms of questions. The second step, Target Recognition, conducts a set prediction of abnormalities as triplets of  $\langle \text{name, attribute, confidence score} \rangle$ . Both qualitative and categorical attributes are predicted with a DETR-style Carion et al. (2020) ViT-DET model Li et al. (2022). The final step, Report Generation, answers each question based on the image feature and recognized targets. We extract question-answer pairs and abnormal targets in a structured format from existing report generation datasets Johnson et al. (2023) to supervise these three modules. Mathematically, we refactor medical report generation from  $P(Y | X)$  into Equation (1):

$$P(y | X) = P_a(y | X, t, q) \cdot P_t(t | X, q) \cdot P_q(q | X), \tag{1}$$

where  $X, y$  denote the image and one sentence in the text output  $Y$ , respectively.  $q$  and  $t$  are the interim outputs of questions and abnormal targets in text, respectively.  $P_q, P_t, P_a$  represents three decoder modules. Equation (1) shows the generation process of one sentence, and it is repeated autoregressively until all sentences are generated.

As radiology reports are open-vocabulary, it is not feasible to define a closed set of classes and train a conventional object detector to recognize abnormal targets. Furthermore, substantial measurements are not convertible to simple numerical values, e.g., “millimetric”, “2 to 3 cm”, “multiple”. Therefore, we modify the ViT-DET implementation of DETR and replace the classification and regression heads with phrase generation. As each output phrase (the name of the target, the name of the attribute and the value) is short, generating the textual phrase still enjoys the benefit of avoiding major language prior biases from the full report. In the common case that the phrase contains only one token (e.g., “2”), phrase generation is essentially a classification over the tokens. The vocabulary of the measurements is much smaller than the vocabulary of radiology reports, making this task easier to train. The output probability is produced based on the logits of the whole phrase.

To summarize, we propose QuantRad, a novel radiology report generation approach which improves quantitative reliability by using three decoders operating in a cascaded fashion. QuantRad combines the strengths of perception and generation models, addressing the limitations of using the plain language modeling loss. Our method achieves state-of-the-art (SOTA) performance on MIMIC-CXR Johnson et al. (2019b), the largest radiology report generation dataset, with a smaller model size. Our study contributes to the broader field of image-to-text generation. Besides healthcare, the proposed methods are applicable in accuracy-critical scenarios, such as legal and finance, where similar challenges could also exist.

## 2 RELATED WORK

### 2.1 IMAGE CAPTIONING

Image captioning is to generate a sentence which describes a given image. The latest work benefit from large scale vision-language pre-training Chen et al. (2020a); Dou et al. (2021); Wang et al. (2021); Kim et al. (2021). Encoder-decoder architectures Li et al. (2023); Wang et al. (2022a); Nguyen et al. (2022) provide a unified implementation for various vision-language tasks.

While many radiology report generation methods are based on image captioning Cornia et al. (2020); Vinyals et al. (2015); Xu et al. (2015); You et al. (2016), there are key differences in the tasks including (1) radiology reports are much longer than generic image captions e.g. as in COCO Captions Lin et al. (2014), and have multiple sentences covering a different medical concepts; (2) radiology reports are highly templated for individual sentences, while are diverse in paraphrasing a paragraph of multiple sentences.

### 2.2 RADIOLOGY REPORT GENERATION

Chest X-ray radiology reports lack a standardized order for presenting medical findings Burbridge (2017). For instance, the inside-out order Smithuis & Otto (2022) and the ABCDE order (each letter represents an anatomical region) Lopez-Cardona (2023) are two approaches from clinical guidelines. Additionally, medical conditions can be omitted from the report Irvin et al. (2019). These inconsistencies lead to loss-evaluation mismatch problems, creating challenges for both training and evaluation Gu et al. (2018b); Yi et al. (2020); Gu et al. (2018a). Previous studies have demonstrated the value of generating reports using a two-step approach Nooralahzadeh et al. (2021); Liu et al. (2019), which are conceptually similar to ours. However, due to the lack of sentence-level concept labels (which clarifies the ambiguity) in existing work, they motivate their approaches from different perspectives.

Specifically, Liu et al. (2019) adopts a hierarchical framework which predicts sentence-level topics as the first step. However, their topic generation module is not supervised with any labels, leaving uncertainty in their actual meaning. Another similar work is Nooralahzadeh et al. (2021), which first generates high-level context sentences and then refine them to the reports. The first step is trained to generate medical keywords per sentence extracted with a text processing model. We differ from them on the supervision of the first step. Our method specifically tackles the label ambiguity problem (uncertainties of topic omission and ordering). The significance of the label ambiguity issue has been primarily discussed in the context of image recognition Rajeswar et al. (2022); Chung et al. (2023); Ridnik et al. (2021), but has not been adequately addressed in the medical imaging domain due to a lack of solutions. Our method converts multi-sentence report generation into a sequence of visual question answering tasks, which not only reduces the ambiguity during training, but also facilitates responding to new questions asked by the user when a particular aspect is omitted from the output.

### 2.3 VISION-LARGE LANGUAGE MODELS

Latest advancements in large language models (LLMs) OpenAI (2023); Touvron et al. (2023); Chowdhery et al. (2022) provide a unified interface for a wide range of tasks.

Researchers built multi-modal large models by adding a vision head to the language model. As the LLMs are mostly decoder-only, they can add visual tokens as additional input if the feature space of image and text modalities are aligned. Among these work, LLaVA Liu et al. (2023a) proposes generating visual question-answering training data by prompting GPT-4, and then use it to train a model based on the open source language model, LLaMA Touvron et al. (2023). LLaVA-Rad Chaves et al. (2024) and MAIRA Hyland et al. (2023) models leverage large text decoders to improve report generation performance.

