Direct Preference Optimization for Suppressing Hallucinated Prior Exams in Radiology Report Generation

Oishi Banerjee[1,2], Hong-Yu Zhou[1,3], Subathra Adithan[4], Stephen Kwak[5], Kay Wu[3], and Pranav Rajpurkar[3]

[1]Equal contribution.
[2]Department of Computer Science, Harvard University
[3]Department of Biomedical Informatics, Harvard University
[4]Department of Radiodiagnosis, Jawaharlal Institute of Postgraduate Medical Education & Research
[5]Department of Radiology, Johns Hopkins University

Abstract

Recent advances in generative vision-language models (VLMs) have exciting potential implications for AI in radiology, yet VLMs are also known to produce hallucinations, nonsensical text, and other unwanted behaviors that can waste clinicians' time and cause patient harm. Drawing on recent work on direct preference optimization (DPO), we propose a simple method for modifying the behavior of pretrained VLMs performing radiology report generation by suppressing unwanted types of generations. We apply our method to the prevention of hallucinations of prior exams, addressing a long-established problem behavior in models performing chest X-ray report generation. Across our experiments, we find that DPO fine-tuning achieves a 3.2-4.8x reduction in lines hallucinating prior exams while maintaining model performance on clinical accuracy metrics. Our work is, to the best of our knowledge, the first work to apply DPO to medical VLMs, providing a data- and compute- efficient way to suppress problem behaviors while maintaining overall clinical accuracy.

1. Introduction

Recent advances in generative vision-language models (VLMs) have exciting potential implications for AI in radiology. Typically pretrained through self-supervised learning on large image-text datasets, VLMs can achieve high performance on complex multimodal tasks, interpreting radiology images and describing them in sophisticated medical text. Unfortunately VLMs are also known to produce hallucinations, nonsensical text, and other unwanted behaviors associated with large generative AI models. In the context of radiology, such behaviors are likely to increase clinician workloads, forcing them to spend time searching for and fixing AI errors, and may lead to patient harm if mistakes are left uncorrected.

To address these challenges, multiple techniques for reducing errors in generative models have been proposed. One approach is to preemptively prevent unwanted behaviors by carefully modifying pretraining datasets, such as by removing references to an off-limits topic. However, preprocessing large pretraining datasets in this fashion can become prohibitively

expensive or time-consuming; furthermore, modifications to pretraining data cannot resolve issues found later on, after pretraining is already complete. In contrast, methods based on reinforcement learning with human or AI feedback (RLHF) can be applied after pretraining, improving already strong models by making their outputs better align with human preferences. Direct preference optimization (DPO), a technique that recently grew out of RLHF, offers a particularly simple and stable way to reduce unwanted behaviors in pretrained models, as it dispenses with the explicit reward model typically needed for reinforcement learning (Rafailov et al., 2023). RLHF-based methods have shown considerable promise in the general domain, but they are not yet well-explored in the context of medicine.

In this paper, we propose and validate a DPO-inspired method for suppressing unwanted types of generations from VLMs performing chest X-ray (CXR) report generation. Radiology report generation is a challenging task for VLMs, requiring them to identify all diseases or other features of interest in a radiology image and fluently describe them in a short essay, ideally while taking into account the particular patient's history. We specifically apply our method to reducing the hallucination of prior exams, which has long raised concerns in radiology report generation (Hyland et al., 2024; Yu et al., 2023; Ramesh et al., 2022; Miura et al., 2021). When engaging in this unwanted behavior, models generate hallucinatory statements such as "The lungs are hyperinflated with emphysematous changes as seen on prior CT" or "There has been interval increase in right-sided opacity [i.e. compared to a previous image]," despite not having access to any prior exam. These hallucinations can reduce the utility of AI-generated radiology reports for multiple reasons. First, they require radiologists to spend extra effort checking and editing these references, which could prove misleading if left uncorrected. Furthermore, references to prior exams can reduce the amount of clinically useful information in a report; for example, a report that claims "cardiac size is unchanged" fails to make clear whether the heart size is abnormal or not. We show that our method makes a pretrained VLM substantially less likely to hallucinate prior exams, while simultaneously maintaining the model's clinical accuracy.

2. Main Insights

We make the following specific contributions:

1. We introduce and evaluate several DPO methods for radiology report generation, including both standard and weighted DPO losses. We generally find that DPO methods can be used to selectively suppress unwanted behaviors with little impact on clinical accuracy.

