VIGOR: ENHANCING RELIABILITY AND REASONING IN MEDICAL REPORT GENERATION THROUGH REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

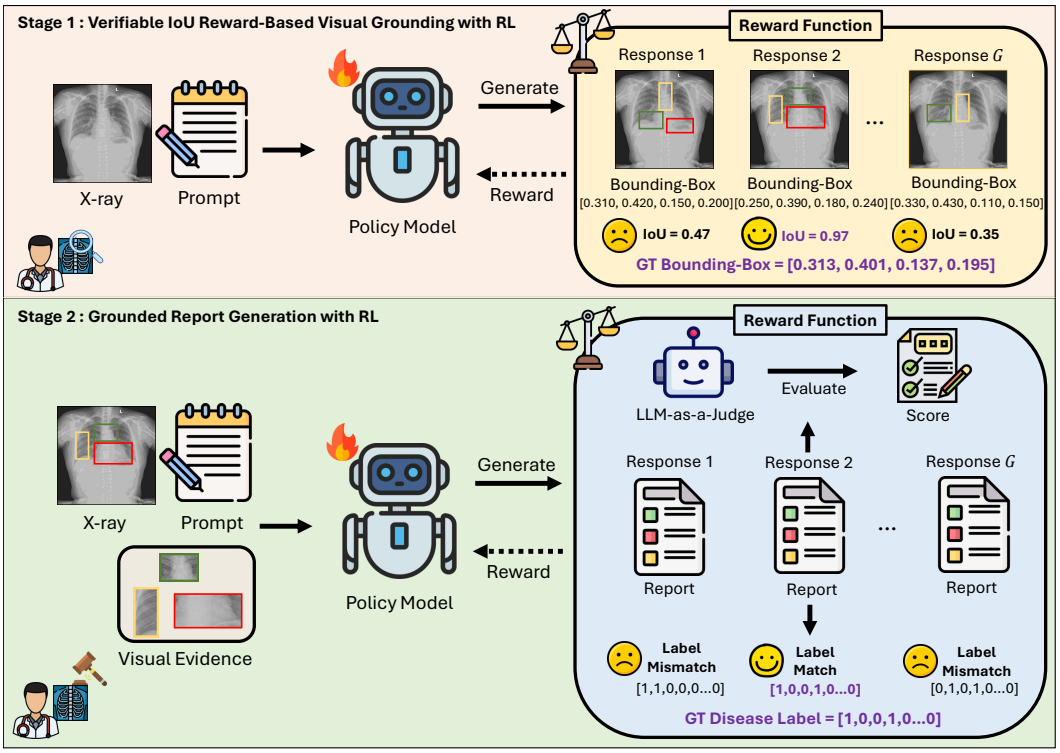


Figure 1: **Overview of VIGOR**, which consists of two stages: (1) Verifiable IoU Reward-based Visual Grounding with RL and (2) Grounded Report Generation with RL. In Stage 1, we endow the visual grounding capability to the base policy. In Stage 2, we continually post-train the policy for grounded radiology report generation, leveraging the information from not only the original X-ray, but also from the zoomed-in region.

ABSTRACT

Multimodal Large Language Models have shown strong performance in vision–language tasks, yet their application to medical report generation remains limited by insufficient visual grounding, opaque reasoning, and reliance on supervised fine-tuning, which primarily learns to imitate reference reports rather than explicitly optimizing clinically grounded objective. We propose **VIGOR**[1], a two-stage reinforcement learning framework for radiology report generation. In the first stage, we introduce a verifiable IoU-based reward and optimize the model using GRPO, enabling the trained model to explicitly align textual descriptions with region-level visual evidence, thereby enforcing accurate localization of abnormalities and improving interpretability. In the second stage, we extend GRPO to report generation by incorporating disease label and LLM-as-a-Judge reward to

[1] **VI**sual **G**rounding for **O**ptimized **R**eport

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

promote diagnostic accuracy and coherent reasoning. This two-stage optimization enables the generation of reports that are visually faithful, diagnostically consistent, and grounded in interpretable reasoning. Extensive experiments on the resulting model, VIGOR-4B, demonstrate reduced hallucinations, improved disease identification, and stronger alignment with verifiable evidence. Our results highlight the importance of explicit visual grounding and reward-guided optimization for developing reliable and faithful medical report generation.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) (Comanici et al., 2025; Bai et al., 2023; Liu et al., 2024; 2023; Achiam et al., 2023; Li et al., 2026; Compagnoni et al., 2026; He et al., 2026) have recently achieved strong performance across multimodal tasks, including image captioning, visual question answering, and instruction-following generation. Their visual understanding and reasoning capabilities have motivated increasing interest in medical vision-language applications (Li et al., 2023; Moor et al., 2023; Chen et al., 2024b; Huang et al., 2024; Bannur et al., 2024; Li et al., 2025; Sellergren et al., 2025; Kumar et al., 2026; Chen et al., 2026). Among these, medical report generation (MRG) aims to generate radiology reports from medical images by translating visual findings into clinically meaningful language. This task requires accurate visual interpretation, precise localization of findings, and reliable generation consistent with radiological reporting standards.

