

CheXalign: Preference fine-tuning in chest X-ray interpretation models without human feedback

Dennis Hein¹, Zhihong Chen¹, Sophie Ostmeier¹, Justin Xu^{1,2}, Maya Varma¹,
Eduardo Pontes Reis¹, Arne Edward Michalson¹, Christian Bluethgen^{1,3},
Hyun Joo Shin⁴, Curtis Langlotz¹, Akshay S Chaudhari¹

¹Stanford University, USA ²University of Oxford, UK

³University Hospital Zurich, University of Zurich, Switzerland

⁴Yongin Severance Hospital, Yonsei University, South Korea

heind@stanford.edu

Abstract

Radiologists play a crucial role in 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. Meanwhile, additional preference fine-tuning in the post-training pipeline has become standard practice in the general domain. The challenge in radiology lies in the prohibitive cost of obtaining radiologist feedback at scale. To address this challenge, we propose an automated pipeline for preference feedback, focusing on chest X-ray radiology report generation (RRG). Specifically, our method leverages publicly available datasets containing pairs of images and radiologist-written reference reports with reference-based metrics, or Judges, eliminating the need for *additional radiologist feedback*. We investigate reward overoptimization via length exploitation in this setting and introduce a length-controlled version of the GREEN score. Our best-performing setup achieves state-of-the-art CheXbert scores on the MIMIC-CXR dataset for the RRG task while on average maintaining robust performance across six additional image perception and reasoning tasks.

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, 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 (Ruutinen et al., 2013; Hanna et al., 2017; Bruls and 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., 2024a; Radford et al., 2021). Typically pre-trained using image-text contrastive learning and subsequently fine-tuned, recent VLMs have started to demonstrate promising performance in CXR interpretation and radiology report generation (RRG) (Chen et al., 2024a; 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), such as Direct Preference Optimization (DPO) (Rafailov et al., 2023), effectively alleviate this problem by employing a negative gradient to lower probabilities of “bad” completions (Tajwar et al., 2024). In fact, most recent large language models (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 pref-

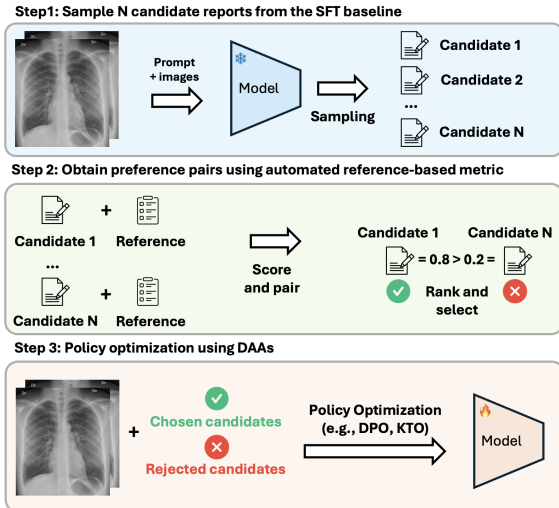


Figure 1: Overview of the CheXalign pipeline. For a given dataset containing CXRs and radiologist written references (e.g., MIMIC-CXR), we obtain preference pairs using some automated reference-based metric (e.g., GREEN or the BERTScore), and then optimize the policy using light-weight DAAs (e.g., DPO or KTO).

erence fine-tuning in their post-training pipeline. Yet, this approach has not yet been investigated within the medical vision-language domain.

Yet, this approach has been only sparsely investigated within the medical vision-language domain (e.g., Xiao et al. (2024); Zhu et al. (2025)), concurrent to our work.

The primary challenge hindering the application of preference fine-tuning in the post-training of VLMs in radiology is the *prohibitive cost* of obtaining radiologist preferences at scale. To overcome this obstacle, we introduce CheXalign, an automated pipeline for generating preference data for the crucial RRG task. 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 CheXpert Plus (Chambon et al., 2024). This allows us to use reference-based, uni-modal, metrics, such as 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. An overview of our preference fine-tuning pipeline is available in Fig. 1. Our approach enables us to obtain high-quality preference datasets in a fully automated and scalable manner. Using our proposed method, we systematically study how DAAs can be used to enhance the clinical efficacy of medical VLMs

without any additional radiologist feedback. Our contributions are as follows:

1. We introduce an automated pipeline for preference pair generation in RRG models, circumventing the *prohibitively expensive* task of obtaining preference feedback from radiologists at scale.
2. We systematically evaluate and benchmark the proposed pipeline using different reference-based metrics, DAAs, and RRG models. Our findings indicate that the RRG performance can be improved even when using inexpensive, general domain, natural language generation (NLG) metrics for preference pair generation.
3. Using our proposed pipeline, we obtain new state-of-the-art CheXbert scores on the MIMIC-CXR data for the RRG task.
4. We study reward overoptimization via length exploitation, and introduce the length-controlled GREEN score.
5. We benchmark our models post alignment on set of diverse additional image perception and reasoning tasks to assess whether there is an alignment tax in this setting.

Code for this project is available in the following repository: <https://github.com/StanfordMIMI/CheXalign>.

2 Related Works

VLMs (Radford et al., 2021; Li et al., 2021, 2022, 2023a; Liu et al., 2024a) 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., RRG in the field of radiology). There is also a line of works to extend VLMs to the medical domain (Thawkar et al., 2023; Hyland et al., 2023; Chaves et al., 2024; Tu et al., 2024; Banur et al., 2024; Chen et al., 2024a,b; Lee et al., 2024b; Jin et al., 2024) which mainly focus on CXR interpretation and RRG 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. As noted above, preference fine-tuning, although now a standard part of the post-training pipeline for general domain LLMs, remains largely unexplored in the context of medical VLMs.

RLHF (Ziegler et al., 2020; Stiennon et al., 2020; Ouyang et al., 2022) is a very effective technique for aligning LLMs and VLMs with human preferences. It was instrumental in the development of frontier models such as ChatGPT and Bard (Lee et al., 2024a). RLHF involves two stages: reward modeling, where a reward model is learned on the preference data, and alignment, where reinforcement learning (RL) algorithms are used to optimize the proxy reward. While extremely effective, RLHF is very computationally heavy and can be finicky for non-experts. 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-form solution to the RLHF objective, effectively bypassing both the reward modeling and RL stages. This yields alignment algorithms that remain performant yet computationally more lightweight and significantly easier to implement. DPO (Rafailov et al., 2023) was the first in this category and remains one of the most popular versions. After the advent of DPO, a large variety of DAAs have been suggested (Azar et al., 2023; Park et al., 2024; Ethayarajh et al., 2024; Hong et al., 2024). A brief introduction to RLHF and DAAs is available in §A. Another dimension for improved efficiency is reinforcement learning from AI feedback (RLAIF), first introduced in Bai et al. (2022b). Replacing humans with LLMs leads to significant reductions in cost, making it much more scalable, while maintaining high quality (Lee et al., 2024a). In this work, we will leverage these advancements for scalable, fully automated, preference data generation and computationally lightweight alignment.

Since the reward model in the RLHF objective is learned, it is an imperfect proxy of the ground truth reward. 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. Despite not fitting an explicit reward model, similar behavior has been observed empirically for DAAs (Rafailov et al., 2024). In particular, 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 minor verbosity bias embedded in the preference data. In this work, we explore this phenomenon in the context of preference fine-tuning of RRG models.

3 Methodology

3.1 RRG Preference Fine-tuning without Human Feedback

Expert human feedback from radiologists is the gold standard for preference data generation and evaluation for the RRG task. However, scaling is impractical, if not unfeasible, due to the limited availability of radiologists for large-scale annotation tasks. In the general domain, it is common to leverage LLMs for cost effective preference data generation (Bai et al., 2022b; Dubois et al., 2023; Lee et al., 2024a). Zheng et al. (2023) categorized “LLM-as-a-Judge” evaluation methods into pairwise, single answer, and reference-guided grading. Pairwise grading being the most common in the general domain both for preference data generation (Dubois et al., 2023; Lee et al., 2024a) 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 RRG to ensure clinical reliability. To overcome these challenges, we propose using reference-based 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 metrics, or “Judge”, setting our approach apart from prior studies of preference alignment of VLMs, such as Sun et al. (2024). A comparison of reference-free and reference-based metrics, or “Judges”, in this setting is available in Fig. 2. For a given Judge, we obtain preference pairs by repeat sampling from the SFT baseline. Canonical alignment algorithms, such as DPO, can

¹Used more loosely than in Rafailov et al. (2024).

²As per Goodhart’s law: “When a measure becomes a target, it ceases to be a good measure.” (Gao et al., 2023).

then be used to preference fine-tune the model.

3.2 Evaluation

Since we have radiologist-written reference available, it is possible to employ standard, general domain, NLG metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and the BERTScore (Zhang et al., 2019). However, these NLG metrics may not be able to differentiate between subtle nuances that are clinically relevant. Thus, we include two validated, radiology-specific, and clinically relevant, metrics: GREEN (Ostmeier et al., 2024) and CheXbert scores (Smit et al., 2020). GREEN is a state-of-the-art metric for radiology report evaluation, based on a single-answer reference-guided LLM-as-a-Judge mechanism. CheXbert scores are a clinical efficacy metrics based on extracting 14 labels³ using the CheXbert labeler (Smit et al., 2020) from the candidate and reference reports.

Initial experimentation illustrated that reward hacking via length exploitation might be a concern in our setting when using the GREEN score as Judge. To counteract this, we propose a simple heuristic approach to explicitly control for length:

$$\text{LC-GREEN} := \text{GREEN} / \max(\text{rel_verbosity}, 1),$$

where `rel_verbosity` denotes the relative verbosity (length in words) of the candidate report compared to the reference report. We call this metric length-controlled GREEN (LC-GREEN).

4 Experimental Details

4.1 SFT Baselines

We adopt CheXagent (Chen et al., 2024a) (8B) as a representative example of a state-of-the-art, open source, English language, VLM for the RRG task. It has been trained in a canonical way by first adapting an LLM, Mistral-7B (Jiang et al., 2023), to medical text by continued pre-training. Second, a vision encoder, EVA-CLIP-g (Sun et al., 2023), 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 for a range of tasks, including RRG. Concurrent to this work, CheXagent

³Enlarged Cardiomeastinum, Cardiomegaly, Lung Opacity, Lung Lesion, Edema, Consolidation, Pneumonia, Atelectasis, Pleural Effusion, Pneumothorax, Pleural Other, Fracture, Support Devices, No Finding.

has been greatly improved upon with CheXagent-2 (Chen et al., 2024b) (3B). CheXagent-2 adopted a fine-tuned SigLIP (Zhai et al., 2023) model⁴ as the vision encoder and a fine-tuned Phi-2 (Li et al., 2023b) model⁵ (2.7B) as the language decoder. For the image-text connector, instead of using the attention layer as in CheXagent, it uses LLaVA-style MLP connector (Liu et al., 2024b). CheXagent-2 additionally use indications, which provide clinical context to the patient being imaged, to aid in the RRG task.