2. We create a subset of MIMIC-CXR's training set, where GPT-4 has removed the majority of references to prior exams from the findings and impression sections of each report. These can be used in future research on reducing references to prior exams.

3. We also create subsets of MIMIC-CXR's validation and test sets, where GPT-4 has removed the majority of references to prior exams from the findings and impression sections. These may be broadly useful when evaluating models, potentially providing an alternative to MIMIC-CXR that better penalizes models for hallucinating prior exams.

Figure 1: Three examples of reports, generated before and after DPO training. Our method can decrease references to prior exams (red), without significantly reducing clinical accuracy.

Our work is, to the best of our knowledge, the first work to apply direct preference optimization to medical VLMs, successfully suppressing a problematic behavior while maintaining overall clinical accuracy. Compared to typical pretraining processes, our method is relatively compute- and data-efficient, making it a practical option even under resource constraints.

3. Related Work

Generative AI for Radiology Report Generation. In the past, some methods for AI-based radiology reporting have avoided training VLMs to produce free-text reports, relying on other approaches such as retrieval (Jeong et al., 2023; Endo et al., 2021), template-based reporting (Kale et al., 2023), or the use of GPT to reformat outputs from image models in text (Yan et al., 2023; Ranjit et al., 2023). However, powerful VLMs that directly generate free-text radiology reports based on input chest X-rays are rapidly growing in popularity (Chaves et al., 2024; Tu et al., 2023; Xu et al., 2023; Hyland et al., 2024). Some works have used reinforcement learning to improve the performance of such generative models (Zhou et al., 2024b; Qin and Song, 2022; Miura et al., 2021). These methods have historically required the construction of explicit reward models and focused on overall clinical accuracy, while our method does not require an extra reward model and aims to target and suppress more specific unwanted behaviors.

Direct Preference Optimization. Direct preference optimization (DPO) simplifies RLHF by training generative models directly on preferred and dispreferred responses for a given

prompt, without requiring a new reward model (Rafailov et al., 2023). DPO can effectively suppress hallucinations (Castricato et al., 2024), including in image captions generated by VLMs (Zhou et al., 2024a; Gunjal et al., 2024). Importantly, DPO has been shown to align pretrained models with human preferences without compromising their existing domain knowledge, making them a promising choice for radiology report generation (Ivison et al., 2023). While the standard DPO loss typically places equal importance on all parts of a preferred or dispreferred response, a few variants prioritize certain response pairs (Liu et al., 2024; Amini et al., 2024) or certain parts of an individual response (Gunjal et al., 2024); inspired by this work, we experiment with weighted DPO losses that prioritize certain parts of a response. A handful of works have used DPO to fine-tune medical LLMs (Ahn et al., 2024; Ye et al., 2023; Dou et al., 2024), but to the best of our knowledge, ours is the first DPO method for medical VLMs.

4. Methods

4.1. Direct Preference Optimization

To perform DPO, we require a reference model π_{ref}, which has already been pretrained for the overall task of radiology report generation. We also require a preference dataset, where every datapoint consists of a prompt x, a preferred response y_w, and a dispreferred response y_l. In the context of radiology report generation, a prompt might contain the image(s) from a study, as well as available background information such as the indication. The preferred response would be the desired report, while the dispreferred response would be an undesirable version of the report that demonstrates the unwanted behavior that should be suppressed. Using this dataset, we train a policy model, π_θ, which is initially set to equal π_{ref}. Given prompt x, the standard DPO loss essentially rewards the π_θ for assigning high probabilities to y_w and low probabilities to y_l, relative to the reference model:

$$\mathcal{L}_{\text{DPO}} = -\log \sigma \bigg(\beta \big(\log \pi_\theta(y_w|x) - \log \pi_{\text{ref}}(y_w|x) \big) - \beta \big(\log \pi_\theta(y_l|x) - \log \pi_{\text{ref}}(y_l|x) \big) \bigg) \quad (1)$$

where $\sigma(\cdot)$ denotes the sigmoid function.

Notably, this standard DPO loss assigns equal weight to all tokens in each response, regardless of whether those tokens relate to the behavior we wish to suppress. This property may be useful when suppressing broad behaviors such as harmfulness, which may affect all parts of a dispreferred response. However, many clinical use cases for DPO will focus on suppressing far more specific behaviors, which affect only a small self-contained part of a dispreferred response. For example, in our use case, a dispreferred response may refer to a prior exam in one line and otherwise be identical to the preferred response. Concerningly, the standard DPO loss would harshly penalize our model for producing any of the tokens in y_l, even though most lines in y_l contain unobjectionable, correct content.