### 3 METHOD

#### 3.1 CASCADED DECODERS PIPELINE FOR RADIOLOGY REPORT GENERATION

We compose radiology reports by separated sentence-level topic planning, clinical abnormality target recognition, and report generation steps. The sentence-level topic planning is disentangled from report generation to isolate the uncertainties of the omission and ordering of the topics, a form of label ambiguity. The target recognition is separated from text generation to leverage the quantitative reliability of perception models, which predict targets without the shortcut of referencing language priors. Three decoders were trained progressively: first, the question generation decoder is trained with the image encoder. Then, we train the question generation decoder and the target recognition decoder. Finally, we train the full model with three decoders jointly. This section introduces the three decoder modules in detail.

#### 3.2 GENERATING VQA DATASET

```
[
  {
    "q": "Are there any support devices visible?",
    "a": "Sternotomy wires are intact. Right Swan-Ganz catheter is close to pulmonic valve. Mitral valve replacement is in correct position."
  },
  {
    "q": "Are there any signs of atelectasis or pleural effusion?",
    "a": "Mild interval increase in retrocardiac opacity from moderate atelectasis and left pleural effusion."
  },
  {
    "q": "Is there any evidence of pneumothorax?",
    "a": "No pneumothorax and right lung is clear."
  },
  {
    "q": "What is the condition of the heart size and mediastinum?",
    "a": "Heart is mildly enlarged and there is a post op appearance to mediastinum."
  },
  {
    "q": "Are the hila normal?",
    "a": "Hila are normal."
  },
  {
    "q": "Are there any bony abnormalities?",
    "a": "No bony abnormality."
  }
],
```

Figure 3: Question-answer dataset converted from radiology reports.

We use ChatGPT to convert all radiology reports to question-answer pairs as is shown in Figure 3

#### 3.3 QUESTION GENERATION DECODER

Given an input image, the decoder generates a sequence of questions, which essentially serves as soft prompts for generating corresponding sentences. We implement it with a six-layer transformer decoder, and train it with questions generated by ChatGPT based on the original report. The question generation decoder conducts sequence generation autoregressively with reference to the previously generated questions. Concretely, it generates  $m$  sentences  $Z = (c_1, c_2, \dots, c_m)$  by modeling Equation (2):

$$P(C | X) = \prod_{i=0}^{m+1} P(c_i | X, c_0, c_1, \dots, c_{i-1}), \quad (2)$$

where  $c_0$  is a [BOS] token denoting the beginning of a sequence, and  $c_{m+1}$  is [EOS] appended after the sequence to signal the end of token generation. With the ground truth being the concatenated questions, we train decoder with the language modeling loss:

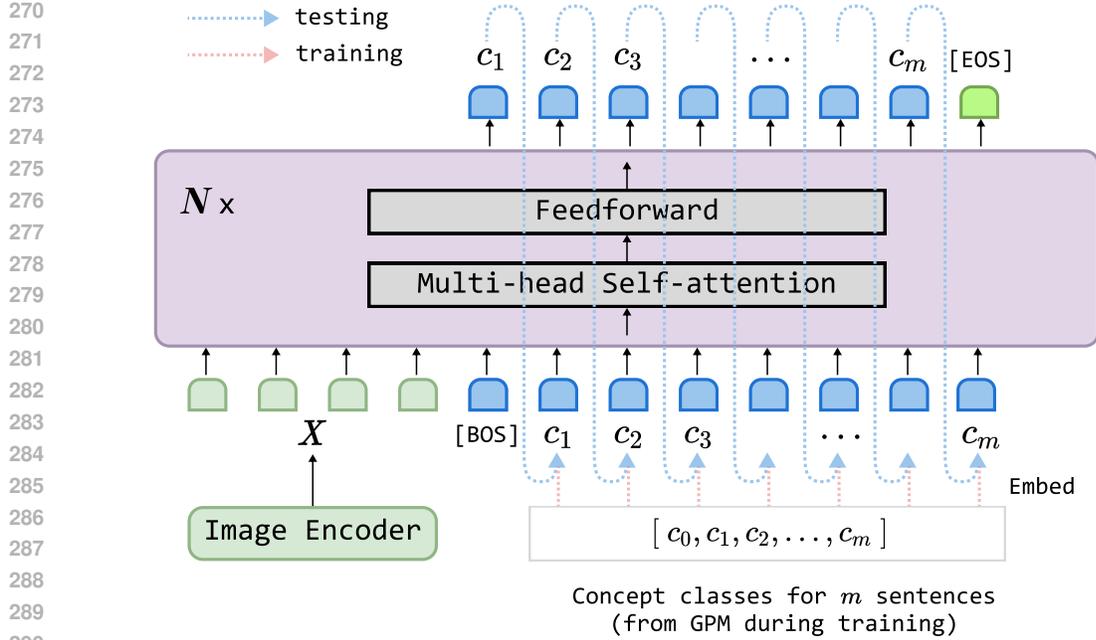


Figure 4: Question Generation Decoder. The model is a multi-modal decoder of  $N$  transformer layers. Each output token  $c_i$  is predicted with  $X$  and its prefixes  $[c_0, \dots, c_{i-1}]$  as the input. Questions are separated by a special  $[EOQ]$  token and the generation process ends when a  $[EOS]$  token is produced.

$$L = \frac{1}{m+1} \sum_{i=1}^{m+1} \text{CE}(c_i, p(c_i | X, c_0, c_1, \dots, c_{i-1})), \quad (3)$$

where CE is the categorical cross-entropy loss with label smoothing of 0.1.

### 3.4 REPORT GENERATION DECODER

The report generation decoder is also a multi-modal text generation decoder. It learns to generate a sentence of  $n$  tokens  $Y_i = (y_{i_1}, y_{i_2}, \dots, y_{i_n})$  conditioned on the image, one topic question  $c_i$  and corresponding targets  $t$ . Mathematically, CGM models the following:

$$P(Y_i | X, c_i, t) = \prod_{j=0}^{m+1} P(y_{i_j} | X, c_i, t, y_{i_0}, y_{i_1}, \dots, y_{i_{j-1}}), \quad (4)$$

where  $Y_i$  denotes the  $i^{\text{th}}$  sentence from the radiology report. Similar to Equation (2),  $y_{i_0}$  and  $y_{i_{n+1}}$  are special  $[BOS]$  and  $[EOS]$  tokens, respectively. By iterating  $c_i$  from  $(c_1, c_2, \dots, c_m)$ , the Question Generator generates  $m$  sentences  $Y = (Y_1, Y_2, \dots, Y_m)$  and composes the whole radiology report.