4.2 Datasets

We use the MIMIC-CXR (Johnson et al., 2019) dataset for training, validation and testing. The image-report pairs consist of one or two CXRs and the corresponding free-text findings section. CheXagent-2 additionally uses the indications. For CheXagent, we randomly sample 80k examples as our training data. For CheXagent-2, we opted to use the full MIMIC-CXR training set⁶ (148k examples). To test robustness for the RRG task, we additionally include test data from the CheXpert Plus (Chambon et al., 2024) dataset. Moreover, to evaluate whether there is an alignment tax, we additionally evaluate our aligned models on six additional CXR tasks: view classification, binary 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-Reconstruct (Pellegrini et al., 2023) datasets. All datasets are in English.

4.3 Preference Data

We evaluate two reference-based Judges: GREEN (Ostmeier et al., 2024) and the BERTScore (Zhang et al., 2019). 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 the score 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 scores, omitting the observation if all

⁴https://huggingface.co/StanfordAIMI/XraySigLIP__vit-1-16-siglip-384__webli.

⁵<https://huggingface.co/StanfordAIMI/RadPhi-2>.

⁶As can be seen below, CheXagent-2 is a strong SFT baseline and we hypothesized that a larger preference dataset would be required to obtain significant performance gains.

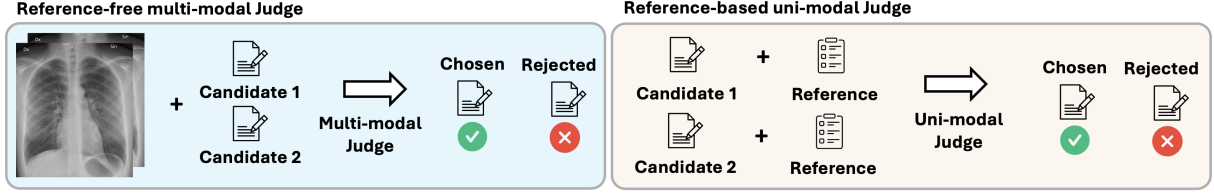


Figure 2: Illustration of preference pairs generation using a reference-free multi-modal Judge (left) and a reference-based uni-modal Judge (right). Human (radiologist) feedback is multi-modal in nature but prohibitively expensive at scale. On the other hand, using LLM, or metrics-based, Judges is highly scalable. However, a high-quality multi-modal Judge is difficult to obtain. For the RRG task, we propose leveraging large publicly available datasets, containing CXRs and radiologist written reference reports, as this enables scalable, factually grounded, preference data generation using reference-based metrics.

Algorithm	Objective	Preference pairs	Reference	Length controlled	Relative wall-clock time
DPO	$-\log \sigma \left(\beta \log \frac{\pi_{\theta}(y_c x)}{\pi_{\text{ref}}(y_c x)} - \beta \log \frac{\pi_{\theta}(y_r x)}{\pi_{\text{ref}}(y_r x)} \right)$	✓	✓	×	1.0
LC-DPO	$-\log \sigma \left(\beta \log \frac{\pi_{\theta}(y_c x)}{\pi_{\text{ref}}(y_c x)} - \beta \log \frac{\pi_{\theta}(y_r x)}{\pi_{\text{ref}}(y_r x)} + \alpha (y_c - y_r) \right), \alpha > 0$	✓	✓	✓	1.0
IPO	$\left(\log \frac{\pi_{\theta}(y_c x)}{\pi_{\text{ref}}(y_c x)} - \log \frac{\pi_{\theta}(y_r x)}{\pi_{\text{ref}}(y_r x)} - \frac{1}{2r} \right)^2$	✓	✓	×	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 $z_{\text{ref}} = \mathbb{E}_{(x,y) \sim \mathcal{D}} [\beta \mathbb{D}_{\text{KL}}(\pi_{\theta}(y x) \pi_{\text{ref}}(y x))]$	×	✓	×	2.2
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 1: Overview of the DAAs considered in this paper. Preference pairs indicates whether the method requires paired data of chosen/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 reward hacking via length exploitation. Relative wall-clock time is measured as the total time to train one epoch, relative to DPO.

$N = 4$ scores are equivalent. For CheXagent, this rejection rule results in the rejection of 1,246 (1.6%) examples for GREEN and 31 (0.04%) examples for the BERTScore. This is similar for CheXagent-2, with 3744 (2.53%) examples and 147 (0.10%) examples rejected for GREEN and the BERTScore, respectively. Summary statistics for the chosen and rejected subsets are available in Table 8. We also report summary statistics of the length (in words) of the generated reports. Notably, the spread in average length of the chosen and rejected subsets is slightly more pronounced for GREEN than for the BERTScore, 6.9% compared to 5.8% and much more significant overall for CheXagent-2 with a difference of 19.3% and 17.1%, respectively. Examples from the chosen and rejected subsets are available in Fig. 5 and 6.

4.4 Alignment Algorithms

We opt for a representative subset of available, offline, DAAs. DPO is the original DAA and serves as our baseline. In addition to DPO, we consider: 1) Length-controlled DPO (LC-DPO) (Park et al., 2024), as an example of a DAA with explicit length

regularization. LC-DPO is an elegant extension of DPO, with an additional hyperparameter α which controls the strength of an additional length regularization term. Setting $\alpha = 0$ reverts the objective to that in DPO. 2) 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 used in DPO. The authors argue that this helps mitigate overfitting issues 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). 3) 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. 4) Odds-Ratio Preference Optimization (ORPO) (Hong et al., 2024), almost outside of the definition of DAAs, is not based on the RLHF objective but instead appends an additional penalty

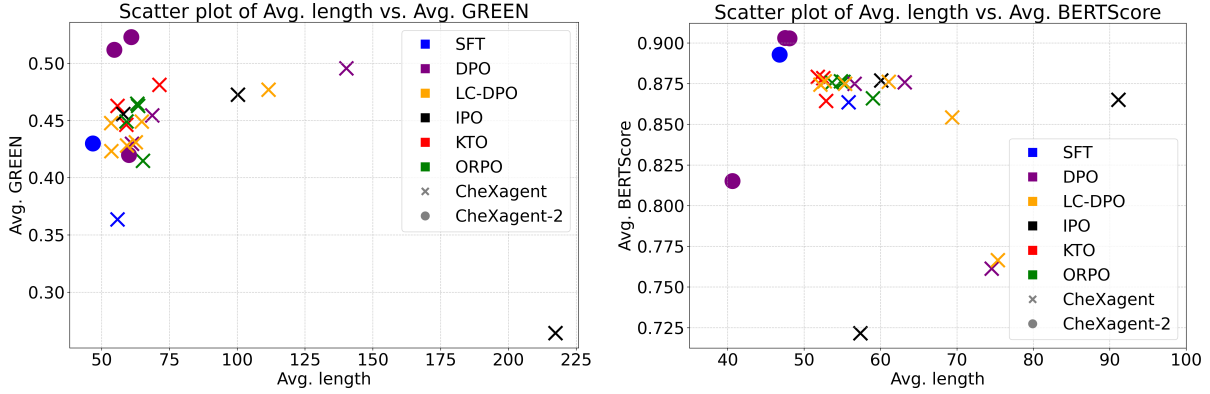


Figure 3: Average length against average GREEN and BERTScore for all aligned policies on the MIMIC-CXR validation set using GREEN (left) and the BERTScore (right) as Judges.

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.

An overview of all DAAs considered in this paper is available in Table 1. More implementation details are available in §B.2.⁷

5 Results and Analyses

5.1 Length Exploitation

We investigate the issue of length exploitation by plotting the average GREEN and BERTScore against average lengths of the resulting radiology reports using GREEN and the BERTScore as Judges for preference data generation. Results for single runs on the MIMIC-CXR validation set, including all configurations considered, are available in Fig. 3. Save for a IPO run in the lower right corner, there seems to be a positive correlation between average GREEN and average length for CheXagent. In fact, for all DAAs, except for ORPO, there is an indication for a trade-off between added verbosity and GREEN. In addition, such a trade-off is not observed for the BERTScore, shown to the right in Fig. 3. Except for a very tight clustering around 0.87-0.88, there are no clear trends.

To further emphasize this issue, we show the best configurations according to GREEN, and LC-GREEN, in Table 2. For CheXagent, using GREEN to guide the hyperparameter search leads to substantially more verbose reports for DPO, LC-DPO ($\alpha = 0.001$), IPO, and KTO. Qualitative evaluation of the resulting reports indicate that the added

⁷Due to compute constraints, we only include results for the baseline alignment algorithm, DPO, for CheXagent-2.

Method	GREEN		LC-GREEN	
	Avg. length	Rel. verbosity	Avg. length	Rel. verbosity
CheXagent	55.8		55.8	
+DPO	140.2	2.51	68.7	1.23
+LC-DPO	111.5	2.00	53.5	0.96
+IPO	100.2	1.80	58.0	1.04
+KTO	71.3	1.28	55.9	1.00
+ORPO	63.1	1.13	63.1	1.13
CheXagent-2	46.8		46.8	
+DPO	60.9	1.30	54.7	1.17
Reference	58.4	1.05	58.4	1.05

Table 2: Average length of reports on the MIMIC-CXR validation set of best performing configurations according to GREEN and GREEN-LC, respectively. Relative verbosity is relative to the SFT baseline.

verbosity was mostly due to exact, semantically or syntactically, repetitions. This is very likely a manifestation of reward hacking via length exploitation. To mitigate this issue, we use LC-GREEN instead of GREEN to guide the hyperparameter search.⁸ As shown in Table 2, this leads to substantially less added verbosity. For CheXagent-2, it is less clear-cut whether length exploitation is present, as the highest and lowest scores are obtained with a similar verbosity. Note that this is in spite of a much more significant spread in average length of the reports in the chosen and rejected subsets. One possible explanation is that the SFT baseline produces substantially shorter reports than the references, and the increase in verbosity actually pushes the average closer to that of the references, effectively closing the gap rather than extending it by overshooting the average length of the references.