Despite recent progress, directly applying MLLMs to medical report generation currently remains limited, and we focus on the following three limitations:

(1) MLLMs lack reliable visual grounding capabilities. Chest X-rays often contain subtle pathological cues—small consolidations, faint opacities, or localized effusions—that appear only in specific anatomical regions (Xie et al., 2024; Li et al., 2025). Many existing MLLMs rely heavily on global image-level supervision, which provides limited guidance for learning fine-grained visual–textual correspondences (Li et al., 2025). Consequently, they often fail to align textual descriptions with localized pathological regions, leading to hallucinated findings, anatomically inconsistent statements, or irrelevant observations.

(2) MLLMs offer limited interpretability and exhibit opaque reasoning processes. The black-box reasoning in current MLLMs often leads to inaccuracies, misreadings (Gu et al., 2024; Shao et al., 2024a), and hallucinations (Pal et al., 2023; Shao et al., 2024a). Although these models demonstrate strong generative capabilities, they frequently fail to capture nuanced symptom–lesion–disease relationships (Li et al., 2025), resulting in outputs that may appear plausible yet are not reliably grounded in the underlying medical image. Even when a model provides seemingly coherent interpretations, the rationale behind its conclusions remains opaque, making it difficult to determine whether the generated findings genuinely reflect valid visual evidence. This lack of transparent and verifiable reasoning undermines trust and significantly restricts the deployment of MLLMs in safety-critical clinical environments, where explainability and evidence-based decision support are essential (Hettikankanamage et al., 2025; Salimparsa et al., 2025).

(3) Supervised fine-tuning (SFT) relies on imitation rather than explicitly optimizing clinical correctness. SFT is commonly used to apply MLLMs to MRG (Bannur et al., 2024; Hyland et al., 2023; Jiang et al., 2025; Chen et al., 2026; Kumar et al., 2026), learning to imitate reference reports via token-level cross-entropy loss. While this approach can produce fluent and well-structured text, it does not explicitly enforce accurate disease identification, precise anatomical localization, or clinically coherent reasoning. Consequently, models trained solely with SFT may achieve high lexical similarity scores yet still generate reports that omit critical findings or contain clinically inconsistent interpretations. These limitations highlight the need for explicit objectives that directly optimize clinical correctness beyond surface-level imitation.

To address these limitations, we propose VIGOR, a two-stage Reinforcement Learning (RL) framework for visually grounded radiology report generation (See Fig. 1 for an illustration). Our approach decomposes the problem into visual grounding and reward-guided report generation, enabling explicit optimization of both localization and diagnostic correctness. In the first stage, we leverage Group Relative Policy Optimization (GRPO) (Shao et al., 2024b) to explicitly optimize visual grounding using verifiable region-level IoU rewards through RL. Unlike SFT, which guides

grounding behavior only implicitly through language imitation, our reward-guided optimization directly enforces accurate localization of abnormalities. This allows the model to learn not only *where* abnormalities are located but also *why* specific findings should be described, thereby establishing interpretable links between visual regions and textual reasoning. In the second stage, we extend GRPO to report generation by incorporating clinical reward signals. Specifically, we introduce a disease label–based reward to explicitly enforce diagnostic correctness by aligning generated reports with underlying clinical labels. In addition, we employ an LLM-as-a-Judge reward (Li et al., 2024; Hein et al., 2024; Wang et al., 2024; Yuan et al., 2024; Gunjal et al., 2026) to evaluate clinical validity and reasoning coherence beyond surface-level language imitation. Importantly, this reward-based formulation enables effective optimization even when ground-truth radiology reports are unavailable due to costly expert annotation or privacy-related constraints on access to patient data. As a scalable proxy for expert assessment, the LLM judge provides fine-grained reward signals that are difficult to manually hand-craft, enabling RL at scale (Lee et al., 2024; Bai et al., 2022; Yuan et al., 2024; Wang et al., 2024; Hein et al., 2024; Gunjal et al., 2026). By jointly optimizing visual grounding and report generation, our framework produces radiology reports that are visually faithful, diagnostically consistent, and grounded in interpretable reasoning.

Our contributions can be summarized as follows:

- We propose VIGOR, a unified two-stage RL framework for radiology report generation, jointly optimizing visual grounding and clinically informed report generation to enhance interpretability and diagnostic reliability.
- We design a region-level reward for visual grounding, leveraging verifiable IoU reward to explicitly enforce accurate localization of abnormalities, thereby establishing interpretable links between visual regions and textual reasoning.
- We design clinically guided RL rewards for report generation, incorporating a disease label reward and an LLM-as-a-Judge reward to directly optimize diagnostic correctness and coherent reasoning.
- VIGOR-4B, trained with our framework, achieves state-of-the-art performance among open 4B-scale models and is competitive with larger/proprietary systems on multiple metrics.

2 Related Works

Medical Multimodal Large Language Models Medical MLLMs extend large language models to the medical domain by jointly processing medical images and text within a unified architecture (Li et al., 2023; Moor et al., 2023; Saab et al., 2024; Sellergren et al., 2025). Unlike general-domain MLLMs, medical MLLMs must operate under domain-specific constraints, including subtle visual patterns, specialized medical terminology, and the need for accurate and reliable outputs (Zhang et al., 2023; Chen et al., 2024a; Sellergren et al., 2025). To address these challenges, recent studies commonly focus on adapting to the medical domain by aligning visual and textual information using paired medical data (Moor et al., 2023), followed by instruction tuning (Li et al., 2023; Zhang et al., 2023) to inject task-specific knowledge and clinical reasoning capabilities. As a result, medical MLLMs can be applied to a wide range of medical vision–language tasks, including disease classification (Chen et al., 2024b), visual question answering (VQA) (Zhang et al., 2023; Huang et al., 2024), and report generation (Li et al., 2025; Chen et al., 2026; Kumar et al., 2026).

