

OraPO: Oracle-educated Reinforcement Learning for Data-efficient and Factual Radiology Report Generation

Zhuoxiao Chen^{1,2†}, Hongyang Yu^{1*}, Ying Xu¹, Yadan Luo², Long Duong¹, Yuan-Fang Li^{1*}

¹Oracle Health & AI, ²The University of Queensland

{ivan.chen, hongyang.yu, ying.x.xu, long.duong, yuanfang.li}@oracle.com

{zhuoxiao.chen, y.luo}@uq.edu.au

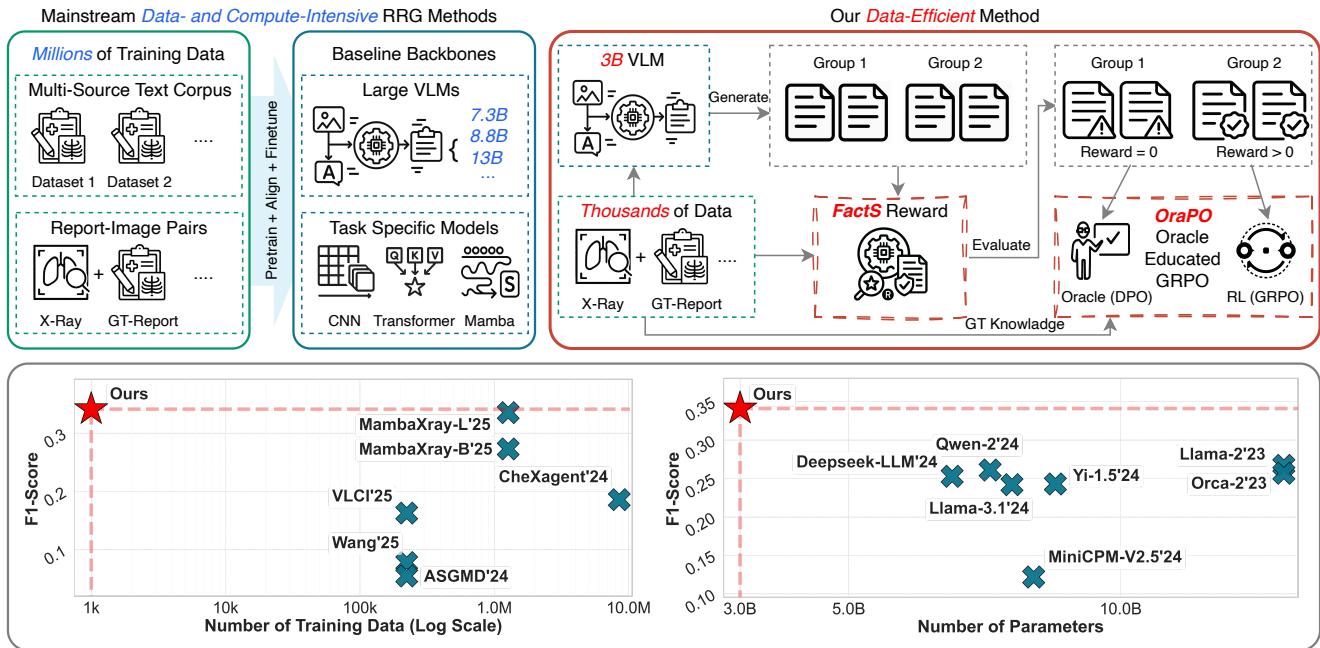


Figure 1. Illustration of mainstream data/compute-intensive pipelines (upper-left) versus our data-efficient pipeline (upper-right). Bottom: on CheXpert Plus [7], our method achieves the **SOTA performance for RRG** with **less than 0.1%** of the training samples (vs. 1.27 M) used by best-performing baselines and a **much smaller model**, demonstrating strong performance under tight data and compute budgets.

Abstract

Radiology report generation (RRG) aims to automatically produce clinically faithful reports from chest X-ray images. Prevailing work typically follows a **scale-driven paradigm**, by multi-stage training over large paired corpora and oversized backbones, making pipelines highly data- and compute-intensive. In this paper, we propose Oracle-educated GRPO (**OraPO**) with a FactScore-based reward (**FactS**) to tackle the RRG task under constrained budgets. OraPO enables single-stage, RL-only training by converting failed GRPO explorations on rare or difficult

studies into direct preference supervision via a lightweight oracle step. FactS grounds learning in diagnostic evidence by extracting atomic clinical facts and checking entailment against ground-truth labels, yielding dense, interpretable sentence-level rewards. Together, OraPO and FactS create a compact and powerful framework that significantly improves learning efficiency on clinically challenging cases, setting the **new SOTA** performance on the CheXpert Plus dataset (0.341 in F1) and MIMIC-CXR dataset (0.357 in F1) with **2–3 orders of magnitude less training data** using a **small** base vision-language model on modest hardware.

[†] Work done during an internship at Oracle Health & AI. ^{*} Joint senior authors.

1. Introduction

Radiology report generation (RRG) from chest X-rays aims to turn imaging evidence into a clinically usable narrative. This is a complex multimodal task that couples disease prediction with free-text report generation, linking detected visual findings (*e.g.* edema, pleural effusion, and pneumothorax) to clear sentences that describe their presence, details, location, and severity. Automating this task matters in practice: the large imaging volumes and backlogs place serious strains on healthcare services and delay patient care, while radiologist staffing remains insufficient. For example, England reports a 29% radiology consultant shortfall alongside sustained vacancies [61], while in the United States there are about 14,000 openings but only 1,150 resident graduates annually [70]. In this context, AI-assisted drafting has begun to show measurable efficiency gains. A recent study [1] reported that using draft reports generated by large language models (LLMs) reduced reporting time by 24%. Given its practical importance, automated radiology report generation has rapidly become an active research area aimed at easing radiologists’ workload and streamlining workflows in hospitals and clinics.

Mainstream RRG methods typically rely on supervised fine-tuning (SFT) via multi-stage training on large corpora and/or scaling up backbone size (Fig. 1, upper-left). The multi-stage training paradigm usually follows three phases: domain-specific pre-training, image-text alignment, and task-specific fine-tuning, where the first two are usually trained on combinations of large-scale, multi-source datasets such as medical textbooks, real radiology reports including but not limited to MIMIC-CXR [37], IU X-ray [20] and CheXpert Plus [7], often for dozens of epochs [83, 85]. Notably, both CheXpert Plus and MIMIC-CXR include training splits **exceeding 200K** image-report pairs. In addition, there is a recent trend of adopting large vision-language models (VLMs) with >13B parameters for SFT, demanding meticulous data curation and substantial GPU budgets [38, 41].

Recently, *Group Relative Policy Optimisation* (GRPO) has emerged as a strong and compute-efficient RL-based learning algorithm for many hard problems with verifiable rewards. Its efficiency stems from two fronts: (i) lower memory/compute by dispensing with the trainable value critic and instead computing group-normalised advantages from the relative rewards of multiple samples [40, 56]; and (ii) a simplified training recipe that enables “RL-only” regimes (*i.e.*, without costly SFT/alignment), avoiding massive pre-training/alignment corpora [9, 19, 72, 104].

However, applying GRPO to the RRG task reveals two challenges. Challenge (i): we observe that vanilla GRPO often produces all-zero reward groups early in training (*e.g.* 30% in the first 50 steps, as shown in Fig. 2 left), causing vanishing gradients and wasted rollouts. Recent vari-

ants attempt to fix this by resampling until a non-zero reward appears [100] or by enlarging the group size [73], both of which are still suboptimal due to *increased cost* of more rollouts. Another line interleaves SFT with RL to maintain gradient flow, yet still leaves low-quality rollouts unused, wasting costly sampling [10, 48, 50, 51, 102]. Challenge (ii): vanilla GRPO requires readily verifiable scalar rewards. Unlike in multiple-choice QA, maths, or coding, where the reward is typically a binary final-answer check, radiology reports are long-form, multi-fact, and must be consistent across sentences, making reward design both critical and difficult. In practice, RLHF-style RRG work often uses reference-overlap metrics that mainly capture *fluency/surface similarity* (*e.g.* BLEU/CIDEr [42, 60]) or a *report-level clinical metric* [55, 97, 110]. These proxies under-penalise sentence-level factual errors and cross-sentence contradictions, yielding fluent yet clinically misleading narratives (missed positives, unsupported claims). Hence, a reward must be able to assess factual correctness of each statement of a radiology report in an efficient manner.

