# MedQPA-Gen: Medical Question Proposing and Answering for Report Generation

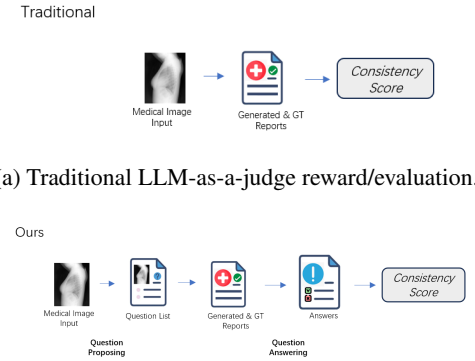
**Weijie Liang**[1*], **Xiyue Zhu**[1*†], **Ruike Zhu**[1], **Chenhao Li**[1], **Cheng Tang**[1], **Zhiyu Liu**[1], **Zhihua Gong**[1], **Shirui Luo**[2], **Yudu Li**[1], **Volodymyr Kindratenko**[1,2]
[1]University of Illinois Urbana-Champaign (UIUC), [2]National Center for Supercomputing Applications (NCSA)
**Correspondence:** xiyuez2@illinois.edu

## Abstract

Medical report generation from medical images is a vital AI task that helps doctors with diagnosis and marks a significant step toward creating general AI-powered medical systems. However, previous methods either fail to optimize factual accuracy or heavily depend on expert preference data. To overcome these challenges, we propose MedQPA, an automatic and generalizable report evaluation technique that uses question proposing and answering to enable controllable, structured reasoning grounded in medical domain knowledge and the factual correctness of the report. Additionally, we design MedQPA-Gen, a medical report generation pipeline that maximizes the MedQPA score through prompt engineering and reinforcement learning with MedQPA as a reward signal. We demonstrate that MedQPA is an accurate evaluation metric that closely correlates with human preferences. More importantly, MedQPA-Gen achieves higher human preference scores and better performance on downstream tasks. We open-source code at this repo https://github.com/MedQPA-gen/MedQPA-gen.

## 1 Introduction

Automated medical report generation from imaging modalities such as X-ray and CT has the potential to substantially improve diagnostic workflows and clinical efficiency (Demner-Fushman et al., 2015; Zhang et al., 2024). However, ensuring factual correctness remains a critical challenge, as even minor inaccuracies—such as omitted findings or hallucinated diagnoses—can lead to serious patient safety risks (Ostmeier et al., 2024). Although recent advances have produced powerful medical report generation models (Saab et al., 2024; Singhal et al., 2025; Jin et al., 2021), existing training paradigms still struggle to explicitly target factual accuracy.




(a) Traditional LLM-as-a-judge reward/evaluation.

(b) Our QPA process enables better reasoning in evaluation.

Figure 1: Comparison between traditional LLM as a judge and our QPA process. Our QPA method can better align with human preference when evaluating reports and can help model generate better reports when used as a reward signal in RLHF.

Supervised fine-tuning (SFT) mainly optimizes surface-level similarity to reference reports, often yielding fluent but factually incorrect statements. Reinforcement learning from human feedback (RLHF) partially addresses this issue but relies heavily on costly and difficult-to-scale human annotations. Meanwhile, using large language models as evaluators ("LLM as a Judge") provides automation but often lacks reliable, domain-specific reasoning and can exhibit inconsistent or self-reinforcing judgments in medical settings.

To address these limitations, we propose MedQPA, a Question Proposing and Answering (QPA) framework that evaluates medical reports through structured, domain-informed reasoning. Unlike surface-level metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), or unconstrained LLM-based evaluators (Ostmeier et al., 2024), MedQPA probes report content by generating targeted questions and verifying whether answers are supported by image evidence and medical knowledge. This design aligns evaluation more

[*]These authors contributed equally to this work.
[†]Corresponding author.



*Findings of the Association for Computational Linguistics: ACL 2026*, pages 43120–43132

closely with clinical expert judgment. Importantly, MedQPA is dual-purpose: it can guide generation through reflective prompting and serve as a reward signal in reinforcement learning, enabling models to explicitly optimize for factual correctness rather than textual similarity.

Recent NLP research has shown that reinforcement learning is most effective when driven by reliable, reasoning-aware reward functions (Xu, 2024; Zhang et al., 2025b). Similarly, medical QA systems such as Med-PaLM (Singhal et al., 2025) demonstrate that structured question answering can elicit high-quality reasoning. Motivated by these advances, we integrate MedQPA into an agentic reinforcement learning framework, where reasoning is decomposed into modular, goal-directed steps (Bandi et al., 2025). This allows the model to internalize clinically grounded reasoning processes during training, resulting in reports that are both factually accurate and logically consistent. 234234 Empirically, MedQPA shows stronger alignment with human evaluation on both the Indiana X-ray and CT-RATE datasets, outperforming existing automatic metrics in capturing semantic correctness. Moreover, when used to guide reinforcement learning, MedQPA consistently improves report generation quality without relying on additional human feedback. Beyond report-level evaluation, we further demonstrate that these improvements translate to downstream clinical benefits. On the CheXpert dataset (Irvin et al., 2019), reports generated under MedQPA supervision lead to improved multi-label disease classification performance over 14 thoracic conditions, reflecting real-world clinical workflows where diagnostic decisions are derived from written reports (Boag et al., 2020). Together, these results show that explicitly optimizing factual correctness in report generation yields practical clinical value beyond surface-level text quality.

## 2 Related Work

**Medical report evaluation** Medical report evaluation has evolved from early rule-based or template-based methods (Irvin et al., 2019) to large language model (LLM)-based evaluators trained on medical corpora (Saab et al., 2024). While LLM-based approaches improve fluency and coverage, prior work shows that they often fail to reliably assess factual correctness and clinical relevance (Fan et al., 2024), largely due to the lack of explicit reasoning structures. Traditional automatic metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) remain widely used but mainly capture surface-level overlap, making them insufficient for evaluating clinical validity. In contrast, MedQPA introduces a question-based reasoning framework that explicitly verifies factual consistency using domain-informed evaluation.

**Medical report generation** Medical report generation aims to produce clinically meaningful narratives from medical images to assist radiologists. Most existing methods rely on supervised fine-tuning (SFT) of vision–language models using token-level objectives (Izhar, 2025), which do not explicitly enforce diagnostic correctness and often lead to hallucinations or omissions. To address this limitation, recent work explores reinforcement learning (RL) to better align generation with clinical objectives (Zhang et al., 2025a; Wang, 2025; Jin, 2024). However, progress is constrained by the lack of reliable, clinically grounded evaluation and reward signals.