Visual Grounding in Medical Report Generation Recent studies have explored visual grounding to improve the interpretability of radiology report generation by linking findings to corresponding image regions. BiRD (Huang et al., 2024) introduces refer-and-ground capabilities, enabling models to associate linguistic references with visual regions, particularly for medical question answering. MAIRA-2 (Bannur et al., 2024) generates sentence-level findings with spatial annotations represented as textual coordinates. AOR (Li et al., 2025) proposes structured anatomical reasoning using region-level representations and anatomy-focused instructions. Unlike prior works that represent localized regions primarily through textual coordinates or region descriptors, our method directly conditions report generation on cropped visual evidence from localized regions. We optimize the grounding stage through reinforcement fine-tuning with verifiable region-level rewards,

explicitly encouraging accurate localization of disease-relevant findings. By providing localized regions as explicit visual inputs to report generation, each generated finding can be directly associated with its corresponding visual evidence.

Reinforcement Fine-Tuning Reinforcement learning (RL) has become a critical component in modern LLM post-training, often referred to as Reinforcement Fine-Tuning (RFT) (Ouyang et al., 2022; Jaech et al., 2024; Lambert et al., 2024; Guo et al., 2025; Lambert, 2025). Unlike supervised fine-tuning (SFT), which incentivizes the model to memorize the ground-truth trajectory, RFT incentivizes the model to get *correct answers*. Compared to SFT, RFT was shown to generalize better across different domains (Kirk et al., 2023; Chu et al., 2025), including changes to stylistic behaviors and reasoning abilities. RL from Human Feedback initially popularized RFT to distil human preferences to a model when paired preference data is available (Ouyang et al., 2022). More recently, after the development of o1 (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025), RFT was applied in a more general context when the reward of the model can be computed with a verifiable program—for instance, in math and coding—showing that the model can learn new behaviors in a data-efficient manner. Later, RFT was extended to MLLMs for visual tasks including classification, object detection, and visual grounding (Liu et al., 2025; Xu et al., 2025a; Ren et al., 2026).

3 Our contribution: VIGOR

3.1 Preliminary

3.1.1 Group Relative Policy Optimization

We adopt GRPO (Guo et al., 2025) as a critic-free RL objective to optimize multimodal generation without relying on an explicit value function. Unlike conventional actor–critic methods that require value estimation, GRPO updates the policy through relative comparisons among multiple sampled outputs for the same input. For each input query q, GRPO samples a group of candidate outputs $\{o_1, \ldots, o_G\}$ from the old policy $\pi_{\theta_{\text{old}}}$ and optimizes the current policy π_θ by maximizing the following objective. Each output is assigned a scalar reward $\{r_1, \ldots, r_G\}$ computed by task-specific, verifiable reward functions. Instead of estimating absolute advantages, GRPO computes a group-normalized relative advantage:

$$A_i = \frac{r_i - \mu_r}{\sigma_r}, \tag{1}$$

where μ_r and σ_r denote the mean and standard deviation of rewards within the sampled group. The GRPO objective is formulated as:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(o|q)} \left[\frac{1}{G} \sum_{i=1}^{G} \min \left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)} A_i,\ \text{clip}\Big(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1-\epsilon, 1+\epsilon \Big) A_i \right) \right]. \tag{2}$$

By maximizing this objective, GRPO favors outputs with higher relative advantages within each sampled group without requiring an auxiliary critic network. In our framework, GRPO optimizes both visual grounding and report generation by exploring multiple candidate outputs and directly maximizing task-specific rewards for localization accuracy and clinical correctness.

3.2 Two-stage Reinforcemenat Learning

The two-stage framework of VIGOR is shown in Fig. 1. Our approach decomposes chest X-ray report generation into (1) Verifiable IoU reward-based visual grounding with RL and (2) grounded report generation with RL. In the first stage, we train the model to localize abnormal regions within the image through reward-guided visual grounding, enabling explicit alignment between textual descriptions and region-level visual evidence. This stage establishes interpretable spatial grounding that serves as a foundation for report generation. In the second stage, we leverage the trained grounding model to generate radiology reports conditioned on the grounded visual evidence, the global X-ray image, and the textual prompt. RL is employed to optimize report generation using two clinically guided reward signals: a disease label–based classification reward and an LLM-as-a-Judge reward, promoting diagnostic correctness and coherent clinical reasoning.