### 3.5 IMPLEMENTATION

**Module Architectures.** For the image encoder, we use a ViT-B/16 Dosovitskiy et al. (2020) pre-trained with MAE He et al. (2022); Xiao et al. (2023) on medical images. For decoders, we use a six-layer, randomly initialized text decoder from GIT Wang et al. (2022a). A linear projection layer is added between the image encoder and decoders for feature space alignment.

**Three-Stage Training.** The model is trained in three stages. In Stage 1, we train the Question Generation decoder to generate sentence topics. In Stage 2, we add Target Recognition decoder to the training and pair with the same image encoder. Ground truth questions are used to train the Target

Pred.	B-4	METEOR	ROUGE-L	CIDEr	Human	GPT-4
#1	0.508	0.312	0.667	3.756	wrong	wrong
#2	0.000	0.185	0.167	0.239	correct	correct

**Reference:** *“The heart size is top normal.”*  
**Prediction #1:** *“The heart size is mildly enlarged.”*  
**Prediction #2:** *“Borderline size of the cardiac silhouette.”*

Table 1: **Limitations of  $n$ -gram metrics for medical texts.** Both the reference and predictions are extracted from real radiology reports. The reference (ground truth) makes a negative diagnose of cardiomegaly (enlarged heart). Prediction #2 restates the same meaning with different words. Prediction #1 is contradictory to the reference, although they have more overlapping words. While existing metrics based on  $n$ -grams fail in this scenario, GPT-4 can be prompted to make consistent judgements with human. We provide the details in the supplementary material. The issue demonstrated here is common in medical texts.

Recognizer. Finally, we add Report Generation decoder to the training, supervised by ground truth questions and targets.

## 4 PERFORMANCE EVALUATION

### 4.1 FROM N-GRAM TO LLM-BASED METRICS

As radiology report generation is treated as an image captioning task, natural language generation (NLG) metrics are commonly used in existing studies as the major evaluation protocol. In this section, we first show that these metrics are not reliable for evaluating medical text. We then propose a new evaluation method based on large language models (LLMs) like GPT-4 OpenAI (2023). Working with radiologists, we show that the LLM-based metric is considerably more consistent with human judgements.

#### 4.1.1 EXISTING METRICS

Existing NLG metrics include BLEU Papineni et al. (2002), METEOR Banerjee & Lavie (2005), ROUGE Lin (2004) and CIDEr Vedantam et al. (2015), all based on  $n$ -grams. The statistics of overlapping words is important to achieve better scores. When applied on medical reports, they are less reliable due the increased length, synonyms and paraphrasing in medical texts. Please find Table 1 for an example.

#### 4.1.2 GPT-4 FOR MEDICAL REPORT EVALUATION

Latest studies from Natural Language Processing show that GPT-4 achieves state-of-the-art correlation with human judgments in most NLG tasks Liu et al. (2023b); Wang et al. (2023); Chiang & Lee (2023). However, the applicability on the medical domain is not yet explored. Working with radiologists, we fill this gap by implementing a GPT-4-based evaluation metric for radiology report generation, and rigorously validating its robustness on medical texts. We hope our work can serve as an effective benchmark for future studies.

**Implementation.** We iterate on the prompt to facilitate robust evaluation with GPT-4. In the prompt, we instruct GPT-4 to give a 0-10 star rating for the predicted radiology report with the ground truth as the reference. We defined the criterion for the rating of 0 and 10, including the factors to be considered and to be ignored based on the characteristics of existing datasets. Our final prompt template is provided in Figure 5, which is validated against human judges. Alternative implementations include asking GPT-4 to give a 0-100 score rating, or selecting a better prediction from the output of two models. Their correlations with human judgements are compared in the ablation studies.

**Human Alignment.** We assess the reliability of GPT-4 and existing metrics by comparing their alignment with human judgements. For this purpose, we form a group of three radiologists, with each one having at least five years experience in Chest X-ray interpretation. We randomly sample 50

378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431

**System:** You are a human evaluator who can assess the quality of a natural language generation model.

**User:** Score the following generated chest x-ray radiology report given the human written report as reference with one to ten stars, where one star means “totally wrong” and ten stars means “perfect”. Note that perfect measures the factual correctness of the diagnose, relevance of two reports, and fluency (well-written and grammatically correct). Ignore sentences about comparisons with previous studies in both the reference and the generated report.

Human Reference:  
[Ground Truth]

Generated Report:  
[Model Prediction]

Stars:

Figure 5: Prompting GPT-4 for robust radiology report evaluation.

radiology studies, and for each study, we prepare the ground truth report and two machine-generated reports using XPRONET Wang et al. (2022b) and our method. Totally, 100 machine-generated reports are rated by radiologists, each report is rated by two radiologists. When rating a report, we present to the radiologist the Chest X-ray image, the ground truth report and the machine generated reports. The radiologists are asked to give a 1-10 rating based on the image and the ground truth. When the machine generated report diverges from the ground truth, the radiologists are asked to make their professional judgement on the points to deduct. Results in Table 2 show that the GPT-4-based metric outperforms all existing metric in terms of both Spearman Zar (2005) and Kendall’s Tau Kendall (1938) correlations.

	BLEU-4	METEOR	ROUGE-L	CIDEr	GPT-4
<b>Spear.</b>	0.158	0.311	0.191	0.154	<b>0.455</b>
<b>Kand.</b>	0.113	0.214	0.133	0.105	<b>0.346</b>

Table 2: Spearman and Kendall-Tau correlations of different metrics with human judgements for radiology report. GPT-4 evaluation outperforms existing  $n$ -gram metrics on human alignment by a clear margin. Please find Appendix E for additional analysis.