In this paper, we propose *Oracle-educated Group Relative Policy Optimisation*: **OraPO**, a novel RL-based learning algorithm that tackles RRG’s challenges in a data- and compute-efficient manner. OraPO addresses Challenge (i) of GRPO by dynamically injecting oracle supervision (*i.e.* ground-truth) for groups with *all-zero* rewards and teaching the policy to avoid repeating low-quality generations. Specifically, When a sampled group yields *all-zero* rewards (group 1 in Fig. 1, upper-right), OraPO triggers a lightweight Direct Preference Optimisation (DPO) update that directly prefers the ground-truth report over the zero-reward rollouts, effectively reusing those failed rollouts as *ready-made negative* examples, with no extra compute or annotation needed. This converts failed exploration in difficult cases (hard or low-prevalence studies) into useful gradients, moving the policy away from low-quality modes and toward the ground-truth. We also introduce a simple adaptive mixing weight to increase DPO’s influence when zero-reward groups are frequent, which gradually tapers once GRPO provides informative advantages, thereby allowing OraPO to salvage otherwise wasted batches, stabilise training, and accelerate convergence. .

Moreover, to provide OraPO with clinically faithful supervision, we introduce a **FactScore**-based reward [53] designed to address Challenge (ii). We treat the generated report as its own rationale, extract *atomic clinical facts*, and test their *entailment* against the ground-truth label set. Facts that support a label contribute positively; unsupported or contradictory labels incur penalties, yielding a dense, interpretable per-label reward. This stands in sharp contrast to the original GRPO reward, which assigns reward solely to the final answer while treating the reasoning trace largely as

a byproduct of generation, without ensuring it is factually grounded or consistently aligned with the answer. Our reward design ensures the generated report (serving in lieu of reasoning trace) not only produces a correct cognitive reasoning trace similar to that of an experienced radiologist but also yields correct diagnostic findings, which is crucial in domains such as healthcare in which safety is of paramount importance.

Extensive experiments are conducted on CheXpert Plus [7] and MIMIC-CXR [7] datasets, where OraPO sets the new SOTA performance for RRG, surpassing the previous best model with F1 score and notably achieving **160.8% improvement in recall**. This higher recall is clinically meaningful. In automated RRG tasks, recall is often prioritised over precision, as false negatives (missed abnormalities) can delay or prevent timely treatment, whereas false positives generally result only in additional clinical review [55, 62, 69, 71]. Previous studies [39, 97] emphasise maintaining high recall to avoid critical finding omissions, recognising that such oversights carry more serious clinical consequences than reduced precision. The proposed method achieves this clinical effectiveness while being exceptionally data-efficient, using **2–3 orders of magnitude less training data** (*i.e.* trained on only **1K** samples compared to the current best baseline MambaXray-L’s **1.27M** samples and other baselines’ minimum of 223K samples). At the same time, OraPO is compute-efficient: it is based on Qwen2.5-VL-3B, a small-sized VLM on modest hardware (**4× A10 GPUs**).

In summary, our contributions are **threefold**:

1. We present **OraPO**, an Oracle (DPO) educated GRPO that turns failed explorations into direct preference supervision significantly improving data- and compute-efficiency. To the best of our knowledge, this is the **first** work to integrate direct preference learning with GRPO-based RL for efficient training.
2. We design a **Facts Reward** for RL, performing fact-level entailment against ground-truth labels, yielding dense, interpretable feedback that aligns reports with diagnostic facts while avoiding long reasoning traces.
3. We show that OraPO achieves new SOTA performance for RRG, surpassing the previous best model (MambaXray-L) in clinical effectiveness F1 and recall while using **2–3 orders of magnitude less training data**.

2. Related Work

Automatic Radiology Report Generation (RRG) aims to turn imaging studies (*e.g.*, chest X-rays) into clinically usable free-text reports. Early approaches adopted seq2seq networks (*e.g.*, CNN–RNN and Transformer) [12, 36, 43, 98, 99] for training on large paired corpora such as MIMIC-CXR [37]. Recent work advances three fronts: (i)

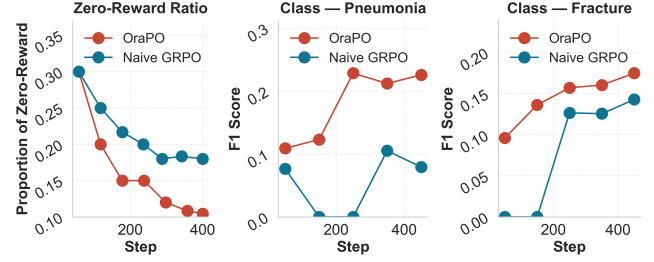


Figure 2. *Left*: Cumulative proportion of *zero-reward* batches (reward batch mean = 0) vs. training step on CheXpert Plus [83]. **OraPO** suppresses zero-reward frequency faster than naïve GRPO. *Centre/Right*: Class-level F1 on the CheXpert Plus validation set [83] across checkpoints for two *clinically challenging* and *rare* classes: **Pneumonia** (2.70%) and **Fracture** (4.05%). **OraPO** learns earlier and maintains higher F1 than naïve GRPO.

knowledge- and rule-guided generation, which injects medical priors to curb hallucination [5] and performs disease-aware image–text alignment for trustworthy text [34, 65]; (ii) *leveraging prior studies and multi-view images* to improve coverage and consistency [17, 28, 47]; and (iii) *scaling/standardizing with pre-training*, releasing multi-stage pre-training recipes [14, 18, 28, 33, 83]. In parallel, LLM-driven RRG explores instruction-tuned or preference-aligned training [41, 46, 76]. Despite progress, these approaches often rely on large paired corpora and/or costly computation resources, leading to deficiencies in data efficiency and practical scalability.

Group Relative Policy Optimisation (GRPO) is a critic-free, reinforcement learning *objective* for generative LLMs, introduced in *DeepSeekMath* [72]. It samples multiple completions per prompt, computes group-normalised advantages from relative rewards, and updates the policy with clipped ratios to a frozen reference; removing the critic reduces memory/compute versus PPO while preserving stability [27, 40, 56, 72, 103]. Recent variants mainly target optimisation stability: DR-GRPO [49] mitigates length bias and skewed group rewards; CISPO [8] clips importance-sampling weights; GSPO [105] adopts sequence-level clipping for MoE; DAPO [100] resamples until a non-zero reward appears. However, these methods were not designed to improve data efficiency: most target optimisation stability while leaving data volume and training epochs essentially unchanged, and some of them often increase sampling/compute [16, 73, 80, 100].

Direct Preference Optimisation (DPO) trains a policy from *preference pairs* to prefer the chosen response over the rejected one [109]. Recent variants [101] streamline or improve the robustness of this recipe: SimPO [52] applies length-normalised mean log-probabilities with a simplified reference; Cal-DPO [89] calibrates the implicit reward to human scales; ORPO [30] drops the reference via an odds-

ratio objective integrated into SFT; and KTO [24] learns from unary thumbs-up/down signals instead of pairs. In our RRG framework, DPO naturally serves as an *oracle step* inside **OraPO** to convert those failed GRPO rollouts into useful preference updates against the ground-truth, leading to data- and compute-efficient learning.

3. Methodology

In this section, we first introduce the background of Group Relative Policy Optimisation (GRPO) and Direct Preference Optimisation (DPO) (§3.1). We then identify a failure mode of GRPO in radiology report generation (RRG), when a pretrained model produces many low-quality groups (*e.g.*, zero rewards) that yield little or no gradient. This observation motivates our proposed **OraPO** algorithm, which efficiently converts such failed explorations into preference supervision (§3.2). Finally, we design the **Facts Reward**, a fact-level entailment signal that aligns reports with diagnostic labels and provides dense, interpretable feedback for reinforcement learning (§3.3).

