PREFERENCE FINE-TUNING FOR FACTUALITY IN CHEST X-RAY INTERPRETATION MODELS WITHOUT HUMAN FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Radiologists play a crucial role by translating medical images into actionable reports. However, the field faces staffing shortages and increasing workloads. While automated approaches using vision-language models (VLMs) show promise as assistants, they require exceptionally high accuracy. Most current VLMs in radiology rely solely on supervised fine-tuning (SFT). Meanwhile, in the general domain, additional preference fine-tuning has become standard practice. The challenge in radiology lies in the prohibitive cost of obtaining radiologist feedback. To address this challenge, we propose an automated pipeline for preference feedback, focusing on chest X-ray (CXR) report generation. Our method leverages publicly available datasets containing pairs of images and radiologist-written reference reports with an LLM-as-a-Judge mechanism, eliminating the need for *additional radiologist feedback*. We evaluate and benchmark five direct alignment algorithms. Our results show up to a 57.4% improvement in average GREEN scores, a LLM-based metric for evaluating CXR reports, compared to the SFT baseline. We study reward overoptimization via length exploitation, with reports lengthening by up to 3.2x. To assess a potential alignment tax, we benchmark on six additional diverse tasks, finding no significant degradations. A reader study involving four board-certified radiologists indicates win rates of up to 0.62 over the SFT baseline, and macro-averaged F1 scores improve by up to 6.7%, highlighting the utility of our approach.

1 INTRODUCTION

X-rays are one of the most frequently collected imaging studies in clinical practice, with the advantages of wide availability, cost-effectiveness, and low radiation dose. Chest X-rays (CXR) are used for diverse purposes in clinical practice, with approximately 1.4 billion diagnostic X-ray examinations collected per year in the world (PAHO, 2012; Organization et al., 2016; Cid et al., 2024). The amount and significance of CXRs can pose a burden for radiologists and a potential negative impact for patients without timely interpretation, especially for those containing critical lesions (Ruutiainen et al., 2013; Hanna et al., 2017; Bruls & Kwee, 2020; Bhargavan et al., 2002; Lyon et al., 2015; Rimmer, 2017).

Recent strides in generative vision-language models (VLMs) hold promising implications for this high-stakes and low-data field (Liu et al., 2024; Radford et al., 2021). Typically pre-trained using image-text contrastive learning and supervised fine-tuned using causal language modeling (a.k.a. next-token prediction), recent VLMs have started to demonstrate promising performance in CXR interpretation (Chen et al., 2024; Bannur et al., 2024). In high-stakes fields like radiology, where accurate medical descriptions directly influence disease diagnosis and treatment decisions, the generated outputs must maintain high factual accuracy to ensure patient safety.

However, recent studies have shown that supervised fine-tuning (SFT) might be insufficient in the post-training process. For example, Hong et al. (2024) illustrate the limitation of SFT by training on a preference dataset, containing "good" and "bad" completions. By tracking the log probabilities of each during the course of training, they show that the log probabilities of the bad completions inadvertently increase alongside the good completions. Preference fine-tuning methods, such as reinforcement learning from human feedback (RLHF) (Ziegler et al., 2020; Stiennon et al., 2020; Ouyang et al., 2022), using Proximal Policy Optimization (PPO) (Schulman et al., 2017) or REINFORCE (Williams, 1992), and direct alignment algorithms (DAAs), effectively alleviate this problem by employing a negative gradient to lower probabilities of "bad" completions (Tajwar et al., 2024). In fact, most recent LLMs (Ouyang et al., 2022; Bai et al., 2022a; Touvron et al., 2023; Jiang et al., 2024; Team et al., 2024) include some form of preference fine-tuning in their post-training pipeline. Yet, this approach has not yet been investigated within the medical vision-language domain.

The primary challenge hindering the application of preference fine-tuning in the post-training of VLMs in fields such as radiology is the prohibitive cost of obtaining radiologist preferences at scale. To overcome this obstacle, we introduce an automated pipeline for generating preference data, focusing on the critical task of CXR report generation. Specifically, we leverage the availability of reference reports written by radiologists in a clinical setting within large, publicly available, datasets such as MIMIC-CXR (Johnson et al., 2019), and GREEN (Ostmeier et al., 2024), a recent state-of-the-art LLM-based metric for evaluating CXR reports, to annotate generated reports in a factually grounded fashion. Our approach enables us to obtain high-quality preference datasets in a fully automated and scalable manner. Adding to previous works, we also incorporate information from prior images when available, mirroring how radiologists use priors to dictate the ground truth reports. Using our proposed method, we systematically study how DAAs can be used to enhance the factual correctness in generative VLMs *without any additional radiologist feedback*, by rigorously benchmarking a representative subset of DAAs. An overview of our preference fine-tuning pipeline is available in Fig. 1. To structure our paper, we formulate it around the following research questions: (i) How do different alignment algorithms compare on the CXR report generation task? (ii) Are there any degradations in performance, to tasks other than the one being aligned, as a result of the alignment (i.e. an alignment tax (Askell et al., 2021; Ouyang et al., 2022))? (iii) How do the resultant policies compare from a clinical perspective? Our contributions are as follows:

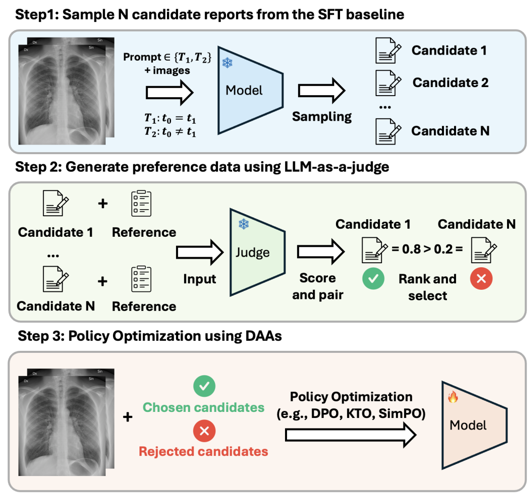

Figure 1: Overview of our preference fine-tuning pipeline. $t_0 = t_1$ and $t_0 \neq t_1$ indicate whether a comparison is made to a prior image.

- We introduce an automated pipeline for preference data generation, focusing on CXR report generation, circumventing the prohibitively expensive task of obtaining preference feedback from radiologist at scale.
- We systematically evaluate five representative DAAs on the CXR report generation task. To the best of our knowledge, this is the first time that a systematic analysis of DAAs has been performed in this setting. Our findings show significant performance gains, over the SFT baseline (CheXagent (Chen et al., 2024)), in terms of average GREEN (Ostmeier et al., 2024), 26.4-42.3% and 17.5-57.4% on the MIMIC-CXR (Johnson et al., 2019) and CheXpert Plus (Chambon et al., 2024) datasets, respectively, with top performance achieved by Direct Preference Optimization (DPO) (Rafailov et al., 2023).
- We study reward overoptimization in terms of length exploitation in the context of CXR report generation. Significant reward overoptimization, or hacking, is observed for some DAAs. The average length of the generated reports increase by approximately a factor of 2.5 on the MIMIC-CXR data and 3.2 on the CheXpert Plus data in the worse case (DPO).
- We benchmark our models post alignment on set of diverse tasks to assess whether there is an alignment tax. We observe no performance degradations, that are statistically significant, on six tasks: view classification, coarse-grained image classification, single disease identification, multi disease identification, VQA, and image-text reasoning.