## 4.2 EXPERIMENTAL RESULTS

Addressing label ambiguity results in a more effective learning process and therefore, improve the overall evaluation metrics. To quantify the gain from alleviating label ambiguity, we conduct ablation studies by comparing the performance of our proposed method with and without the sentence-level concepts as conditions.

We evaluate our method on two public datasets, MIMIC-CXR Johnson et al. (2019b;a). Following existing works, we use the *findings* section from the report as the ground truth.

**MIMIC-CXR** is the largest datasets for Chest X-ray (CXR) report generation. The dataset contains 227,835 radiographic studies, where each study is a pair of a radiology report and corresponding CXR images. We use the official training/validation/testing split.

### 4.2.1 RESULTS

**Automatic Metrics.** Section 4.2.1 shows the comparison of existing methods. Our method, QuantRad outperforms the existing state of the art by a clear margin of 8.5%.

**Human Evaluation.** We conduct human evaluation with 50 randomly selected reports from the MIMIC test set. The predictions of two methods, including an existing SOTA Wang et al. (2022b) and QuantRad (ours) are provided per study (totally 100 reports), and each report is rated by two radiologists (totally 200 data points). To remove bias, the reports are presented randomly with the names of methods blind to the radiologists. Results in Table 3 shows that reports generated by our method is better than or equal to XPRONET on 71.0% of the studies.

432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485

Method	Mean	Std.	Comparative	
XProNet	7.12	2.17	29.0% better,	8.0% tie
<b>QuantRad (ours)</b>	<b>8.16</b>	<b>1.65</b>	63.0% better,	8.0% tie

Table 3: Human evaluation of two methods. The rating follows a 0-10 star scale. Mean and Std. are the mean and standard deviation of all ratings per method, and Comparative measures the percentage of one method being rated better (or tie) than the other method. Our QuantRad method outperforms XPRONET, an existing SOTA, by a clear margin.

Table 4: Report Generation Performance on MIMIC-CXR

Model	CheXbert								R <sub>GER</sub>	BLEU	ROUGE	
	("uncertain" as <i>negative</i> )				("uncertain" as <i>positive</i> )							
	Micro-avg		Macro-avg		Micro-avg		Macro-avg					
F1-14	F1-5	F1-14	F1-5	F1-14	F1-5	F1-14	F1-5	(1)	(4)	(L)		
<i>Model size &gt;7B</i>												
LLaVA-Rad Chaves et al. (2024)	<b>57.3</b>	57.4	39.5	47.7	<b>57.3</b>	<b>60.2</b>	<b>44.0</b>	<b>53.3</b>	29.4	38.1	15.4	30.6
Med-PaLM M Tu et al. (2024)	53.6	<b>57.9</b>	<b>39.8</b>	<b>51.6</b>	-	-	-	-	-	32.3	11.3	27.3
MAIRA Hyland et al. (2023)	55.7	56.0	38.6	47.7	55.3	58.8	42.3	51.7	<b>29.6</b>	39.2	14.2	28.9
CheXagent Chen et al. (2024)	39.3	41.2	24.7	34.5	39.4	42.1	27.3	35.8	20.5	16.9	4.7	21.5
LLaVA-Med Li et al. (2024)	27.2	22.0	15.5	16.6	27.3	24.4	18.7	20.5	6.5	22.2	1.0	13.3
LLaVA Liu et al. (2024)	22.9	23.4	15.4	17.5	23.7	26.9	17.0	20.3	4.5	21.0	1.3	13.8
<i>Model size &lt;1B</i>												
Flamingo-CXR Alayrac et al. (2022)	-	-	-	-	<b>51.9</b>	56.5	-	-	-	-	10.1	<b>29.7</b>
CvT2Dist. Nicolson et al. (2023)	<b>44.2</b>	-	<b>30.7</b>	-	-	-	-	-	-	<b>39.3</b>	<b>12.7</b>	28.6
$\mathcal{M}^2$ trans Miura et al. (2020)	-	-	-	-	-	<b>56.7</b>	-	-	-	-	11.4	-
RGRG Tanida et al. (2023)	-	-	-	-	-	54.7	-	-	-	37.3	12.6	26.4
R2Gen Chen et al. (2020b)	-	-	-	-	22.8	34.6	-	-	-	35.3	8.6	27.7
TieNet Wang et al. (2018)	-	-	-	-	-	27.1	-	-	-	-	8.1	-
<i>Ours – Model size 0.8B</i>												
QuantRad	<b>58.6</b>	<b>58.8</b>	<b>39.9</b>	<b>48.6</b>	<b>57.3</b>	<b>59.5</b>	<b>43.4</b>	<b>52.5</b>	<b>30.1</b>	<b>39.6</b>	<b>15.9</b>	<b>31.9</b>
Ablative Baseline	49.1	53.9	31.0	43.9	49.7	56.2	35.2	49.2	27.1	33.5	13.4	29.9

## 5 ABLATION STUDIES

### 5.1 MEDICAL CONCEPTS AS CONDITIONS

This ablative study validates our assumption that our Concept-to-Content method improves the learning by reducing label ambiguities. Results in Table 5 shows that there is a clear performance degradation when removing the medical concept which was introduced to alleviate the ambiguity.

Method	BLEU-4	METEOR	ROUGE-L	GPT-4
Baseline	0.103	0.144	0.276	0.538
Clustering	0.111	0.154	0.282	0.546
GPM	<b>0.113</b>	<b>0.164</b>	<b>0.287</b>	<b>0.564</b>

Table 5: Ablation study of latent prompts. Baseline: remove the latent prompts. Clustering: cluster sentences by the text embeddings, and use the cluster ID as the pseudo-label. GPM: our learnable approach of training a generative proxy model.

## 5.2 SIZE OF CODEBOOK

The size of the codebook  $K$  in the GPM is a hyper-parameter, which represents the granularity of the medical concepts. We searched  $K$  in [32, 128], a range informed by medical knowledge on the number of medical concepts as well as anatomical regions in radiology reports. We searched  $K$  in Table 6 and validate the choice by both the quantitative model performance and qualitative visualizations as shown in Appendix D.