3.1. Preliminaries

Group Relative Policy Optimisation (GRPO). GRPO [27, 72] is a computation- and memory-efficient RL objective for aligning generative models (LLMs/VLMs), introduced for reasoning-heavy settings and now used broadly because of its efficient critic-free design and proven effectiveness. It replaces the value critic with *group-normalized* advantages computed among multiple completions sampled for the same input. Given a study-prompt pair (x, p) , the policy π_θ draws a group of K sequences $\{\hat{y}_j\}_{j=1}^K \sim \pi_\theta(\cdot | x, p)$ and receives scalar rewards $r_j = r(x, \hat{y}_j)$. GRPO forms a group baseline and scale,

$$\bar{r} = \frac{1}{K} \sum_{j=1}^K r_j, \quad \sigma = \sqrt{\frac{1}{K} \sum_{j=1}^K (r_j - \bar{r})^2 + \varepsilon}, \quad (1)$$

and defines the group-normalised advantage:

$$A_j = \frac{r_j - \bar{r}}{\sigma}. \quad (2)$$

Let the proximal policy optimisation (PPO) ratio be:

$$\rho_j = \exp(\log \pi_\theta(\hat{y}_j | x, p) - \log \pi_{\theta_{\text{old}}}(\hat{y}_j | x, p)), \quad (3)$$

where θ_{old} denotes the parameters of the behaviour policy used to sample the current group (*i.e.*, the parameters before the update). Then the GRPO loss over a batch is defined as:

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E} \left[\min(\rho_j A_j, \text{clip}(\rho_j, 1 - \epsilon, 1 + \epsilon) A_j) \right] + \lambda_{\text{KL}} \mathbb{E} [\text{KL}(\pi_\theta(\cdot | x, p) \| \pi_{\text{ref}}(\cdot | x, p))], \quad (4)$$

where ϵ is the clipping threshold, $\lambda_{\text{KL}} \geq 0$ controls regularization toward a frozen reference policy π_{ref} . In practice, GRPO samples K completions per prompt, computes r_j , normalizes within-group to obtain A_j , and updates π_θ with the clipped-ratio objective in (4), thereby achieving critic-free policy optimisation.

Direct Preference Optimisation (DPO). DPO [109] is a widely used, preference-alignment objective for generative LLMs/VLMs. It updates the policy to prefer *chosen* completions over *rejected* ones relative to a frozen reference model. Given a preference dataset $\mathcal{P} = \{(x, p, y^+, y^-)\}$, where y^+ is preferred to y^- for the same input-prompt pair (x, p) , DPO maximizes the log-probability margin of the preferred sample after subtracting the reference log-probabilities. Let π_θ be the trainable policy and π_{ref} the frozen reference. The expectation below is taken over preference pairs drawn from $\mathcal{P}$. With temperature $\tau > 0$, define

$$\Delta^+ = \log \pi_\theta(y^+ | x, p) - \log \pi_{\text{ref}}(y^+ | x, p), \quad (5)$$

$$\Delta^- = \log \pi_\theta(y^- | x, p) - \log \pi_{\text{ref}}(y^- | x, p), \quad (6)$$

and the DPO loss

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E} \left[\log \sigma(\tau (\Delta^+ - \Delta^-)) \right], \quad (7)$$

where $\sigma(\cdot)$ is the logistic sigmoid.

3.2. OraPO: Oracle-educated GRPO

Exploration Failure of GRPO in RRG. On radiology report generation, applying vanilla GRPO to a base VLM with limited radiology knowledge often fails to explore effectively. The root reason is that the base model carries very limited domain knowledge about radiology. Most base VLMs [107] are pre-trained on general vision-language corpora and tasks, for example, everyday-image captioning on MS COCO [44, 87] and open-domain visual question answering on VQA v2 [25], rather than medical imaging and reporting. Hence, it tends to miss or misstate rare and clinically challenging findings (*e.g.*, small pneumothorax, subtle interstitial edema, or linear atelectasis) when asked to produce clinical narratives. For many chest X-ray inputs, the sampled group of completions is uniformly low quality with near-zero rewards. Thus, the policy receives little learning signal to drive effective exploration.

Empirically, before any domain-specific fine-tuning, the pretrained backbone attains only 0.034 macro-F1 (Tab. 4) on clinical effectiveness. As shown in Fig. 2 (left), in the first 50 optimisation steps approximately 30% of groups have all rewards equal to zero. Under the GRPO construction, such groups give $\bar{r} \approx 0$ and $\sigma \approx \varepsilon$, so $A_j = \frac{r_j - \bar{r}}{\sigma} \approx 0$ for all j . Consequently, the clipped-ratio term in Eq. (4) contributes essentially no gradient, and the update is dominated by the KL regularizer to π_{ref} rather than by task reward. In effect, the optimizer wastes GPU resources by

sampling K completions per prompt without extracting useful signals, which both slows convergence and squanders valuable compute.

Zero-Reward Rate (ZRR). To resolve GRPO’s exploration failure on RRG, we first *detect* when a prompt yields little or no usable signal, then adapt learning accordingly. We introduce a *Zero-Reward Rate* (ZRR) that quantifies, per prompt, how severely GRPO has stalled.

For prompt x_i with a GRPO group of K completions $\{\hat{y}_{i,j}\}_{j=1}^K$ and rewards $r_{i,j} \in [0, 1]$, define the raw zero-reward rate:

$$z_i = \frac{1}{K} \sum_{j=1}^K \mathbf{1}[r_{i,j} = 0], \quad (8)$$

here, $z_i = 1$ indicates an all-zero group (no learning signal) and $z_i = 0$ indicates at least one non-zero reward (GRPO has exploitable signal). To reduce step-to-step noise, we maintain an exponential moving average (EMA):

$$\tilde{z}_i^{(t)} = \alpha \tilde{z}_i^{(t-1)} + (1 - \alpha) z_i^{(t)}, \quad \alpha \in [0, 1), \quad (9)$$

computed per prompt (or per minibatch) at step t .

We then map $\tilde{z}_i^{(t)}$ to a bounded, monotonic mixing weight that will later control how much oracle education to apply:

$$w_i^{(t)} = \text{clip}\left(w_{\min} + (w_{\max} - w_{\min}) [\tilde{z}_i^{(t)}]^\gamma, w_{\min}, w_{\max}\right), \quad (10)$$

where $w_{\min} \leq w_i^{(t)} \leq w_{\max}$ enforces a narrow operating range for stability. The exponent γ shapes sensitivity: larger γ concentrates education on the most severe failures while leaving ordinary prompts to GRPO.

Integrating Two Objectives. With the ZRR-derived weight $w_i^{(t)}$ in hand, we modulate learning per prompt: when GRPO provides usable signal ($w_i^{(t)}$ is small), we emphasise exploration and exploitation; when GRPO stalls ($w_i^{(t)}$ is large), we shift toward direct education via preference alignment. We combine both signals in a single objective over a batch of size B :

$$\mathcal{L}_{\text{OraPO}} = \frac{1}{B} \sum_{i=1}^B \left[(1 - w_i^{(t)}) \mathcal{L}_{\text{GRPO}}(x_i, p_i) + w_i^{(t)} \mathcal{L}_{\text{DPO}}(x_i, p_i) \right], \quad (11)$$

where $\mathcal{L}_{\text{GRPO}}(x_i, p_i)$ is the GRPO loss in Eq. (4) evaluated on the sampled completions (x_i, p_i) , and $\mathcal{L}_{\text{DPO}}(x_i, p_i)$ is the DPO loss in Eq. (7) computed on preference pairs associated with the same prompt (x_i, p_i) . We set the positive to the ground-truth report y_i^* and take negatives from all other model completions:

$$y_i^+ = y_i^*, \quad y_i^- \in \{\hat{y}_{i,j}\}_{j=1}^K \sim \pi_\theta(\cdot | x_i, p_i) \quad (12)$$

where $\{\hat{y}_{i,j}\}_{j=1}^K$ is the entire group sampled from $\pi_\theta(\cdot | x_i, p_i)$. We then instantiate $\mathcal{L}_{\text{DPO}}(x_i, p_i)$ by applying Eq. (7) over pairs $\{(x_i, p_i, y_i^+, y_i^-)\}$. This construction reuses low-reward GRPO generation as a preference negative rather than discarding them. As a result, it enables DPO training without requiring any additional hard-negative mining or auxiliary generators. Generated rollouts with zero or low rewards serve as high-quality, domain-specific negative samples that are otherwise difficult to obtain manually.

Intuitively, a **high ZRR** flags failure cases where preference *education* must lead. OraPO responds by increasing $w_i^{(t)}$ and applying DPO updates that teach the policy to reject the sampled negatives and align with the ground-truth report. This converts groups into labeled preferences and supplies gradients exactly when GRPO has none/few. A **low ZRR** indicates that GRPO is uncovering usable reward. This can happen on simpler studies, and it also emerges after DPO has taught the model the cues for difficult findings (e.g., small apical pneumothorax or subtle interstitial edema), enabling a shift from education to exploration and exploitation. In practice, we observe the intended training dynamic, as shown in Fig. 2: early training is education-heavy with larger ratio of zero-reward groups, higher $w^{(t)}$ is implied, then $w^{(t)}$ gradually declines as the policy learns and GRPO takes over.