**Reinforcement learning for LLMs** Reinforcement learning methods such as RLHF, Direct Preference Optimization (DPO), and Group Relative Policy Optimization (GRPO) have been widely adopted to align LLMs with human preferences (Rafailov et al., 2023; Mroueh, 2025). The effectiveness of these methods critically depends on the quality of the reward model. Recent studies propose agent-based evaluation frameworks, such as Agent-as-a-Judge (Zhuge et al., 2024), where one agent evaluates another's reasoning process. Building on this direction, MedQPA serves as a rule-guided, reasoning-aware reward model, enabling more stable and scalable RL training for complex tasks like medical report generation.

**Medical question answering** Medical question answering (Med-QA) demonstrates the strong reasoning capabilities of LLMs in the healthcare domain. Models such as Med-PaLM 2 (Singhal et al., 2025) and Med-Gemini (Saab et al., 2024) achieve state-of-the-art performance on the MedQA benchmark (Jin et al., 2021), highlighting the effectiveness of domain-specific fine-tuning and structured reasoning. These successes motivate the use of QA-style reasoning as a foundation for both evaluation and optimization in medical report generation.

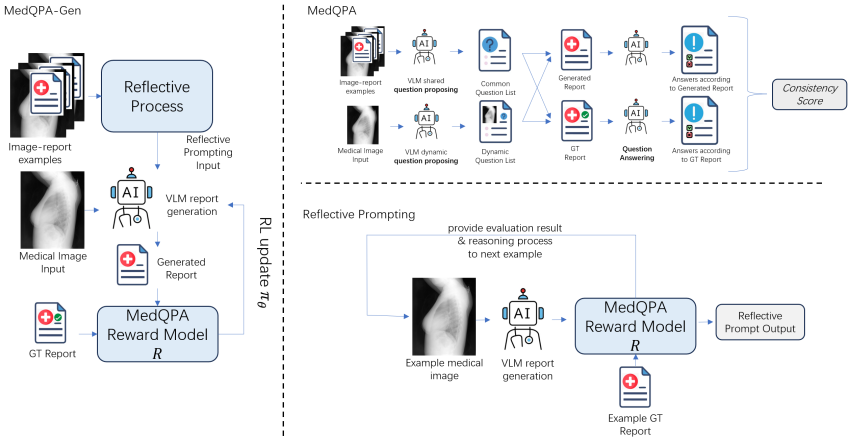



Figure 2: Overview of the generation pipeline and reinforcement learning pipeline used in our method.

## 3 Method

**Overview** We propose MedQPA, an automatic and generalizable medical report evaluation technique that utilizes question proposing and answering to facilitate reasoning related to medical domain knowledge and the factual correctness of the report. Additionally, we design MedQPA-Gen, a medical report generation pipeline that maximizes the MedQPA score through reflective prompting, as well as reinforcement learning with MedQPA as a reward signal.

### 3.1 MedQPA for Report Evaluation

**Overview** MedQPA evaluates medical reports by proposing and answering questions. Given ground truth reports and generated reports, MedQPA aims to output a score to evaluate the similarity between the reports. MedQPA first proposes questions during the **question proposing** stage. Then, during the **question answering** stage, MedQPA answers the proposed questions based on the generated and ground truth reports to give the final evaluation.

**Question Proposing** We now introduce how MedQPA proposes questions to structure the evaluation and generation of medical reports. MedQPA adopts a two-level question proposing strategy that consists of *shared questions* and *dynamic questions*, which together enable both consistency and flexibility in reasoning.

Shared questions are proposed by summarizing patterns across multiple ground truth reports within a dataset. Medical reports typically follow a relatively standardized structure and repeatedly describe a set of clinically important organs and anatomical regions. As a result, shared questions capture common and essential aspects that are expected to be addressed across most reports, such as findings related to major airways, lung fields, pleural regions, mediastinal structures, and the cardiac silhouette. These shared questions provide a stable and interpretable basis for both report evaluation and report generation, encouraging coverage of core clinical content.

In addition to shared questions, MedQPA also proposes dynamic questions to account for case-specific variations. Unlike shared questions, dynamic questions are proposed independently for each individual image–report pair, focusing on abnormalities, localized findings, or unique diagnostic cues that may not be present across the entire dataset. This design allows the evaluation process to remain flexible and adaptive, ensuring that medically salient but uncommon findings are properly assessed. By jointly reasoning over shared and dynamic questions, MedQPA balances structural consistency with instance-level specificity, enabling more accurate and semantically grounded evaluation of generated reports.

**Question Answering for Report Evaluation** When doing report evaluation, given the ground truth report, the generated report, and the question list, MedQPA will answer questions in the question list and get the final accumulated consistency score. More specifically, MedQPA first answers the questions based on the ground truth report to get the ground truth answer list, $A = \{a_i\}$. Each answer $a_i$ can be positive, negative, or not mentioned. MedQPA then answers the same set of questions using the generated report to get the draft answer list, $A = \{\hat{a}_i\}$. We then calculate the final score based on the consistency of each answer in the lists with a rule-based scoring method. For each question and answer in the shared and dynamic question list, the scoring follows these rules:

• If $a_i$ and $\hat{a}_i$ report the same outcome, we assign 2 points.

• If $a_i$ mentions the question but $\hat{a}_i$ does not, we assign 1 point, due to missing information.

• If $\hat{a}_i$ mentions the question but $a_i$ does not, we assign 0.5 points due to redundant and potentially fake information, which can be more harmful than missing information.

• If $a_i$ and $\hat{a}_i$ provide contradictory outcomes (one is positive and the other is negative), we assign 0 points due to fake information.

Given a total of $N$ questions, the maximum score is $2N$. This scheme rewards accurate reproduction of clinical findings while still giving partial credit for partial coverage.

### 3.2 MedQPA-Gen for Report Generation

**Overview** Given a medical image (e.g., chest X-ray or CT slice), medical report generation aims to obtain a free-form radiology report describing clinically relevant findings in natural language. MedQPA-Gen adopts a standard pretrained vision-language architecture, which performs cross-modal reasoning and auto-regressive report generation.

Built upon the MedQPA evaluation framework, MedQPA-Gen improves report generation quality by optimizing the MedQPA score. First, through reflective prompting, the generator is explicitly informed how the generated reports will be evaluated. To do that, the model is provided with multiple report examples that consist of an image, a ground truth report, a generated report, and the evaluation process of MedQPA. Second, MedQPA is further used as a reward model in reinforcement learning, where the generator is optimized to produce reports that achieve higher MedQPA scores. By integrating question-based reasoning into both evaluation and training, MedQPA-Gen produces reports that are better aligned with the MedQPA metric and exhibit improved factual correctness and clinical relevance.