$K$	BLEU-4	METEOR	ROUGE-L	GPT-4
32	0.106	0.154	0.281	0.552
64	<b>0.113</b>	<b>0.164</b>	<b>0.287</b>	<b>0.564</b>
128	0.109	0.158	0.285	0.561

Table 6: Ablation of codebook size  $K$ . We train three proxy models with codebook sizes of 32, 64 and 128 and generate three sets of pseudo labels. Results are evaluated on the final radiology report on the MIMIC dataset.  $k = 64$  achieves the best result, which aligns with our empirical understanding of radiology reports.

## 6 LIMITATIONS

The evaluation results may not represent real-world performance because the images are mostly collected from the intensive care unit (ICU) of hospitals. It creates a bias in terms of the category and severity of the diseases our model was exposed to.

## 7 CONCLUSION

In this paper, we introduce two methods to enhance the learning of long paragraph generation when factual correctness is crucial. We identify the label ambiguity issue and address it by a Concept-to-Content (QuantRad) approach. To facilitate QuantRad, we propose Generative Proxy Modeling to learn sentence concept classification without labels. Furthermore, we validate the use of GPT-4 as a state-of-the-art metric for evaluating radiology reports. We hope our proposed modeling and evaluation methods will advance future research in image-to-text generation for correctness-sensitive scenarios.

## 8 BROADER IMPACT

Our study contributes to the broader field of image-to-text generation. When applied in the medical domain, the proposed approach has the potential to reduce radiologists' workload and improve patient care by providing efficient diagnostic reports. However, it is essential to establish clear guidelines and safeguards to ensure the responsible use of AI in healthcare settings, as the reliability and safety of such systems have not been sufficiently tested in clinical environments. Specifically, AI-generated content should not be used to replace expert judgments but rather as a supplement to ensure safety. Furthermore, providing AI-generated medical reports directly to patients who have limited medical knowledge may lead to misunderstandings and cause risky situations.

The proposed machine learning methods, including Generative Proxy Modeling and the Concept-to-Content approach for long text generation, are generic and can be applied beyond the medical domain. These methods can be employed in other accuracy-critical scenarios, such as legal and finance, where label ambiguity exists due to insufficient training data not covering the full variations of correct outputs. The successful application of these methods in various domains could lead to increased efficiency and improved decision-making processes. Meanwhile, the risks discussed in this section also apply to these other domains. As with any AI system, it is crucial to thoroughly validate and monitor the performance of these methods in real-world settings to mitigate risks and ensure their safe and responsible use.

## REFERENCES

- 540  
541  
542 Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel  
543 Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language  
544 model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736,  
545 2022.
- 546 Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with  
547 improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin,  
548 and Clare Voss (eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation*  
549 *Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan,  
550 June 2005. Association for Computational Linguistics. URL <https://aclanthology.org/W05-0909>.
- 551  
552 Brent Burbridge. *Undergraduate diagnostic imaging fundamentals*. Distance Education Unit,  
553 University of Saskatchewan, 2017.
- 554 Daniel J Cao, Casey Hurrell, and Michael N Patlas. Current status of burnout in canadian radiology.  
555 *Canadian Association of Radiologists Journal*, 74(1):37–43, 2023.
- 556  
557 Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey  
558 Zagoruyko. End-to-end object detection with transformers. In *European conference on computer*  
559 *vision*, pp. 213–229. Springer, 2020.
- 560  
561 Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng  
562 Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Hassan Awadalla, Julia Gong,  
563 Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu-Hsin Wei, Tristan  
564 Naumann, Muhao Chen, Matthew P. Lungren, Serena Yeung-Levy, Curtis P. Langlotz, Sheng  
565 Wang, and Hoifung Poon. Towards a clinically accessible radiology foundation model: open-access  
566 and lightweight, with automated evaluation. 2024. URL <https://api.semanticscholar.org/CorpusID:268379244>.
- 567  
568 Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and  
569 Jingjing Liu. UNITER: universal image-text representation learning. In *ECCV*, 2020a.
- 570  
571 Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via  
572 memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in*  
573 *Natural Language Processing*, November 2020b.
- 574  
575 Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave  
576 Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis,  
577 et al. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint*  
*arXiv:2401.12208*, 2024.
- 578  
579 Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human  
580 evaluations? *arXiv preprint arXiv:2305.01937*, 2023.
- 581  
582 Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam  
583 Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm:  
Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
- 584  
585 Hyunhee Chung, Kyung Ho Park, Taewon Seo, and Sungwoo Cho. Phantom of benchmark dataset:  
586 Resolving label ambiguity problem on image recognition in the wild. In *Proceedings of the*  
*IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 444–453, 2023.
- 587  
588 Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory trans-  
589 former for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and*  
*Pattern Recognition*, pp. 10578–10587, 2020.
- 590  
591 Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas  
592 Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An  
593 image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint*  
*arXiv:2010.11929*, 2020.