- We study the aligned policies from a clinical perspective. First, we elicit feedback on our aligned policies from human experts via a reader study including four board-certified radiologist. One key finding is that verbosity, resulting from length exploitation, is significantly penalized. In particular, DPO and Identity Preference Optimization (IPO) (Azar et al., 2023), the two DAAs yielding the most significant length exploitation, received win rates well below 0.5. Odds-Ratio Preference Optimization (ORPO) (Hong et al., 2024), on the other hand, achieves a win rate of 0.62. Second, we extract 14 categories using the CheXbert labeler (Smit et al., 2020) and compute the F1 score. The clinical efficacy performance is well aligned with the reader study, clearly emphasizing limitations of the policies aligned by DPO and IPO, while illustrating the clinically relevant gains obtained by ORPO and Kahneman-Tversky optimization (KTO) (Ethayarajh et al., 2024), with up to 8.4% and 6.7% increase in micro and macro averages, respectively.

2 Preliminaries and Related Work

In this section, we provide an overview of vision-language models in both general and medical domains, DAAs, and reward overoptimization.

2.1 Vision-language models

Vision-language models (Radford et al., 2021; Li et al., 2021; 2022; 2023; Liu et al., 2024) are a multi-modal extension to LLMs. In this setting, the prompt x contains images and/or text. Typical tasks include Vision Question Answering (VQA) and image captioning (e.g., report generation in the field of radiology). There are also a line of works to extend VLMs to the medical domain (Thawkar et al., 2023; Hyland et al., 2023; Chaves et al., 2024) which mainly focus on CXR interpretation due to the wide availability of public datasets (Johnson et al., 2019; Chambon et al., 2024). However, even with strong LLMs and vision-backbones, VLMs have been observed to "hallucinate" and produce outputs that are not factually grounded in the image (Zhou et al., 2024). Such hallucinations represent a significant risk in high-stakes healthcare fields such as radiology. Similar to Zhou et al. (2024), we pose the problem of hallucinations as an alignment problem and propose tackling it via preference fine-tuning.

2.2 Direct alignment algorithms

RLHF (Ziegler et al., 2020; Stiennon et al., 2020; Ouyang et al., 2022) is based on the constrained reward maximization objective

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(y|x)}[R_\psi(x,y)] - \beta\mathbb{D}_{\text{KL}}\left[\pi_\theta(y|x)||\pi_{\text{ref}}(y|x)\right], \tag{1}$$

where $\mathbb{D}_{\text{KL}}$ is the Kullback-Leibler (KL) divergence and π_{ref} is the reference policy. R_ψ is the proxy reward model learned on a dataset of human preferences $\mathcal{D} = \{x^{(n)}, y_c^{(n)}, y_r^{(n)}\}_{n=1}^N$, where y_c and y_r denote the chosen and rejected completions for the prompt x, such that $y_c \succ y_r|x$.

Whilst extremely powerful, RLHF is computationally heavy, involves several steps, and can be tricky to implement in practice. Relatively recently, a new class of algorithms called DAAs (Rafailov et al., 2024) have become increasingly popular.[1] This class of algorithms re-parameterize the reward model via a change-of-variables using the closed-formed solution to the objective in equation 1, effectively bypassing both the reward modeling and reinforcement learning (RL) stages. Resulting in algorithms that remain performant yet computationally more light weight and easier to implement. DPO (Rafailov et al., 2023) was the first in this category and remains one of the most popular versions.

DPO exploits the closed-formed solution to equation 1, $\pi(y|x) \propto \pi_{\text{ref}}(y|x)\exp(R(x,y)/\beta)$ and the Bradley-Terry (BT) model (Bradley & Terry, 1952) of human preferences $p^*(y_1 \succ y_2|x) = \sigma(\exp(R^*(x,y_1)) - \exp(R^*(x,y_2)))$, where R^* is the latent reward model, $\exp$ is the exponential function, and σ is the logistic function. The reward can be isolated and written as a function of the policy $R(x,y) = \beta\log\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}$. This re-parametrization can be applied to the latent reward R^* and substituted into the BT model, $p^*(y_1 \succ y_1|x) = \sigma\left(\beta\log\frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)} - \beta\log\frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)}\right)$, where

[1] In this paper, we use this terminology more loosely than in Rafailov et al. (2024).

π^* is the optimal policy corresponding to the latent reward. Crucially, the probability of human preferences is now in terms of the policy instead of the reward model. A parameterized policy π_θ can then be learned via a simple classification loss over the preference data

$$\mathcal{L}_{\text{DPO}}(\theta) = -\log \sigma \left(\beta \log \frac{\pi_\theta \left(y_c \mid x\right)}{\pi_{\text{ref}} \left(y_c \mid x\right)} - \beta \log \frac{\pi_\theta \left(y_r \mid x\right)}{\pi_{\text{ref}} \left(y_r \mid x\right)} \right).$$

Hence, this change-of-variables has transformed a loss over rewards into a loss over policies.

2.3 Reward overoptimization

The reward model in equation 1 is learned, and therefore an imperfect proxy of the ground truth reward R^*. As this proxy is optimized, ground truth performance might saturate or even deteriorate.[2] This reward overoptimization, or hacking, phenomena was first studied in Gao et al. (2023) for RLHF. The KL divergence term in equation 1 is included explicitly to mitigate this issue, but has proven insufficient (Gao et al., 2023). Despite not fitting an explicit reward model, similar behavior has been observed empirically for DAAs (Rafailov et al., 2024).

Length exploitation, the tendency to learn to produce excessively verbose completions, is one common dimension of reward overoptimization, observed in both RLHF and for DAAs. For instance, Park et al. (2024) showed that DPO amplifies verbosity bias embedded in the preference data. In this work, we explore this phenomenon in the context of preference fine-tuning of medical VLMs.

3 Experimental setup

In this section, we present our experimental setup. We structure the paper around the following research questions: (i) How do different alignment algorithms compare on the CXR report generation task? (ii) To what extent is there an alignment tax? (iii) How does the aligned policy compare to the SFT baseline from a clinical perspective?

3.1 Dataset and Baseline

Dataset We use the MIMIC-CXR (Johnson et al., 2019) dataset for training, validation and testing. The image-report pairs consists of one or two CXRs and the corresponding free-text findings section. The reports describe the findings in the image at a static timepoint (a single image or two images from the same timepoint) or describe findings using also a prior image (two images from different timepoints). Radiology reports usually include information from prior timepoints in the clinic, but this remains understudied in the context of automated CXR report generations using VLMs.

To limit the computational burden, we randomly sample 80k examples as our training data. To test robustness for the CXR report generation task, we additionally include test data from the CheXpert Plus (Chambon et al., 2024) dataset. To evaluate whether there is an alignment tax, we additionally evaluate our aligned models on six tasks different from CXR report generation: view classification, coarse-grained image classification, single disease identification, multi disease identification, VQA, and image-text reasoning, using test data from five additional datasets RSNA (Shih et al., 2019), SIIM (American College of Radiology, 2019), OpenI (Demner-Fushman et al., 2016), SLAKE (Liu et al., 2021), and Rad-Restruct (Pellegrini et al., 2023) datasets.

Baseline We adopt CheXagent (Chen et al., 2024) as a representative example of a state-of-the-art, open source, VLM for CXR interpretation. It has been trained in the canonical way by first adapting the LLM to medical text by continued pre-training. Second, a vision encoder was adapted via vision pre-training, using contrastive learning on CXR image-text pairs. Third, the two modalities were merged by training a vision-language bridger, or adapter network, keeping the LLM and vision encoder frozen. Finally, the model was instruction tuned. In addition, CheXagent is of average size, 8B, for an open source model, providing a good balance between computational complexity and performance.