⁸We opt for this approach, in lieu of obtaining new preference data and re-running the alignment due to computational constraints.

Model	Judge	MIMIC-CXR			CheXpert Plus		
		GREEN (↑)	LC-GREEN (↑)	BERTScore(↑)	GREEN (↑)	LC-GREEN (↑)	BERTScore(↑)
CheXagent		0.249	0.218	0.856	0.248	0.202	0.851
+DPO	GREEN	0.325 (30.6)	0.263 (20.5)	0.862 (0.65)	0.322 (29.9)	0.222 (9.69)	0.859 (0.91)
+LC-DPO		0.320 (28.7)	0.288 (32.0)	0.864 (0.90)	0.330 (32.8)	0.268 (32.3)	0.861 (1.26)
+IPO		0.326 (30.9)	0.282 (29.3)	0.863 (0.84)	0.334 (34.6)	0.267 (31.9)	0.861 (1.15)
+KTO		0.328 (31.9)	0.293 (34.1)	0.867 (1.27)	0.341 (37.2)	0.266 (31.4)	0.863 (1.42)
+ORPO		0.322 (29.4)	0.275 (26.2)	0.862 (0.69)	0.326 (31.4)	0.232 (14.6)	0.856 (0.64)
+DPO	BERTScore	0.285 (14.4)	0.242 (11.0)	0.869 (1.56)	0.278 (12.0)	0.208 (2.84)	0.867 (1.89)
+LC-DPO		0.283 (13.5)	0.259 (18.7)	0.871 (1.75)	0.299 (20.3)	0.248 (22.5)	0.871 (2.34)
+IPO		0.283 (13.6)	0.247 (13.1)	0.870 (1.62)	0.282 (13.6)	0.218 (7.88)	0.866 (1.75)
+KTO		0.304 (21.9)	0.279 (28.2)	0.872 (1.88)	0.308 (24.0)	0.249 (23.1)	0.869 (2.11)
+ORPO		0.291 (16.9)	0.265 (21.4)	0.869 (1.57)	0.298 (19.9)	0.240 (18.4)	0.870 (2.21)
CheXagent-2		0.326	0.297	0.888	0.349	0.304	0.892
+DPO	GREEN	0.387 (18.9)	0.339 (14.1)	0.891 (0.30)	0.387 (10.9)	0.320 (5.34)	0.888 (-0.38)
+DPO	BERTScore	0.352 (8.11)	0.326 (9.58)	0.896 (0.95)	0.359 (2.88)	0.310 (1.78)	0.893 (0.15)

Table 3: Results on the MIMIC-CXR and CheXpert Plus test sets (percentage change compared to SFT baseline in brackets). Best results in bold, separate for CheXagent and CheXagent-2.

Model	Judge	F1-14		F1-5		Avg. (↑)
		Macro (↑)	Micro (↑)	Macro (↑)	Micro (↑)	
GPT-4V		20.4	35.5	19.6	25.8	25.3
MAIRA-1		38.6	55.7	47.7	56.0	49.5
MAIRA-2		41.6	58.1	50.4	59.1	52.3
Med-PaLM M (12B)		37.3	51.4	50.6	56.5	49.0
Med-PaLM M (84B)		39.8	53.6	51.6	57.9	50.7
Med-PaLM M (562B)		37.8	51.6	49.9	56.3	48.9
LLaVA-Rad		39.5	57.3	47.7	57.4	50.5
CheXagent		38.9	50.9	47.6	54.1	47.9
+DPO	GREEN	41.5* (6.87)	54.1* (6.18)	51.8* (8.83)	58.3* (7.78)	51.4* (7.43)
+LC-DPO		37.6 (-3.32)	52.1 (2.20)	47.5 (-0.02)	55.2* (1.92)	48.1 (0.45)
+IPO		39.0 (0.39)	52.9* (3.87)	48.9 (2.88)	56.5* (4.36)	49.3 (3.06)
+KTO		40.7 (4.81)	55.0* (8.07)	51.6* (8.52)	59.3* (9.54)	51.7* (7.93)
+ORPO		41.4* (6.64)	55.0* (8.03)	51.5* (8.20)	58.3* (7.69)	51.5* (7.69)
+DPO	BERTScore	42.5* (9.37)	56.1* (10.1)	54.7* (15.0)	61.6* (13.8)	53.7* (12.2)
+LC-DPO		42.3* (8.75)	56.3* (10.6)	52.7* (10.8)	60.8* (12.4)	53.0* (10.8)
+IPO		43.1* (10.8)	57.0* (11.9)	53.6* (12.6)	61.5* (13.6)	53.8* (12.4)
+KTO		44.0* (13.3)	58.0* (13.9)	54.0* (13.5)	62.3* (15.1)	54.6* (14.0)
+ORPO		42.4* (9.16)	56.7* (11.4)	52.5* (10.5)	60.7* (12.2)	53.1* (10.9)
CheXagent-2		44.6	57.8	55.5	62.4	55.1
+DPO	GREEN	45.8 (2.6)	59.7* (3.3)	56.0 (0.8)	64.1 (2.7)	56.4 (2.4)
+DPO	BERTScore	43.0 (-3.6)	59.7* (3.3)	53.5 (-3.6)	63.8 (2.3)	55.0 (-0.1)

Table 4: CheXbert scores on the MIMIC-CXR test set. Percentage change compared to SFT baseline in brackets. Best results in bold. *statistically significantly different from SFT baseline at a 10% level based on confidence intervals in Table 11.

5.2 Judge Optimization Results

Results, using single runs, for GREEN, LC-GREEN, and the BERTScore on the MIMIC-CXR and CheXpert Plus test sets are available in Table 3. Additional results for ROUGE-L and BLEU-4 are available in Table 10. As expected, using GREEN as Judge results in the largest boost in GREEN, and using the BERTScore in the BERTScore. In particular, the top performing configuration on the MIMIC-CXR test set for CheXagent according to GREEN and LC-GREEN is KTO, using GREEN as Judge, boosting these metrics by 31.9% and 34.1% percent, respectively. For CheXagent, the top performing configuration according to the BERTScore is obtained by KTO, using the BERTScore as Judge. We observe similar trends for CheXagent-2 with

DPO yielding an increase of 18.9% and 14.1% for GREEN and LC-GREEN, respectively when using GREEN as Judge. We observed overall similar trends for the CheXbert Plus, despite representing two different distributions: MIMIC-CXR was collected in an emergency department and CheXpert Plus was collected from in- and out-patient centers.

5.3 Generalization to CheXbert Scores

Although GREEN is a high quality and clinically relevant metric, the observed performance gains might be inflated due to the fact that we used GREEN as Judge for preference data generation. Thus, we instead turn to the CheXbert scores to be our silver-standard, a low-cost approximation of expert human (radiologists) judgment. CheXbert

Model	Judge	View Classification	Accuracy (\uparrow)					Avg.
			Binary Image Classification	Single Disease Identification	Multi Disease Identification	Visual Question Answering	Image-Text Reasoning	
CheXagent		98.5	83.8	62.8	69.0	75.5	66.3	76.0
+DPO	GREEN	98.5	84.0	63.2	68.7	75.5	66.1	76.0
+LC-DPO		98.5	83.7	62.2	68.3	75.5	65.3	75.6
+IPO		98.3	83.8	62.7	68.3	76.8	65.5	75.9
+KTO		98.3	83.8	63.2	68.7	74.6	65.0	75.6
+ORPO		98.2	83.5	63.5	68.6	74.3	64.2	75.4
+DPO	BERTScore	98.5	84.3	62.8	68.4	75.1	66.8	76.0
+LC-DPO		98.5	84.1	63.2	68.3	75.1	65.5	75.8
+IPO		98.5	83.8	62.9	68.2	76.0	66.3	76.0
+KTO		98.5	84.2	63.2	68.5	76.3	66.3	76.2
+ORPO		98.3	84.1	62.7	68.8	73.9	66.3	75.7
CheXagent-2		99.0	83.0	65.5	83.9	83.2	78.7	82.2
+DPO	GREEN	99.2	84.8	66.7	84.4	80.5	77.9	82.2
+DPO	BERTScore	99.2	83.0	65.9	84.5	82.3	78.2	82.2

Table 5: Performance on six image perception and reasoning tasks different from RRG on the datasets listed in §4.2.

scores are ubiquitous in this setting and we include results from a representative sample of recent state-of-the-art medical VLMs for the RRG tasks: GPT-4V⁹, MAIRA-1 (Hyland et al., 2023), MAIRA-2 (Bannur et al., 2024), Med-PaLM (Tu et al., 2024), and LLaVA-Rad (Chaves et al., 2024). We report macro and micro averaged F1-scores for the full 14 categories as well as the 5 categories subset. In addition, we provide the average across these scores. Results for our method are averages based on 1000 bootstrap samples, with confidence intervals available in Table 11.

For CheXagent, which is comparable to LLaVA-Rad and MAIRA-1 prior to alignment, our method boosts the average CheXbert scores by up to 7-8% and 14% using GREEN and the BERTScore, respectively. Very interestingly, and unexpectedly, the top performing setup for CheXagent using the BERTScore as Judge is better than that for GREEN as Judge. One possible explanation for this is reward overoptimization, as this seems to be a more prominent issue for GREEN than for the BERTScore. For CheXagent-2, which is the overall state-of-the-art prior to alignment and second only to MAIRA-2¹⁰ in terms of micro F1-14 scores, we can see that it is possible to improve performance even further using our proposed method. We note, however, that only the micro F1-14 scores are statistically significantly different from CheXbert-2, at a 10% level, for both GREEN and the BERTScore as Judge. Thus, we depict that it is possible to improve upon the already very strong micro averaged F1-scores, even when employing a general

domain NLG metric like the BERTScore. Granular results, for each of the 14 categories, are available in Table 12.

5.4 Alignment Tax Analysis

While RLHF is powerful, it has been observed that it might lead to performance degradations or, forgetting (Askeff et al., 2021; Ouyang et al., 2022). Ouyang et al. (2022) assessed such an alignment tax by evaluating the aligned policies on several natural language processing (NLP) benchmarks. Inspired by this, we benchmark the SFT baseline and the aligned policies on six diverse image perception and reasoning tasks using datasets listed in §4.2. Although there are some minor variations, on average, the performance matches that of the SFT baselines. Thus, our method substantially improves the quality of generated reports without compromising the quality of other image understanding tasks.