#### 3.2.1 Reflective Prompting

To integrate MedQPA into the report generation process, we design a reflective prompting strategy that allows the generation model to iteratively improve by conditioning on prior evaluation feedback.

Concretely, given an image–report example, the model first generates a report based on the input image. This generated report is then evaluated by MedQPA, which produces structured reasoning in the form of proposed questions, intermediate answers, and a final assessment of factual correctness. Instead of discarding this evaluation, we retain the full reasoning trace produced by MedQPA. When moving to the next referencing example, the generation prompt is combined with previous MedQPA evaluation traces, demonstrating how generated reports were assessed, where errors occurred, and how they were identified through question-driven reasoning.

By repeating this process over multiple examples, the generation model is exposed to a growing set of reflective signals that explicitly link report content to evaluation outcomes. As a result, each subsequent report generation is conditioned not only on the current input image, but also on accumulated reasoning patterns derived from prior successes and failures. This design enables consistent and progressive improvement, as the model implicitly learns to anticipate evaluation criteria, avoid previously identified mistakes, and align its outputs with the reasoning principles encoded in MedQPA. In this way, reflective prompting serves as a lightweight yet effective mechanism for self-refinement, bridging evaluation and generation without introducing additional trainable components.

#### 3.2.2 Reinforcement Learning (RL)

Through MedQPA-informed prompting, we try to optimize the MedQPA score only at inference time. It does not directly update model parameters. To explicitly optimize the report generator toward MedQPA, we further incorporate reinforcement learning (RL) into MedQPA-Gen. In this work, we explore two complementary but independent RL formulations: (1) iterative Direct Preference Optimization (DPO). and (2) Generative Reinforcement Policy Optimization (GRPO).

**DPO:** Direct Preference Optimization (DPO) (Rafailov et al., 2023) is a preference-based fine-tuning method that eliminates the need for an explicit reward model by directly optimizing the policy using preference pairs.

Inspired by previous half-online variants of DPO(Chen, 2024; Xu, 2024), we adopt an **iterative DPO** strategy driven by the LLM-based MedQPA score. Training proceeds in an epoch-wise manner. At the beginning of each epoch, we generate candidate reports for all training samples using the current policy. These generated reports are then evaluated by the MedQPA pipeline to be la-

beled as preferred or non-preferred samples. The model is subsequently trained for one epoch using this newly constructed preference dataset. After the epoch completes, preference data are regenerated using the updated policy, and the process repeats until convergence. The DPO objective is defined as: $\mathcal{L}_{\text{DPO}} = \mathbb{E}_{(x,y^+,y^-)}\Big[\log \sigma\big(\beta\, \Delta s_\theta\big)\Big]$, where $\Delta s_\theta = \log \frac{\pi_\theta(y^+|x)}{\pi_{\text{ref}}(y^+|x)} - \log \frac{\pi_\theta(y^-|x)}{\pi_{\text{ref}}(y^-|x)}$. Here, $x$ denotes the input image and context, $y^+$ and $y^-$ are the preferred and non-preferred reports, $\pi_\theta$ is the policy model, $\pi_{\text{ref}}$ is a frozen reference model, and $\beta$ is a temperature parameter. This loss encourages the policy to assign higher relative likelihood to preferred reports while remaining close to the reference distribution.

Compared to standard offline DPO, this iterative formulation more closely resembles online preference learning and leads to stronger alignment and improved final performance. Compared to GRPO, which directly optimizes scalar rewards, iterative DPO provides a more stable and sample-efficient optimization process by relying on relative preferences rather than high-variance reward signals. As a result, iterative DPO offers a favorable trade-off between alignment quality and training efficiency in our setting.

**GRPO:** In addition to preference-based optimization, we also apply Generative Reinforcement Policy Optimization (GRPO) by directly using the LLM-based MedQPA score as a scalar reward. Specifically, the MedQPA score is normalized to the range $[0, 1]$ and treated as the reward $r(x, y)$ for each generated report $y$ given input $x$. This enables direct optimization toward factual correctness and semantic alignment without requiring human feedback.

The GRPO objective is defined as: $\mathcal{L}_{\text{GRPO}} = -\mathbb{E}_{(x,y)\sim\pi_\theta}\Big[w(x, y)\,\big(r(x, y) - b(x)\big)\Big]$, where $w(x, y) = \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$, $\pi_\theta$ is the policy model, $\pi_{\text{ref}}$ is the frozen reference model, $b(x)$ is a baseline for variance reduction, and $r(x, y)$ is the normalized MedQPA reward. Intuitively, GRPO increases the probability of reports that achieve higher factual correctness scores while maintaining stability through reference regularization.

### 3.3 Downstream Task

We further evaluate the practical utility of our report generation framework by using disease classification as a downstream task, following previous work(Irvin et al., 2019). This task aims to predict a discrete disease label or a set of disease indicators given a generated or ground truth medical report.

We select disease classification tasks for downstream evaluation for several reasons. First, disease classification directly impacts clinical decision-making and diagnostic workflows, making it a meaningful measure of clinical utility. Second, successful classification requires the report to accurately capture and reason about image-specific findings, rather than relying on generic or fluent descriptions, thus serving as a strong test of whether the generated reports encode substantive medical information. Third, these tasks are supported by well-defined ground-truth labels and standard evaluation metrics, enabling reliable and reproducible assessment.

## 4 Experiment

We evaluate both MedQPA for medical report evaluation and MedQPA-Gen for medical report generation in Sec. 4.2. We show that MedQPA can better evaluate medical reports by testing its ability to accurately rank reports with different levels of semantic corruption and its alignment with human preference in Tab. 2 and Tab. 4. We also evaluate MedQPA-Gen using a comprehensive set of metrics and a downstream task in Tab. 3 and Tab. 6. We provide key ablation on reflective prompting and RL in Tab. 4 and Tab. 5. In addition, we include a benchmark study of MedQPA across a range of state-of-the-art general-purpose VLMs in Tab. 1.

### 4.1 Experiment setup

**Datasets.** We conduct experiments on three publicly available medical imaging datasets to evaluate both report generation quality and downstream clinical utility. The Indiana X-ray dataset contains paired chest X-ray images and radiology reports and is widely used for benchmarking medical report generation. CT-RATE consists of chest CT images with corresponding diagnostic reports, providing a complementary imaging modality with richer anatomical detail. For downstream evaluation, we use the CheXpert dataset (Irvin et al., 2019), which provides chest X-ray images annotated with 14 thoracic disease labels and is commonly adopted for multi-label disease classification. There are nearly 4000 training samples and 800 test samples in X-ray dataset. We also select 4000 training samples and 800 test samples from CT-RATE dataset. For

the CheXpert dataset (Irvin et al., 2019), we select 4000 training samples from the Chexpert Plus dataset (Chambon et al., 2024) and 800 samples for testing. Together, these datasets enable evaluation across different imaging modalities, report styles, and clinical tasks.