[2]As per Goodhart's law: "When a measure becomes a target, it ceases to be a good measure." (Gao et al., 2023).

3.2 PREFERENCE DATA

Expert human feedback from radiologists is the gold standard for preference data generation and evaluation in CXR report generation. However, scaling is impractical due to the limited availability of radiologists for large-scale annotation tasks. In the general domain, leveraging LLMs for cost effective preference data generation has been proposed (Bai et al., 2022b; Dubois et al., 2023; Lee et al., 2024). Zheng et al. (2023), focusing on the related task of automated evaluation, introduced the terminology of "LLM-as-a-Judge" and categorized evaluation methods into pairwise, single answer, and reference-guided grading. In the general domain, pairwise grading is the most common both for preference data generation Dubois et al. (2023); Lee et al. (2024) and evaluation Zheng et al. (2023); Dubois et al. (2024).

These existing methods, however, are tailored for uni-modal, general-domain LLMs and do not directly apply to our multi-modal setting, which involves both visual and textual data. Moreover, factual grounding is essential in medical report generation to ensure clinical reliability. To overcome these challenges, we propose using reference-guided grading, leveraging publicly available datasets that contain paired prompts—including images—and *radiologist-written* reference reports. This abundance of high-quality references allows us to provide factually grounded annotations without the need for a multi-modal Judge, setting our approach apart from prior studies with multi-modal Judges, or reward models, such as Sun et al. (2024).

GREEN (Ostmeier et al., 2024) is a state-of-the-art metric for radiology report evaluation, based on a single answer reference-guided LLM-as-a-Judge mechanism. While no metric is perfect, GREEN better reflect radiologist preferences than general domain metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and the BERTScore (Zhang et al., 2019), as well as radiology specific metrics such as F1RadGraph (Jain et al., 2021). Hence, we treat GREEN as the silver standard, employing it as a low-cost approximation of expert human judgment.

We obtain our preference data as follows: 1) for each example in the training data, we prompt the SFT baseline $N = 4$ times; 2) we get a GREEN reward for each of the generated reports, compared with the corresponding singular reference; 3) we set the chosen and rejected completions as the highest and lowest rewards, omitting the observation if all $N = 4$ scores are equivalent. This rejection rule results in the rejection of 1,246 (1.6%) examples. Summary statistics for the chosen and rejected subsets are available in Table 1. We also report summary statistics of the length (in words) of the generated reports as verbosity-bias is a well-known issue in preference fine-tuned LLMs evaluation (Park et al., 2024; Dubois et al., 2024). Notably, there is a slight verbosity bias in the chosen subset. We additionally illustrate the distributions of GREEN reward and report length in Fig. 4 in the Appendix.

	GREEN Reward			Report Length		
Subset	**Mean**	**Median**	**Std.**	**Mean**	**Median**	**Std.**
Chosen	0.629	0.600	0.248	56.3	54.0	20.2
Rejected	0.263	0.222	0.191	52.7	51.0	24.9

Table 1: Summary statistics of GREEN reward and report length in the chosen and rejected subsets.

3.3 ALIGNMENT ALGORITHMS

Due to compute constraints, we only consider offline DAAs and leave their online counterpart as well as on-policy RL algorithms to future work. However, even when restricting the focus to offline DAAs, there are more methods available than would be feasible to include. Hence, we choose representative algorithms from different categories. DPO is the original DAA and serves as our baseline. In addition to DPO, we consider:

- Identity Preference Optimization (IPO) (Azar et al., 2023) as an example of a DAA with generalized preference, relaxing the assumption of the Bradley-Terry model. The authors argue that this helps mitigate over-fitting issues observed in DPO even when preferences are transitive. Relatively recent work has shown that IPO indeed seems to be less prone to reward overoptimization (Rafailov et al., 2024).
- Kahneman-Tversky optimization (KTO) (Ethayarajh et al., 2024) as an example of a DAA that does not require preference pairs, but instead only binary feedback on whether a completion is

desirable or undesirable. This type of data is much more ubiquitous in practice. In addition, for any given dataset of preference pairs KTO provides twice the number of examples.

- SimPO (Meng et al., 2024) as an example of a DAA that does not require a reference policy, meaning that it is computationally lighter weight. SimPO suggests directly using the average log probabilities as implicit reward function, as this is what is relevant for generation. Taking the average over the generated tokens means that the objective is "length controlled" and has the potential of mitigating length exploitation.
- Odds-Ratio Preference Optimization (ORPO) (Hong et al., 2024), almost outside of the definition of DAAs, is not based on the RLHF objective and instead appends an additional penalty directly to the negative log likelihood used in SFT. This adds a "negative gradient", using the terminology in Tajwar et al. (2024), which will help reduce the log probabilities of rejected completions.

These algorithms also differ in their dependence of a strong SFT baseline, capable of producing high quality completions. For instance, DPO require a strong baseline, whereas KTO has been shown to work well even without a prior SFT phase (Ethayarajh et al., 2024). ORPO takes this to the extreme as it in principle combines the SFT and alignment phases. An overview of all DAAs considered in this paper is available in Table 2. Implementation details are available in §A.2.

Algorithm	**Objective**	**Preference pairs**	**Reference**	**Length controlled**	**Relative wall-clock time**
DPO	$-\log\sigma\left(\beta\log\frac{\pi_\theta(y_c\|x)}{\pi_{\text{ref}}(y_c\|x)} - \beta\log\frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)}\right)$	✓	✓	×	1.0
KTO	$-\lambda_c\sigma\left(\beta\log\frac{\pi_\theta(y_c\|x)}{\pi_{\text{ref}}(y_c\|x)} - z_{\text{ref}}\right) + \lambda_r\sigma\left(z_{\text{ref}} - \beta\log\frac{\pi_\theta(y_r\|x)}{\pi_{\text{ref}}(y_r\|x)}\right)$, where $z_{\text{ref}} = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\beta\mathbb{D}_{\text{KL}}\left(\pi_\theta(y\|x)\|\|\pi_{\text{ref}}(y\|x)\right)]$	×	✓	×	2.2
IPO	$\left(\log\frac{\pi_\theta(y_c\|x)}{\pi_{\text{ref}}(y_c\|x)} - \log\frac{\pi_\theta(y_r\|x)}{\pi_{\text{ref}}(y_r\|x)} - \frac{1}{2\tau}\right)^2$	✓	✓	×	1.0
SimPO	$\log\sigma\left(\frac{\beta}{\|y_c\|}\log\pi_\theta\left(y_c \mid x\right) - \frac{\beta}{\|y_r\|}\log\pi_\theta\left(y_r \mid x\right) - \gamma\right)$	✓	×	✓	0.7
ORPO	$-\log p_\theta(y_c\|x) - \lambda\log\sigma\left(\log\frac{p_\theta(y_c\|x)}{1-p_\theta(y_c\|x)} - \log\frac{p_\theta(y_r\|x)}{1-p_\theta(y_r\|x)}\right)$, where $p_\theta(y\|x) = \exp\left(\frac{1}{\|y\|}\log\pi_\theta(y\|x)\right)$	✓	×	×	0.7

Table 2: Overview of the DAAs considered in this paper. Preference pairs indicates whether the method requires paired data of accepted/rejected or only binary feedback indicating whether a completion is desirable/undesirable. Reference indicates whether an additional reference model is loaded during training. Length controlled indicates whether the objective directly controls for the length of the completions in order to mitigate over-optimization/reward hacking via verbosity bias. Wall-clock time, measured as total time to train one epoch, is relative to DPO.