To suppress specific behaviors without penalizing unrelated correct statements, we also explore weighted DPO losses that selectively pay extra attention to certain parts of each response. First, we assume every token in the responses is labeled as either relevant or irrelevant to our behavior of interest. We next introduce a new hyperparameter γ, where $0 \leq \gamma \leq 1$. When calculating $\log \pi(y|x)$, we multiply the log-probabilities of irrelevant tokens by γ, effectively reducing their importance in comparison to the relevant tokens.

Setting $\gamma = 0$ essentially skips irrelevant tokens, while $\gamma = 1$ gives the standard DPO loss where all tokens are weighted equally. To revisit the previous example where y_l contained one reference to a prior exam, a weighted DPO loss would allow us to penalize the one problematic line in y_l more heavily than the other prior-free lines.

To formally define the loss, let us assume y consists of substrings y_1 through y_n and that there are corresponding binary variables r_1 through r_n, where $r_i = 1$ if y_i is relevant to the behavior of interest and 0 otherwise. The loss can be written as follows:

$$\mathcal{L}_{\text{WDPO}} = -\log \sigma \bigg(\beta \big(f(\pi_\theta, y_w, x) - f(\pi_{\text{ref}}, y_w, x) \big) - \beta \big(f(\pi_\theta, y_l, x) - f(\pi_{\text{ref}}, y_l, x) \big) \bigg), \quad (2)$$

where the function f provides the weighted log probability that a model π will produce a response y, given a prompt x:

$$f(\pi, y, x) = \sum_{i=1}^{n} \big(\mathbb{1}[r_i = 1] + \gamma \mathbb{1}[r_i = 0] \big) \log \pi \big(y_i | (x, y_1...y_{i-1}) \big), \quad (3)$$

where $\mathbb{1}(\cdot)$ stands for the indicator function.

4.2. Pretrained Model

We pretrain a VLM through next-token prediction on MIMIC-CXR (Johnson et al., 2019). We use this pretrained model as the reference model when calculating DPO losses and as the initial model checkpoint for all our fine-tuning experiments. Additionally, we use this pretrained model as a baseline model and report its performance.

The VLM mainly consists of three components: a vision encoder, a vision-language adapter, and a LLM. The vision encoder first converts the input image into visual tokens, which are mapped to the language space by the adapter. The processed visual tokens are then forwarded through the LLM along with the background information and the prompt to generate the CXR report. Specifically, we use the base model of Swin Transformer (Liu et al., 2021) as the vision encoder. The adapter design consists of three layers. The first layer reduces visual tokens to manage GPU memory using adaptive pooling. Then, layer normalization is applied, followed by a linear projection to map visual representations to the language space. We initialize the LLM using Llama2-Chat-7b (Touvron et al., 2023) and apply parameter-efficient tuning using LoRA (Hu et al., 2021).

4.3. Dataset Creation

We build and release ReXPref-Prior, a preference dataset available on PhysioNet (Banerjee et al., 2024). We use CXRs and reports from MIMIC-CXR, following MIMIC-CXR-JPG's standard train/validation/test split(Johnson et al., 2024). In the training set, we include reports that contain a "findings" section and/or an "impression" section marked under clear headers (i.e. "FINDINGS:" and "IMPRESSION:", "CONCLUSION:" or "SUMMARY:"). In the validation and test sets, we only include reports that contain both a "findings" section and an "impression" section marked under clear headers. Where available, we also extract the indication and comparison sections and include these with the image in our prompts.
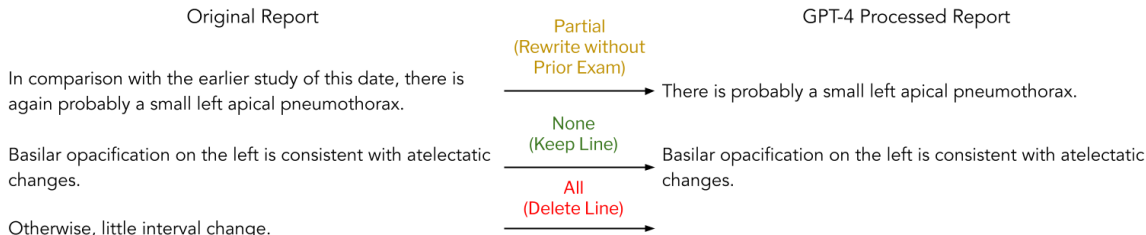
Figure 2: GPT-4 labels each line in a report based on how much of it depends on a prior exam ("none", "partial", "all") and rewrites "partial" lines to omit references to prior exams.