3.2.1 STAGE 1: VERIFIABLE IOU REWARD-BASED VISUAL GROUNDING WITH RL

In Stage 1, we warm up the visual grounding model with SFT and subsequently optimize it using GRPO. Given an image (and its query), the model generates a set of predicted bounding boxes $\mathcal{B} = \{b_1, b_2, \ldots, b_K\}$, while the ground-truth annotations are $\mathcal{B}^* = \{b_1^*, b_2^*, \ldots, b_M^*\}$. Each predicted box b_k is represented by normalized coordinates relative to the image size:

$$b_k = (x_k, y_k, w_k, h_k) \in [0, 1]^4, \tag{3}$$

where (x_k, y_k) denotes the top-left corner and (w_k, h_k) denotes width and height.[2]

To measure localization accuracy, we adopt the Intersection-over-Union (IoU) metric, which quantifies spatial overlap between predicted and ground-truth boxes. For a predicted box b and a ground-truth box b^*, IoU is defined as

$$\mathrm{IoU}(b, b^*) = \frac{\mathrm{area}(b \cap b^*)}{\mathrm{area}(b \cup b^*)}. \tag{4}$$

This IoU-based reward provides a verifiable signal that directly reflects localization accuracy, enabling optimization of the grounding policy. We also use a format reward to penalize invalid bounding boxes, encouraging structured outputs and stable reinforcement learning. Finally, the visual grounding reward is defined as

$$R_v = R_{\mathrm{IoU}} + \lambda_{\mathrm{format}} R_{\mathrm{format}}. \tag{5}$$

We set $\lambda_{\mathrm{format}} = 1$. We sample a group of candidate outputs (box sets) per input and update the policy to increase the likelihood of candidates with higher relative advantages under R_v.

3.2.2 STAGE 2: GROUNDED REPORT GENERATION WITH RL

In stage 2, we leverage the visual grounding model trained to perform visually grounded radiology report generation. Predicted abnormal regions are cropped from the input image and provided as additional visual inputs, allowing report generation to leverage localized evidence.

Similar to stage 1, we formulate report generation as a reinforcement learning using two complementary rewards. We introduce a disease-label reward that evaluates the diagnostic consistency between the generated report and the ground-truth clinical labels. Specifically, given a generated report $\hat{r}$, we extract a set of disease labels $\hat{\mathbf{y}} \in \{0, 1\}^K$ using a CheXbert (Smit et al., 2020) model. Let $\mathbf{y} \in \{0, 1\}^K$ denote the corresponding ground-truth disease label vector. The disease-label reward is defined as

$$R_{\mathrm{label}} = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}(\hat{y}_k = y_k), \tag{6}$$

where $\mathbb{I}(\cdot)$ is the indicator function, and $K = 14$. This reward encourages coherent and aligned with known disease annotations, thereby grounding textual descriptions in clinical outcomes. In addition, we employ an LLM-as-a-Judge reward to assess the overall quality of the generated report. The judge evaluates a generated report $\hat{r}$ conditioned on the global chest X-ray image I, the ground-truth disease labels $\mathbf{y}$, and a detailed evaluation prompt c_{judge} that specifies the criteria for clinical correctness and reasoning quality. The resulting scalar score is defined as

$$R_{\mathrm{LLM}} = \mathcal{J}(\hat{r}, I, \mathbf{y}, c_{\text{judge}}), \tag{7}$$

where $\mathcal{J}(\cdot) \in [0, 1]$ denotes a normalized score for a better report, evaluated by an external large language model based on clinical validity and coherence. The final reward for report generation is defined as the sum of the two complementary signals:

$$R_{\mathrm{report}} = R_{\mathrm{label}} + \lambda_{\mathrm{LLM}} R_{\mathrm{LLM}}. \tag{8}$$

We set $\lambda_{\mathrm{LLM}} = 1$. This combined reward enables the model to jointly optimize diagnostic correctness and overall clinical quality within the GRPO framework.

[2]For example, $b_k = (0.432, 0.221, 0.182, 0.142)$ represents a box defined in normalized image coordinates.

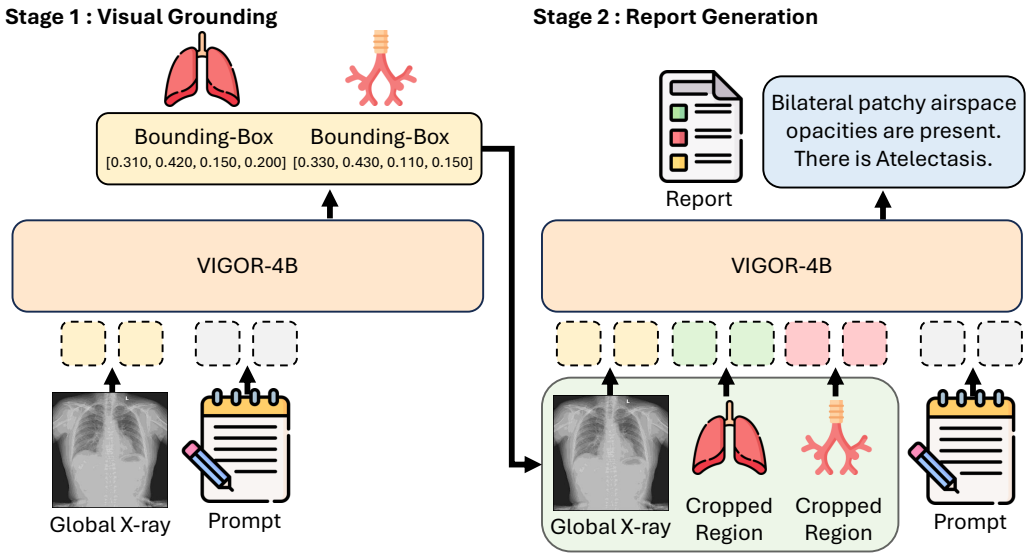


Figure 2: **Inference of VIGOR**. Given the global X-ray image and textual prompt, the model first performs visual grounding to localize abnormal regions and then generates reports conditioned on the grounded visual evidence, the global X-ray image, and the textual prompt.