**Evaluation metrics.** We evaluate our framework from two complementary perspectives: assessing MedQPA as a reasoning-aware evaluation metric, and assessing report generation quality under the MedQPA-gen pipeline.

To evaluate MedQPA as an evaluation metric, we consider two criteria. First, we examine its ability to capture semantic degradation in reports through a noisy report ranking task. Specifically, we systematically inject different levels of noise into ground-truth radiology reports. The injected noise includes factual errors (e.g., incorrect findings), omission of clinically important observations, and semantic inconsistencies. We use a large language model to generate 5 noisy reports with different noise levels within one dialog. Higher noise levels correspond to more severe degradation in factual correctness and clinical coherence. Given a set of reports with known noise levels, each evaluation metric assigns a scalar score to every report, and the reports are ranked accordingly. A reliable metric should rank reports with lower noise (i.e., higher quality) above those with higher noise.

We quantify ranking quality using Spearman rank correlation between the metric-induced ranking and the ground-truth noise-level ordering. Spearman correlation measures the agreement between two ranked lists. Given $n$ reports, let $r_i$ be the rank assigned by a metric and $g_i$ the rank defined by noise level. The Spearman correlation is computed as: $\rho = 1 - \frac{6\sum_{i=1}^{n}(r_i - g_i)^2}{n(n^2-1)}$.

Second, we evaluate alignment with human judgment on the Indiana X-ray dataset, where expert annotations are available. We sample a subset of generated reports and collect pairwise preference judgments from clinical experts, who are asked to choose the report that is more factually correct and clinically appropriate. Each report is evaluated by multiple experts, and we record the percentage of times a report is preferred over alternatives as the human preference score. This provides an external reference for assessing whether automatic metrics, including MedQPA, are consistent with expert evaluation. Overall, we have 8 experts evaluating 400 reports. For each image, we generate four candidate reports from different models and include the corresponding ground-truth (GT) report as a reference. Clinical experts are asked to score each generated report based on its similarity to the GT report in terms of factual correctness and clinical relevance.

For evaluating report generation quality under the MedQPA-gen pipeline, we primarily rely on human preference scores as the main evaluation metric, as they provide an independent and clinically grounded assessment of report quality.

We additionally report the MedQPA score as a complementary metric, which reflects how well the generated reports satisfy the question-based factual criteria defined by MedQPA. Standard automatic metrics such as BLEU, METEOR, ROUGE-L, and BERTScore are also reported for completeness.

**Model architecture and training setup.** For all experiments, we use Qwen-7B (Bai et al., 2025) as the backbone language model for both our method and all baselines to ensure a fair comparison. The input to the model consists of a medical image together with a textual prompt, and the output is a free-form radiology report.

All models are trained using a unified three-stage pipeline. The first stage consists of supervised fine-tuning (SFT) on each dataset using paired image–report data. In the second stage, reinforcement learning is applied with MedQPA using only shared questions, which provides a stable and structured reward signal. The third stage uses a mixture of shared and dynamically proposed questions, enabling more flexible, sample-specific reasoning during training. Across all experiments, we use the same optimizer and backbone configuration. This unified architecture and training scheme allows us to isolate the impact of MedQPA-based supervision and reinforcement learning from other confounding factors.

**Downstream tasks.** To assess whether improvements in report generation translate into clinically meaningful signals, we evaluate our models on downstream disease classification tasks using generated reports as the sole input.

Prior work has shown that evaluating report generation via a downstream clinical task by parsing generated reports into disease labels and comparing them against reference labels, as this label-level evaluation better reflects clinical correctness than surface-level text similarity metrics. (Boag et al., 2020) On the CheXpert dataset, we evaluate

a multi-label classification task over 14 thoracic diseases. Following prior work, we feed the generated reports into Chexpert labeler (Irvin et al., 2019) to get labels. These predicted labels are then compared against ground-truth CheXpert labels.

For both datasets, we compare four model variants: Original (no report generation), SFT, SFT + DPO, and SFT + GRPO. Performance is evaluated using standard classification metrics, including Precision, Recall, and F1 score, aggregated following the official CheXpert evaluation protocol. By fixing the downstream classifier and varying only the quality of the generated reports, this setup allows us to directly measure whether MedQPA-guided training produces reports that encode more accurate and discriminative clinical information.

### 4.2 Experiment results

#### 4.2.1 MedQPA Results for Report Evaluation

We provide the experiment results for MedQPA for report evaluation. We first benchmark existing state-of-the-art vision-language models using MedQPA. We then show that MedQPA is a reasonable metric by testing its ability to rank noise reports and its alignment with human evaluation.

**Benchmark across existing VLMs.** We first benchmark a set of state-of-the-art VLMs using the proposed MedQPA metric, including both open-source and closed-source models. As shown in Table 1, closed-source VLMs such as GPT-4o and Gemini-1.5-Pro consistently outperform strong open-source models like Qwen-72B-VL across most metrics and MedQPA, indicating a persistent performance gap in general-purpose multimodal reasoning for medical report generation. However, the domain-specialized model MedGemma-27B achieves the highest MedQPA score, surpassing all general-purpose models, including closed-source ones. This result highlights the importance of domain-specific modeling in achieving factual correctness in medical contexts.

Importantly, this benchmark provides, to our knowledge, one of the first systematic evaluations of modern VLMs under a structured, question-driven factuality metric. By revealing both the strengths of proprietary models and the advantages of domain specialization, it establishes a strong and meaningful testbed for future research in medical report generation and evaluation.

**MedQPA performance on ranking noisy reports.** We first evaluate whether MedQPA can reliably assess report quality by ranking reports in the Indiana X-ray dataset with controlled levels of injected noise. As shown in Table 2, MedQPA achieves the highest Spearman correlation with noise levels. substantially outperforming traditional metrics such as BLEU, ROUGE, and BERTScore. It also achieves improvement compared to other prompting methods, such as LLM-Judge and GREEN. This result indicates that MedQPA more accurately captures semantic degradation caused by factual and structural noise, rather than surface-level textual similarity. Also, to validate the rationality of the scoring weights of MedQPA, we fix the maximum match reward (2) and the contradiction penalty (0), and perform a grid search over the weights assigned to missing information ($w_m$) and hallucinated information ($w_h$). Results are shown in Table 7 and the adopted configuration (2, 1.0, 0.5, 0) achieves the best correlation under both criteria. We also verify that fake information can be more harmful than missing information, so it should be assigned a lower score.