4.4. Removing References to Prior Exams with GPT-4

We use GPT-4 to edit MIMIC-CXR's reports so they avoid references to prior exams, using few-shot prompting. In collaboration with radiologists, we developed a list of keywords often associated with references to prior exams, and we customize the examples in our few-shot prompt for each report, depending on which keywords are present. The prompt instructs GPT-4 to examine every line in the original MIMIC-CXR report and label how much of the line depends on references to prior exams: "none", "partial", or "all". For this list of keywords and further details on the prompt, please see Appendix A. Depending on the label, we take the following actions (Figure 2):

1. If the label is "none", we copy the line into the edited report without making any changes. When calculating the weighted DPO loss, we treat this line as an "irrelevant" part of both the dispreferred and preferred responses.

2. If the label is "partial", we prompt GPT-4 to rewrite the line without references to prior exams, and we use the rewritten line in the edited report (e.g. "Consolidation has worsened" becomes "Consolidation is present"). When calculating the weighted DPO loss, we treat the original line as a "relevant" part of the dispreferred response. In addition, we treat the rewritten line as a "relevant" part of the preferred response.

3. If the label is "all", we omit the line from the edited report. When calculating the weighted DPO loss, we treat this line as a "relevant" part of the dispreferred response; this line is absent from the preferred response.

To preserve all possible information from the ground-truth report, we instruct GPT-4 to generally prefer the label of "partial" over "all". We reserve the label of "all" for extreme cases where a line offers no meaningful information without access to the prior exam (e.g. "No change from prior image", "Cardiac size is unchanged"). Sometimes, a report's findings or impression section is entirely deleted when we remove references to prior exams because all its lines are labeled "all". If even one section is missing, either because it is removed during processing or because it is not found in the original report, we exclude the report

6

from the validation and test datasets. If both sections end up missing, we exclude the report from the training dataset.

4.5. Training Dataset

We create a DPO training dataset, consisting of dispreferred reports that reference prior exams and preferred reports that avoid such references. To identify dispreferred reports, we identify 20,000 reports in MIMIC-CXR's training set. To improve dataset diversity, we select reports with unique values when we concatenate the findings and impression section. Additionally, we only select reports that contain the substring "compar" in the findings or impression section, as reports containing "compar" are highly likely to contain at least one reference to a prior exam. To produce the preferred version of each report, we remove references to prior exams using GPT-4, as described above. We filter out 162 reports that receive malformed responses from GPT-4 (e.g. outputs for hallucinated extra lines, invalid JSON syntax). After removing reports where both the findings and impression section are missing, we arrive at a final training dataset with 19,806 studies.

4.6. Validation Dataset

To produce our validation dataset, we use GPT-4 to remove references to prior exams from 1,718 reports in the MIMIC-CXR validation set that contain a clearly marked findings or impression section. When creating the validation set, we filter out 13 reports that receive malformed responses from GPT-4. After also filtering out reports where the findings or impression section is missing, we arrive at a final validation set with 915 studies.

4.7. Test Datasets

We evaluate all models using two versions of our test set. The first is an original set of reports taken directly from MIMIC-CXR. The second is a processed set of these reports where GPT-4 has removed references to prior exams, following the procedure described above.

We first use GPT-4 to remove references to prior exams from 2,919 reports in the MIMIC-CXR test set that contain a clearly marked findings or impression section. Unlike in the training dataset, we include reports regardless of whether they contain the substring "compar". When creating the processed test set, a research assistant manually fixes GPT-4's malformed responses, rather than excluding these reports. After removing reports where the findings or impression section is missing in one or both versions of the dataset, we arrive at our final test sets, each representing the same 1,383 studies.

We expect that a model that frequently refers to prior exams may better match the original set of reports, while an otherwise identical model that rarely refers to prior exams would score higher on the processed set. Since some of our models are likely to hallucinate prior exams far more frequently than others, we provide results from evaluating reports on both distributions.