- 594 Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu,  
595 Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of  
596 training end-to-end vision-and-language transformers. *arXiv preprint arXiv: 2111.02387*, 2021.  
597
- 598 Dhakshinamoorthy Ganeshan, Andrew B Rosenkrantz, Roland L Bassett Jr, Lori Williams, Leon  
599 Lenchik, and Wei Yang. Burnout in academic radiologists in the united states. *Academic radiology*,  
600 27(9):1274–1281, 2020.
- 601 Jiuxiang Gu, Jianfei Cai, Gang Wang, and Tsuhan Chen. Stack-captioning: Coarse-to-fine learning  
602 for image captioning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32,  
603 2018a.
- 604 Jiuxiang Gu, Shafiq Joty, Jianfei Cai, and Gang Wang. Unpaired image captioning by language  
605 pivoting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 503–519,  
606 2018b.
- 607
- 608 Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked  
609 autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer  
610 vision and pattern recognition*, pp. 16000–16009, 2022.  
611
- 612 Stephanie L Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Mercy Ranjit, Anton  
613 Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, et al.  
614 MAIRA-1: A specialised large multimodal model for radiology report generation. *arXiv preprint  
615 arXiv:2311.13668*, 2023.
- 616 Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik  
617 Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest  
618 radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI  
619 conference on artificial intelligence*, volume 33, pp. 590–597, 2019.  
620
- 621 Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren,  
622 Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available  
623 database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019a.
- 624 Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng,  
625 Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a  
626 large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*,  
627 2019b.
- 628 Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng,  
629 Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible  
630 electronic health record dataset. *Scientific data*, 10(1):1, 2023.  
631
- 632 Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.  
633
- 634 Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolu-  
635 tion or region supervision. In *ICML*, 2021.
- 636 Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan  
637 Naumann, Hoifung Poon, and Jianfeng Gao. LLava-Med: Training a Large Language-and-Vision  
638 Assistant for Biomedicine in One Day. *Advances in Neural Information Processing Systems*, 36,  
639 2024.
- 640 Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-  
641 training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*,  
642 2023.  
643
- 644 Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. @inproceedingsli2022exploring, ti-  
645 tle=Exploring plain vision transformer backbones for object detection, author=Li, Yanghao and  
646 Mao, Hanzi and Girshick, Ross and He, Kaiming, booktitle=European conference on computer  
647 vision, pages=280–296, year=2022, organization=Springer . In *European conference on computer  
vision*, pp. 280–296. Springer, 2022.

- 648 Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization*  
649 *Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.  
650 URL <https://aclanthology.org/W04-1013>.  
651
- 652 Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr  
653 Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–*  
654 *ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings,*  
655 *Part V 13*, pp. 740–755. Springer, 2014.
- 656 Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter  
657 Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. In *Machine*  
658 *Learning for Healthcare Conference*, pp. 249–269. PMLR, 2019.  
659
- 660 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv*  
661 *preprint arXiv:2304.08485*, 2023a.
- 662 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in*  
663 *neural information processing systems*, 36, 2024.  
664
- 665 Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg  
666 evaluation using gpt-4 with better human alignment, may 2023. *arXiv preprint arXiv:2303.16634*,  
667 2023b.
- 668 Hector Lopez-Cardona. Chest x-ray review: Abcde. *Radiopaedia*, 2023. URL <https://radiopaedia.org/articles/chest-x-ray-review-abcde?lang=us>.  
669  
670
- 671 Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P Langlotz, and Dan Jurafsky. Improving  
672 factual completeness and consistency of image-to-text radiology report generation. *arXiv preprint*  
673 *arXiv:2010.10042*, 2020.
- 674 Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. Grit: Faster and better image  
675 captioning transformer using dual visual features. In *European Conference on Computer Vision*,  
676 pp. 167–184. Springer, 2022.  
677
- 678 Aaron Nicolson, Jason Dowling, and Bevan Koopman. Improving chest X-ray report generation by  
679 leveraging warm starting. *Artificial Intelligence in Medicine*, 144:102633, 2023. ISSN 0933-3657.  
680 doi: <https://doi.org/10.1016/j.artmed.2023.102633>. URL <https://www.sciencedirect.com/science/article/pii/S0933365723001471>.  
681
- 682 Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael  
683 Krauthammer. Progressive transformer-based generation of radiology reports. *arXiv preprint*  
684 *arXiv:2102.09777*, 2021.  
685
- 686 OpenAI. Gpt-4 technical report, 2023.  
687
- 688 Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic  
689 evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.),  
690 *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp.  
691 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.  
692 doi: 10.3115/1073083.1073135. URL <https://aclanthology.org/P02-1040>.
- 693 Jay R Parikh, Darcy Wolfman, Claire E Bender, and Elizabeth Arleo. Radiologist burnout according  
694 to surveyed radiology practice leaders. *Journal of the American College of Radiology*, 17(1):78–81,  
695 2020.
- 696 Sai Rajeswar, Pau Rodriguez, Soumye Singhal, David Vazquez, and Aaron Courville. Multi-label  
697 iterated learning for image classification with label ambiguity. In *Proceedings of the IEEE/CVF*  
698 *Conference on Computer Vision and Pattern Recognition*, pp. 4783–4793, 2022.  
699
- 700 Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi  
701 Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF*  
*international conference on computer vision*, pp. 82–91, 2021.

- 702 Robin Smithuis and Delden Otto. Chest x-ray - basic interpretation. *Radiology As-*  
703 *stant*, 2022. URL [https://radiologyassistant.nl/chest/chest-x-ray/](https://radiologyassistant.nl/chest/chest-x-ray/basic-interpretation)  
704 [basic-interpretation](https://radiologyassistant.nl/chest/chest-x-ray/basic-interpretation).
- 705
- 706 Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-  
707 guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer*  
708 *Vision and Pattern Recognition*, pp. 7433–7442, 2023.
- 709 Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée  
710 Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and  
711 efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- 712
- 713 Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang,  
714 Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards Generalist Biomedical AI.  
715 *NEJM AI*, 1(3):AIoa2300138, 2024.
- 716 Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image  
717 description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern*  
718 *recognition*, pp. 4566–4575, 2015.
- 719 Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural  
720 image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern*  
721 *recognition*, pp. 3156–3164, 2015.
- 722
- 723 Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and  
724 Jie Zhou. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*,  
725 2023.
- 726 Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and  
727 Lijuan Wang. UFO: A unified transformer for vision-language representation learning. *arXiv*  
728 *preprint arXiv:2111.10023*, 2021.
- 729
- 730 Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu,  
731 and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv*  
732 *preprint arXiv:2205.14100*, 2022a.
- 733 Jun Wang, Abhir Bhalerao, and Yulan He. Cross-modal prototype driven network for radiology report  
734 generation. In *European Conference on Computer Vision*, pp. 563–579. Springer, 2022b.
- 735
- 736 Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. Tienet: Text-image  
737 embedding network for common thorax disease classification and reporting in chest x-rays. In  
738 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9049–9058,  
739 2018.
- 740 Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. Delving into masked autoencoders for  
741 multi-label thorax disease classification. In *Proceedings of the IEEE/CVF Winter Conference on*  
742 *Applications of Computer Vision*, pp. 3588–3600, 2023.
- 743 Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich  
744 Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual  
745 attention. In *International conference on machine learning*, pp. 2048–2057. PMLR, 2015.
- 746
- 747 Yanzhi Yi, Hangyu Deng, and Jinglu Hu. Improving image captioning evaluation by considering  
748 inter references variance. In *Proceedings of the 58th Annual Meeting of the Association for*  
749 *Computational Linguistics*, pp. 985–994, 2020.
- 750 Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with  
751 semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern*  
752 *recognition*, pp. 4651–4659, 2016.
- 753
- 754 Jerrold H Zar. Spearman rank correlation. *Encyclopedia of biostatistics*, 7, 2005.
- 755