4 EMPIRICAL ANALYSIS

In this section, we present our empirical analysis. We examine the performance on the CXR report generation tasks using five different DAAs (Question 1), investigate the presence of an alignment tax (Question 2), and explore our aligned policies from a clinical perspective (Question 3). Code to run all experiments, as well as all the examples in the reader study and their preference, will be made publicly available.

4.1 QUESTION 1: ALIGNMENT ALGORITHM

We report results for GREEN (Ostmeier et al., 2024), F1RadGraph (Jain et al., 2021), and the BERTScore (Zhang et al., 2019) on the MIMIC-CXR and CheXpert Plus test datasets in Table 3. Additional metrics are available in Table 11. DPO provides substantially higher GREEN scores on the MIMIC-CXR data yielding an improvement over the SFT baseline of 42.3%. In addition, the aligned policy generalizes well to data unseen in the preference fine-tuning stage, achieving an improvement of 57.4% on the CheXpert Plus data. However, according to the BERTScore, DPO is actually worse than the SFT baseline. IPO follows similar trends as DPO, with slightly lower improvements over the baseline. Only KTO and ORPO improve all metrics on both datasets. Between those, KTO is better in terms of GREEN: leading to a 36.7% improvement on the MIMIC-CXR data and 37.1% on the CheXpert Plus dataset, compared to 29.4% and 31.4% for ORPO. In addition, KTO is the top performer according to F1RadGraph, leading to a 20.6% and a 19.4% improvement on the MIMIC-CXR and CheXpert Plus datasets, respectively. ORPO also yields substantial improvements in F1RadGraph, 13.2% and 11.6% on the two datasets respectively. SimPO yields smaller improvements, achieving a 26.4% and a 17.5% increase in average GREEN on the MIMIC-CXR and CheXpert Plus datasets, respectively. Notably, we see very similar trends in two datasets despite representing two different distributions: MIMIC-CXR was collected in a emergency department (ED) and CheXpert Plus was collected from in- and out-patient centers.

	MIMIC-CXR			CheXpert		
Method	**GREEN** (↑)	**F1RadGraph** (↑)	**BERTScore**(↑)	**GREEN** (↑)	**F1RadGraph** (↑)	**BERTScore**(↑)
CheXagent	0.249	0.215	0.856	0.248	0.222	0.851
+DPO	**0.354 (0.105)**	0.247 (0.032)	0.830 (-0.026)	**0.391 (0.142)**	0.246 (0.024)	0.821 (-0.030)
+KTO	0.340 (0.091)	**0.260 (0.045)**	**0.862 (0.006)**	0.340 (0.092)	**0.265 (0.043)**	**0.859 (0.008)**
+IPO	0.349 (0.100)	0.252 (0.037)	0.846 (-0.010)	0.358 (0.110)	0.248 (0.026)	0.844 (-0.007)
+SimPO	0.315 (0.066)	0.225 (0.010)	0.854 (-0.002)	0.292 (0.043)	0.205 (-0.017)	0.844 (-0.006)
+ORPO	0.322 (0.073)	0.244 (0.029)	**0.862 (0.006)**	0.326 (0.078)	0.248 (0.026)	0.856 (0.005)

Table 3: Results on the MIMIC-CXR and CheXpert Plus test sets (with Δ compared to SFT baseline in brackets). Green shades for improvements and red shades for degradations. Shades are separated into bins of 10%, running from $> 0\%$ and $\leq 10\%$ up to $> 50\%$. Best results in bold.

One possible reason for the deteriorating performance observed in the BERTScore is verbosity bias, a common form of reward overoptimizaiton. The mean and standard deviation of report length, in addition to the relative added verbosity are available in Table 4. Indeed, DPO and IPO are both excessively verbose, resulting in an increase of average length by a factor 2.50 and 1.79 on the MIMIC-CXR dataset and 3.15 and 1.88 on the CheXpert dataset, respectively. KTO and ORPO also increase the average length, but significantly less so. SimPO, which is the only length controlled method considered, stays very close to the average length of the SFT baseline, or even decreases it.

	MIMIC-CXR		CheXpert	
Method	**Mean**	**Relative verbosity**	**Mean**	**Relative verbosity**
CheXagent	63.2 (23.5)	1.00	56.1 (28.2)	1.00
+DPO	157.6 (84.0)	2.50	176.5 (68.5)	3.15
+KTO	77.7 (33.0)	1.23	83.6 (46.4)	1.49
+IPO	113.2 (62.4)	1.79	105.3 (56.0)	1.88
+SimPO	63.6 (23.4)	1.01	51.0 (25.4)	0.91
+ORPO	69.0 (27.6)	1.09	82.8 (43.4)	1.48
Reference	66.2 (23.4)		58.4 (24.9)	

Table 4: Average length (with standard deviation in brackets). Relative verbosity is relative to the SFT baseline.

We plot average lengths against average GREEN for all aligned policies in Fig. 2. There is a very clear positive correlation. As was shown in Park et al. (2024) for DPO, we surmise that excessively verbose completions are a result of reward overoptimization in the form of length exploitation, due to verbosity biases embedded in the preference dataset. More results on verbosity are available in §A.4.

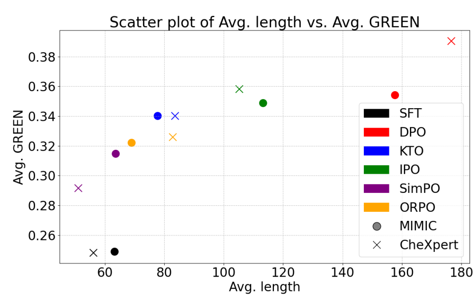

Figure 2: Scatter plot of average lengths against average GREEN.

The GREEN metric reported in Table 3 is an aggregate over six subcategories: (a) False report of a finding in the candidate, (b) Missing a finding present in the reference, (c) Misidentification of a finding's anatomic location/position, (d) Misassessment of the severity of a finding, (e) Mentioning a comparison absent in the reference, (f) Omitting a comparison detailing a change from a prior study. We report average error counts, considering clinically significant errors, for each of these subcategories in Table 5 on the MIMIC-CXR data. Interestingly, across all methods, only the first four subcategories (a-d) decrease on average, whereas for the last two (e-f), the frequency of errors actually increases compared to the SFT baseline. Since both (e) and (f) pertain to "comparisons", these errors may have been exacerbated by our setup, which treated both the task of generating reports for exams at a static timepoint (a single image or two images from the same timepoint) and exams using a prior image (images from different timepoints). This may have led to forgetting, resulting in more errors of type (e) and (f).

	Error subcategories in GREEN (↓)					
Method	**(a)**	**(b)**	**(c)**	**(d)**	**(e)**	**(f)**
CheXagent	1.82	2.40	0.241	0.342	0.071	0.040
+DPO	**1.16 (-0.668)**	2.43 (0.030)	**0.140 (-0.100)**	**0.286 (-0.056)**	0.123 (0.052)	**0.053 (0.013)**
+KTO	1.47 (-0.353)	**2.07 (-0.328)**	0.190 (-0.051)	0.385 (0.043)	0.093 (0.022)	0.055 (0.016)
+IPO	1.23 (-0.592)	2.37 (-0.025)	0.157 (-0.084)	0.292 (-0.050)	0.114 (0.043)	0.059 (0.019)
+SimPO	1.28 (-0.547)	2.39 (-0.013)	0.172 (-0.068)	0.302 (-0.040)	**0.087 (0.016)**	0.066 (0.026)
+ORPO	1.46 (-0.369)	2.14 (-0.256)	0.207 (-0.034)	0.378 (0.036)	0.092 (0.021)	0.054 (0.014)

Table 5: Average error counts for each subcategory in GREEN on the MIMIC-CXR test set (with Δ compared to SFT baseline in brackets). The subcategories are: (a) False report of a finding in the candidate, (b) Missing a finding present in the reference, (c) Misidentification of a finding's anatomic location/position, (d) Misassessment of the severity of a finding, (e) Mentioning a comparison that isn't in the reference, (f) Omitting a comparison detailing a change from a prior study. Green shades for improvements and red shades for degradations. Shades are separated into bins of 10%, running from $> 0\%$ and $\leq 10\%$ up to $> 50\%$. Best results in bold.