What does OraPO Buy Us? As shown in Fig. 2, **OraPO** drives down the zero-reward ratio faster, and to a lower plateau than naïve GRPO, implying fewer wasted rollouts and more gradient-carrying updates. Corroborating this, the class-level performance curves for *Pneumonia* and *Fracture* (clinically challenging, low-prevalence labels) show that **OraPO** improves *earlier* and sustains *higher* F1 across checkpoints. The coupled trend (rapid ZRR decline $\rightarrow$ earlier F1 lift on hard classes) indicates that DPO is converting failed exploration into effective supervision rather than merely stabilising optimisation, yielding both better sample efficiency and stronger performance where it matters.

This initiates a **self-reinforcing data flywheel**: as the model learns through Oracle education and GRPO’s exploration–exploitation process, it produces increasingly informative negative samples. These, in turn, carry higher reward signals, pushing the model to achieve even greater rewards and completing a positive, compounding feedback cycle.

3.3. FactScore-based Reward Design

OraPO needs a reward that is reliable and efficient to compute. Many RL-based RRG approaches optimize report-level proxies or embedding similarity, which favors fluent paraphrases over verifiable clinical facts and can yield plausible but clinically incorrect narratives. We introduce a FactScore-based reward (**FactS**) that treats the report as its

own rationale: extract atomic clinical facts and test each against the ground-truth label set to produce a dense, fact-entailment signal for reinforcement learning.

Step 1: Extract atomic clinical facts. Inspired by FactScore [53], we define a *fact* as an atomic, verifiable clinical statement (e.g. finding, attribute, and optionally location), for instance, “no pleural effusion”, “linear atelectasis at the left base”. Given a generated report $\hat{y}_i$, we prompt GPT-4.1 [63] to extract a set of atomic facts:

$$\mathcal{F}(\hat{y}_i) = \{s_{i,k}\}_{k=1}^{K_i}, \quad (13)$$

ensuring each $s_{i,k}$ is concise and verifiable.

Step 2: Per-label entailment against ground-truth. Let $\mathcal{L}$ be the fixed label set (e.g. 14 CheXpert labels) and let $z_{i,\ell}^* \in \{0, 1\}$ be the ground-truth for label $\ell \in \mathcal{L}$ on study x_i . We then run an LLM-based entailment check that maps the fact set to label-level predictions:

$$\hat{z}_{i,\ell} = \mathbf{1}[\exists s \in \mathcal{F}(\hat{y}_i) \text{ s.t. } \text{entails}(s, \ell) = \text{true}] \quad (14)$$

for all $\ell \in \mathcal{L}$. Explicit contradictions to ℓ are treated as false positives for that label (discouraging unsupported claims).

Step 3: A dense, interpretable reward. Based on $(\hat{z}_{i,\ell}, z_{i,\ell}^*)_{\ell \in \mathcal{L}}$, we compute standard per-instance precision P_i and recall R_i , and the reward as an F_β score

$$r(x_i, \hat{y}_i) = F_{\beta,i} = \frac{(1 + \beta^2) P_i R_i}{\beta^2 P_i + R_i + \xi}, \quad (15)$$

with $\beta > 1$ placing more weight on recall (penalizing missed positive findings) and $\beta < 1$ emphasizing precision; ξ is a small constant for stability. The reward is dense and label-wise interpretable, directly supervising diagnostic correctness while remaining token-efficient. It requires no critic, no auxiliary reward model, nor long reasoning traces. By anchoring each sentence to specific clinical facts and aligning them with ground-truth labels, the FactS Reward reduces unsupported claims and improves coverage of required findings, providing a faithful signal for OraPO.

4. Experiments

4.1. Experimental Setup

Datasets. CheXpert Plus [7] is a recent large-scale paired image–text resource that links chest X-ray DICOMs to fully de-identified radiology reports, organized into report subsections and annotated for 14 target pathologies (i.e. disease labels). The public release includes $\sim 223\text{K}$ image–report pairs across $\sim 187.7\text{K}$ studies and $\sim 64.7\text{k}$ patients, totaling $\sim 36\text{M}$ report tokens. MIMIC-CXR [37] is a large, publicly available paired image–text resource of chest X-rays in DICOM format with contemporaneous, de-identified radiology reports. It contains 377,110 images from 227,943 imaging studies covering 65,379 patients. Each study includes one or more views (typically frontal and/or lateral)

Table 1. Experimental results on the CheXpert Plus dataset [7] using **task-specific RRG algorithms**. We report macro-averaged **Precision, Recall, and F1** across 14 CheXpert pathologies. **Train Size** is the number of training samples. Some baselines train on multiple corpora and/or in multiple stages. Best and second best are in **bold** and underline, respectively.

Algorithm	Venue	Precision, Recall, F1	Train Size
R2GenRL [67]	ACL22	0.193, 0.229, 0.196	223K
XProNet [78]	ECCV22	0.314, 0.247, 0.259	223K
MSAT [84]	MICCAI22	0.044, 0.142, 0.057	223K
ORGan [31]	ACL23	0.288, 0.287, 0.277	223K
M2KT [93]	MIA21	0.044, 0.142, 0.058	223K
TIMER [88]	CHIL23	<u>0.345</u> , 0.238, 0.234	223K
CvT2DistilGPT2 [58]	AIM23	0.285, 0.252, 0.246	223K
R2Gen [12]	EMNLP20	0.318, 0.200, 0.181	223K
R2GenCMN [13]	ACL21	0.329, 0.241, 0.231	223K
Zhu et al. [111]	MICCAI23	0.217, 0.308, 0.205	223K
CAMANet [79]	IEEE JBHI23	0.328, 0.224, 0.216	223K
ASGMD [91]	ESWA24	0.146, 0.108, 0.055	223K
Token-Mixer [94]	IEEE TMI23	0.309, 0.270, 0.288	223K
PromptMRG [35]	AAAI24	0.258, 0.265, 0.281	223K
R2GenGPT [86]	Meta-Rad.23	0.315, 0.244, 0.260	223K
WCL [92]	EMNLP21	0.335, 0.259, 0.256	223K
R2GenCSR [81]	arXiv24	0.315, 0.247, 0.259	223K
Wang et al. [82]	VIS INT25	0.175, 0.099, 0.078	223K
CheXagent [15]	arXiv24	0.373, 0.194, 0.186	8.5M
MambaXray-B [83]	CVPR25	0.333, 0.264, 0.273	1.27M
MambaXray-L [83]	CVPR25	0.377 , 0.319, 0.335	1.27M
VLCI [11]	IEEE TIP25	0.341, 0.175, 0.163	223K
Ours	–	0.237, 0.832 , 0.341	1K

Table 2. Experimental results on the CheXpert Plus benchmark [7] across diverse **LLMs/VLMs**, all supervised fine-tuned with R2GenGPT [86] on CheXpert Plus. We report macro-averaged **Precision, Recall, and F1** across 14 CheXpert pathologies. The **Params** listed denotes the parameters that need to be tuned in the training phase. **FT-Size** indicates the number of samples used for fine-tuning. Best and second best are in **bold** and underline, respectively.