## A ETHICAL CONSIDERATIONS

Medical datasets often contain sensitive patient information. To ensure the ethical use of such data, this study adheres to strict guidelines. All participants who accessed the MIMIC-CXR dataset, including the authors and radiologists involved in this research, completed the required onboarding process through PhysioNet<sup>2</sup>. For the IU X-ray dataset, we complied with the license<sup>3</sup>.

To maintain compliance with PhysioNet’s policy on the use of large language model APIs during the automatic evaluation, we utilized a secure, private, in-house deployment of GPT-4. This approach guarantees that no sensitive information is shared with external parties.

Furthermore, to protect patient privacy, X-ray images presented in this paper were carefully selected from open, compliance-free sources, ensuring that no identifiable patient information is disclosed.

## B QUALITATIVE RESULT OF CPM AND CGM

We present a qualitative result of our method in Figure 6. We show the process that medical concept classes  $c_i$  are generated by CPM, and are then converted to sentences in CGM, ensuring clinical accuracy compared with the ground truth. In the illustrated case, our method adheres to a clinically accurate structure for medical reporting.

	Human Annotation	C2C (ours)	Existing Work
	Single AP upright portable view of the chest was obtained. The patient is rotated to the right. The patent esophageal stent has migrated in position with the superior portion now projecting over the right lung apex. Also, since the prior study, there has been development of significant opacity projecting over the right hemithorax which may be due to a combination of pleural effusion and consolidation. The left lung is clear. The cardiac silhouette is not enlarged. ED aware at the time of the dictation.	CPM: [ $c_{15}$ , $c_{29}$ , $c_{39}$ , $c_{28}$ , $c_{100}$ , $c_{43}$ , $c_{52}$ ] CGM: A single portable frontal upright view of the chest was obtained. There is a large opacity in the right mid-to-lower hemithorax with air bronchograms. The left lung is relatively clear. The left heart border is partially obscured and there is a right basilar opacity. The left lung is clear. There is a possible small right pleural effusion. There is no pneumothorax.	As compared to the previous radiograph the patient has undergone right thoracocentesis. The extent of the right pleural effusion has substantially decreased. There is no evidence of pneumothorax. The right lung is unremarkable. Unchanged appearance of the cardiac silhouette.

Figure 6: Qualitative results from our method, C2C. CPM generates discrete medical concept classes. Texts highlighted in the same color indicate matching meanings. Red-colored text in the existing work shows a factually incorrect prediction, despite using the same words with the ground truth, e.g., “pleural effusion”.

## C QUALITATIVE EXAMPLES OF LABEL AMBIGUITY AS A SIGNIFICANT ISSUE

Take the human annotation in Figure 6 as an example, the radiology report is still correct when the order of some sentences are shuffled. For example, pleural effusion (highlighted in yellow) and left lung (highlighted in green) have no correct orders. Besides, diagnosis of healthy findings, such as the left lung (highlighted in green) are sometimes omitted in other ground truth annotations. As the loss function demands an exact match with the ground truth, the learning process become less efficient in leveraging the training data.

Our approach first produce a sequence of the medical concepts with CPM and then generate each sentence based on CPM’s outputs as a condition. In the example, CPM’s output doesn’t perfectly fit the ground truth, but the final output from CGM is still considered correct in real applications. CPM captures the ambiguity making the training of CGM ambiguity-free.

## D EXAMINATION OF LEARNED MEDICAL CONCEPTS

We compared with the Generative Proxy Model (GPM), which leverages text generation as a proxy task for learning medical concept classification per sentence. Rather than manually curating a list

<sup>2</sup>MIMIC-CXR on PhysioNet: <https://physionet.org/content/mimic-cxr/2.0.0/>

<sup>3</sup>IU X-ray dataset license: <https://creativecommons.org/licenses/by-nc-nd/4.0/>

810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863

(a) Concept class 6.  
Empirical meaning: "medical devices, tubes"

- $c = 6$
- "There is an enteric tube which extends to the distal esophagus and must be advanced."
  - "A nasogastric tube courses into the stomach, with the last side port at the GE junction."
  - "A right-side PICC line appears to terminate in the mid SVC."
  - "The left PICC line ends in the mid SVC, unchanged."
  - "An NG tube ends in the stomach and its last side port is near the EG junction."
  - "A nasoenteric tube enters the stomach."
  - "Tip of feeding tube continues to terminate in the proximal stomach."

(b) Concept class 15.  
Empirical meaning: "views of the x-ray"

- $c = 15$
- "PA and lateral views of the chest provided."
  - "Frontal and lateral views of the chest."
  - "AP and lateral views of the chest."
  - "Semi-erect AP and lateral images of the chest were obtained."
  - "Frontal and lateral views of the chest with dedicated views of the ribs were obtained for a total of 7 images."

(c) Concept class 28.  
Empirical meaning: "heart conditions"

- $c = 28$
- "The heart is normal in size."
  - "The heart size is within normal limits."
  - "The heart size is normal."
  - "The heart size is top normal."
  - "Cardiac silhouette is top normal for technique."
  - "Cardiac silhouette is not enlarged."
  - "Cardiac silhouette size is normal."