4.2 Question 2: Alignment tax

While RLHF is powerful, it has been observed that it might lead to performance degradations or, forgetting (Askell et al., 2021; Ouyang et al., 2022). Any degradations in performance due to alignment is loosely referred to as an alignment tax. Ouyang et al. (2022) assessed such an alignment tax by evaluating the aligned policies on several NLP benchmarks. Inspired by this, we benchmark the SFT baseline and the aligned policies on six different tasks: view classification, coarse-grained image classification, single disease identification, multi disease identification, VQA, and image-text reasoning using datasets listed in §3.1. Interestingly, despite fairly large gains in the CXR report generation tasks, there are no statistically significant degradations in these additional tasks.

Model	**View Classification**	**Binary Image Classification**	**Single Disease Identification**	**Multi Disease Identification**	**Visual Question Answering**	**Image-Text Reasoning**	**Avg.**
CheXagent	$98.6_{[97.7,99.4]}$	$83.1_{[79.7,86.6]}$	$61.1_{[57.9,64.2]}$	$67.8_{[65.2,70.2]}$	$62.4_{[59.8,64.8]}$	$66.6_{[61.8,71.1]}$	73.3
+DPO	$98.4_{[97.6,99.3]}$	$82.4_{[78.8,85.9]}$	$61.2_{[58.1,64.6]}$	$67.3_{[64.7,69.8]}$	$61.8_{[59.2,64.2]}$	$66.1_{[61.3,70.5]}$	72.9
+KTO	$98.6_{[97.7,99.4]}$	$82.1_{[78.5,85.7]}$	$61.8_{[58.6,64.9]}$	$68.3_{[65.8,70.8]}$	$62.5_{[60.0,65.1]}$	$66.6_{[61.6,71.6]}$	73.3
+IPO	$98.4_{[97.4,99.3]}$	$82.3_{[78.7,85.7]}$	$61.1_{[58.0,64.4]}$	$67.4_{[64.9,69.8]}$	$61.8_{[59.3,64.4]}$	$66.7_{[61.8,71.3]}$	73.0
+SimPO	$98.4_{[97.4,99.3]}$	$82.0_{[78.5,85.9]}$	$60.8_{[57.5,64.0]}$	$67.1_{[64.7,69.6]}$	$62.1_{[59.5,64.8]}$	$65.2_{[60.5,69.5]}$	72.6
+ORPO	$98.3_{[97.3,99.1]}$	$83.2_{[79.7,86.8]}$	$61.3_{[58.1,64.4]}$	$67.5_{[64.8,70.0]}$	$61.8_{[59.1,64.0]}$	$65.3_{[60.5,69.7]}$	72.9

Table 6: Performance on six tasks CheXagent is capable of other than CXR report generation, the task used for alignment. 95% confidence intervals in subscripts.

4.3 Question 3: Clinical perspective

Finally, we ask board-certified radiologists to analyze the generated reports both qualitatively and quantitatively. The key thing we strive for with the qualitative analysis is to understand how the verbosity materializes, and how this relates to GREEN. We found a particularly interesting example, where both DPO and SimPO achieve GREEN=1 but DPO is significantly more verbose. We show this example in Fig. 3. For brevity, we show only DPO and SimPO here. Results for all DAAs are available in §A.5. The generated text has been color coded as correct, incorrect, and repeated (i.e. exact repetition or semantically equivalent repetition). There is an

Method	**Win rate**	**SEP**
CheXagent		
+DPO	0.17	0.05
+KTO	0.55	0.06
+IPO	0.23	0.05
+SimPO	0.48	0.06
+ORPO	0.62	0.06

Table 7: Win rates against the SFT baseline and standard error of a proportions (SEP). Win rates are according to human experts (radiologist).

uncompleted sentence for DPO due to truncation. Both DPO and SimPO achieve an significant improvement over the SFT baseline. However, the difference in verbosity is stark. In addition, the added verbosity is mainly in terms of repetitions. This example emphasize the need for length regularization, currently employed in SimPO and but not DPO.

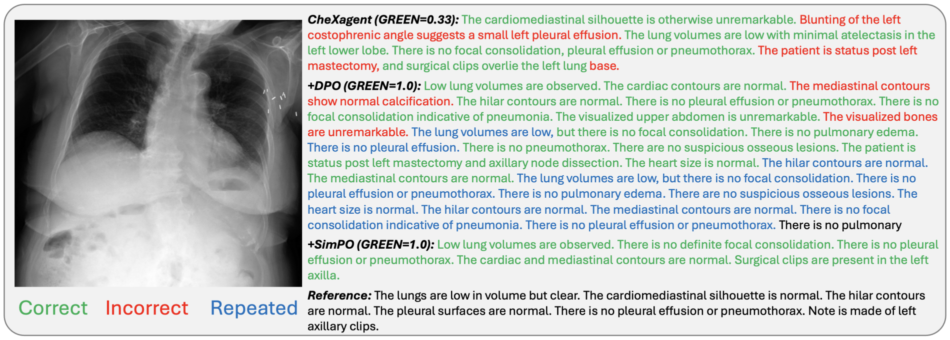

Figure 3: Qualitative results on one example from the MIMIC-CXR test set. The text in the generated reports is color coded as correct, incorrect, and repeated (i.e. exact repetitions or semantically equivalent).

Win rates, with respect to the SFT baseline, were obtained from a random subset of 60 examples in the MIMIC-CXR test data, yielding a total of 300 cases due to the five DAAs considered. These examples were read by four radiologist who were asked to indicate a preference between the SFT baseline and a generated report by one of the DAAs. We opted to elicit preferences instead of rankings using a Likert scale due to the higher variance of the latter. In addition to preferences, the ranker can also optionally give a reason why the choice is made. One stark difference in Table 7 compared to Table 3 is that DPO and IPO are now the worst performing alternatives. As shown in Table 8, the two most common reasons for preferring the SFT baseline over DPO and IPO were: *Selected report contains LESS repeated Information* and *Selected report is of a MORE preferable length*. Hence, the excessive verbosity produced by DPO and IPO was heavily penalized by the radiologists. KTO, SimPO, and ORPO, all of which maintain average lengths close to the reference, fared considerably better. With ORPO showing an improvement over the SFT baseline with a win rate of 0.62. Now, if we consider why ORPO and KTO were chosen over the SFT baseline, then the most common reason, by far, was *Selected report contains LESS false information*. In other words, factuality was improved.

	Aligned preferred				**Baseline preferred**			
Method	**(a)**	**(b)**	**(c)**	**(d)**	**(a)**	**(b)**	**(c)**	**(d)**
CheXagent								
+DPO	7	1	2	4	18	34	24	4
+KTO	22	3	8	3	10	9	8	5
+IPO	8	1	3	3	15	29	20	1
+SimPO	18	2	12	3	20	4	10	7
+ORPO	27	3	12	7	12	6	8	4

Table 8: Counts of why a preferred report was chosen in the reader study. The categories are: (a) Selected report contains LESS false information, (b) Selected report contains LESS repeated information, (c) Selected report is of a MORE preferable length, (d) Other. Note that indicting why report was preferred was optional.