Figure 3: An overview of our pretraining and fine-tuning methods. Models are rewarded for producing prior-free responses (green) and penalized for generating responses referencing prior exams (red). Depending on the DPO weighting scheme, lines unrelated to prior exams may receive less weight ($\gamma = .5$) or be entirely ignored ($\gamma = 0$).

4.8. DPO Experiments

As our first baseline, we examine our original pretrained model, without fine-tuning it further. As a second baseline, we perform supervised fine-tuning on our processed reports; this experiment essentially tests whether the DPO loss is necessary to avoid hallucinating prior exams or whether exposure to our GPT-4-processed data is sufficient. Finally, we fine-tune the model on our preference dataset using 3 DPO losses:

1. $\gamma = 1$, or the standard DPO loss. In this setting, we give equal weight to all parts of a response.

2. $\gamma = .5$. In this setting, we double the weight of relevant lines, compared to irrelevant ones.

3. $\gamma = 0$. We effectively ignore irrelevant lines, which are the same in both reports. Instead we focus only on the differences between reports.

We fine-tune all models for 12,000 iterations. Similar to past work on DPO (Rafailov et al., 2023), we warm up the learning rate from 0 by 300 iterations, reaching a maximum value of 5e-7, and set $\beta = .1$. We use the RMSprop optimizer with a batch size of 16, where the weight decay is set to 0.05. In the training phase, we apply a random cropping method to process input images, with the cropped region varying between 50% and 100% of the original image size. Subsequently, these cropped images are resized to a standardized dimension of 224×224 pixels. We use LoRA (Hu et al., 2021) for parameter-efficient tuning. We saved a

8

checkpoint every 2,000 iterations. For each checkpoint, we calculated both the RadCliq-v1 score and the average number of lines with priors on the validation set. We then ranked all the checkpoints based on these two metrics. The checkpoint with the highest average ranking was selected for testing.

4.9. Evaluation Approach

When evaluating our fine-tuning approaches, we are interested in two questions:

1. Does the model produce fewer references to prior exams?

2. Does the model maintain its original level of clinical accuracy without overfitting?

To estimate the number of references to prior exams, we again collaborate with radiologists to produce an extensive list of key terms associated with prior exams. We measure the average number of lines in each report containing at least one of these terms to estimate the frequency of references of prior exams. We also measure the proportion of responses from each model that contains at least one of these terms.

To determine whether the model maintains or improves on its original level of clinical accuracy, we rely on RadCliq-v1 (Yu et al., 2023) and RadGraph-F1 Jain et al. (2021), two metrics specialized for measuring radiology report quality. We note that RadCliq-v1 is of primary importance, as it provides the best indication of radiologist opinion (Yu et al., 2023). We also report results on BLEU-1 and BLEU-2, two general-purpose lexical metrics. To assess significance, we calculate bootstrapped 95% confidence intervals using scipy.stats's bootstrap function.

In addition to computing automated metrics across our entire dataset, we perform expert evaluation on a random subset of 33 studies. The 5 model responses for each study are presented in random order and annotated by either a radiologist or a radiology resident, who do not know which model produced each response. When evaluating a response, the annotator counts how many lines referred to a prior exam. They also compare the response to the original MIMIC ground-truth report, including the indication or comparison where available, and determine whether the model response contains a clinically significant discrepancy that would likely impact patient management.

5. Results

5.1. GPT-4 Annotation Quality

We find that GPT-4 is a noisy annotator, removing approximately 60% of references to prior exams from our training dataset. A more detailed breakdown of GPT-4's annotation performance is available in Appendix C.

5.2. Reduction of References to Prior Exams

Based on our search for key terms related to prior exams, we find that all DPO methods significantly reduce the frequency of hallucinated prior exams (Table 1 and Figure 4a). On average, our original pretrained model refers to prior exams in 1.34 lines per report,

Experiment	# Lines w/ Prior	% Reports w/ Prior
Pretrained	1.34 (1.25, 1.44)	51.63 (48.95, 54.23)
SFT	1.28 (1.20, 1.37)	55.97 (53.36, 58.64)
$\gamma = 1$	0.42 (0.38, 0.48)	27.40 (25.09, 29.79)
$\gamma = .5$	**0.28 (0.25, 0.33)**	**20.90 (18.80, 23.07)**
$\gamma = 0$	0.39 (0.35, 0.44)	25.23 (22.99, 27.62)

Table 1: Effect of our fine-tuning methods on hallucinated prior exams. We estimate the average number of lines referring to prior exams per report, as well as the proportion of reports referring to a prior exam at least once.