**Alignment with human evaluation.** Moreover, we examine how well MedQPA aligns with expert human judgment. For each image in the Indiana X-ray dataset, we generate four candidate reports using different generation strategies. A total of eight human evaluators are asked to assess these AI-generated reports and score them according to factual correctness and clinical appropriateness. We collect the average score and use it to rank the reports. This results in a human preference ordering for generated reports.

To quantify alignment, we compute the Spearman rank correlation between the rankings induced by each automatic metric and the corresponding human ranking across all evaluated samples. As shown in Table 2, MedQPA achieves the highest correlation with human judgments, outperforming traditional surface-level metrics as well as LLM-as-a-judge baselines. These results indicate that MedQPA's question-based reasoning process more faithfully captures the criteria used by clinicians when evaluating medical reports, confirming its effectiveness as a human-aligned evaluation metric.

#### 4.2.2 MedQPA-Gen Results for Report generation

We then provide MedQPA-Gen results for report generation. We present results on reflective prompting and RL with human evaluation, MedQPA, and

Table 1: Benchmark comparison of MedQPA across vision-language models.

| **Model** | **BLEU-4** | **METEOR** | **ROUGE-L** | **BERTScore** | **LLM-Judge** | **GREEN** | **MedQPA** |
|---|---|---|---|---|---|---|---|
| Qwen-72B-VL | 0.1935 | 0.1237 | 0.1642 | 0.8236 | 0.4231 | 0.6524 | 0.6731 |
| GPT-4o | 0.2038 | 0.1142 | 0.1526 | 0.8348 | 0.4524 | 0.6872 | 0.6957 |
| Gemini-1.5-Pro | 0.1823 | 0.1349 | 0.1622 | 0.8268 | 0.4357 | 0.6678 | 0.6848 |
| Grok-3 Vision | 0.1838 | 0.1256 | 0.1702 | 0.8272 | 0.4437 | 0.6642 | 0.6908 |
| MedGemma-27B | 0.1724 | 0.1625 | 0.1758 | 0.8527 | 0.4672 | 0.6962 | 0.7027 |

Table 2: Correlation between metrics and noise levels/human preference. Higher values indicate better alignment.

| Spearman correlation $\rho$ | BLEU-4 | METEOR | ROUGE-L | BERTScore | LLM-Judge | Green | MedQPA |
|---|---|---|---|---|---|---|---|
| Noise Levels | 0.1023 | 0.1802 | 0.1923 | 0.3452 | 0.5821 | 0.6212 | 0.6542 |
| Human Preference | 0.1452 | 0.2034 | 0.1748 | 0.4155 | 0.5961 | 0.6135 | 0.6328 |

downstream task performance as metrics.

**Quantitative results.** We present quantitative results for MedQPA-Gen by analyzing two key components of the framework: reflective prompting and MedQPA-guided reinforcement learning. We first study reflective prompting as an ablation within MedQPA-Gen, where the generation process is guided by reasoning-aware reference examples. Specifically, we compare a simple prompt, a one-shot baseline using the same prompt, and reflective prompting with an increasing number of reference samples. As shown in Table 3, reflective prompting consistently outperforms simple and one-shot prompting across most metrics. More importantly, we observe a clear monotonic trend under the MedQPA metric: providing more reasoning-aware reference samples leads to progressively higher MedQPA scores. This result indicates that reflective prompting enables the model to better internalize MedQPA's evaluation criteria and generate reports with improved factual correctness and semantic alignment. In contrast, traditional automatic metrics exhibit relatively smaller variations, suggesting that MedQPA is more sensitive to improvements brought by reasoning-aware generation.

We further evaluate the effectiveness of RL in MedQPA-Gen on the Indiana X-ray and CT-RATE datasets. Both DPO and GRPO consistently improve MedQPA scores over the SFT baseline (Tables 4 and 5), demonstrating that MedQPA provides a stable and effective optimization signal. Compared with DPO, GRPO leverages an online RL strategy and achieves slightly better performance. In contrast, iterative DPO decouples training and evaluation at the epoch level, offering greater flexibility. Given that MedQPA is usually performed

Table 3: Comparison of different prompting strategies for medical report generation. Higher values indicate better performance.

| Method | BLEU-4 | METEOR | ROUGE-L | BERTScore | LLM-Judge | MedQPA |
|---|---|---|---|---|---|---|
| Simple prompt | 0.1762 | 0.1342 | 0.1534 | 0.7321 | 0.3055 | 0.5421 |
| One-shot (simple prompt) | 0.1654 | 0.1359 | 0.1510 | 0.7334 | 0.3082 | 0.6134 |
| MedQPA (1 sample) | 0.1783 | 0.1427 | 0.1732 | 0.7345 | 0.3342 | 0.6345 |
| MedQPA (2 samples) | 0.1832 | 0.1352 | **0.1691** | 0.7425 | 0.3523 | 0.6428 |
| MedQPA (5 samples) | 0.1932 | **0.1603** | 0.1562 | 0.7413 | 0.3647 | 0.6552 |
| MedQPA (10 samples) | **0.2031** | 0.1587 | 0.1687 | **0.7458** | **0.3687** | **0.6723** |

Table 4: Comparison of RL on the X-ray dataset.

| Model | MedQPA | Human Score | Human Win Rate |
|---|---|---|---|
| Original | 0.6652 | 0.42 | 0.10 |
| SFT | 0.6834 | 0.46 | 0.21 |
| SFT + DPO | 0.6945 | 0.52 | 0.35 |
| SFT + GRPO | 0.7028 | 0.51 | 0.34 |

using LLM API calls and training is usually on local GPUs, GRPO can suffer from GPU underutilization while waiting for the MedQPA evaluation during training. To assess whether these gains translate to human judgment, we additionally conduct a human preference study on the Indiana X-ray dataset. We randomly select 100 images and generate four reports per image using the Original, SFT, SFT+DPO, and SFT+GRPO models. A total of 8 clinical evaluators independently score and rank the reports based on factual correctness and clinical appropriateness. The results show that reports generated by MedQPA-guided RL models are more frequently preferred by human experts, confirming that improvements measured by MedQPA align with human evaluation.