4 Experiments

Datasets We conduct experiments on five public chest X-ray benchmarks. For Stage 1 (visual grounding), we use bounding-box annotated datasets: MS-CXR (Boecking et al., 2022), VinDr-CXR (Nguyen et al., 2022), and NIH Chest X-ray14 (Summers, 2019). MS-CXR provides $1,162$ phrase–region pairs, VinDr-CXR includes $18,000$ expert-annotated images, and NIH Chest X-ray14 contains disease labels with bounding boxes. Excluding No Finding cases, we use Stage 1 using $5,839$ abnormal cases. For Stage 2, we train the model using MIMIC-CXR (Johnson et al., 2019), MS-CXR (Boecking et al., 2022), VinDr-CXR (Nguyen et al., 2022), and NIH Chest X-ray14 (Summers, 2019) with their provided disease annotations. We further incorporate Open-I (Demner-Fushman et al., 2016) by extracting disease labels from its ground-truth reports using the CheXbert labeler (Smit et al., 2020). In total, we use $309,077$ cases for Stage 2.

Implementation Details. We use MedGemma-4B (Sellergren et al., 2025) as the pre-trained medical vision–language model and fine-tune it using LoRA adapters (Hu et al., 2022) with rank=16. For GRPO, 4 candidate model outputs are sampled per input to encourage output diversity and compute relative advantages, without using KL divergence regularization. For the LLM-as-a-Judge reward, we employ Gemini-2.5-Pro (Comanici et al., 2025) to assess clinical correctness and hallucination. We use the AdamW optimizer (Loshchilov & Hutter, 2017) with a learning rate of 5×10^{-6} for both stage 1 (visual grounding) and stage 2 (report generation).

Evaluation Metric. For visual grounding evaluation, we assess localization performance using Intersection-over-Union (IoU) between predicted bounding boxes and ground-truth annotations. Specifically, we report Mean IoU and Median IoU to measure overall region-level alignment, along with IoU-based hit rates to quantify the proportion of correctly localized abnormalities under different overlap thresholds. These IoU-based metrics provide an objective and verifiable measure of how accurately abnormal regions are grounded in the image. We evaluate the generated report using a structured LLM-as-a-judge framework based on Gemini-2.5-Pro (Comanici et al., 2025). The chest X-ray image is treated as the primary source of truth, while disease labels serve as reference information, and a ground-truth report, when available, is used as reference. Inspired by (Xie et al., 2024; Wang et al., 2024), we design a structured LLM evaluation in which the judge scores each report across four criteria: disease–label consistency, localization accuracy, severity/state description, and hallucinations and over-reporting, with detailed rubrics provided in the supplementary material.

Table 1: **Quantitative evaluation of stage 1 visual grounding performance.** Accurate localization at this stage provides the visual evidence used by downstream report generation in stage 2.

Method	Mean IoU	Median IoU	IoU@0.5	IoU@0.3
MedGemma (Sellergren et al., 2025)	0.114	0.011	0.026	0.159
MedGemma + SFT	0.290	0.288	0.284	0.452
MedGemma + SFT + RL (Ours)	**0.406**	**0.408**	**0.437**	**0.601**

Table 2: **Quantitative evaluation of report generation performance.** The left block reports disease label consistency, localization accuracy, severity/state description, and hallucination avoidance (higher is better). The right block reports the CheXbert (Smit et al., 2020) F1, Precision, and Recall scores and the RadGraph-XL (Delbrouck et al., 2024) score (higher is better).

Method	Disease	Local	Severe	Halluc	F1	Prec	Rec	Rad
Closed Models								
Gemini 2.5 Flash (Comanici et al., 2025)	3.202	3.548	3.176	2.810	0.399	0.386	0.412	0.160
Gemini 2.5 Pro (Comanici et al., 2025)	3.352	3.986	**3.486**	3.324	0.397	0.383	0.412	0.101
GPT-5 (Singh et al., 2025)	**3.444**	**4.090**	3.474	3.356	0.440	0.413	0.470	0.301
Open Models								
LLaVA1.5-7B (Liu et al., 2024)	0.815	0.790	0.610	0.606	0.199	0.283	0.153	0.058
LLaVA-Med-7B (Li et al., 2023)	1.288	1.303	0.656	1.004	0.219	0.316	0.167	0.069
MAIRA-2 (Bannur et al., 2024)	3.044	3.221	2.499	3.543	0.453	0.450	0.458	0.253
Qwen3-VL-4B (Yang et al., 2025)	2.080	1.810	1.621	1.266	0.344	0.414	0.293	0.224
HuluMed-4B (Jiang et al., 2025)	1.725	1.425	1.130	3.353	0.313	0.326	0.300	0.215
HuluMed-7B (Jiang et al., 2025)	1.792	1.570	1.314	3.295	0.343	0.360	0.326	0.227
Lingshu-7B (Xu et al., 2025b)	2.795	3.130	2.469	3.792	0.421	0.434	0.408	0.249
MedGemma-4B (Sellergren et al., 2025)	2.630	2.333	2.140	2.670	0.464	0.475	0.453	0.333
MedGemma-4B (Sellergren et al., 2025)+SFT	2.680	2.355	2.320	2.890	0.468	0.473	0.463	0.244
VIGOR-4B (ours)	3.138	3.436	2.724	**3.942**	**0.503**	**0.490**	**0.517**	**0.352**

Each criterion is rated on a 0–5 scale, yielding a total score ranging from 0 to 20. We also adopt CheXbert (Smit et al., 2020) and RadGraph-XL (Delbrouck et al., 2024) metrics. CheXbert (Smit et al., 2020) measures disease-level performance by extracting clinical labels from generated reports and comparing them against ground-truth labels. RadGraph-XL (Delbrouck et al., 2024) evaluates structured clinical entity and relation consistency by comparing predicted medical entities and their relations with those extracted from the ground-truth report, providing a graph-level F1 score.