Figure 7: Qualitative results of the Generative Proxy Model for sentence medical concept classification. We select three classes as examples to illustrate that the codebook indices  $c$  can roughly cluster sentences of the same medical concept.  $c$  is sent to the text decoder (CGM) to generate a sentence for the corresponding medical concept. Details (e.g. position, severity) of the medical condition will be determined in CGM by attending to the image features, instead of provided by  $c$ .

of medical concepts, a task that can be challenging in achieving comprehensive coverage, we allow the Codebook to identify soft concept classes in a learnable manner. This enables us to set a hyperparameter  $K$  for the codebook size to determine the granularity of the classes. Following empirical examination after training, we observe that the discrete classes, determined by the codebook indices  $c$ , can approximately cluster sentences by their meaning. We present qualitative results in Figure 7.

## E ADDITIONAL ANALYSIS OF HUMAN EVALUATION

We provide human evaluation in Table 2 and Table 3 to validate the GPT-4 based metric and the performance of our method. In this section, we offer a detailed analysis.

### E.1 VARIANCE AMONG RADIOLOGISTS AND MODELS

While the GPT-4 based metric statistically outperforms  $n$ -gram metrics by a significant margin, we find it varies for different models.

As shown in Table 7,  $n$ -gram metrics tend to be more robust on XPRONET Wang et al. (2022b) than our method, C2C. This variation is likely attributable to each model’s methodology. XPRONET, with its prototype-driven approach, generally yields more word overlap with the ground truth. Conversely, for C2C, we occasionally observe a negative correlation between  $n$ -gram metrics and human evaluation. These instances could indicate overfitting: the output might have more overlapping words, which favors  $n$ -gram metrics, but the meaning is incorrect.

In connection to Section 1, where we discussed the diverse styles of radiologists, we observe further variations in the evaluation styles of radiologists. For instance, Radiologist R3 demonstrates more extreme ratings.

(a) Spearman Correlation					
	BLEU-4	METEOR	ROUGE-L	CIDEr	GPT-4
<b>R1/XPro</b>	0.192	0.280	0.265	0.292	<b>0.544</b>
<b>R1/C2C</b>	-0.210	-0.130	-0.123	-0.003	<b>0.184</b>
<b>R2/XPro</b>	0.071	0.256	0.113	0.019	<b>0.434</b>
<b>R2/C2C</b>	0.023	0.082	-0.008	0.019	<b>0.377</b>
<b>R3/XPro</b>	<b>0.454</b>	0.427	0.389	0.362	0.413
<b>R3/C2C</b>	0.247	0.427	0.288	0.132	<b>0.832</b>

(b) Kendall-Tau Correlation					
	BLEU-4	METEOR	ROUGE-L	CIDEr	GPT-4
<b>R1/XPro</b>	0.155	0.209	0.217	0.217	<b>0.417</b>
<b>R1/C2C</b>	-0.158	-0.102	-0.093	-0.007	<b>0.151</b>
<b>R2/XPro</b>	0.048	0.212	0.084	0.022	<b>0.348</b>
<b>R2/C2C</b>	0.017	0.068	-0.004	0.017	<b>0.305</b>
<b>R3/XPro</b>	<b>0.334</b>	0.313	0.287	0.253	0.312
<b>R3/C2C</b>	0.191	0.324	0.207	0.082	<b>0.718</b>

Table 7: Spearman and Kendall-Tau correlations between metrics and human evaluation. Each row shows the correlation analysis for the data provided by one radiologist for one method. R1, R2, R3 represent three radiologists. XPro and C2C are the two methods being evaluated. These results demonstrate variability with both radiologists and the method being evaluated.

## F ADDITIONAL ANALYSIS OF GPT-4 EVALUATION

### F.1 INTERPOLATION OF MEDICAL TEXT

Given two sentences in medical texts, the statistics of overlapping words do not reliably evaluate the alignment of their meaning. In Table 1, we present two test cases that challenge the  $n$ -gram metrics, but GPT-4 can be effectively prompted to make the correct judgement. We include the prompt and GPT’s output in Figure 8.

918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971

Below are two sentences from a Chest X-ray radiology report. Reply “correct” if you think the reference and the prediction sentences have the same meaning. Reply “incorrect” if the reference and the prediction have different meaning. Reply “NAN” if the two sentences are not related. Make sure you apply medical knowledge and try to be as correct as possible, and explain why.

Reference:

[Reference]

Prediction:

[Prediction]

Reference:

The heart size is top normal.

Prediction:

The heart size is mildly enlarged.

GPT-4 Response :

incorrect

Explanation:

The phrase “top normal” suggests that the heart size is at the upper limit of what is considered normal but not exceeding it. On the other hand, “mildly enlarged” indicates that the heart size is slightly beyond the normal range. These are two different clinical observations where “top normal” represents the highest size within the normal range, and “mildly enlarged” represents a size just outside the upper limit of the normal range, signifying a possible abnormal enlargement. Therefore, the meanings of the two sentences are different in a clinical context.

Reference:

The heart size is top normal.

Prediction:

Borderline size of the cardiac silhouette.

GPT-4 Response :

correct

Explanation:

In radiological terms, “top normal” implies that the heart size is at the upper limit of the normal range, while “borderline” suggests that the size is at the boundary between normal and potentially abnormal. Both phrases indicate that the heart is neither clearly within normal size parameters nor clearly enlarged but is at the high end of what would be considered normal. Thus, in the context of a radiology report, the two sentences can be interpreted to have the same meaning, signaling that the heart size is at the upper threshold of normal.

Figure 8: GPT-4 understands medical text. The three blocks of text show the prompt used to ask GPT-4 whether two sentences align in their meanings, and two test cases from Table 1, where the  $n$ -gram metrics fail but GPT-4 is correct.

972 F.2 RANDOMNESS IN GPT-4'S RESPONSE  
973

974 There exists a degree of randomness in the evaluation result from GPT-4. For the same input, the  
975 output rating from GPT-4 may vary when we call the API multiple times. Existing work in natural  
976 language processing Liu et al. (2023b) suggests requiring GPT-4 to return 20 candidate responses  
977 and take the average. We find that the average rating on a test set is statistically stable for medical  
978 reports. We think the reason is due to the extra length in the medical texts.

979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1020  
1021  
1022  
1023  
1024  
1025