To further evaluate the generated reports from a clinical perspective, we consider clinical efficacy by extracting labels (14 categories) using the CheXbert labeler (Smit et al., 2020) from the generated and reference reports. We then compute the F1 score. Results are available in Table 9. These results seem to correlate well with the reader study, as the macro averages for DPO and IPO are actually worse than for the SFT baseline. Moreover, KTO and ORPO are the top performer in terms of micro and macro averages. For F1, we observe 8.4% and 5.9% increase in micro and macro averages, for KTO and a 8.1% and 6.7% increase for ORPO. While we observe an overall improvement in macro and micro averages for KTO and ORPO, we also observe deteriorating performance in certain categories, for instance *Facture*, indicating that performance is not improved uniformly across categories.

In sum, GREEN appears somewhat susceptible to length exploitation—a weakness that DPO and IPO heavily exploited, leading to no clinical improvements just increased verbosity. ORPO and

	MIMIC-CXR															
F1 (↑)	**ECm.**	**Cmgl.**	**LOpac.**	**LLes.**	**Edema**	**Cnsl.**	**Pna.**	**Atel.**	**Pmtx.**	**PEff.**	**POth**	**Frac.**	**SuDev.**	**noF.**	**Micro**	**Macro**
CheXagent	0.347	0.620	0.461	0.171	0.493	0.158	0.227	0.453	0.444	0.655	0.092	**0.240**	0.787	0.304	0.509	0.389
+DPO	0.383	**0.688**	0.257	0.144	0.352	**0.254**	0.087	0.349	0.268	0.625	0.149	0.219	0.815	0.333	0.500	0.352
+KTO	0.400	0.683	0.425	**0.240**	**0.554**	0.167	0.164	0.441	0.500	**0.724**	0.130	0.158	**0.840**	0.340	**0.552**	0.412
+IPO	**0.423**	0.675	0.307	0.178	0.433	0.189	0.111	0.335	0.261	0.643	**0.185**	0.146	0.819	0.326	0.513	0.359
+SimPO	0.381	0.668	0.398	0.150	0.320	0.167	0.178	0.332	0.456	0.669	0.152	0.078	0.812	0.351	0.506	0.365
+ORPO	0.348	0.684	**0.479**	0.201	0.492	0.224	**0.247**	**0.475**	**0.511**	0.698	0.072	0.177	0.835	**0.365**	0.550	**0.415**

Table 9: F1 scores on the MIMIC-CXR test set using 14 categories from the CheXbert labeler Smit et al. (2020): Enlarged Cardiomediastinum (ECm.), Cardiomegaly (Cmgl.), Lung Lesion (LLes.), Lung Opacity (LOpac.), Edema, Consolidation (Cnsl.), Pneumonia (Pna.), Atelectasis (Atel.), Pneumothorax (Pmtx.), Pleural Effusion (PEff.), Pleural Other (POth.), Fracture (Frac.), Support Devices (SuDev.), no Findings (NoF.). Best results in bold.

KTO, on the other hand, seem less prone to this bias, leading to clinical improvements by reducing the prevalence of false information and thus enhancing factual accuracy.

5 Limitations and discussion

Due to compute constraints, our work focuses on a single model, CheXagent (Chen et al., 2024). Other VLMs, from different families and sizes, may behave differently. For example, it is possible that DPO yielded nonsensical results, increased verbosity with no clinical utility, due to a insufficiently strong baseline–despite being a state-of-the-art model. We do counteract this point, however, by including a range of DAAs with varying sensitivity to the strength of the SFT baseline. Nonetheless, to further validate the results in this study, evaluating another VLM is warranted.

Moreover, based on previous work (Ostmeier et al., 2024), we treat GREEN as the silver standard, effectively a low-cost approximation of expert human judgment. However, we have observed that verbosity bias is a significant issue. Thus, further work, including considering length-controlled metrics, is necessary. One simple updated to the GREEN metric worth exploring is as follows: GREEN-LC = GREEN / max(length of generated report/length of reference report,1). Where LC refers to it being length-controlled. Intuitively, this down weights GREEN when the length of the generated report is larger than that of the reference. If this is not the case, then the correction does nothing. Although very simplistic, such a correction will allow us to deal with the apparent trade off between average GREEN and verbosity. In addition, further investigation of length-controlled alignment algorithms, such as length-controlled DPO (Park et al., 2024), would be helpful to decouple length and quality of the generated reports. The issue of potential biases extends beyond verbosity, as there might be other societal biases, with regards to for instance race and sex, embedded in the data or the Judge. These biases should be carefully studied and mitigated.

In addition, our hyperparameter search in non-exhaustive and it is possible that the relative ranking of the methods considered would change with a more extensive search. Finally, we restrict ourselves to only offline DAAs. This leaves out a range of very competitive alignment algorithms, including on-policy RL algorithms, such as PPO (Schulman et al., 2017) and REINFORCE Williams (1992), as well as the online, or iterative, counterparts to the DAAs considered.

6 Conclusion

Our study highlights the significant potential of including preference fine-tuning in the post-training pipeline of medical VLMs. Using our approach to preference data generation, we have shown that DAAs can substantially improve AI-generated reports in clinically meaningful ways *without additional radiologist feedback*. Results indicate maintained performance on diverse tasks, suggesting no alignment tax. The preference of aligned policies by board-certified radiologists and improvements in clinical efficacy metrics, highlight the clinical value of our method. Our systematic analysis yields actionable insights for preference alignment of medical VLMs, paving the way for more accurate AI assistance in radiology, potentially addressing workforce shortages and improving patient care.

REFERENCES

American College of Radiology. Siim-acr pneumothorax segmentation 2019. 2019. URL `https://www.kaggle.com/competitions/siim-acr-pneumothorax-segmentation/data`.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. A general language assistant as a laboratory for alignment, 2021. URL `https://arxiv.org/abs/2112.00861`.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences, 2023.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022a. URL `https://arxiv.org/abs/2204.05862`.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022b. URL `https://arxiv.org/abs/2212.08073`.

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Fabian Falck, Ozan Oktay, Anja Thieme, Matthew P. Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. Maira-2: Grounded radiology report generation, 2024.

Mythreyi Bhargavan, Jonathan H Sunshine, and Barbara Schepps. Too few radiologists? *American Journal of Roentgenology*, 178(5):1075–1082, 2002.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. Biomedlm: A 2.7b parameter language model trained on biomedical text, 2024. URL `https://arxiv.org/abs/2403.18421`.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444, 14643510. URL `http://www.jstor.org/stable/2334029`.

RJM Bruls and RM Kwee. Workload for radiologists during on-call hours: dramatic increase in the past 15 years. *Insights into imaging*, 11:1–7, 2020.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. Chexpert plus: Hundreds of thousands of aligned radiology texts, images and patients. *arXiv preprint arXiv:2405.19538*, 2024.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, et al. Training small multimodal models to bridge biomedical competency gap: A case study in radiology imaging. *arXiv preprint arXiv:2403.08002*, 2024.

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. *arXiv preprint arXiv:2401.12208*, 2024.