LLM/VLM	Year	Precision, Recall, F1	Params	FT-Size
Vicuna-V1.5 [108]	2023	0.334, 0.258, 0.276	6.7B	223K
Qwen-1.5 [2]	2024	0.303, 0.233, 0.241	7.7B	223K
Qwen-2 [2]	2024	0.313, 0.269, 0.261	7.6B	223K
InternLM [6]	2024	0.307, <u>0.274</u> , <u>0.284</u>	7.3B	223K
Llama-2 (7B) [75]	2023	0.315, 0.244, 0.260	6.7B	223K
Llama-2 (13B) [75]	2023	0.321, 0.254, 0.267	13.0B	223K
Llama-3 [22]	2024	0.306, 0.232, 0.222	8.0B	223K
Llama-3.1 [22]	2024	0.295, 0.251, 0.242	8.0B	223K
GPT2-Medium [68]	2019	0.358 , 0.186, 0.165	354M	223K
Orca-2 (7B) [54]	2023	0.330, 0.251, 0.271	6.7B	223K
Orca-2 (13B) [54]	2023	0.317, 0.242, 0.257	13.0B	223K
Deepseek-LLM [4]	2024	<u>0.336</u> , 0.256, 0.253	6.9B	223K
Yi-1.5 (6B) [96]	2024	0.322, 0.229, 0.226	6.1B	223K
Yi-1.5 (9B) [96]	2024	<u>0.336</u> , 0.241, 0.243	8.8B	223K
InternVL-2 [6]	2023	0.196, 0.127, 0.132	8.0B	223K
MiniCPM-V2.5 [95]	2024	0.254, 0.152, 0.122	8.4B	223K
Ours	–	0.237, 0.832 , 0.341	3B	1K

and a free-text report that often follows Findings and Impression sections. For data-efficient fine-tuning, we sample 1,000 studies per dataset (i.e. 1,000 from CheXpert-Plus and 1,000 from MIMIC-CXR) with a balanced ground-truth

Table 3. Experimental results on the MIMIC-CXR dataset [7]. We report macro-averaged **Precision**, **Recall**, and **F1** across 14 CheXpert pathologies. **Train Size** is the number of training samples. Some baselines train on multiple corpora and/or in multiple stages. Best are in **bold**.

Algorithm	Venue	Precision, Recall, F1	Train Size
R2Gen [12]	EMNLP20	0.333, 0.273, 0.276	227K
CXR-RePaiR [23]	ML4H21	–, –, 0.274	227K
MRG/SILC [45]	TIP24	0.457 , 0.337, 0.330	227K
CXRMate [59]	IMU24	0.277, 0.351, 0.283	227K
HERGen [77]	ECCV24	0.415, 0.301, 0.317	145K
ChEX [57]	ECCV24	–, –, 0.326	227K
MambaXray-L [83]	CVPR25	0.371, 0.321, 0.340	1.27M
Git-CXR [74]	INFO25	–, –, 0.327	227K
MCA-RG [90]	MICCAI25	0.443, 0.306, 0.335	227K
Ours	–	0.242, 0.891 , 0.357	1K

Table 4. Ablation on CheXpert Plus [7]: Impact of the **FactS** reward and **OraPO**. The first row indicates the base Qwen2.5-VL-3B-Instruct [3] direct test results, without any fine-tuning. The second row corresponds to a naïve GRPO-based RLHF baseline using accuracy as the reward [27]. We report per-label **Precision**, **Recall**, and **F1** across 14 findings. Best are in **bold**.

FactS	GRPO	DPO	SFT	Train Size	Precision, Recall, F1
×	×	×	×	0	0.097, 0.104, 0.034
×	✓	×	×	1,000	0.026, 0.162, 0.089
✓	✓	×	×	1,000	0.204, 0.605, 0.291
✓	✓	✓	×	400	0.217, 0.732, 0.296
✓	✓	×	✓	1,000	0.171, 0.176, 0.106
✓	✓	✓	×	1,000	0.237 , 0.832 , 0.341

Table 5. Experimental results on CheXpert validation set with gold labels from certified radiologists [32].

Method	Train Size	Precision, Recall, F1
GPT4.1 [63]	–	0.292, 0.305, 0.253
GPT5 (Thinking) [64]	–	0.261, 0.421, 0.301
CheXagent [15]	8.5M	0.462, 0.226, 0.254
MambaXray-B [83]	1.27M	0.362, 0.254, 0.238
MambaXray-L [83]	1.27M	0.395, 0.326, 0.280
Ours	1K	0.234, 0.641, 0.288

label set via CheXpert labeller [32]. We also exclude reports with missing sections. Following recent works [83], we report results on the test split and extract labels from both generated and GT reports by the CheXpert labeller. We further evaluate on the CheXpert Plus human gold set, whose expert-verified labels provide a more reliable basis for evaluation and stronger conclusions.

Baselines. Following the recent benchmarking study [83], we compare our proposed approach against a broad suite of methods: **28 task-specific RRG models** spanning diverse architectures and **16 open-source LLMs/VLMs**, as listed in Tables 1, 2 and 3. All task-specific algorithms are fine-tuned on at least the full training set of the dataset

used in each experiment (CheXpert Plus or MIMIC-CXR), with some methods additionally incorporating multi-source corpora. Thus, our 1K training set sampled from CheXpert Plus, represents $\sim 0.45\%$ of the standard 223K split; and only $\sim 0.079\%$ of the 1.27M used by MambaXray [11], which is the current SOTA model. A similar ratio applies to MIMIC-CXR.

Evaluation Metric. We follow [83] to report *clinical effectiveness* as macro-averaged over labels: for each of the 14 conditions we compute Precision, Recall, and F1 on the test set, then take the unweighted mean across labels so frequent and rare conditions contribute equally.

Implementation Details. We fine-tune a **3B** Qwen-2.5-VL policy [3] on the **1,000** CheXpert Plus studies in an **RL-only** setting (no SFT/alignment), using Hugging Face TRL’s GRPOTrainer with transformers and accelerate. Training runs on four NVIDIA A10 GPUs (24 GB) with PyTorch 2.7.1/CUDA 12.1. We empirically set batch size $B = 16$. For each prompt, we sample $K = 8$ completions, compute the **FactS** reward per completion, and optimize the objective in Eq. (11). More detailed description of the hyper-parameters for OraPO can be found in the Supplementary.

4.2. Main Results and Analysis on CheXpert Plus

Against RRG task-specific algorithms. As shown in Table 1, **OraPO+FactS** sets the new SOTA **F1 = 0.341 (best)** and **Recall = 0.832 (best)** on CheXpert Plus, significantly surpassing best-performing baselines by **+160.8%** in recall. When compared with the most recent baseline, VLCl, our proposed method outperforms VLCl by **+109.2%** in F1 and **+375.4%** in recall. While our precision trails MambaXray-L, this is a deliberate trade-off between emphasizing per-finding coverage and accurate statements of detected findings. In practice, recall-oriented drafts are preferred. Since radiologists perform the final review, they tend to favour AI assistance that has higher sensitivity (even at the cost of precision) [29]. Overall, our method sets the new SOTA performance in F1 and recall, with **three orders of magnitude** less training data.

Against Alignment-based SFT for LLM/VLM. In this experiment, we compare our framework against R2GenGPT [86], a popular alignment-based SFT approach that is widely used for training large LLM/VLM. We report the performance of R2GenGPT [86] with different LLM/VLM backbones in Table 2. Our approach is trained with a compact **3B** policy on **1K** CheXpert Plus samples. Under this small-data setup, **OraPO** attains **F1 = 0.341** and **Recall = 0.832**, significantly outperforming larger aligned backbones trained on 223K. These results validate that our proposed **OraPO** with fact-level supervision is a superior learning paradigm to standard vision–language alignment, achieving higher performance with **less than**

half backbone parameters and markedly less training data.

Evaluation on Human Gold Label Set. To obtain a more faithful evaluation, we assess models on the CheXpert validation set with human gold labels: 200 studies from 200 patients independently annotated by three board-certified radiologists, where a majority vote defines the ground truth [32]. Table 5 reports macro Precision/Recall/F1 across the 14 findings. With only 1K training cases, our method reaches 0.234/0.641/0.288, surpassing MambaXray-B, MambaXray-L and CheXagent in both Recall and F1. These gains arrive while using 0.1% of MambaXray’s data and around 0.01% of CheXagent’s.

We also evaluate two commercial APIs (GPT-4.1 [63] and GPT-5 Thinking [64]) on the CheXpert validation set with human gold labels, using the identical image input and report-generation prompt. GPT-5 is configured with a medium reasoning level. As reported in Table 5, our model outperforms GPT-4.1 by +110.2% in recall and +13.8% in F1. Relative to GPT-5 Thinking, our model achieves +52.3% higher recall, though F1 is -4.3% lower. Critically, compared to GPT-5 Thinking, our compact 3B model delivers real-time inference (3.3 s/image vs. 25.2 s/image), a lightweight and transparent footprint (3B vs. trillions of parameters), and cheaper deployment that eliminates per-token API fees. These advantages make our method substantially more cost-efficient and suitable for time-critical clinical use than calling commercial APIs. Overall, the results confirm that our improvements persist when measured against expert gold labels, while being dramatically more data-efficient.