and roughly 50% of its reports mention a prior exam at least once. In contrast, the best-performing DPO model on this metric, trained with a weighted loss where $\gamma = .5$, refers to prior exams in .28 lines per report, significantly outperforming all of our other models. Even our worst-performing DPO model on this metric, trained with a standard DPO loss, refers to prior exams in .42 lines per report, significantly improving on both non-DPO models. Furthermore, DPO generally halves the proportion of reports that mention prior exams to 20-25%. We find that supervised fine-tuning on processed reports does not significantly reduce hallucinated prior exams.
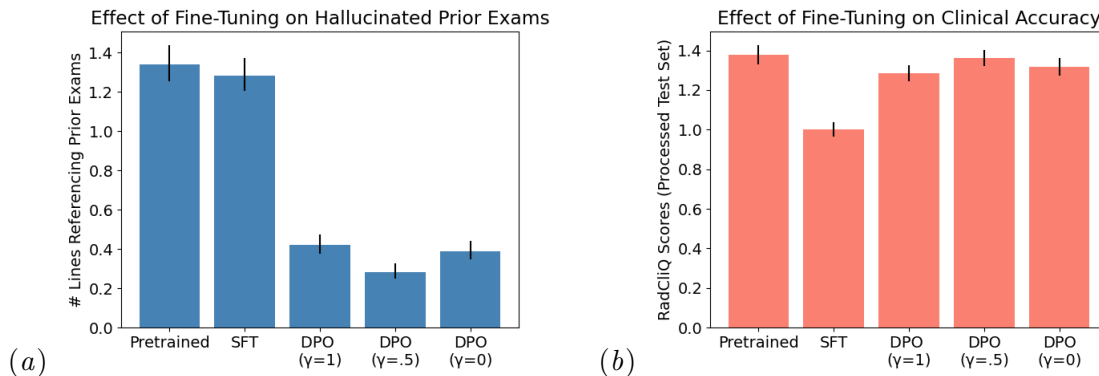

Figure 4: DPO fine-tuning dramatically reduces the average number of lines referring to hallucinated prior exams (left), while maintaining clinical accuracy (right). Supervised fine-tuning (SFT) improves clinical accuracy but does not meaningfully prevent the hallucination of prior exams. For both metrics, lower scores indicate higher performance.

5.3. Effect of Fine-Tuning on Clinical Accuracy

We next measure the accuracy our models using specialized clinical metrics and find that the original pretrained model and DPO models perform similarly, while supervised fine-tuning consistently achieves the highest clinical accuracy (Table 2 and Figure 4b). For RadGraph-F1, higher scores indicate better performance. For RadCliq-V1, lower scores indicate better performance. For performance on lexical metrics, please see Appendix D.

RadGraph-F1. On the original, unprocessed dataset, the original pretrained model outperforms the DPO models, though the difference is only significant when $\gamma = .5$. On the processed dataset with fewer references to prior exams, the gap in performance shrinks, and two DPO models ($\gamma = 0$ or 1) outperform the original pretrained model, though the differences between the DPO models and the pretrained model are again non-significant.

RadCliQ-V1. On the original, unprocessed dataset, two DPO models ($\gamma = 0$ or 1) outperform the original pretrained model, while the third DPO model ($\gamma = .5$) performs worse; these differences are not significant. On the processed dataset, all three DPO models outperform the original pretrained model. For one of these models ($\gamma = 1$), the improvement is significant.

Experiment	RadGraph-F1 (Original)	RadCliQ-V1 (Original)	RadGraph-F1 (Processed)	RadCliQ-V1 (Processed)
Pretrained	0.2005 (0.1906, 0.2115)	1.3914 (1.3413, 1.4412)	0.1924 (0.1827, 0.2032)	1.3782 (1.3295, 1.4257)
SFT	0.2604 (0.2521, 0.2694)	1.0147 (0.9769, 1.0515)	0.2521 (0.2437, 0.2610)	1.0010 (0.9629, 1.0371)
$\gamma = 1$	0.1937 (0.1857, 0.2016)	1.3352 (1.2951, 1.3769)	0.1932 (0.1850, 0.2012)	1.2847 (1.2455, 1.3263)
$\gamma = .5$	0.1746 (0.1671, 0.1824)	1.4187 (1.3780, 1.4603)	0.1755 (0.1679, 0.1834)	1.3623 (1.3225, 1.4036)
$\gamma = 0$	0.1968 (0.1879, 0.2057)	1.3601 (1.3156, 1.4049)	0.1952 (0.1864, 0.2041)	1.3158 (1.2721, 1.3603)