In addition to automatic evaluation, we conduct a human preference study on the Indiana X-ray dataset to assess the clinical quality. For each image, we generate reports using four model variants and ask experts to score them. The reported human score is the average expert score normalized to the range $[0, 1]$, while the win rate measures the proportion of cases in which a model's report receives

Table 5: Comparison of RL improvements on the CT-RATE dataset.

| Model | MedQPA Score |
|---|---|
| Original | 0.7132 |
| SFT | 0.7242 |
| SFT + DPO | 0.7354 |
| SFT + GRPO | 0.7385 |

Table 6: CheXpert 14-disease classification results via generated reports and VisualCheXbert.

| Model | Positive F1 | Negative F1 | Uncertain F1 |
|---|---|---|---|
| Original | 0.61 | 0.54 | 0.52 |
| SFT | 0.63 | 0.56 | 0.49 |
| SFT + DPO | 0.65 | 0.55 | 0.46 |
| SFT + GRPO | 0.64 | 0.57 | 0.46 |

the highest human score among all candidates for the same image.

As shown in Table 4, both RL–based models significantly outperform the Original and SFT baselines in human evaluation, achieving substantially higher normalized human scores and win rates. These results provide strong evidence that MedQPA not only aligns well with human judgment but also serves as an effective reward signal for improving the clinical quality of generated medical reports.

**Downstream task performance.** We further assess whether improved report quality benefits downstream clinical classification. As shown in Table 6, MedQPA-guided RL (DPO and GRPO) consistently improves F1 scores over the baselines across positive, negative, and uncertain labels. These results indicate that MedQPA-optimized reports encode more clinically discriminative information, and that gains in MedQPA and human evaluation translate into tangible downstream improvements.

## 5 Discussion and Conclusion

We propose MedQPA, a reasoning-aware evaluation framework for medical report generation based on structured question proposing and answering. By decomposing evaluation into shared and sample-specific questions with explicit scoring rules, MedQPA provides an interpretable and semantically grounded alternative to conventional metrics. Building on this framework, we introduce MedQPA-Gen, which improves report generation through reflective prompting and reinforcement learning using MedQPA as a reward signal.

Experiments on Indiana X-ray and CT-RATE show that MedQPA aligns better with human preference and provides more reliable ranking under semantic corruption. Furthermore, MedQPA-guided training improves downstream clinical classification performance, demonstrating that optimizing factual correctness yields practical benefits.

A key factor behind the effectiveness of MedQPA lies in its structured, reasoning-aware design. Instead of relying on surface-level metrics or unconstrained LLM judgments, MedQPA decomposes evaluation into explicit question–answering steps with clearly defined checkpoints, encouraging the evaluator to verify factual consistency and semantic alignment in a step-by-step manner. This results in a more stable and discriminative evaluation signal that focuses on clinically meaningful semantics rather than superficial textual similarity.

Using an LLM as a rule-guided reward model further amplifies these benefits. Compared to neural reward models that require costly annotation and retraining, the LLM-based reward leverages pretrained semantic knowledge together with explicit scoring rules to produce flexible, interpretable, and task-aligned feedback. As a result, reinforcement learning methods such as DPO and GRPO can effectively optimize report generation toward factual correctness and clinical relevance.

Beyond medical report generation, this reasoning-aware, rule-guided evaluation paradigm is inherently generalizable to domains with well-defined criteria but complex semantic requirements, such as legal writing and scientific reporting. Looking forward, future work may explore adaptive question generation, multi-agent evaluation, and vision-grounded verification, as well as scaling the framework to larger models and broader medical domains.

We hope this work encourages further research on interpretable and reasoning-aware supervision for domain-specific text generation.

## 6 Acknowledgments

This work used the DeltaAI computing resource, supported by the National Science Foundation (award OAC 2320345) and the State of Illinois, and jointly operated by the University of Illinois Urbana-Champaign and the National Center for Supercomputing Applications, with access provided in part by the Illinois Computes project. We thank the human evaluation participants for their time and valuable feedback.

## 7 Limitation

Despite its effectiveness, our approach has several limitations that warrant further investigation. First, our experiments are conducted using relatively small to medium-sized backbone models (e.g., Qwen-7B) to ensure efficient experimentation and controlled analysis. While this setting allows us to clearly isolate the impact of MedQPA-based supervision, larger and more powerful vision–language models may benefit even more from reasoning-aware evaluation and reinforcement learning, potentially leading to stronger absolute performance.

Second, the datasets used in this work are limited in scope, primarily focusing on chest X-ray and CT imaging. Although these datasets are widely used benchmarks, they represent a narrow subset of medical imaging modalities and clinical scenarios. Extending MedQPA to more diverse modalities (e.g., MRI, ultrasound), anatomical regions, and diagnostic tasks would be necessary to fully assess the robustness and generalizability of the framework.

Third, while MedQPA provides a structured and interpretable evaluation signal, its effectiveness still depends on the reasoning capabilities of the underlying LLM. In rare or ambiguous cases, the evaluator may exhibit biases or reasoning errors, which could propagate into the reward signal during training.

Finally, we do not directly compare our method against state-of-the-art medical report generation systems trained with large-scale proprietary data or specialized architectures. Our primary goal is to demonstrate that MedQPA-based supervision consistently improves report quality when applied to a strong but general-purpose backbone. Given the observed gains on Qwen, we expect that integrating MedQPA into state-of-the-art medical report generation models could yield further improvements—a direction we leave for future work.

## References


Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, and Jun et al. Tang. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Ajay Bandi, Bhavani Kongari, Roshini Naguru, Sahitya Pasnoor, and Sri Vidya Vilipala. 2025. The rise of agentic ai: A review of definitions, frameworks, architectures, applications, evaluation metrics, and challenges. *Future Internet*, 17(9):404.

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.

William Boag, Tzu-Ming Harry Hsu, Matthew McDermott, Gabriela Berner, Emily Alesentzer, and Peter Szolovits. 2020. Baselines for chest x-ray report generation. In *Machine learning for health workshop*, pages 126–140. PMLR.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. 2024. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *arXiv preprint arXiv:2405.19538*.

Yifan et al. Chen. 2024. Iterative preference optimization for language model alignment. *arXiv preprint arXiv:2402.04733*.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Yijian Fan, Zhenbang Yang, Rui Liu, Mingjie Li, and Xiaojun Chang. 2024. Medical report generation is a multi-label classification problem. *arXiv preprint arXiv:2409.00250*.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, and Katie et al. Shpanskaya. 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.

Amaan et al. Izhar. 2025. Medical radiology report generation: A systematic review of current deep learning methods, trends, and future directions. *Artificial Intelligence in Medicine*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Haibo et al. Jin. 2024. Promptmrg: Diagnosis-driven prompts for medical report generation. In *AAAI*.

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

Youssef Mroueh. 2025. Reinforcement learning with verifiable rewards: Grpo's effective loss, dynamics, and success amplification. *arXiv preprint arXiv:2503.06639*.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, and Akshay S et al. Chaudhari. 2024. Green: Generative radiology report evaluation and error notation. *arXiv preprint arXiv:2405.03595*.