4.3. Main Results and Analysis on MIMIC-CXRs

On MIMIC-CXR, our approach attains $F1 = 0.357$ and $Recall = 0.891$ using just 1K training samples, as shown in Table 3. This yields a +5.0% F1 gain over the best baseline (0.340, MambaXray-L) and a +153.8% recall gain over the strongest baseline recall (0.351, CXRMate). The markedly higher recall under macro averaging indicates substantially fewer omissions across both common and rare pathologies, aligning with clinical preference for high sensitivity and omission avoidance [55, 62, 69, 71, 106]. At the same time, the overall F1 is the best among all baselines, despite a precision of 0.242. Together with the CheXpert-Plus results, these findings show that OraPO generalises across datasets and achieves state-of-the-art clinical effectiveness with 2–3 orders-of-magnitude less data than prior approaches.

4.4. Ablation Study

To study the effectiveness and efficiency of our proposed method, we evaluate four model variants on CheXpert Plus using macro **Precision**, **Recall**, and **F1**. As reported in Table 4, the second row is a naïve GRPO-based RLHF baseline that uses *accuracy* as reward (as in DeepSeek-R1

[27]), which yields only marginal gains over the base model without any domain-specific training (row 1). Adding **FactS** to GRPO (row 3) lifts F1 from 0.089 to 0.291 (+227%), Recall from 0.162 to 0.605 (+274%), and Precision from 0.026 to 0.204 (+685%), showing that fact-level entailment provides dense, clinically aligned supervision. Layering **OraPO** on top (row 6) further improves F1 to 0.341 (+17.2% vs. **FactS**-only), with Recall improved by (+37.5%) and Precision improved by (+16.2%). Notably, **OraPO+FactS** with only **400** studies (row 4) attains $F1 = 0.296$, $Precision = 0.217$ and $Recall = 0.732$ —surpassing the **FactS**-only 1K setting ($F1 = 0.291$, $Precision = 0.204$, and $Recall = 0.605$)—providing evidence of OraPO’s superb *data efficiency*. We further observe that the **GRPO+SFT** variant (with FactS) collapses to very low recall (0.176) and F1 (0.106), a **-70.9%** and **-63.6%** drop vs. FactS+GRPO, respectively. The core issue here is that plain SFT only imitates ground-truth prose: the model learns “what to say when it’s right,” but not “what not to say,” so reports that omit all true positives aren’t explicitly discouraged, leading to recall and F1 collapses. Beyond that, SFT on small datasets is known to be unstable and sample-inefficient: fine-tuning with limited data often over-fits, degrading generalisation [21, 66]. OraPO fixes this by explicitly pushing probability mass away from “no-TP” outputs such that the recall is effectively boosted when trained with limited set. Overall, **FactS** supplies faithful, sentence-level rewards, while **OraPO** converts zero-reward groups into preference updates, yielding higher precision and recall and faster convergence on scarce yet clinically important signals.

5. Conclusion

In this work, we present **OraPO**, an oracle-educated RL training algorithm for high-quality radiology report generation under tight data and compute budgets. To our knowledge, we are the *first* to *couple DPO with GRPO*: when GRPO exploration fails, we recycle those rollouts as negatives and apply a lightweight DPO step to recover useful gradients. In parallel, we design the **FactS** reward to provide dense, fact-level, clinically faithful supervision tailored to RRG. Comprehensive experiments show that **OraPO sets the new SOTA performance while using 2–3 orders of magnitude fewer training samples**. In future work, we target four directions: (1) Extend OraPO beyond RRG to sparse-signal tasks (*e.g.* multi-step math, code) where vanilla GRPO struggles. (2) Study scaling across model size and data/training budgets beyond our current 3B/1K setup. (3) Tackle non-verifiable rewards via rubric-driven signals (*e.g.* Rubrics as Rewards [26]). (4) Enrich FactS with fine-grained attribute entailment (location, laterality, size/severity, temporal change) to reduce diagnostic attribute misalignment.

References

- [1] Julián Nicolás Acosta, Siddhant Dogra, Subathra Adithan, Kay Wu, Michael Moritz, Stephen Kwak, and Pranav Rajpurkar. The impact of AI assistance on radiology reporting: A pilot study using simulated AI draft reports. *CoRR*, abs/2412.12042, 2024. 2
- [2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, and et al. Qwen technical report. *CoRR*, abs/2309.16609, 2023. 6
- [3] Shuai Bai, Keqin Chen, and et al. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025. 7
- [4] Xiao Bi, Deli Chen, and et al. Deepseek LLM: scaling open-source language models with longtermism. *CoRR*, abs/2401.02954, 2024. 6
- [5] Shenshen Bu, Taiji Li, Yuedong Yang, and Zhiming Dai. Instance-level expert knowledge and aggregate discriminative attention for radiology report generation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14194–14204, 2024. 3
- [6] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, and et al. Internlm2 technical report. *CoRR*, abs/2403.17297, 2024. 6
- [7] Pierre J. Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven Q. H. Truong, Chu The Chuong, and Curtis P. Langlotz. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *CoRR*, abs/2405.19538, 2024. 1, 2, 3, 6, 7
- [8] Aili Chen, Aonian Li, , and et al. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *CoRR*, abs/2506.13585, 2025. 3
- [9] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. SFT or rl? an early investigation into training r1-like reasoning large vision-language models. *CoRR*, abs/2504.11468, 2025. 2
- [10] Liang Chen, Xueting Han, Li Shen, Jing Bai, and Kam-Fai Wong. Beyond two-stage training: Cooperative sft and rl for llm reasoning. *CoRR*, abs/2509.06948, 2025. 2
- [11] Weixing Chen, Yang Liu, Ce Wang, Jiarui Zhu, Guanbin Li, Cheng-Lin Liu, and Liang Lin. Cross-modal causal representation learning for radiology report generation. *IEEE Transactions on Image Processing*, 34:2970–2985, 2025. 6, 7
- [12] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, 2020. 3, 6, 7
- [13] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. In *Proc. Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing, (ACL/IJCNLP)*, pages 5904–5914, 2021. 6
- [14] Zhuoxiao Chen, Yadan Luo, Zheng Wang, Mahsa Baktashmotlagh, and Zi Huang. Revisiting domain-adaptive 3d object detection by reliable, diverse and class-balanced pseudo-labeling. In *Proc. International Conference on Computer Vision (ICCV)*, pages 3691–3703, 2023. 3
- [15] 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, Emily Bao Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S. Chaudhari, and Curtis P. Langlotz. Chexagent: Towards a foundation model for chest x-ray interpretation. *CoRR*, abs/2401.12208, 2024. 6, 7
- [16] Zhuoxiao Chen, Zixin Wang, Yadan Luo, Sen Wang, and Zi Huang. DPO: dual-perturbation optimization for test-time adaptation in 3d object detection. In *Proc. International Conference on Multimedia (MM)*, pages 4138–4147, 2024. 3
- [17] Zhuoxiao Chen, Yadan Luo, Zixin Wang, Zijian Wang, and Zi Huang. Open-crb: Toward open world active learning for 3d object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(10):8336–8350, 2025. 3
- [18] Zhuoxiao Chen, Junjie Meng, Mahsa Baktashmotlagh, Yonggang Zhang, Zi Huang, and Yadan Luo. MOS: model synergy for test-time adaptation on lidar-based 3d object detection. In *Proc. International Conference on Learning Representations (ICLR)*, 2025. 3
- [19] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, and et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. 2
- [20] Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer K. 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. 2
- [21] Yupei Du and Dong Nguyen. Measuring the instability of fine-tuning. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6209–6230, 2023. 8
- [22] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. 6
- [23] Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y. Ng, and Pranav Rajpurkar. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Machine Learning for Health, (ML4H@NeurIPS)*, pages 209–219, 2021. 7
- [24] Kavin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. KTO: model alignment as prospect theoretic optimization. *CoRR*, abs/2402.01306, 2024. 4
- [25] Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision*, 127(4):398–414, 2019. 4
- [26] Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath,

- Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains, 2025. 8
- [27] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shiron Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081): 633–638, 2025. 3, 4, 7, 8
- [28] Ibrahim Ethem Hamamci, Sezgin Er, and Bjoern H. Menze. Ct2rep: Automated radiology report generation for 3d medical imaging. In *Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 476–486, 2024. 3
- [29] Nathaniel Hendrix, Kathryn P Lowry, Joann G Elmore, William Lotter, Gregory Sorensen, William Hsu, Geraldine J Liao, Sana Parsian, Suzanne Kolb, Arash Naeim, et al. Radiologist preferences for artificial intelligence-based decision support during screening mammography interpretation. *Journal of the American College of Radiology*, 19(10):1098–1110, 2022. 7
- [30] Jiwoo Hong, Noah Lee, and James Thorne. ORPO: monolithic preference optimization without reference model. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 11170–11189, 2024. 3
- [31] Wenjun Hou, Kaishuai Xu, Yi Cheng, Wenjie Li, and Jiang Liu. ORGAN: observation-guided radiology report generation via tree reasoning. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8108–8122, 2023. 6
- [32] Jeremy Irvin, Pranav Rajpurkar, and et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proc. Conference on Artificial Intelligence (AAAI)*, pages 590–597, 2019. 7, 8
- [33] Songtao Jiang and et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *CoRR*, abs/2510.08668, 2025. 3
- [34] Songtao Jiang, Tuo Zheng, Yan Zhang, Yeying Jin, Li Yuan, and Zuozhu Liu. Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models. In *Findings of the Association for Computational Linguistics (EMNLP Findings)*, 2024. 3
- [35] Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proc. Conference on Artificial Intelligence*, pages 2607–2615, 2024. 6
- [36] Baoyu Jing, Pengtao Xie, and Eric P. Xing. On the automatic generation of medical imaging reports. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2577–2586, 2018. 3
- [37] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *CoRR*, abs/1901.07042, 2019. 2, 3, 6
- [38] Yunsoo Kim, Jinge Wu, Yusuf Abdulle, Yue Gao, and Honghan Wu. Enhancing human-computer interaction in chest x-ray analysis using vision and language model with eye gaze patterns. In *Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 184–194. Springer, 2024. 2
- [39] Paras Lakhani, Woojin Kim, and Curtis P Langlotz. Automated detection of critical results in radiology reports. *Journal of digital imaging*, 25(1):30–36, 2012. 3
- [40] Alan Lee and Harry Tong. Reinforcement learning for LLM reasoning under memory constraints. *CoRR*, abs/2504.20834, 2025. 2, 3
- [41] 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. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 3
- [42] Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 1537–1547, 2018. 2
- [43] Jia Syuen Lim, Zhuoxiao Chen, Zhi Chen, Mahsa Baktashmotlagh, Xin Yu, Zi Huang, and Yadan Luo. Dipex: Dispersing prompt expansion for class-agnostic object detection. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3
- [44] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Proc. European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 4
- [45] Aohan Liu, Yuchen Guo, Jun-Hai Yong, and Feng Xu. Multi-grained radiology report generation with sentence-level image-language contrastive learning. *IEEE Transactions on Image Processing*, 43(7):2657–2669, 2024. 7
- [46] Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. Bootstrapping large language models for radiology report generation. In *Proc. Conference on Artificial Intelligence (AAAI)*, pages 18635–18643, 2024. 3
- [47] Kang Liu, Zhuoqi Ma, Xiaolu Kang, Yunan Li, Kun Xie, Zhicheng Jiao, and Qiguang Miao. Enhanced contrastive learning with multi-view longitudinal data for chest x-ray report generation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10348–10359, 2025. 3
- [48] Mingyang Liu, Gabriele Farina, and Asuman E. Ozdaglar. UFT: unifying supervised and reinforcement fine-tuning. *CoRR*, abs/2505.16984, 2025. 2
- [49] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *CoRR*, abs/2503.20783, 2025. 3
- [50] Xingtai Lv, Yuxin Zuo, and et al. Towards a unified view of large language model post-training. *CoRR*, abs/2509.04419, 2025. 2
- [51] Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, Xiaochen Ma, Zhen Hao Wong, Junbo Niu, Chengyu Shen, Runming He, Bin Cui, and Wentao Zhang. Learning what reinforcement learning can’t: Interleaved online fine-tuning for hardest questions. *CoRR*, abs/2506.07527, 2025. 2

- [52] Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3
- [53] Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12076–12100, 2023. 2, 6
- [54] Arindam Mitra, Luciano Del Corro, and et al. Orca 2: Teaching small language models how to reason. *CoRR*, abs/2311.11045, 2023. 6
- [55] Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P. Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT)*, pages 5288–5304, 2021. 2, 3, 8
- [56] Youssef Mroueh, Nicolas Dupuis, Brian Belgodere, Apoorva Nitsure, Mattia Rigotti, Kristjan H. Greenewald, Jirí Navrátil, Jerret Ross, and Jesus Rios. Revisiting group relative policy optimization: Insights into on-policy and off-policy training. *CoRR*, abs/2505.22257, 2025. 2, 3
- [57] Philip Müller, Georgios Kaissis, and Daniel Rueckert. Chex: Interactive localization and region description in chest x-rays. In *Proc. European Conference on Computer Vision (ECCV)*, pages 92–111, 2024. 7
- [58] Aaron Nicolson, Jason Dowling, and Bevan Koopman. Improving chest x-ray report generation by leveraging warm starting. *Artificial Intelligence*, 144:102633, 2023. 6
- [59] Aaron Nicolson, Jason Dowling, Douglas Anderson, and Bevan Koopman. Longitudinal data and a semantic similarity reward for chest x-ray report generation. *Informatics in Medicine Unlocked*, 50:101585, 2024. 7
- [60] Aaron Nicolson, Jinghui Liu, Jason Dowling, Anthony N. Nguyen, and Bevan Koopman. e-health CSIRO at RRG24: entropy-augmented self-critical sequence training for radiology report generation. In *Proc. Workshop on Biomedical Natural Language Processing (BioNLP@ACL)*, pages 99–104, 2024. 2
- [61] John Nightingale, Stephen Etty, Beverley Snaith, Tracy Stevens, Richard Appleyard, and Stephen Kelly. Establishing the size and configuration of the imaging support workforce: a census of national workforce data in england. *BJR—Open*, 6(1):tzae026, 2024. 2
- [62] Tomoyuki Noguchi, Koji Yamashita, Shuji Matsuura, Ryotaro Kamei, Junki Maehara, Kiyomi Furuya, Shino Harada, Saki Adachi, and Yasushi Okada. Analysis of “visible in retrospect” to monitor false-negative findings in radiological reports. *Japanese Journal of Radiology*, 41(2): 219–227, 2023. 3, 8
- [63] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. 6, 7, 8
- [64] OpenAI. GPT-5 System Card. Technical report, OpenAI, 2025. 7, 8
- [65] Sang-Jun Park, Keun-Soo Heo, Dong-Hee Shin, Young-Han Son, Ji-Hye Oh, and Tae-Eui Kam. DART: disease-aware image-text alignment and self-correcting re-alignment for trustworthy radiology report generation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15580–15589, 2025. 3
- [66] Branislav Pecher, Ján Cegin, Róbert Belanec, Jakub Simko, Ivan Srba, and Mária Bielíková. Fighting randomness with randomness: Mitigating optimisation instability of fine-tuning using delayed ensemble and noisy interpolation. In *Findings of Empirical Methods in Natural Language Processing (EMNLP Findings)*, pages 11005–11044, 2024. 8
- [67] Han Qin and Yan Song. Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics (ACL Findings)*, pages 448–458, 2022. 6
- [68] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, San Francisco, CA, 2019. Technical report. 6
- [69] Vishwanatha M Rao, Serena Zhang, Julian N Acosta, Subathra Adithan, and Pranav Rajpurkar. Rexerr: Synthesizing clinically meaningful errors in diagnostic radiology reports. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pages 70–81. World Scientific, 2024. 3, 8
- [70] James V. Rawson, Dana Smetherman, and Eric Rubin. Short-term strategies for augmenting the national radiologist workforce. *American Journal of Roentgenology*, 222(6):e2430920, 2024. 2
- [71] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12):2176–2182, 2021. 3, 8
- [72] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. 2, 3, 4
- [73] Vaishnavi Shrivastava, Ahmed Awadallah, Vidhisha Balachandran, Shivam Garg, Harkirat Behl, and Dimitris Papailiopoulos. Sample more to think less: Group filtered policy optimization for concise reasoning. *CoRR*, abs/2508.09726, 2025. 2, 3
- [74] Iustin Sirbu, Iulia-Renata Sirbu, Jasmina Bogojeska, and Traian Rebedea. Git-cxr: End-to-end transformer for chest x-ray report generation. *Information*, 16(7):524, 2025. 7
- [75] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, and et al. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023. 6
- [76] Weijie Tu, Weijian Deng, Dylan Campbell, Yu Yao, Jiyang Zheng, Tom Gedeon, and Tongliang Liu. Ranked from within: Ranking large multimodal models without labels. In *Proc. Conference on Machine Learning, (ICML)*, 2025. 3
- [77] Fuying Wang, Shenghui Du, and Lequan Yu. Hergen: Elevating radiology report generation with longitudinal