4.1 Result

4.1.1 Visual Grounding.

In Tab.1, MedGemma-4B (Sellergren et al., 2025) shows limited localization capability, while SFT substantially improves alignment with ground-truth abnormalities. Our SFT+RL approach further achieves consistent gains across all metrics, with similar mean and median IoU values indicating robust improvements across cases. These results demonstrate that reinforcement learning effectively refines localization after SFT, providing more reliable visual evidence for downstream report generation in Stage 2.

4.1.2 Report Generation.

VIGOR-4B achieves strong overall performance. As shown in Tab. 2, VIGOR-4B achieves the strongest overall performance, outperforming LLaVA1.5-7B (Liu et al., 2024), LLaVA-Med-7B (Li et al., 2023), and Qwen3-VL-4B (Yang et al., 2025).

Although closed models achieve higher scores on most LLM-as-a-Judge criteria, VIGOR remains competitive and achieves the best hallucination score while surpassing them in CheXbert (Smit et al., 2020) and RadGraph-XL (Delbrouck et al., 2024). These gains indicate improved disease-level consistency and more accurate clinical entity-relation alignment. We attribute this performance to the complementary optimization of the two stages: Stage 1 provides reliable localized visual

Table 4: **Ablation study on Stage 1 and Stage 2 components.** Each report generation criterion is rated on a 0–5 scale, and we report the average score for disease label consistency, localization accuracy, severity/state description, and hallucination avoidance (higher is better).

Stage 1	Stage 2		Criteria			
Visual Grounding	LLM-Judge	Disease	Disease	Local	Severe	Halluc
✓	×	×	2.780	2.890	2.340	3.060
✓	✓	×	2.810	2.960	2.630	3.350
✓	×	✓	2.960	3.120	2.690	3.300
×	✓	✓	2.750	2.780	2.224	2.794
✓	✓	✓	**3.318**	**3.436**	**2.724**	**3.942**

evidence, while Stage 2 aligns generated reports with clinical findings using disease label and LLM-as-a-Judge rewards.

Clinical evaluation. To evaluate VIGOR in a real-world clinical setting, we conduct a board-certified physician evaluation under a strictly blinded setting. fig. 3 shows that VIGOR generates more reliable reports than MedGemma-4B. These results suggest that the four clinical rubrics used in our LLM-as-a-Judge are aligned with real-world clinical assessment criteria.

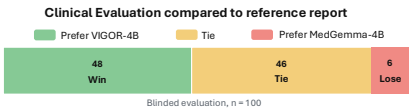


Figure 3: Clinical evaluation of VIGOR vs. MedGemma-4B.

Cross-judge evaluation. To assess whether the observed gains generalize across different evaluator models, we conduct a cross-judge evaluation, where VIGOR is trained using Gemini-2.5-Pro as the LLM-as-a-Judge reward model and evaluated using different judge models. Tab. 3 shows consistent performance improvements across evaluators, indicating that the gains are not simply due to overfitting to the specific evaluation preferences of the reward model used during training.

Table 3: Cross-judge evaluation using different evaluator models.

Evaluator	Method	Disease	Local	Severe	Halluc
Gemini-2.5-Flash [10]	MedGemma-4B [40]	2.560	2.434	2.144	2.470
	VIGOR-4B (Ours)	**3.106**	**3.554**	**2.720**	**3.798**
GPT-5 [43]	MedGemma-4B [40]	2.810	2.178	2.200	2.390
	VIGOR-4B (Ours)	**3.312**	**3.224**	**2.716**	**3.446**

4.2 Ablation Study.

In Tab. 4, we analyze the individual and joint contributions of Stage 1 visual grounding and Stage 2 rewards. Visual grounding alone improves over the base MedGemma (Sellergren et al., 2025), showing that spatial grounding provides meaningful benefits even without Stage 2 optimization. Adding either the disease label or LLM-as-a-Judge reward further improves performance in complementary ways: the former strengthens pathology-level consistency, while the latter enhances overall clinical coherence. Removing visual grounding while retaining both Stage 2 rewards degrades performance, indicating that reward-based optimization alone is insufficient without reliable spatial evidence. The full configuration achieves the most balanced performance across evaluation criteria.

4.3 Qualitative Analysis

Normal Case Figure 4 presents a qualitative comparison between the ground-truth report, MedGemma-4B, and VIGOR-4B. In the normal case, MedGemma-4B correctly describes clear lungs but hallucinates "the cardiac silhouette is mildly enlarged", contradicting the normal cardiomediastinal silhouette in the reference report. In contrast, VIGOR-4B correctly preserves the absence of pathology without introducing unsupported findings. We attribute this improvement to the Stage 2 rewards: the LLM-as-a-Judge reward penalizes hallucinated or inconsistent statements, while the disease label reward encourages consistency with "No Finding". Together, these rewards reduce false-positive findings and promote accurate reporting of normal case.