5.5 Qualitative Analysis

To further shed light on policy behavior pre and post alignment, a qualitative study was conducted by selecting three interesting cases from the MIMIC-CXR test set, covering a range of positive and negative findings and the presence and absence of indications. A board-certified radiologist was then asked to color-code candidate reports from the SFT baselines and aligned policies. Results for CheXagent-2 using GREEN as Judge for the first example are available in Fig. 4. For this case, despite being very strong, CheXagent-2 exhibited two errors: 1) it incorrectly reported clear lungs and 2) it failed to detect small bilateral pleural effusions. Notably, the aligned version addressed both of these errors. A similar conclusion is true for

⁹<https://openai.com/index/gpt-4v-system-card/>.

¹⁰Notably, MAIRA-2 use additional information at train and test time, including the *radiologist-written* prior report, when available.

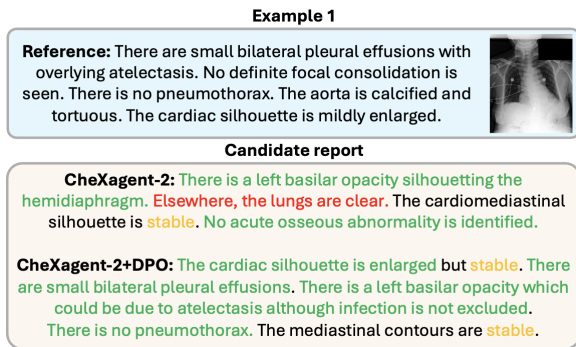


Figure 4: Color-coded candidate reports. Green and red represent correct and incorrect. Orange highlights references to a prior imaging study.

the other two cases. In addition, GREEN score as Judge seems to be superior to BERTScore and we observe exact, and semantically equivalent, repetitions in some cases. Full results, along with an extensive discussion, are available in §C.4.

6 Conclusion

Our study highlights the significant potential of including preference fine-tuning in the post-training pipeline of medical VLMs. By developing an automated pipeline we effectively address the prohibitive cost of obtaining radiologist preferences at scale. Using our approach, we have shown that DAAs can substantially improve AI-generated radiology reports in clinically meaningful ways *without additional radiologist feedback*. Our approach achieves state-of-the-art performance on the MIMIC-CXR dataset while maintaining robust capabilities across diverse visual reasoning tasks. The surprising effectiveness of, inexpensive, general-domain NLG metrics for preference pair generation suggests a promising path forward for computationally efficient exploration of online alignment algorithms in this setting.

7 Limitations

Our work focuses on only two VLMs, CheXagent (Chen et al., 2024a) and CheXagent-2 (Chen et al., 2024b). While different in terms of underlying architecture, information used, and baseline performance, it would be of interest to additionally study VLMs from other families and sizes. In addition, our study of CheXagent-2 was more limited in nature and should be considered more preliminary.

Moreover, we treat clinically relevant metrics such as GREEN and CheXbert scores as the silver standard. While these metrics are highly relevant,

a thorough evaluation involving clinical experts, radiologists, should be conducted for this study to be considered complete. Our qualitative analysis is a first step in this direction. However, conducting larger-scale reader studies remains an important direction for future work. We acknowledge potential biases beyond verbosity, such as societal biases related race, sex, or other demographic factors that may be embedded either in the underlying data or in the Judge. These potential biases warrant thorough investigation and mitigation in future work.

In addition, our hyperparameter search is 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, as well as the online, or iterative, counterparts to the DAAs considered. The recent success of DeepSeek-R1 (DeepSeek-AI et al., 2025) has led to a resurgence of interest in on-policy RL alignment algorithms. In particular, developing “verification functions” for the RRG task and employing reinforcement learning with verifiable rewards (RLVR) (Lambert et al., 2025) is an very interesting avenue for future work.

8 Acknowledgements

We would like to acknowledge compute support from Microsoft, Google, and Stanford Marlowe (Kapfer et al., 2025). D.H. is supported by MedTechLabs, the Göran Gustafsson foundation, and the foundation for Technical Scientific Research and Education. J.X. acknowledges support from the Canadian Institutes of Health Research (CIHR) and Oxford University Press. S.O. acknowledges funding from the DFG (German Research Foundation, ID: 517316550). M.V. is supported by graduate fellowship awards from the Department of Defense (NDSEG), the Knight-Hennessy Scholars program at Stanford University, and the Quad program. C.B. receives research support from the Promedica Foundation, Chur, CH. A.C. receives research support from NIH grants R01 HL167974, R01HL169345, R01 AR077604, R01 EB002524, R01 AR079431, P41 EB027060; ARPA-H grants AY2AX000045 and 1AYSAX0000024-01; and NIH contracts 75N92020C00008 and 75N92020C00021.

References

- American College of Radiology. 2019. [Siim-acr pneumothorax segmentation 2019](#).
- Amanda Askeff, 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. 2021. [A general language assistant as a laboratory for alignment](#). *Preprint*, arXiv:2112.00861.
- Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. [A general theoretical paradigm to understand learning from human preferences](#). *Preprint*, arXiv:2310.12036.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askeff, 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. 2022a. [Training a helpful and harmless assistant with reinforcement learning from human feedback](#). *Preprint*, arXiv:2204.05862.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askeff, 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. 2022b. [Constitutional ai: Harmlessness from ai feedback](#). *Preprint*, arXiv: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. 2024. [Maira-2: Grounded radiology report generation](#). *Preprint*, arXiv:2406.04449.
- Mythreyi Bhargavan, Jonathan H Sunshine, and Barbara Schepps. 2002. Too few radiologists? *American Journal of Roentgenology*, 178(5):1075–1082.
- Ralph Allan Bradley and Milton E. Terry. 1952. [Rank analysis of incomplete block designs: I. the method of paired comparisons](#). *Biometrika*, 39(3/4):324–345.
- RJM Bruls and RM Kwee. 2020. Workload for radiologists during on-call hours: dramatic increase in the past 15 years. *Insights into imaging*, 11:1–7.
- Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. 2024. Chexpert plus: Hundreds of thousands of aligned radiology texts, images and patients. *arXiv preprint arXiv:2405.19538*.
- 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. 2024. Training small multimodal models to bridge biomedical competency gap: A case study in radiology imaging. *arXiv preprint arXiv:2403.08002*.
- 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. 2024a. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208v1*.
- Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Andrew Johnston, Alaa Youssef, Louis Blankemeier, Christian Bluethgen, Stephan Altmayer, Jeya Maria Jose Valanarasu, Mohamed Siddig Eltayeb Muneer, Eduardo Pontes Reis, Joseph Paul Cohen, Cameron Olsen, Tanishq Mathew Abraham, Emily B. Tsai, Christopher F. Beaulieu, Jlenia Jitsev, Sergios Gatidis, Jean-Benoit Delbrouck, Akshay S. Chaudhari, and Curtis P. Langlotz. 2024b. [A vision-language foundation model to enhance efficiency of chest x-ray interpretation](#). *Preprint*, arXiv:2401.12208.
- Yashin Dicente Cid, Matthew Macpherson, Louise Gervais-Andre, Yuanyi Zhu, Giuseppe Franco, Ruggero Santeramo, Chee Lim, Ian Selby, Keerthini Muthuswamy, Ashik Amlani, et al. 2024. 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.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang,

- Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanxia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yudian Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.
- Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2023. [Alpacafarm: A simulation framework for methods that learn from human feedback](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 30039–30069. Curran Associates, Inc.
- Yann Dubois, Percy Liang, and Tatsunori Hashimoto. 2024. [Length-controlled alpacaeval: A simple debiasing of automatic evaluators](#). In *First Conference on Language Modeling*.
- Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. [Kto: Model alignment as prospect theoretic optimization](#). *Preprint*, arXiv:2402.01306.
- Leo Gao, John Schulman, and Jacob Hilton. 2023. [Scaling laws for reward model overoptimization](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10835–10866. PMLR.
- Tarek N Hanna, Haris Shekhani, Christine Lamoureux, Hanna Mar, Refky Nicola, Clint Sliker, and Jamlik-Omari Johnson. 2017. Emergency radiology practice patterns: shifts, schedules, and job satisfaction. *Journal of the American College of Radiology*, 14(3):345–352.
- Jiwoo Hong, Noah Lee, and James Thorne. 2024. [Orpo: Monolithic preference optimization without reference model](#). *Preprint*, arXiv:2403.07691.
- 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. 2023. [Maira-1: A specialised large multimodal model for radiology report generation](#). *arXiv preprint arXiv:2311.13668*.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. [Mistral 7b](#). *arXiv preprint arXiv:2310.06825*.
- 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, L elio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th eophile Gervet, Thibaut Lavril, Thomas Wang, Timoth ee Lacroix, and William El Sayed. 2024. [Mistral of experts](#). *Preprint*, arXiv:2401.04088.
- Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. 2024. [Promptmrg: Diagnosis-driven prompts for medical report generation](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(3):2607–2615.
- Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chihying Deng, Roger G Mark, and Steven Horng. 2019. [Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports](#). *Scientific data*, 6(1):317.
- Craig Kapfer, Kurt Stine, Balasubramanian Narasimhan, Christopher Mentzel, and Emmanuel Candes. 2025. [Marlowe: Stanford’s gpu-based computational instrument](#).

- Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahma, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2025. [Tulu 3: Pushing frontiers in open language model post-training](#). *Preprint*, arXiv:2411.15124.
- Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024a. [RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 26874–26901. PMLR.
- Suhyeon Lee, Won Jun Kim, Jinho Chang, and Jong Chul Ye. 2024b. [LLM-CXR: Instruction-finetuned LLM for CXR image understanding and generation](#). In *The Twelfth International Conference on Learning Representations*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.
- Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.
- Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023b. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Matthew Lyon, LaShon Sturgis, Darren Lendermon, Ann Marie Kuchinski, Taylor Mueller, Patrick Loeffler, Hongyan Xu, and Robert Gibson. 2015. Rural ed transfers due to lack of radiology services. *The American journal of emergency medicine*, 33(11):1630–1634.
- World Health Organization et al. 2016. Communicating radiation risks in paediatric imaging: information to support health care discussions about benefit and risk.
- Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, et al. 2024. Green: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- WHO PAHO. 2012. World radiography day: Two-thirds of the world’s population has no access to diagnostic imaging. *Pan American Health Organization*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. [Disentangling length from quality in direct preference optimization](#). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4998–5017, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Chantal Pellegrini, Matthias Keicher, Ege Özsoy, and Nassir Navab. 2023. Rad-restruct: A novel vqa benchmark and method for structured radiology reporting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 409–419. Springer.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, Bradley Knox, Chelsea Finn, and Scott Niekum. 2024. [Scaling laws for reward](#)

- model overoptimization in direct alignment algorithms. *Preprint*, arXiv:2406.02900.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. [Direct preference optimization: Your language model is secretly a reward model](#). *Preprint*, arXiv:2305.18290.
- Abi Rimmer. 2017. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ: British Medical Journal (Online)*, 359.
- Alexander T Ruutiainen, Daniel J Durand, Mary H Scanlon, and Jason N Itri. 2013. Increased error rates in preliminary reports issued by radiology residents working more than 10 consecutive hours overnight. *Academic radiology*, 20(3):305–311.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- 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. 2019. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041.
- Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. 2020. 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)*, pages 1500–1519.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. [Learning to summarize with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.
- Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. [Eva-clip: Improved training techniques for clip at scale](#). *Preprint*, arXiv:2303.15389.
- 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. 2024. [Aligning large multimodal models with factually augmented RLHF](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110, Bangkok, Thailand. Association for Computational Linguistics.
- Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Stefano Ermon, Chelsea Finn, and Aviral Kumar. 2024. [Preference fine-tuning of LLMs should leverage suboptimal, on-policy data](#). In *Forty-first International Conference on Machine Learning*.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, et al. 2024. [Gemini: A family of highly capable multimodal models](#). *Preprint*, arXiv:2312.11805.
- Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mulpally, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. 2023. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*.
- 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. 2023. [Llama: Open and efficient foundation language models](#). *Preprint*, arXiv:2302.13971.
- Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, Anil Palepu, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Sementur, S. Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Aguerre y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan. 2024. [Towards generalist biomedical ai](#). *NEJM AI*, 1(3):A10a2300138.
- R. J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256.
- Ting Xiao, Lei Shi, Peng Liu, Zhe Wang, and Chen-jia Bai. 2024. [Radiology report generation via multi-objective preference optimization](#). *Preprint*, arXiv:2412.08901.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- 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. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.
- Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024. [Aligning modalities in vision large language models via preference fine-tuning](#). *arXiv preprint arXiv:2402.11411*.

Kangyu Zhu, Peng Xia, Yun Li, Hongtu Zhu, Sheng Wang, and Huaxiu Yao. 2025. [Mmedpo: Aligning medical vision-language models with clinical-aware multimodal preference optimization](#). *Preprint*, arXiv:2412.06141.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. [Fine-tuning language models from human preferences](#). *Preprint*, arXiv:1909.08593.

A RLHF and DAAs

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

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

where \mathbb{D}_{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.¹¹ This class of algorithms re-parameterize the reward model via a change-of-variables using the closed-form solution to the objective in (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-form solution to (1), $\pi(y|x) \propto \pi_{\text{ref}}(y|x) \exp(R(x, y)/\beta)$ and the Bradley-Terry (BT) model (Bradley and 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, $p^*(y_1 \succ y_2 | x) = \sigma \left(\beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)} - \beta \log \frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)} \right)$, where 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

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

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

B Implementation Details

B.1 Data Details

We use the official train, validation, and test splits for the MIMIC-CXR (Johnson et al., 2019) and CheXpert Plus (Chambon et al., 2024) datasets. A key difference between CheXagent and CheXagent-2 is that CheXagent-2 additionally use indications, which provide clinical context, to aid in the RRG task. This means that CheXagent-2 imposes a stricter data requirement (i.e. each image-report pair must have a corresponding indication). Due to this, CheXagent-2 has slightly fewer examples than CheXagent for the RRG task. The number of examples in each split is available in Table 6. For MIMIC-CXR, the difference in test and validation is so small that it is negligible for the RRG task. On the other hand, for CheXpert Plus the difference is quite substantial. However, this is immaterial for any conclusions in this paper as we do not directly compare the models with each other on the CheXpert Plus data.

The view classification, binary image classification, single disease identification, multi disease identification, VQA, and image-text reasoning test data from the RSNA (Shih et al., 2019), SIIM (American College of Radiology, 2019), OpenI (Demner-Fushman et al., 2016), SLAKE (Liu et al., 2021), and Rad-Reconstruct (Pellegrini et al., 2023) datasets are processed as in Chen et al. (2024b).

B.2 Training Details

All models are trained using a machine with 4xA100, 4xA6000 or 8xH100 GPUs using learning rate learning rate 10^{-6} . We set global batch size to 32 for CheXagent (Chen et al., 2024a) and 64 for

¹¹In this paper, we use this terminology more loosely than in Rafailov et al. (2024).

Dataset	CheXagent			CheXagent-2		
	Train	Validation	Test	Train	Validation	Test
MIMIC-CXR	148463	1164	2309	148090	1162	2300
CheXpert Plus			208			151

Table 6: Number of examples in each split of MIMIC-CXR and CheXpert plus for CheXagent and CheXagent-2.

CheXagent-2 (Chen et al., 2024b). Each model is trained for one epoch. The image encoder is frozen while we train the LLM. 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 7, including optimal configurations. This non-exhaustive search was based on previous work and our initial experiments. For CheXagent, 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. The same issue emerged for some λ for CheXagent-2.

Algorithm	Objective	Hyperparameters
DPO (Rafailov et al., 2023)	$-\log \sigma \left(\beta \log \frac{\pi_{\theta}(y_c x)}{\pi_{\text{ref}}(y_c x)} - \beta \log \frac{\pi_{\theta}(y_r x)}{\pi_{\text{ref}}(y_r x)} \right)$	$\beta \in [0.01, 0.05^{*\dagger}, 0.1]$
LC-DPO (Park et al., 2024)	$-\log \sigma \left(\beta \log \frac{\pi_{\theta}(y_c x)}{\pi_{\text{ref}}(y_c x)} - \beta \log \frac{\pi_{\theta}(y_r x)}{\pi_{\text{ref}}(y_r x)} + \alpha (y_c - y_r) \right)$	$\beta \in [0.01, 0.05^{*\dagger}, 0.1], \alpha \in [0.001, 0.01^{*\dagger}]$
IPO (Azar et al., 2023)	$\left(\log \frac{\pi_{\theta}(y_c x)}{\pi_{\text{ref}}(y_c x)} - \log \frac{\pi_{\theta}(y_r x)}{\pi_{\text{ref}}(y_r x)} - \frac{1}{2\tau} \right)^2$	$\tau \in [0.1, 0.5, 1.0^{*\dagger}]$
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 $z_{\text{ref}} = \mathbb{E}_{(x,y) \sim \mathcal{D}} [\beta \mathbb{D}_{\text{KL}}(\pi_{\theta}(y x) \pi_{\text{ref}}(y x))]$	$\beta \in [0.01, 0.05^*, 0.1^{\dagger}], \lambda_c = \lambda_r$
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 $p_{\theta}(y x) = \exp \left(\frac{1}{ y } \log \pi_{\theta}(y x) \right)$	$\lambda \in [0.5, 1.0, 4.0^{\dagger}, 5.0^*]$

Table 7: Hyperparameter search for all direct alignment algorithms (DAAs) considered in this paper. For CheXagent, * and † denotes best for GREEN and the BERTScore as Judge, respectively. For CheXagent-2, $\beta = 0.1$ was found optimal for both GREEN and the BERTScore.

B.3 Sampling Details

In this paper, we set up CheXagent to treat the cases where the two CXRs are a frontal and a lateral image (i.e. from the same point in time) and a frontal and a prior frontal image (i.e. from two points in time) as separate cases with a separate prompt. This is in contrast to CheXagent-2, which employs the same prompt for both cases. Moreover, we employed stochastic sampling with temperature 1.0 for CheXagent at test time, while CheXagent-2 is based on greedy sampling, both with beam size set to 1. For preference pairs generation, both employ stochastic sampling with temperature 1.0 and beam size set to 1. These differences will explain parts of the performance differences between CheXagent and CheXagent-2. However, this is of minor concern to this study as we are mainly interested in the effect of preference fine-tuning, comparing the same setup prior and post alignment.

B.4 Evaluation Details

GREEN (Ostmeier et al., 2024) and GREEN-LC are based on the official implementation¹² using StanfordAIMI/GREEN-radllama2-7b. The BERTScore (Zhang et al., 2019) used is from evaluate (v0.4.0) with distilbert-base-uncased as BERT model. The F1 CheXbert (Smit et al., 2020) scores use f1chexbert (v0.0.2) with default configurations. ROUGE (Lin, 2004) is from rouge-score (v0.1.2) using rougeL. Finally, the BLEU (Papineni et al., 2002) is BLEU-4, based on a custom code, available in the code base corresponding to this project.

CheXbert scores for GPT-4V¹³, MAIRA-1 (Hyland et al., 2023), Med-PaLM (Tu et al., 2024), and LLaVA-Rad (Chaves et al., 2024) are borrowed directly from Supplementary Table 1 in Chaves et al.

¹²<https://github.com/Stanford-AIMI/GREEN>.

¹³<https://openai.com/index/gpt-4v-system-card/>.

(2024). The CheXbert scores for MAIRA-2 (Bannur et al., 2024) are from Table D.1 in Bannur et al. (2024).

C Additional results

C.1 Preference data

Summary statistics for the chosen and rejected subsets are available in Table 8. We also report summary statistics of the length (in words) of the generated reports. Notably, the difference, or spread, in average length of the chosen and rejected subsets is slightly more pronounced for GREEN than for the BERTScore, 6.9% compared to 5.8% and much more significant overall for CheXagent-2 with a difference of 19.3% and 17.1%, respectively. To build further build intuition on the resulting chosen and rejected subsets we report CheXbert scores in Table 9. Scores for CheXagent and CheXagent-2 are not directly comparable since the training data is different. As expected, these scores indicate that the chosen subset is considerably better than the rejected. The average CheXbert scores for the chosen subset are more or less the same for GREEN and BERTScore as Judge. Similarly is true for the rejected subset, though now BERTScore as Judge results in slightly lower scores—meaning that the spread is marginally larger. Finally, as a qualitative study, we include a couple of examples of rejected and chosen candidates for CheXagent and CheXagent-2 is Fig. 5 and Fig. 6, respectively. We include the score assigned by the respective Judge in brackets.