- data. In *Proc. European Conference on Computer Vision (ECCV)*, pages 183–200, 2024. 7
- [78] Jun Wang, Abhir Bhalerao, and Yulan He. Cross-modal prototype driven network for radiology report generation. In *Proc. European Conference on Computer Vision (ECCV)*, pages 563–579, 2022. 6
- [79] Jun Wang, Abhir Bhalerao, Terry Yin, Simon See, and Yulan He. Camanet: Class activation map guided attention network for radiology report generation. *IEEE Journal of Biomedical and Health Informatics*, 28(4):2199–2210, 2024. 6
- [80] Liangyu Wang, Huanyi Xie, Xinhai Wang, Tianjin Huang, Mengdi Li, and Di Wang. Infinite sampling: Efficient and stable grouped RL training for large language models. *CoRR*, abs/2506.22950, 2025. 3
- [81] Xiao Wang, Yuehang Li, Fuling Wang, Shiao Wang, Chuanfu Li, and Bo Jiang. R2gencsr: Retrieving context samples for large language model based x-ray medical report generation. *CoRR*, abs/2408.09743, 2024. 6
- [82] Xiao Wang, Yuehang Li, Wentao Wu, Jiandong Jin, Yao Rong, Bo Jiang, Chuanfu Li, and Jin Tang. Pre-training on high-resolution x-ray images: an experimental study. *Visual Intelligence*, 3(1), 2025. 6
- [83] Xiao Wang, Fuling Wang, Yuehang Li, Qingchuan Ma, Shiao Wang, Bo Jiang, and Jin Tang. Cxpmrg-bench: Pre-training and benchmarking for x-ray medical report generation on chexpert plus dataset. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5123–5133, 2025. 2, 3, 6, 7
- [84] Zhanyu Wang, Mingkan Tang, Lei Wang, Xiu Li, and Luping Zhou. A medical semantic-assisted transformer for radiographic report generation. In *Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 655–664, 2022. 6
- [85] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 11558–11567, 2023. 2
- [86] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033, 2023. 6, 7
- [87] Zixin Wang, Dong Gong, Sen Wang, Zi Huang, and Yadan Luo. Is less more? exploring token condensation as training-free test-time adaptation. In *Proc. International Conference on Computer Vision (ICCV)*, pages 144–154, 2025. 4
- [88] Yuexin Wu, I-Chan Huang, and Xiaolei Huang. Token imbalance adaptation for radiology report generation. In *Proc. Conference on Health, Inference, and Learning (CHIL)*, pages 72–85, 2023. 6
- [89] Teng Xiao, Yige Yuan, Huaisheng Zhu, Mingxiao Li, and Vasant G. Honavar. Cal-dpo: Calibrated direct preference optimization for language model alignment. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3
- [90] Qilong Xing, Zikai Song, Youjia Zhang, Na Feng, Junqing Yu, and Wei Yang. MCA-RG: enhancing llms with medical concept alignment for radiology report generation. In *Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 380–390, 2025. 7
- [91] Youyuan Xue, Yun Tan, Ling Tan, Jiaohua Qin, and Xuyu Xiang. Generating radiology reports via auxiliary signal guidance and a memory-driven network. *Expert Systems With Applications*, 237(Part B):121260, 2024. 6
- [92] An Yan, Zexue He, Xing Lu, Jiang Du, Eric Y. Chang, Amilcare Gentili, Julian J. McAuley, and Chun-Nan Hsu. Weakly supervised contrastive learning for chest x-ray report generation. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4009–4015, 2021. 6
- [93] Shuxin Yang, Xian Wu, Shen Ge, Zhuozhao Zheng, S. Kevin Zhou, and Li Xiao. Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, 86:102798, 2023. 6
- [94] Yan Yang, Jun Yu, Zhenqi Fu, Ke Zhang, Ting Yu, Xianyun Wang, Hanliang Jiang, Junhui Lv, Qingming Huang, and Weidong Han. Token-mixer: Bind image and text in one embedding space for medical image reporting. *IEEE Transactions on Medical Imaging*, 43(11):4017–4028, 2024. 6
- [95] Yuan Yao, Tianyu Yu, and et al. Minicpm-v: A GPT-4V level MLLM on your phone. *CoRR*, abs/2408.01800, 2024. 6
- [96] Alex Young, Bei Chen, and et al. Yi: Open foundation models by 01.ai. *CoRR*, abs/2403.04652, 2024. 6
- [97] Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9):100802, 2023. 2, 3
- [98] Hongyang Yu and Hongjiang C. Yu. Powerful molecule generation with simple convnet. *Bioinformatics*, 38(13):3438–3443, 2022. 3
- [99] Hongyang Yu and Hongjiang C. Yu. Tensorvae: a simple and efficient generative model for conditional molecular conformation generation. *Transactions on Machine Learning Research*, 2024, 2024. 3
- [100] Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xianguang Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: an open-source LLM reinforcement learning system at scale. *CoRR*, abs/2503.14476, 2025. 2, 3
- [101] Suqin Yuan, Xingrui Yu, Jiyang Zheng, Lei Feng, Dadong Wang, Ivor Tsang, and Tongliang Liu. Mitigating mismatch within reference-based preference optimization. In *Proc. International Conference on Learning Representations (ICLR)*, 2026. 3
- [102] Wenhao Zhang, Yuexiang Xie, Yuchang Sun, Yanxi Chen, Guoyin Wang, Yaliang Li, Bolin Ding, and Jingren Zhou.

- On-policy RL meets off-policy experts: Harmonizing supervised fine-tuning and reinforcement learning via dynamic weighting. *CoRR*, abs/2508.11408, 2025. [2](#)
- [103] Yi Zhang, Ruihong Qiu, Jiajun Liu, and Sen Wang. Roler: Effective reward shaping in offline reinforcement learning for recommender systems. In *Proc. International Conference on Information and Knowledge Management (CIKM)*, pages 3269–3278, 2024. [3](#)
- [104] Yi Zhang, Lili Xie, Ruihong Qiu, Jiajun Liu, and Sen Wang. Beyond static LLM policies: Imitation-enhanced reinforcement learning for recommendation. In *Proc. International Conference on Data Mining (ICDM)*, pages 903–912, 2025. [2](#)
- [105] Chujie Zheng, Shixuan Liu, , and et al. Group sequence policy optimization. *CoRR*, abs/2507.18071, 2025. [3](#)
- [106] Jiyang Zheng, Islam Nassar, Thanh Tien Vu, Xu Zhong, Yang Lin, Tongliang Liu, Long Duong, and Yuan-Fang Li. Meddcr: Learning to design agentic workflows for medical coding. *CoRR*, abs/2511.13361, 2025. [8](#)
- [107] Jiyang Zheng, Jialiang Shen, Yu Yao, Min Wang, Yang Yang, Dadong Wang, and Tongliang Liu. Chain-of-focus prompting: Leveraging sequential visual cues to prompt large autoregressive vision models. In *Proc. International Conference on Learning Representations (ICLR)*, 2025. [4](#)
- [108] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2023. [6](#)
- [109] Han Zhong, Guhao Feng, Wei Xiong, Xinle Cheng, Li Zhao, Di He, Jiang Bian, and Liwei Wang. DPO meets PPO: Reinforced token optimization for RLHF. In *International Conference on Machine Learning Workshop on Models of Human Feedback for AI Alignment (ICML Workshop)*, 2024. [3](#), [4](#)
- [110] Zijian Zhou, Miaojing Shi, Meng Wei, Oluwatosin Alabi, Zijie Yue, and Tom Vercauteren. Large model driven radiology report generation with clinical quality reinforcement learning. *CoRR*, abs/2403.06728, 2024. [2](#)
- [111] Qingqing Zhu, Tejas Sudharshan Mathai, Pritam Mukherjee, Yifan Peng, Ronald M. Summers, and Zhiyong Lu. Utilizing longitudinal chest x-rays and reports to pre-fill radiology reports. In *Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 189–198, 2023. [6](#)