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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, and Elahe et al. Vedadi. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, and Heather et al. Cole-Lewis. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Pengyu et al. Wang. 2025. Mrg-r1: Reinforcement learning for clinically aligned medical report generation. *arXiv preprint arXiv:2512.16145*.

Rui et al. Xu. 2024. Online direct preference optimization for language models. *arXiv preprint arXiv:2406.01704*.

Kai Zhang, Christopher Malon, Lichao Sun, and Martin Renqiang Min. 2025a. Editgrpo: Reinforcement learning with post-rollout edits for clinically accurate chest x-ray report generation. *arXiv preprint arXiv:2509.22812*.

Shimao Zhang, Xiao Liu, Xin Zhang, Junxiao Liu, Zheheng Luo, Shujian Huang, and Yeyun Gong. 2025b. Process-based self-rewarding language models. *arXiv preprint arXiv:2503.03746*.

Zhenyu Zhang, Benlu Wang, Weijie Liang, Yizhi Li, Xuechen Guo, Guanhong Wang, Shiyan Li, and Gaoang Wang. 2024. Sam-guided enhanced fine-grained encoding with mixed semantic learning for medical image captioning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1731–1735. IEEE.

Mingchen Zhuge, Changsheng Zhao, Dylan Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, and Yuandong et al. Tian. 2024. Agent-as-a-judge: Evaluate agents with agents. *arXiv preprint arXiv:2410.10934*.

# A Appendix

## A.1 Potential Risks and Ethical Considerations

Our work focuses on automated medical report generation, which may pose potential risks if misused in clinical practice. Inaccurate or hallucinated findings could lead to inappropriate clinical decisions if reports are interpreted without professional oversight. To mitigate these risks, we emphasize that our system is intended solely as a decision-support tool rather than a replacement for clinicians. Moreover, MedQPA explicitly evaluates factual correctness, which helps reduce hallucinations and unsupported statements in generated reports.

All datasets used in this work are publicly available and fully de-identified. No personally identifiable information is accessed, stored, or generated by our models.

## A.2 Human Evaluation Protocol

To assess the alignment between automatic metrics and expert judgment, we conduct a human evaluation study on the Indiana X-ray dataset. We randomly sample 100 images and generate four reports for each image using four models: Original, SFT, SFT+DPO, and SFT+GRPO. For each image, the corresponding ground-truth report is also provided as a reference.

We recruit a total of eight human evaluators with background knowledge in medical imaging. For each image, evaluators are presented with the four generated reports together with the ground-truth report, and are asked to score each generated report based on its factual correctness and clinical appropriateness relative to the reference. Specifically, evaluators assign an integer score on a five-point scale (from 1 to 5), where higher scores indicate better alignment with the ground-truth report in terms of correctness, completeness, and absence of hallucinated findings.

Each report is independently scored by all evaluators. We normalize the scores to the range $[0, 1]$ and report the average normalized human score for each model. In addition, we compute the win rate, defined as the proportion of samples for which a

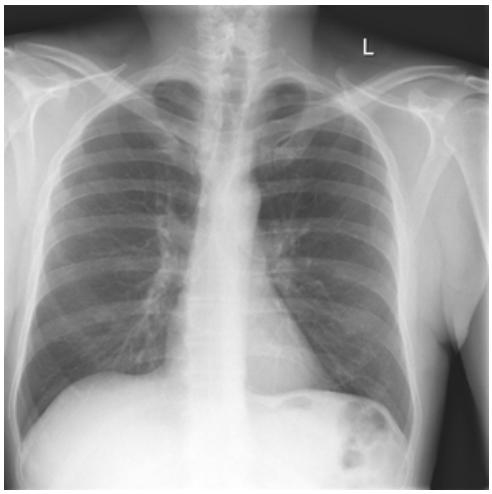



Figure 3: Sample chest X-ray image used in human evaluation (example 1).

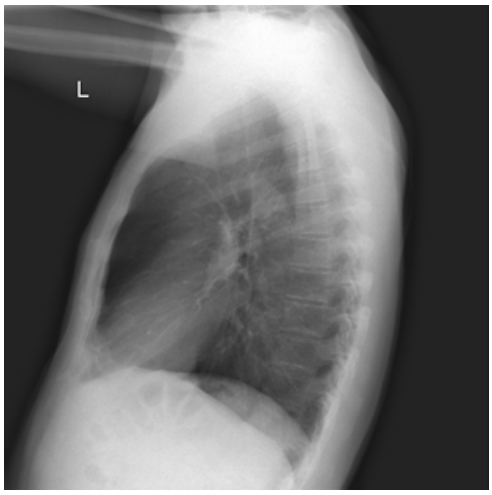



Figure 4: Sample chest X-ray image used in human evaluation (example 2).

model's report receives the highest score among the four candidates. These two metrics provide complementary views of absolute quality and relative preference. The human evaluation results are reported in Table 4.

Evaluators participated voluntarily and were not financially compensated.

The evaluation is conducted solely for research purposes and does not involve any clinical decision-making.

### A.3 Human Evaluation Example

Experts are provided GT reports, medical images like figure.3 and figure.4 and four generated reports and they need to score the reports.

**Ground-Truth Report.**
Lungs are clear bilaterally with no focal infiltrate, pleural effusion, or pneumothoraces. Cardiomediastinal silhouette is within normal limits. XXXX and soft tissues are unremarkable. No acute cardiopulmonary abnormality.

**Report A**
The chest X-ray reveals a patent trachea and no clear evidence of significant mediastinal widening or pulmonary consolidations. The diaphragm appears normal in position and orientation. The lung fields show minimal increase in lung markings with no obvious masses, nodules, pleural effusion, or pneumothorax.

**Report B**
Lungs are clear bilaterally with no consolidation, masses, focal opacities, pleural effusion, or pneumothorax. Cardiac size and silhouette are within normal limits. No pulmonary vascular enlargement.

**Report C**
Lungs are clear bilaterally with no consolidation, nodules, or infiltrates. No pneumothorax or pleural effusion. Cardiac silhouette is within normal limits. No thoracic masses, calcifications, or bony abnormalities. No acute cardiopulmonary abnormality.

**Report D**
No significant cardiomegaly. Lungs show no focal parenchymal abnormality; right lung with slightly decreased density suggesting mild atelectasis or reduced expansion. No pleural effusion. Trachea is midline. Left hemidiaphragm is well defined.

### A.4 Noise Injection Protocol for Ground-Truth Reports

To evaluate the robustness of report evaluation metrics, we construct semantically perturbed versions of ground-truth (GT) medical reports by injecting controlled levels of *semantic noise*. The perturbations are designed to modify medical meaning while preserving fluent radiology-style language, thereby isolating semantic degradation from surface-level wording changes.