Yashin Dicente Cid, Matthew Macpherson, Louise Gervais-Andre, Yuanyi Zhu, Giuseppe Franco, Ruggiero Santeramo, Chee Lim, Ian Selby, Keerthini Muthuswamy, Ashik Amlani, et al. Development and validation of open-source deep neural networks for comprehensive chest x-ray reading: a retrospective, multicentre study. *The Lancet Digital Health*, 6(1):e44–e57, 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.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 30039–30069. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/5fc47800ee5b30b8777fdd30abcaaf3b-Paper-Conference.pdf.

Yann Dubois, Percy Liang, and Tatsunori Hashimoto. Length-controlled alpacaeval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=CybBmzWBX0.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization, 2024.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Tarek N Hanna, Haris Shekhani, Christine Lamoureux, Hanna Mar, Refky Nicola, Clint Sliker, and Jamlik-Omari Johnson. Emergency radiology practice patterns: shifts, schedules, and job satisfaction. *Journal of the American College of Radiology*, 14(3):345–352, 2017.

Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model, 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.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024. URL https://arxiv.org/abs/2401.04088.

Alistair EW 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 de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 26874–26901. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/lee24t.html`.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

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

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1650–1654. IEEE, 2021.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

Matthew Lyon, LaShon Sturgis, Darren Lendermon, Ann Marie Kuchinski, Taylor Mueller, Patrick Loeffler, Hongyan Xu, and Robert Gibson. Rural ed transfers due to lack of radiology services. *The American journal of emergency medicine*, 33(11):1630–1634, 2015.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward, 2024.

World Health Organization et al. Communicating radiation risks in paediatric imaging: information to support health care discussions about benefit and risk. 2016.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, et al. Green: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024.

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.

WHO PAHO. World radiography day: Two-thirds of the world's population has no access to diagnostic imaging. *Pan American Health Organization*, 2012.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 4998–5017, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.297.

Chantal Pellegrini, Matthias Keicher, Ege Özsoy, and Nassir Navab. Rad-restruct: A novel vqa benchmark and method for structured radiology reporting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 409–419. Springer, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2023.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms, 2024. URL https://arxiv.org/abs/2406.02900.

Abi Rimmer. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ: British Medical Journal (Online)*, 359, 2017.

Alexander T Ruutiainen, Daniel J Durand, Mary H Scanlon, and Jason N Itri. Increased error rates in preliminary reports issued by radiology residents working more than 10 consecutive hours overnight. *Academic radiology*, 20(3):305–311, 2013.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1500–1519, 2020.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented RLHF. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.775. URL https://aclanthology.org/2024.findings-acl.775.

Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Stefano Ermon, Chelsea Finn, and Aviral Kumar. Preference fine-tuning of LLMs should leverage suboptimal, on-policy data. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=bWNPx6t0sF.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kepa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali

Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu

Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2024. URL https://arxiv.org/abs/2312.11805.

Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL https://arxiv.org/abs/1909.08593.

A Appendix

A.1 Reward and length distributions

The distributions of reward and length for the chosen and rejected subsets are available in Fig. 4.

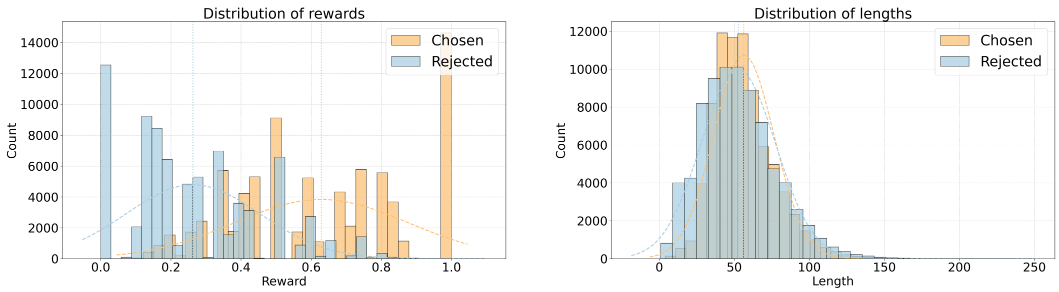

Figure 4: Distribution of GREEN reward and report length in the chosen and rejected subsets.

A.2 Implementation details

All models are trained using either a machine with 4xA100 GPUs or 4xA6000 GPUs using a global batch size of 32 and learning rate 10^{-6}. Each model is trained for one epoch. The image encoder is frozen while we train the LLM. Hyperparameters are important in DAAs. However, tuning large models is very expensive. Due to compute constraints, we only tune hyperparameters that are specific of the DAAs considered while keeping everything else fixed. An overview is given in Table 10. We do a non-exhaustive search, based on previous work and initial experiments, and we only consider GREEN as metrics for hyperparameter tuning. For each $\lambda \in [0.5, 1.0, 4.0, 5.0]$, ORPO resulted in a model which produced a special token at odd places, leading to a crash of our evaluation pipeline. We address this by catching the error and set the special token to the padding token.

Algorithm	**Objective**	**Hyperparameters**
DPO (Rafailov et al., 2023)	$-\log\sigma\left(\beta\log\frac{\pi_\theta(y_c\|x)}{\pi_{\text{ref}}(y_c\|x)} - \beta\log\frac{\pi_\theta(y_r\|x)}{\pi_{\text{ref}}(y_r\|x)}\right)$	$\beta \in [0.01, 0.05, 0.1]$
KTO (Ethayarajh et al., 2024)	$-\lambda_c\sigma\left(\beta\log\frac{\pi_\theta(y_c\|x)}{\pi_{\text{ref}}(y_c\|x)} - z_{\text{ref}}\right) + \lambda_r\sigma\left(z_{\text{ref}} - \beta\log\frac{\pi_\theta(y_r\|x)}{\pi_{\text{ref}}(y_r\|x)}\right)$ where $z_{\text{ref}} = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\beta\mathbb{D}_{\text{KL}}\left(\pi_\theta(y\|x)\|\|\pi_{\text{ref}}(y\|x)\right)]$	$\beta \in [0.01, 0.05, 0.1], \lambda_c = \lambda_r$
IPO (Azar et al., 2023)	$\left(\log\frac{\pi_\theta(y_c\|x)}{\pi_{\text{ref}}(y_c\|x)} - \log\frac{\pi_\theta(y_r\|x)}{\pi_{\text{ref}}(y_r\|x)} - \frac{1}{2\tau}\right)^2$	$\tau \in [0.1, 0.5, 1.0]$
ORPO (Hong et al., 2024)	$-\log p_\theta(y_c\|x) - \lambda\log\sigma\left(\log\frac{p_\theta(y_c\|x)}{1-p_\theta(y_c\|x)} - \log\frac{p_\theta(y_r\|x)}{1-p_\theta(y_r\|x)}\right),$ where $p_\theta(y\|x) = \exp\left(\frac{1}{\|y\|}\log\pi_\theta(y\|x)\right)$	$\lambda \in [0.5, 1.0, 4.0, 5.0]$
SimPO (Meng et al., 2024)	$\log\sigma\left(\frac{\beta}{\|y_c\|}\log\pi_\theta(y_c \mid x) - \frac{\beta}{\|y_r\|}\log\pi_\theta(y_r \mid x) - \gamma\right)$	$\beta \in [2.5, 4.0, 5.0, 10.0], \gamma = 0.1$

Table 10: Hyperparameter search for all direct alignment algorithms (DAAs) considered in this paper.

A.3 Alignment algorithm: Additional metrics

For a more holistic approach, we consider some additional metric to what was included in Table. 3. In particular, we also include the lexical, general domain, metrics BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). In addition, to obtain a "biomedical" metric, we extract the contextual embeddings from BioMedML (Bolton et al., 2024), a 2.7B model trained on biomedical text, and compute the cosine similarity to produce a scalar score–exactly as what is done for BERTScore. We call this metric the BioMedMLScore. Results for the MIMIC-CXR and CheXpert Plus datasets are available in Table 11.