Table 2: Effect of our fine-tuning methods on common clinical accuracy metrics. For RadGraph-F1, higher scores indicate better performance, while for RadCliQ-V1, lower scores indicate better performance

5.4. Human Evaluation

Though less pronounced, the trends from our expert evaluation of 33 studies were largely consistent with our findings on the whole dataset. We again find that DPO consistently reduces the number of hallucinated prior exams. Setting $\gamma = 0$ or 1 resulted in fairly little change to clinical accuracy. Again, setting $\gamma = .5$ led to the largest reduction in hallucinated prior exams but slightly worsened clinical accuracy. Additionally, we find that our automatic metric for counting references to prior exams is generally accurate, agreeing

Experiment	# Lines w/ Prior	% Reports w/ Prior	% Reports with Major Error
Pretrained	.97	33%	64%
SFT	.97	45%	61%
$\gamma = 1$	.52	24%	64%
$\gamma = .5$	.30	18%	85%
$\gamma = 0$	.39	18%	70%

Table 3: We report the average number of lines referring to prior exams per report, the proportion of reports that refer to a prior exam at least once, and the percentage reports with a clinically significant error, based on expert evaluation of 33 studies.

with radiologists on over 80% of responses with a tendency towards overcounting otherwise. More details are provided in Appendix B.

6. Discussion

We find that, across all values of γ, DPO methods are effective in reducing how frequently the model refers to prior exams. Setting $\gamma = .5$ reduces lines with these unwanted references by a factor of 4.8, at the cost of slightly lower clinical accuracy. Our other two DPO models, where $\gamma = 1$ or 0, reduce lines with these references by a factor of 3.2 or 3.4, with higher clinical accuracy. When measured by RadCliQ-V1, a metric designed specifically to reflect radiologist opinions (Yu et al., 2023), these two DPO models outperform the pretrained model on both versions of our dataset. When measured by RadGraph-F1, the clinical accuracy of these two DPO models is comparable to the pretrained model's on the original dataset and slightly better on the processed dataset.

Prior work has raised concerns that DPO's improvements may stem from improvements in fine-tuning data (i.e. from processing by strong models like GPT-4), rather than from the DPO loss itself (Sharma et al., 2024). However, we show that data quality alone cannot reduce hallucinated prior exams. Supervised fine-tuning on our processed training dataset improves clinical accuracy, perhaps because cases mentioning prior exams tend to be especially challenging, but it does not meaningfully reduce the number of lines referencing prior exams. We therefore find that the DPO loss itself is necessary for suppressing unwanted behaviors.

Limitations We use GPT-4 to annotate our preference dataset, identifying and rewriting lines that refer to prior exams. While GPT-4 is generally capable of reducing references to prior exams, it does not eliminate them entirely, with potentially negative implications for our results. We hypothesize references remaining in our training set limit the performance of our DPO methods, so future work may find that DPO is even more effective with higher-quality annotations. Likewise, we hypothesize that performance on our test sets underrepresents the clinical accuracy of our DPO models, since even our processed test set still contains some references to prior exams. To better assess performance on CXR report

generation in the future, it would be useful to develop new benchmark datasets that are entirely free of references to prior exams, better representing the ideal target distribution.

Data Availability Our dataset can be found on PhysioNet at https://doi.org/10.13026/hwzp-zq96.

Acknowledgments

References

Imjin Ahn, Hansle Gwon, Young-Hak Kim, Tae Joon Jun, and Sanghyun Park. Note: Notable generation of patient text summaries through efficient approach based on direct preference optimization, 2024.

Afra Amini, Tim Vieira, and Ryan Cotterell. Direct preference optimization with an offset, 2024.

Oishi Banerjee, Hong-Yu Zhou, Subathra Adithan, Stephen Kwak, Kay Wu, and Pranav Rajpurkar. Rexpref-prior: A mimic-cxr preference dataset for reducing hallucinated prior exams (version 1.0.0). PhysioNet, 2024.