**Ground-Truth Report.** The following chest X-ray report is used as the original reference (GT) in our example:

> Lungs are clear bilaterally with no focal infiltrate, pleural effusion, or pneumothoraces. Cardiomediastinal silhouette is within normal limits. XXXX and soft tissues are unremarkable. No acute cardiopulmonary abnormality.

**Noise Injection Prompt.** You are given the following ground-truth (GT) medical report: "Lungs are clear bilaterally with no focal infiltrate, pleural effusion, or pneumothoraces. Cardiomediastinal silhouette is within normal limits. XXXX and soft tissues are unremarkable. No acute cardiopulmonary abnormality." Your task is to generate five noisy versions of this report by injecting semantic-level noise at different intensities: 5%, 10%, 15%, 20%, and 25%. Definition of Semantic Noise: Noise must affect the meaning of the report and may introduce: 1. Factual modifications – e.g., negation becoming affirmation, adding or removing findings, altering clinical status. 2. Contradictions or misleading statements – statements that conflict with the GT. 3. Irrelevant or redundant information – clinically unrelated content that distracts from the core meaning. 4. Subtle wording changes with semantic impact – phrasing that changes interpretation or introduces ambiguity. 5. Rephrasing with altered focus – shifting attention to different or incorrect aspects of the case. You are NOT restricted to the original wording. You may freely rephrase and restructure the report as long as the semantic noise is introduced. Noise Levels: - 5% noise: Very slight semantic distortion; mostly faithful to GT with minimal misleading or altered meaning. - 10% noise: Small but noticeable semantic changes; minor factual shifts or added ambiguity. - 15% noise: Moderate semantic corruption; clear inaccuracies or contradictions appear.

- 20% noise: High semantic distortion; multiple incorrect or misleading statements. - 25% noise: Very high semantic corruption; substantial misunderstanding of the original medical meaning. You should flexibly combine different noise types (fact changes, contradictions, redundancy, wording shifts) to achieve the desired level. Output Requirements: - Produce exactly five reports, corresponding to the five noise levels. - Output them in the following numbered format: 1. [Report with 5% noise] 2. [Report with 10% noise] 3. [Report with 15% noise] 4. [Report with 20% noise] 5. [Report with 25% noise] - Each report should be a coherent medical-style text. - The reports must be in English. - Do NOT include any explanations or commentary. Only output the five numbered reports. Generate the outputs now.

**Example Outputs at Different Noise Levels.** Below we show an example of the five reports generated using the above protocol:

> **5% noise:** Lungs are largely clear bilaterally with no obvious focal infiltrate, pleural effusion, or pneumothorax identified on this view. The cardiomediastinal silhouette appears within normal limits. Soft tissues are unremarkable. Overall, there is no clear evidence of an acute cardiopulmonary abnormality.
>
> **10% noise:** The lungs are generally clear on both sides, though a very subtle area of haziness cannot be entirely excluded. No definite pleural effusion or pneumothorax is seen. The cardiomediastinal silhouette is within normal limits, and the soft tissues appear unremarkable. No acute cardiopulmonary abnormality is confidently identified.
>
> **15% noise:** Lung fields are mostly clear bilaterally, although a small questionable opacity is noted that could represent early infiltrate. No large pleural effusion or pneumothorax is present. The cardiomediastinal silhouette is described as within normal limits, but mild prominence cannot be ruled out. Soft tissues are otherwise unremarkable. An acute cardiopulmonary process is unlikely but not definitively excluded.
>
> **20% noise:** The lungs show areas of patchy opacity bilaterally, raising concern for possible infiltrates, although no large pleural effusion is identified. A tiny pneumothorax cannot be excluded on the right. The cardiomediastinal silhouette appears borderline enlarged. Soft tissues are without obvious abnormality. These findings suggest a potential acute cardiopulmonary abnormality.
>
> **25% noise:** Bilateral lung fields demonstrate abnormal opacities consistent with infiltrates, with associated pleural effusion and evidence of pneumothorax. The cardiomediastinal silhouette is enlarged and distorted. Soft tissues show questionable abnormal changes. Overall, the findings are indicative of a significant acute cardiopulmonary pathology rather than a normal study.

## A.5 Use of AI Assistants

AI-based writing tools were used solely for language editing, such as improving grammar, readability, and stylistic consistency. They were not used to generate scientific content, experimental procedures, data, results, or interpretations. All substantive contributions, including the design of the method, experiments, and analyses, were produced and validated by the authors.

Table 7: Grid search over missing weight ($w_m$) and hallucination weight ($w_h$). Each cell reports Spearman correlation $\rho$ with Noise Levels / Human Preference. Higher is better.

| $w_m \backslash w_h$ | 0.5 | 0.75 | 1.0 |
|---|---|---|---|
| 0.5 | 0.6245 / 0.6287 | 0.6148 / 0.6217 | 0.6176 / 0.6235 |
| 0.75 | 0.6365 / 0.6312 | 0.6237 / 0.6122 | 0.6288 / 0.6223 |
| 1.0 | **0.6542 / 0.6328** | 0.6452 / 0.6217 | 0.6243 / 0.6308 |

Table 8: Results on Indiana X-ray using MedGemma models.

| Model | Training Method | MedQPA |
|---|---|---|
| MedGemma-4B | Base | 0.6452 |
| | SFT | 0.6634 |
| | SFT+DPO | **0.6849** |
| MedGemma-27B | Base | 0.7238 |
| | SFT | **0.7453** |
| | SFT+DPO | 0.7412 |

## A.6 Grid Search for MedQPA's weight for Report Evaluation

As shown in Table 7, MedQPA is relatively robust to the choice of weights, with consistent performance across a wide range of $(w_m, w_h)$ settings. The best results are achieved when placing slightly higher emphasis on missing errors (e.g., $w_m = 1.0, w_h = 0.5$), but the overall ranking quality remains stable. This suggests that MedQPA does not rely on precise tuning and generalizes well across reasonable weight configurations.

## A.7 Additional experiments on MedGemma.

We further evaluate MedQPA on the Indiana X-ray dataset using MedGemma models with different scales. We follow the same experimental setup as in the main paper and compare Base, SFT, and SFT+DPO training strategies.

For MedGemma-4B, we observe consistent improvements with stronger supervision, where SFT improves over the Base model and SFT+DPO further boosts performance. For MedGemma-27B, SFT provides a noticeable gain, while SFT+DPO does not yield additional improvement. We attribute this to the relatively small size of the Indiana X-ray dataset, which limits the effectiveness of RL-based optimization for larger models. Due to these marginal gains, we do not conduct additional human evaluation for the 27B model.