	MIMIC-CXR						
	Radiology		BioMedical	General			
Method	GREEN (↑)	F1RadGraph (↑)	BioMedMLScore(↑)	BERTScore (↑)	BLEU-4 (↑)	ROUGE-L (↑)	Avg. (↑)
CheXagent	0.249	0.215	0.712	0.856	0.041	0.274	0.391
+DPO	**0.354 (0.105)**	0.247 (0.032)	0.724 (0.012)	0.830 (-0.026)	0.042 (0.001)	0.237 (-0.037)	0.406 (0.015)
+KTO	0.340 (0.091)	**0.260 (0.045)**	0.735 (0.023)	**0.862 (0.006)**	**0.054 (0.013)**	**0.297 (0.023)**	**0.425 (0.034)**
+IPO	0.349 (0.100)	0.252 (0.037)	**0.736 (0.024)**	0.846 (-0.010)	0.051 (0.010)	0.267 (-0.007)	0.417 (0.026)
+SimPO	0.315 (0.066)	0.225 (0.010)	0.711 (-0.001)	0.854 (-0.002)	0.041 (0.000)	0.270 (-0.004)	0.403 (0.012)
+ORPO	0.322 (0.073)	0.244 (0.029)	0.727 (0.015)	**0.862 (0.006)**	0.053 (0.012)	0.290 (0.016)	0.416 (0.025)
	CheXpert						
	Radiology		BioMedical	General			
Method	GREEN (↑)	F1RadGraph (↑)	BioMedMLScore(↑)	BERTScore (↑)	BLEU-4 (↑)	ROUGE-L (↑)	Avg. (↑)
CheXagent	0.248	0.222	0.702	0.851	0.038	0.274	0.389
+DPO	**0.391 (0.142)**	0.246 (0.024)	0.726 (0.024)	0.821 (-0.030)	0.030 (-0.008)	0.217 (-0.057)	0.405 (0.016)
+KTO	0.340 (0.092)	**0.265 (0.043)**	**0.734 (0.032)**	**0.859 (0.008)**	**0.050 (0.013)**	**0.301 (0.027)**	**0.425 (0.036)**
+IPO	0.358 (0.110)	0.248 (0.026)	**0.734 (0.032)**	0.844 (-0.007)	0.041 (0.003)	0.274 (0.000)	0.416 (0.027)
+SimPO	0.292 (0.043)	0.205 (-0.017)	0.693 (-0.009)	0.844 (-0.006)	0.035 (-0.002)	0.269 (-0.005)	0.390 (0.001)
+ORPO	0.326 (0.078)	0.248 (0.026)	0.728 (0.026)	0.856 (0.005)	0.046 (0.008)	0.287 (0.013)	0.415 (0.026)

Table 11: Results on the MIMIC-CXR and CheXpert Plus test sets (with Δ compared to SFT baseline in brackets). Green shades for improvements and red shades for degradations. Shades are separated into bins of 10%, running from $> 0\%$ and $\leq 10\%$ up to $> 50\%$. Best results in bold.

A.4 Alignment algorithm: Additional results for verbosity bias

To further build intuition, we illustrate the resulting distributions of length in Fig. 5. Consistent with the results in Table 4, we can see that the SFT baseline, KTO, SimPO, and ORPO maintain a distribution similar to that for the reference reports. DPO and IPO, on the other hand, results in a significant shift towards more verbose reports. In particular, the distributions of length is bimodal. Closer inspection indicate that the extra verbosity is due to repetition of words or entire sentences. This can be exact repetition or semantically equivalent repetitions. In the mode to the right, almost all examples have exact repeats of sentences. Whereas in the mode to the left this is far less common. Simple heuristics to filter the outputs could be explored in future work.

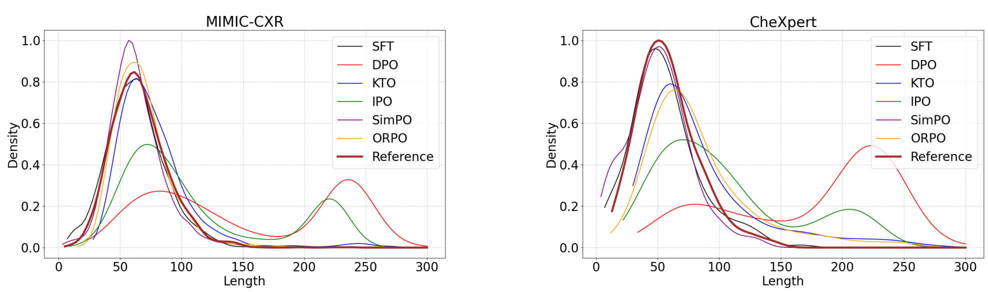

Figure 5: Kernel density of length in the generated and reference reports.

A.5 Clinical perspective: Qualitative analysis

Color coded version of the generated reports for a particular case are available in Fig. 6. This particular example was chosen to build intuition on the length exploitation issue. DPO and SimPO both achieve GREEN=1, but DPO is significantly more verbose than SimPO. ORPO is also very verbose for this partiuclar example.

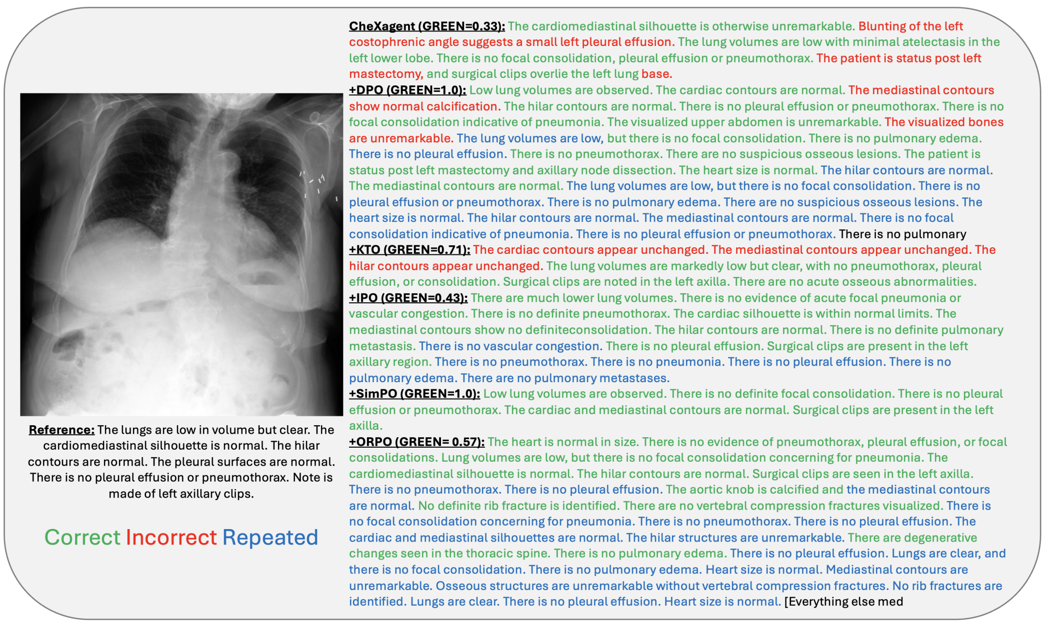

Figure 6: Qualitative results on one example from the MIMIC-CXR test set. The text in the generated reports is color-coded as correct, incorrect, and repeated (i.e. exact repetitions or semantically equivalent).