	GREEN						BERTScore					
	Metric			Report Length			Metric			Report Length		
	Mean	Median	Std.	Mean	Median	Std.	Mean	Median	Std.	Mean	Median	Std.
CheXagent												
Chosen	0.63	0.60	0.24	56.3	54.0	20.2	0.90	0.90	0.03	55.1	53.0	20.2
Rejected	0.26	0.22	0.19	52.7	51.0	24.9	0.85	0.86	0.04	52.1	50.0	25.3
CheXagent-2												
Chosen	0.57	0.50	0.25	54.4	52.0	19.3	0.91	0.90	0.03	52.7	51.0	18.6
Rejected	0.24	0.20	0.20	45.6	44.0	22.8	0.85	0.86	0.05	45.0	43.0	23.8

Table 8: Summary statistics of reference-based metric and report length in the chosen and rejected subsets using GREEN and the BERTScore as Judges.

Model	Judge	Chosen					Rejected				
		F1-14		F1-5		Avg. (†)	F1-14		F1-5		Avg. (†)
		Macro (†)	Micro (†)	Macro (†)	Micro (†)		Macro (†)	Micro (†)	Macro (†)	Micro (†)	
CheXagent	GREEN	0.574	0.668	0.649	0.700	0.648	0.411	0.506	0.455	0.507	0.463
	BERTScore	0.582	0.680	0.647	0.700	0.652	0.395	0.488	0.444	0.493	0.419
CheXagent-2	GREEN	0.488	0.587	0.550	0.621	0.561	0.336	0.416	0.365	0.415	0.373
	BERTScore	0.488	0.594	0.541	0.615	0.560	0.316	0.394	0.346	0.393	0.362

Table 9: CheXbert scores on the MIMIC-CXR on the chosen and rejected subsets.

GREEN		BERTScore		
<p>Reference: The heart size is normal. The mediastinal contours are normal. The hilar contours are normal. Bibasilar atelectasis is present. There is no focal consolidation. There is no pleural effusion or pneumothorax. There is no free intraperitoneal air.</p>	<p>Chosen: The lungs are well expanded, with no evidence of pleural effusion, pulmonary edema, pneumothorax or focal consolidation concerning for pneumonia. There is mild blunting of the left costophrenic angle which may be due to atelectasis but a trace pleural effusion is not excluded. The cardiomeastinal silhouette is stable. The partially imaged upper abdomen is unremarkable. (0.67)</p> <p style="text-align: center;">✓</p>	<p>Rejected: The heart size is normal. Calcifications are noted at the aortic arch. The lungs are clear without focal consolidation. There is no pleural effusion or pneumothorax. There is no evidence of a large hiatal hernia. The air fluid levels seen on the most recent ct represent colonic interposition. (0.5)</p> <p style="text-align: center;">✗</p>	<p>Chosen: Slightly leftward patient rotation limits assessment. The heart size is normal. The mediastinal contours are normal. The hilar contours are normal. There is no focal consolidation, pleural effusion or pneumothorax. Bibasilar opacities are most consistent with atelectasis. No displaced rib fracture is identified. (0.92)</p> <p style="text-align: center;">✓</p>	<p>Rejected: The lungs are well expanded, with no evidence of pleural effusion, pulmonary edema, pneumothorax or focal consolidation concerning for pneumonia. There is mild blunting of the left costophrenic angle which may be due to atelectasis but a trace pleural effusion is not excluded. The cardiomeastinal silhouette is stable. The partially imaged upper abdomen is unremarkable. (0.85)</p> <p style="text-align: center;">✗</p>
<p>An endotracheal tube is in place, ending about 4.0 cm above the carina. An enteric tube is seen below the diaphragm, with the side port at the ge junction and the distal tip in the proximal stomach. Recommend advancement of the enteric tube for better placement within the stomach. A right internal jugular central venous catheter is present, ending in the mid-to-lower svc without signs of pneumothorax. There are extensive left greater than right bilateral alveolar opacities which continue to increase. Differential diagnosis includes pulmonary hemorrhage with quickly moving infection, pulmonary hemorrhage, ards, component of underlying aspiration is not excluded. Possible left pleural effusion is seen.</p>	<p>An endotracheal tube is in place, ending about 4.9 cm above the carina. An enteric tube is seen, ending in the distal esophagus. A right internal jugular central venous catheter is present, ending in the mid svc. There are extensive left greater than right bilateral airspace opacities. A small left pleural effusion is present. No pneumothorax is seen. The cardiac silhouette is stable. The mediastinal silhouette is stable. (0.38)</p> <p style="text-align: center;">✓</p>	<p>An endotracheal tube is in a stable position. An interval placement of a right ij has been done and appears to be in the correct position. There are extensive left greater than right bilateral alveolar opacities which may be due to pulmonary hemorrhage contusion. Mildly displaced left sided rib fractures are seen. (0.17)</p> <p style="text-align: center;">✗</p>	<p>An endotracheal tube is in place, ending about 4.9 cm above the carina. An enteric tube is seen, ending in the distal esophagus. A right internal jugular central venous catheter is present, ending in the mid svc. There are extensive left greater than right bilateral airspace opacities. A small left pleural effusion is present. No pneumothorax is seen. The cardiac silhouette is stable. The mediastinal silhouette is stable. (0.90)</p> <p style="text-align: center;">✓</p>	<p>An endotracheal tube is in a stable position. An interval placement of a right ij has been done and appears to be in the correct position. There are extensive left greater than right bilateral alveolar opacities which may be due to pulmonary hemorrhage contusion. Mildly displaced left sided rib fractures are seen. (0.85)</p> <p style="text-align: center;">✗</p>

Figure 5: First two examples of chosen and rejected candidates for CheXagent in the MIMIC-CXR train set. Number in brackets is assigned score (GREEN or BERTScore).

GREEN			BERTScore	
<p>Reference: There is no focal consolidation. There is no pleural effusion or pneumothorax. Bilateral nodular opacities that most likely represent nipple shadows. The cardiomeastinal silhouette is normal. Clips project over the left lung, potentially within the breast. The imaged upper abdomen is unremarkable. Chronic deformity of the posterior left sixth and seventh ribs are noted.</p>	<p>Chosen: The lungs are clear without focal consolidation, effusion or edema. The cardiomeastinal silhouette is within normal limits. Chronic abnormalities of the left fifth rib are unchanged. Surgical clips are seen overlying the periphery of the left lung and anterior chest wall. No acute osseous abnormalities. (0.6)</p> <p style="text-align: center;">✓</p>	<p>Rejected: Surgical clips are seen over the left breast. There is mild streaky density on the left, likely due to subsegmental atelectasis. Nipple shadows are visible bilaterally. The lungs appear clear with no focal consolidation. The cardiomeastinal silhouette is normal. (0.29)</p> <p style="text-align: center;">✗</p>	<p>Chosen: There is no focal consolidation, effusion, or pneumothorax. The cardiomeastinal silhouette is normal. The imaged osseous structures are intact. No free air is seen below the right hemidiaphragm. There are surgical clips in the left chest wall (0.9)</p> <p style="text-align: center;">✓</p>	<p>Rejected: The lungs are clear of consolidation, effusion, or vascular congestion. The cardiomeastinal silhouette is within normal limits. Surgical clips project over the left chest wall and right axilla. (0.86)</p> <p style="text-align: center;">✗</p>
<p>The lung volumes are low. The cardiac contours appear unchanged. The mediastinal contours appear unchanged. The hilar contours appear unchanged. There are a number of round nodular densities projecting over each upper lung, but more numerous and discretely visualized in the left upper lobe. However, in addition, there is a more hazy widespread opacity projecting over the left mid upper lung which could be compatible with a coinciding pneumonia. Pulmonary nodules in the left upper lobe are also not completely characterized on this study. There is no pleural effusion or pneumothorax. Post-operative changes are similar along the right chest wall.</p>	<p>Lung volumes are low. The cardiac contours appear stable. The mediastinal contours appear stable. The hilar contours appear stable. There is no definite pleural effusion or pneumothorax. The lungs demonstrate perihilar ground-glass opacities with an upper lobe predominance, left greater than right. Findings suggest large airways inflammation or infection. There is a nodular focus projecting over the left upper lobe measuring up to 8mm, perhaps a focus of acute infection or inflammation or potentially a pulmonary nodule. Small nodular focus projecting over the right costophrenic angle suggests a stable nodule measuring 6 mm or less. Mild-to-moderate bibasilar atelectasis appears unchanged on the left and improved on the right. (0.38)</p> <p style="text-align: center;">✓</p>	<p>There are multiple bilateral pulmonary nodules. There are no definite signs of pneumonia in the lower lungs. There is no pleural effusion or pneumothorax. The cardiomeastinal silhouette is stable. The bony structures are intact. (0.18)</p> <p style="text-align: center;">✗</p>	<p>Lung volumes are low. The cardiac contours are within normal limits. The mediastinal contours are within normal limits. The hilar contours are within normal limits. There is crowding of the bronchovascular structures but no overt pulmonary edema is demonstrated. Multiple nodular opacities are noted bilaterally, being more numerous on the left. The lung volumes are low. Patchy opacities are noted in both lung bases, more pronounced within the left lung base. Chronic pleural thickening within the right lung base is observed, with blunting of the lateral costophrenic angle compatible with chronic pleural thickening. No pneumothorax or large pleural effusion is demonstrated. There are no acute osseous abnormalities. (0.90)</p> <p style="text-align: center;">✓</p>	<p>Multiple bilateral lung nodules are present without obvious cavitation. There is increased bibasilar atelectasis. The cardiomeastinal silhouette and hilar contours appear normal allowing for low lung volumes. There is no large pleural effusion or pneumothorax. Chronic right rib deformities are unchanged. (0.88)</p> <p style="text-align: center;">✗</p>

Figure 6: First two examples of chosen and rejected candidates for CheXagent-2 in the MIMIC-CXR train set. Number in brackets is assigned score (GREEN or BERTScore).