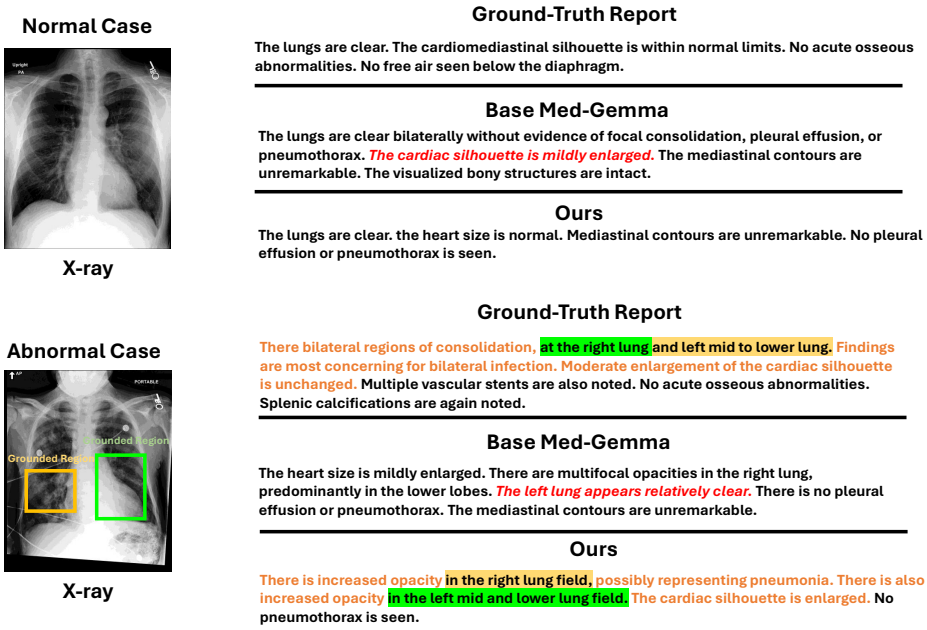


Figure 4: **Qualitative analysis of generated reports for two MIMIC-CXR cases.** For the normal case, VIGOR-4B does not attend to a specific region and directly produces the report. Descriptions aligned with the grounded regions are highlighted in the same color to denote their correspondence to the same anatomical location, while abnormal findings that are consistent across the reports for the patient are highlighted in orange. Incorrect statements produced are marked in red.

Abnormal Case In the abnormal case, the ground-truth report describes bilateral consolidative opacities and cardiomegaly. MedGemma-4B captures the cardiac enlargement and right lung opacities but incorrectly describes the left lung as relatively clear, failing to reflect the bilateral extent of disease. In contrast, VIGOR-4B correctly identifies increased opacities in both lungs and reports an enlarged cardiac silhouette, more closely matching the ground truth. This qualitative analysis shows that visual grounding with cropped regions suppresses hallucinations and unsupported statements, improving the visual faithfulness of generated reports.

5 Conclusion

We introduce VIGOR, a two-stage reinforcement learning framework for grounded and reliable chest X-ray report generation. VIGOR first optimizes visual grounding with IoU-based reward, then improves report generation using disease-label and LLM-as-a-Judge rewards. Together, these objectives yield reports that better match the image evidence, reduce hallucinations, and improve disease identification and reasoning transparency, moving medical report generation closer to systems clinicians can trust and deploy.

Discussion and Limitation Despite the gains from optimizing report quality with an LLM-as-a-judge reward, this signal remains an imperfect proxy for expert assessment. Prior work shows that strong LLM judges can correlate well with human preference, but they exhibit systematic biases (e.g., position/verbosity/self-preference) and can be sensitive to prompt design and evaluation protocol (Zheng et al., 2023). We therefore view the judge reward as scalable AI feedback—useful for shaping high-level coherence beyond label-based objectives, but inherently noisy and model-dependent (Lee et al., 2023). In our experiments we use a high-capacity judge, which likely reduces noise but increases reliance on a closed, evolving evaluator. A promising direction is to reduce reliance on a closed external evaluator by distilling the judge into an open, lightweight reward model, or by exploring self-judging variants where the generator provides the reward under carefully controlled protocols.

AI USE STATEMENT

In this work, we used generative AI tools to design or provide feedback on research methodology or experiments, assist with translation, and clean and reformat datasets. Specifically, this includes our use of an LLM-as-a-Judge as part of the experimental methodology. We did not use generative AI tools to generate synthetic datasets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in writing proofs, propose or refine hypotheses, implement methods, support qualitative or thematic data analysis, or interpret results. We have reviewed all AI-assisted work. We checked LLM-generated research ideas for potential plagiarism through a manual literature survey. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel Coelho de Castro, Anton Schwaighofer, Stephanie Hyland, Maria Teodora Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez Valle, et al. Ms-cxr: Making the most of text semantics to improve biomedical vision-language processing. *URL: http://dx. doi. org/10.13026/b90j-vb87*, 2022.