C.2 Judge Optimization Results

		MIMIC-CXR				
Model	Judge	GREEN (↑)	GREEN-LC (↑)	BERTScore(↑)	BLEU-4 (↑)	ROUGE-L (↑)
CheXagent		0.249	0.218	0.856	0.041	0.274
+DPO	GREEN	0.325 (30.6)	0.263 (20.5)	0.862 (0.65)	0.057 (39.9)	0.293 (6.94)
+LC-DPO		0.320 (28.7)	0.288 (32.0)	0.864 (0.90)	0.056 (38.4)	0.302 (10.4)
+IPO		0.326 (30.9)	0.282 (29.3)	0.863 (0.84)	0.059 (45.0)	0.297 (8.57)
+KTO		0.328 (31.9)	0.293 (34.1)	0.867 (1.27)	0.056 (38.8)	0.305 (11.4)
+ORPO		0.322 (29.4)	0.275 (26.2)	0.862 (0.69)	0.053 (29.5)	0.290 (6.08)
+DPO	BERTScore	0.285 (14.4)	0.242 (11.0)	0.869 (1.56)	0.059 (44.2)	0.309 (12.9)
+LC-DPO		0.283 (13.5)	0.259 (18.7)	0.871 (1.75)	0.057 (41.2)	0.315 (15.1)
+IPO		0.283 (13.6)	0.247 (13.1)	0.870 (1.62)	0.060 (47.2)	0.309 (12.9)
+KTO		0.304 (21.9)	0.279 (28.2)	0.872 (1.88)	0.056 (37.7)	0.310 (13.4)
+ORPO		0.291 (16.9)	0.265 (21.4)	0.869 (1.57)	0.054 (33.8)	0.307 (12.0)
CheXagent-2		0.326	0.297	0.888	0.136	0.350
+DPO	GREEN	0.387 (18.9)	0.339 (14.1)	0.891 (0.301)	0.150 (10.3)	0.357 (1.93)
+DPO	BERTScore	0.352 (8.11)	0.326 (9.58)	0.896 (0.949)	0.153 (12.5)	0.372 (6.37)
		CheXpert Plus				
Model	Judge	GREEN (↑)	GREEN-LC (↑)	BERTScore(↑)	BLEU-4 (↑)	ROUGE-L (↑)
CheXagent		0.248	0.202	0.851	0.038	0.274
+DPO	GREEN	0.322 (29.9)	0.222 (9.69)	0.859 (0.91)	0.053 (40.2)	0.289 (5.25)
+LC-DPO		0.330 (32.8)	0.268 (32.3)	0.861 (1.26)	0.049 (30.2)	0.305 (10.6)
+IPO		0.334 (34.6)	0.267 (31.9)	0.861 (1.15)	0.052 (39.1)	0.295 (7.22)
+KTO		0.341 (37.2)	0.266 (31.4)	0.863 (1.42)	0.052 (36.9)	0.305 (10.9)
+ORPO		0.326 (31.4)	0.232 (14.6)	0.856 (0.64)	0.046 (21.9)	0.287 (4.63)
+DPO	BERTScore	0.278 (12.0)	0.208 (2.84)	0.867 (1.89)	0.054 (43.6)	0.307 (11.3)
+LC-DPO		0.299 (20.3)	0.248 (22.5)	0.871 (2.34)	0.056 (47.6)	0.318 (15.1)
+IPO		0.282 (13.6)	0.218 (7.88)	0.866 (1.75)	0.052 (37.6)	0.309 (12.0)
+KTO		0.308 (24.0)	0.249 (23.1)	0.869 (2.11)	0.054 (43.8)	0.316 (14.6)
+ORPO		0.298 (19.9)	0.240 (18.4)	0.870 (2.21)	0.049 (30.8)	0.314 (14.0)
CheXagent-2-3b		0.349	0.304	0.892	0.123	0.361
+DPO	GREEN	0.387 (10.89)	0.320 (5.34)	0.888 (-0.37)	0.126 (2.74)	0.356 (-1.30)
+DPO	BERTScore	0.359 (2.88)	0.310 (1.78)	0.893 (0.15)	0.130 (2.86)	0.357 (-0.91)

Table 10: Results on the MIMIC-CXR and CheXpert Plus test sets (percentage change compared to SFT baseline in brackets). Best results in bold, separate for CheXagent and CheXagent-2.

C.3 Generalization to CheXbert Scores

Model	Judge	F1-14		F1-5		Avg. (↑)
		Macro (↑)	Micro (↑)	Macro(↑)	Micro (↑)	
GPT-4V		20.4	35.5	19.6	25.8	25.3
MAIRA-1		38.6	55.7	47.7	56.0	49.5
MAIRA-2		41.6	58.1	50.4	59.1	52.3
Med-PaLM M (12B)		37.3	51.4	50.6	56.5	49.0
Med-PaLM M(84B)		39.8	53.6	51.6	57.9	50.7
Med-PaLM M (562B)		37.8	51.6	49.9	56.3	48.9
LLaVA-Rad		39.5	57.3	47.7	57.4	50.5
<hr/>						
CheXagent		38.9 _[37.8,40.0]	50.9 _[50.0,51.8]	47.6 _[46.3,48.8]	54.1 _[52.9,55.3]	47.9 _[46.9,48.8]
+DPO		41.5 _[40.2,42.9]	54.1 _[53.2,54.9]	51.8 _[50.4,53.1]	58.3 _[57.1,59.5]	51.4 _[50.4,52.4]
+LC-DPO		37.6 _[36.3,38.9]	52.1 _[51.1,53.1]	47.5 _[46.1,49.0]	55.2 _[53.9,56.4]	48.1 _[47.1,49.2]
+IPO		39.0 _[37.7,40.3]	52.9 _[52.0,53.8]	48.9 _[47.5,50.4]	56.5 _[55.3,57.7]	49.3 _[48.3,50.3]
+KTO		40.7 _[39.5,41.9]	55.0 _[54.1,55.9]	51.6 _[50.3,52.9]	59.3 _[58.1,60.4]	51.7 _[50.7,52.6]
+ORPO		41.4 _[40.1,42.7]	55.0 _[54.2,55.9]	51.5 _[50.1,52.8]	58.3 _[57.1,59.5]	51.5 _[50.6,52.5]
<hr/>						
+DPO		42.5 _[41.4,43.7]	56.1 _[55.2,56.9]	54.7 _[53.5,55.9]	61.6 _[60.5,62.6]	53.7 _[52.8,54.6]
+LC-DPO		42.3 _[41.0,43.4]	56.3 _[55.5,57.2]	52.7 _[51.3,54.0]	60.8 _[59.6,61.9]	53.0 _[52.1,54.0]
+IPO		43.1 _[41.9,44.2]	57.0 _[56.2,57.8]	53.6 _[52.3,54.8]	61.5 _[60.4,62.5]	53.8 _[52.9,54.6]
+KTO		44.0 _[42.9,45.3]	58.0 _[57.2,58.9]	54.0 _[52.8,55.2]	62.3 _[61.2,63.4]	54.6 _[53.7,55.5]
+ORPO		42.4 _[41.2,43.7]	56.7 _[55.9,57.6]	52.5 _[51.3,53.6]	60.7 _[59.7,61.7]	53.1 _[52.2,53.9]
<hr/>						
CheXagent-2		44.6 _[43.3,45.9]	57.8 _[56.9,58.7]	55.5 _[54.2,57.1]	62.4 _[61.2,63.6]	55.1 _[54.1,56.1]
+DPO		45.8 _[44.5,47.1]	59.7 _[58.9,60.6]	56.0 _[54.6,57.5]	64.1 _[63.0,65.3]	56.4 _[55.4,57.4]
+DPO		43.0 _[41.7,44.4]	59.7 _[58.8,60.6]	53.5 _[52.1,54.9]	63.8 _[62.6,65.0]	55.0 _[54.1,56.0]

Table 11: CheXbert scores on the MIMIC-CXR test set. 90% confidence interval obtained by 1000 bootstrap samples in subscripts. Best results in bold.

		F1-scores (↑)													
Model	Judge	ECm.	Cmgl.	LOpac.	LLes.	Edema	Cnsl.	Pna.	Atel.	PEff.	Pmtx.	POth.	Frac.	SuDev.	NoF.
CheXagent		0.347	0.620	0.461	0.171	0.493	0.158	0.227	0.453	0.655	0.444	0.092	0.240	0.787	0.304
+DPO		0.372	0.675	0.444	0.186	0.525	0.226	0.231	0.452	0.710	0.416	0.178	0.217	0.822	0.365
+LC-DPO		0.385	0.666	0.380	0.178	0.430	0.211	0.158	0.404	0.665	0.407	0.083	0.129	0.835	0.339
+IPO	GREEN	0.402	0.686	0.395	0.160	0.472	0.222	0.206	0.405	0.661	0.483	0.090	0.138	0.822	0.321
+KTO		0.388	0.682	0.422	0.202	0.542	0.192	0.180	0.444	0.721	0.552	0.074	0.118	0.841	0.349
+ORPO		0.348	0.684	0.479	0.201	0.492	0.224	0.247	0.475	0.698	0.511	0.072	0.177	0.835	0.365
<hr/>															
+DPO		0.373	0.688	0.467	0.192	0.580	0.236	0.141	0.503	0.728	0.483	0.157	0.239	0.831	0.336
+LC-DPO		0.301	0.692	0.506	0.208	0.567	0.175	0.107	0.471	0.731	0.567	0.167	0.225	0.832	0.376
+IPO	BERTScore	0.386	0.690	0.500	0.218	0.559	0.190	0.111	0.500	0.739	0.550	0.195	0.220	0.836	0.342
+KTO		0.430	0.701	0.487	0.217	0.577	0.172	0.201	0.514	0.735	0.584	0.166	0.197	0.830	0.360
+ORPO		0.377	0.690	0.491	0.148	0.562	0.157	0.231	0.481	0.736	0.497	0.116	0.218	0.826	0.418
<hr/>															
CheXagent-2		0.324	0.672	0.539	0.244	0.606	0.235	0.284	0.543	0.720	0.527	0.085	0.330	0.815	0.333
+DPO		0.450	0.696	0.493	0.218	0.606	0.214	0.325	0.545	0.738	0.526	0.157	0.255	0.856	0.343
+DPO		0.358	0.703	0.547	0.166	0.622	0.127	0.066	0.478	0.749	0.550	0.143	0.271	0.853	0.396

Table 12: F1 scores on the MIMIC-CXR test set using 14 categories from the CheXbert labeler (Smit et al., 2020): Enlarged Cardiomeadiastinum (ECm.), Cardiomegaly (Cmgl.), Lung Opacity (LOpac.), Lung Lesion (LLes.), Edema, Consolidation (Cnsl.), Pneumonia (Pna.), Atelectasis (Atel.), Pleural Effusion (PEff.), Pneumothorax (Pmtx.), Pleural Other (POth.), Fracture (Frac.), Support Devices (SuDev.), no Findings (NoF.). Best results in bold, separate for CheXagent and CheXagent-2.

C.4 Qualitative Analysis

To build some further intuition of behavior pre and post alignment, we conduct a qualitative study in which a board-certified radiologist was asked to color-code candidate reports from CheXagent, CheXagent-2, as well as their top performing aligned versions for GREEN and BERTScore as Judge, respectively, in terms of cheXbert scores in Table 4. We selected three interesting, and representative, examples from MIMIC-CXR test set, including examples with positive and negative findings, only negative findings, and with and without indications. These were subsequently given to a radiologist to color-code, using the reference reports as well as the CXRs in png format (i.e. exact measurements are not possible). Three cases with six reports each, resulted in a total of 18 reports to be processed. The reports are color-coded as green (correct), red (incorrect), and orange to indicate where an reference to prior report is made though none was presented at the time of inference. For instance, the statement that something is “stable” may or may not be true, depending on the prior imaging study.

Example 1 and 2 are available in Fig. 7. For example 1, CheXagent-2, despite being a very strong model, contains two errors. First, it incorrectly states that lungs are elsewhere clear—marked in red. Secondly, comparing with the reference report, the detection of small bilateral pleural effusions is missing. None of these errors are made by the aligned versions, using both GREEN and BERTScore as Judge. Moreover, for CheXagent with both GREEN and BERTScore as Judge, there are repetitions. For GREEN “no overt edema” is mentioned twice. For BERTScore as Judge, “no pneumothorax” is repeated. These types of exact repetitions may be due to verbosity bias.

Moving on to example 2, we can see that there are exact repetition for CheXagent-2 aligned with DPO. “There is no pneumothorax” and “There is no pleural effusion” are bot repeated as “There is no evidence of ...”. Interestingly, such repetition are not present for BERTScore as Judge, nor for the results from CheXagent. Notably, however, both for CheXagent-2 and CheXagent there are errors in the candidate report using BERTScore as Judge for this case. “Support devices have been removed” from CheXagent cannot cannot be supported nor refuted definitively.

Finally, consider example 3 in Fig. 8. This case includes a lot of details, with a reference report significantly longer than the previous two cases. One thing that is very interesting about this case is that the reference report includes a quantitative measurement: the distance between where the endotracheal tube terminates and the carina, expressed in centimeters. Quantitative results are given in some of the candidate reports. However, due to the set up of this reader study (i.e. that we used CXRs in png format) it is not feasible for the radiologist to make measurements and determine distance definitively. To capture this additional feature, we introduce the additional color blue. Takeaways are similar as to the previous two cases. Errors are made by CheXagent-2 and CheXagent, as well as the aligned versions using BERTScore as Judge. For GREEN as Judge, the performance remains strong, with no direct errors. Moreover, cheXagent makes a partial mistake not capture by our color coding scheme in this case as it states that there is “A small right pleural effusion,” where it is actually bilateral.

Example 1

Reference: There are small bilateral pleural effusions with overlying atelectasis. No definite focal consolidation is seen. There is no pneumothorax. The aorta is calcified and tortuous. The cardiac silhouette is mildly enlarged.



Candidate reports

CheXagent-2: There is a left basilar opacity silhouetting the hemidiaphragm. Elsewhere, the lungs are clear. The cardiomeastinal silhouette is stable. No acute osseous abnormality is identified.

CheXagent-2 + DPO (GREEN): The cardiac silhouette is enlarged but stable. There are small bilateral pleural effusions. There is a left basilar opacity which could be due to atelectasis although infection is not excluded. There is no pneumothorax. The mediastinal contours are stable.

CheXagent-2 + DPO (BERTScore): There is left basilar opacity. Small bilateral pleural effusions are present. The lungs are otherwise clear. The cardiac silhouette is enlarged. The mediastinal contours are stable. There is no pneumothorax.

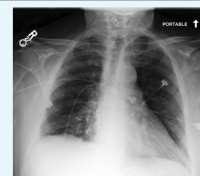
CheXagent: The heart is at the upper limits of normal size. The mediastinal contour is prominent likely due to an unfolded thoracic aorta. The lungs are well expanded with left basilar linear opacity, likely representing atelectasis. There is no focal consolidation concerning for pneumonia. There may be tiny bilateral pleural effusions. No pneumothorax is detected. There is mild rightward convex curvature centered along the mid thoracic spine with small to moderate osteophyte formation.

CheXagent + KTO (GREEN): There are tiny, persistent bilateral pleural effusions. There is no overt edema or pneumothorax. The cardiomeastinal silhouette is stable. There is no mediastinal widening. There are no definite signs of pneumonia. There is no overt edema. The imaged osseous structures appear intact.

CheXagent + KTO (BERTScore): There are small bilateral pleural effusions with overlying mild atelectasis. The cardiac silhouette is mildly enlarged. The cardiomeastinal silhouette is stable. There is no pneumothorax. Tortuosity of the descending thoracic aorta is noted. There is no pneumothorax. No acute osseous abnormalities are seen.

Example 2

Indication: ___-year-old male with esophageal stricture, now status post esophageal dilation and Kenalog injection, requiring assessment for pneumomediastinum and pneumothorax.



Reference: There is no pneumothorax or pneumomediastinum. There is no deep cervical air. The lungs are well expanded. There is no evidence of acute cardiac or pulmonary process. The cardiomeastinal silhouette is unremarkable.

Candidate reports

CheXagent-2: The lungs are well expanded. There is a right lower lobe opacity. There is no pneumothorax or pleural effusion. The heart is top normal in size. The mediastinal contours are normal. The hilar contours are normal. There is a mild dextrosciosis of the thoracic spine.

CheXagent-2 + DPO (GREEN): The lungs are well expanded. There is no focal consolidation. There is no pneumothorax. There is no pleural effusion. The heart size is normal. The mediastinal silhouette is normal. The hilar contours are normal. There is no evidence of pneumomediastinum. There is no evidence of pneumothorax. There is no evidence of pleural effusion.

CheXagent-2 + DPO (BERTScore): The lungs are well expanded. There is mild right basilar atelectasis. There is no pneumothorax. The cardiomeastinal silhouette is normal. The hilar contours are normal. There is no pleural effusion.

CheXagent: The heart size is normal. The hilar and mediastinal contours are within normal limits. Support devices have been removed and lung volumes are improved. The opacity in the right lower lung has improved and may represent atelectasis. There is no opacity, pulmonary edema or pneumothorax. A small right pleural effusion is seen.

CheXagent + KTO (GREEN): There is no pleural effusion or pneumothorax. There is no pulmonary edema. The heart size is at the upper limit of normal. There is no focal consolidation concerning for aspiration or pneumonia. The hilar and mediastinal silhouettes are normal.

CheXagent + KTO (BERTScore): The heart size is normal. The cardiomeastinal silhouette and hilar contours are normal. There is minimal opacification at the right lung base. The lungs are otherwise clear. There is no pleural effusion or pneumothorax.

Figure 7: Color-coded candidate reports: Examples 1 and 2. Green and red represent correct and incorrect. Orange refers to prior imaging study.

Example 3

Reference: The endotracheal tube terminates 4.2 cm above the carina. The side port of a nasogastric tube is below the diaphragm. Pulmonary vasculature is ill-defined, compatible with severe pulmonary edema. Hazy opacity overlying both lungs and blunting of the costophrenic angles are compatible with bilateral pleural effusions. No lobar consolidation or pneumothorax. Mild cardiomegaly is present. The leads of a left chest wall pacer terminate in the right atrium and ventricle. The median sternotomy wires and numerous mediastinal clips are intact.



Candidate reports

CheXagent-2: The endotracheal tube is seen with its tip residing approximately 3.8 cm above the carina. The ng tube courses into the left upper abdomen. A dual-lead pacing device is noted. Midline sternotomy wires and mediastinal clips are noted. There is diffuse pulmonary opacity which could indicate edema, ards, or pneumonia. Bilateral pleural effusions are present. The heart size is top normal. The bony structures are intact.

CheXagent-2 + DPO (GREEN): An endotracheal tube is seen with its tip residing approximately 4.5 cm above the carina. An ng tube is seen passing into the stomach. A dual-lead pacemaker is seen with leads extending to the region of the right atrium and right ventricle. There is severe pulmonary edema with bilateral pleural effusions. The cardiac silhouette is enlarged. The mediastinal contour is stable. The bony structures are intact.

CheXagent-2 + DPO (BERTScore): The endotracheal tube tip is approximately 5 cm from the carina. An enteric tube is seen below the diaphragm. A left-sided pacemaker device is noted with leads terminating in the right atrium and right ventricle. The heart size is mildly enlarged. The mediastinal contours are unchanged. There is moderate pulmonary edema. Small bilateral pleural effusions are present. There is no pneumothorax. No acute osseous abnormalities are seen.

CheXagent: An endotracheal tube has been placed, which terminates in the mid thoracic trachea. An enteric tube extends inferiorly with tip in the stomach. Left sided pacer wires are in appropriate position. Sternotomy wires are intact. Small surgical clips are seen over the mediastinum. The heart is moderately enlarged. Diffuse bilateral opacities are compatible with moderate pulmonary edema. The right diaphragm is elevated with right base atelectasis. Surgical clips project over the left upper mediastinum. A small right pleural effusion is seen. There is no pneumothorax.

CheXagent + KTO (GREEN): An endotracheal tube is seen, terminating above the carina. An enteric tube courses below the diaphragm, the tip of which is not visualized. A left sided pacemaker is seen with leads overlying the right atrium and ventricle, expected locations. The cardiac silhouette is moderately enlarged. Diffuse pulmonary opacities are present, most consistent with moderate to severe pulmonary edema. Pleural effusions may be present. No definitive pneumothorax is seen.

CheXagent + KTO (BERTScore): An endotracheal tube is noted terminating above the carina. A left anterior chest wall pacer defibrillator is seen with leads terminating over the heart. The cardiomeastinal silhouette is enlarged. Diffuse pulmonary opacities are present, likely reflecting moderate pulmonary edema. Layering pleural effusions are also noted. There is no pneumothorax. Median sternotomy wires and mediastinal clips are noted.

Figure 8: Color-coded candidate reports: Example 3. Green and red represent correct and incorrect. Orange refers to prior imaging study. Blue indicates measurements that have not been verified.