Qi Chen, Ruoshan Zhao, Sinuo Wang, Vu Minh Hieu Phan, Anton van den Hengel, Johan Verjans, Zhibin Liao, Minh-Son To, Yong Xia, Jian Chen, et al. A survey of medical vision-and-language applications and their techniques. *arXiv preprint arXiv:2411.12195*, 2024a.

Zailong Chen, Peng Gao, Johan Barthelemy, Luping Zhou, and Lei Wang. Analyzing and enhancing visual learning in llm-based radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9327–9336, 2026.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024b.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Alberto Compagnoni, Marco Morini, Sara Sarto, Federico Cocchi, Davide Caffagni, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Reag: Reasoning-augmented generation for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11901–11911, 2026.

Jean-Benoit Delbrouck, Pierre Chambon, Zhihong Chen, Maya Varma, Andrew Johnston, Louis Blankemeier, Dave Van Veen, Tan Bui, Steven Truong, and Curtis Langlotz. Radgraph-xl: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12902–12915, 2024.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

Bowen Gu, Rishi J Desai, Kueiyu Joshua Lin, and Jie Yang. Probabilistic medical predictions of large language models. *npj Digital Medicine*, 7(1):367, 2024.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In *International Conference on Learning Representations*, volume 2026, pp. 127924–127945, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yicheng He, Chengsong Huang, Zongxia Li, Jiaxin Huang, and Yonghui Yang. Visplay: Self-evolving vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26274–26284, June 2026.

Dennis Hein, Zhihong Chen, Sophie Ostmeier, Justin Xu, Maya Varma, Eduardo Pontes Reis, Arne Edward Michalson, Christian Bluethgen, Hyun Joo Shin, Curtis Langlotz, et al. Preference fine-tuning for factuality in chest x-ray interpretation models without human feedback. 2024.

Nadeesha Hettikankanamage, Niusha Shafiabady, Fiona Chatteur, Robert MX Wu, Fareed Ud Din, and Jianlong Zhou. explainable artificial intelligence (xai): A systematic review for unveiling the black box models and their relevance to biomedical imaging and sensing. *Sensors (Basel, Switzerland)*, 25(21):6649, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Xiaoshuang Huang, Haifeng Huang, Lingdong Shen, Yehui Yang, Fangxin Shang, Junwei Liu, and Jia Liu. A refer-and-ground multimodal large language model for biomedicine. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 399–409. Springer, 2024.

Stephanie L Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, et al. Maira-1: A specialised large multimodal model for radiology report generation. *arXiv preprint arXiv:2311.13668*, 2023.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

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

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*, 2023.

Ashwin Kumar, Robbie Holland, Corey Barrett, Jangwon Kim, Maya Varma, Zhihong Chen, Yunhe Gao, Greg Zaharchuk, Tara Taghavi, Krishnaram Kenthapadi, and Akshay Chaudhari. Chexmix: Unified generative pretraining for vision language models in medical imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 9466–9476, 2026.

Nathan Lambert. Reinforcement learning from human feedback. *arXiv preprint arXiv:2504.12501*, 2025.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback.(2023). In *URL https://openreview. net/forum*, 2023.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. In *International Conference on Machine Learning*, pp. 26874–26901. PMLR, 2024.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564, 2023.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024.

Hengzhuang Li, Xinsong Zhang, Qiming Peng, Bin Luo, Han Hu, Dengyang Jiang, Han-Jia Ye, Teng Zhang, and Hai Jin. Unleashing the intrinsic visual representation capability of multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1771–1786, 2026.

Qingqiu Li, Zihang Cui, Seongsu Bae, Jilan Xu, Runtian Yuan, Yuejie Zhang, Rui Feng, Quanli Shen, Xiaobo Zhang, Junjun He, et al. Aor: Anatomical ontology-guided reasoning for medical large multimodal model in chest x-ray interpretation. *arXiv preprint arXiv:2505.02830*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pp. 353–367. PMLR, 2023.

Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*, 2023.

Xiangxuan Ren, Zhongdao Wang, Liping Hou, Pin Tang, Guoqing Wang, and Chao Ma. Grounding everything in tokens for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41171–41181, 2026.

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

Mozhgan Salimparsa, Kamran Sedig, Daniel J Lizotte, Sheikh S Abdullah, Niaz Chalabianloo, and Flory T Muanda. Explainable ai for clinical decision support systems: Literature review, key gaps, and research synthesis. In *Informatics*, volume 12, pp. 119. MDPI, 2025.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024a.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024b.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*, 2020.

R Summers. Nih chest x-ray dataset of 14 common thorax disease categories. *NIH Clinical Center: Bethesda, MD, USA*, 2019.

Zilong Wang, Xufang Luo, Xinyang Jiang, Dongsheng Li, and Lili Qiu. Llm-radjudge: Achieving radiologist-level evaluation for x-ray report generation. *arXiv preprint arXiv:2404.00998*, 2024.

Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, et al. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. *arXiv preprint arXiv:2408.02900*, 2024.

Huihui Xu, Yuanpeng Nie, Hualiang Wang, Ying Chen, Wei Li, Junzhi Ning, Lihao Liu, Hongqiu Wang, Lei Zhu, Jiyao Liu, et al. Medground-r1: Advancing medical image grounding via spatial-semantic rewarded group relative policy optimization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 391–401. Springer, 2025a.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. Self-rewarding language models. In *Forty-first International Conference on Machine Learning*, 2024.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.