Louis Castricato, Nathan Lile, Suraj Anand, Hailey Schoelkopf, Siddharth Verma, and Stella Biderman. Suppressing pink elephants with direct principle feedback, 2024.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, Tristan Naumann, Muhao Chen, Matthew P. Lungren, Serena Yeung-Levy, Curtis P. Langlotz, Sheng Wang, and Hoifung Poon. Training small multi-modal models to bridge biomedical competency gap: A case study in radiology imaging, 2024.

Chengfeng Dou, Zhi Jin, Wenpin Jiao, Haiyan Zhao, Yongqiang Zhao, and Zhenwei Tao. Integrating physician diagnostic logic into large language models: Preference learning from process feedback, 2024.

Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y. Ng, and Pranav Rajpurkar. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In Subhrajit Roy, Stephen Pfohl, Emma Rocheteau, Girmaw Abebe Tadesse, Luis Oala, Fabian Falck, Yuyin Zhou, Liyue Shen, Ghada Zamzmi, Purity Mugambi, Ayah Zirikly, Matthew B. A. McDermott, and Emily Alsentzer, editors, *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pages 209–219. PMLR, 04 Dec 2021. URL https://proceedings.mlr.press/v158/endo21a.html.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Stephanie L. Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, Noel Codella, Matthew P. Lungren, Maria Teodora Wetscherek, Ozan Oktay, and Javier Alvarez-Valle. Maira-1: A specialised large multimodal model for radiology report generation, 2024.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. Camels in a changing climate: Enhancing lm adaptation with tulu 2, 2023.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, Curtis Langlotz, et al. Radgraph: Extracting clinical entities and relations from radiology reports. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

Jaehwan Jeong, Katherine Tian, Andrew Li, Sina Hartung, Fardad Behzadi, Juan Calle, David Osayande, Michael Pohlen, Subathra Adithan, and Pranav Rajpurkar. Multimodal image-text matching improves retrieval-based chest x-ray report generation, 2023.

Alistair Johnson, Matthew Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. Mimic-cxr-jpg - chest radiographs with structured labels. PhysioNet, 2024.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Kaveri Kale, pushpak Bhattacharyya, and Kshitij Jadhav. Replace and report: Nlp assisted radiology report generation, 2023.

Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, Peter J. Liu, and Xuanhui Wang. Lipo: Listwise preference optimization through learning-to-rank, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021. naacl-main.416. URL https://aclanthology.org/2021.naacl-main.416.

Han Qin and Yan Song. Reinforced cross-modal alignment for radiology report generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 448–458, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022. findings-acl.38. URL https://aclanthology.org/2022.findings-acl.38.

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

Vignav Ramesh, Nathan Andrew Chi, and Pranav Rajpurkar. Improving radiology report generation systems by removing hallucinated references to non-existent priors, 2022.

Mercy Ranjit, Gopinath Ganapathy, Ranjit Manuel, and Tanuja Ganu. Retrieval augmented chest x-ray report generation using openai gpt models, 2023.

Archit Sharma, Sedrick Keh, Eric Mitchell, Chelsea Finn, Kushal Arora, and Thomas Kollar. A critical evaluation of ai feedback for aligning large language models, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Semturs, S Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan. Towards generalist biomedical ai, 2023.

Shawn Xu, Lin Yang, Christopher Kelly, Marcin Sieniek, Timo Kohlberger, Martin Ma, Wei-Hung Weng, Atilla Kiraly, Sahar Kazemzadeh, Zakkai Melamed, Jungyeon Park, Patricia Strachan, Yun Liu, Chuck Lau, Preeti Singh, Christina Chen, Mozziyar Etemadi, Sreenivasa Raju Kalidindi, Yossi Matias, Katherine Chou, Greg S. Corrado, Shravya Shetty, Daniel Tse, Shruthi Prabhakara, Daniel Golden, Rory Pilgrim, Krish Eswaran, and Andrew Sellergren. Elixr: Towards a general purpose x-ray artificial intelligence system through alignment of large language models and radiology vision encoders, 2023.

Benjamin Yan, Ruochen Liu, David Kuo, Subathra Adithan, Eduardo Reis, Stephen Kwak, Vasantha Venugopal, Chloe O'Connell, Agustina Saenz, Pranav Rajpurkar, and Michael Moor. Style-aware radiology report generation with RadGraph and few-shot prompting. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14676–14688, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp. 977. URL https://aclanthology.org/2023.findings-emnlp.977.

Qichen Ye, Junling Liu, Dading Chong, Peilin Zhou, Yining Hua, and Andrew Liu. Qilin-med: Multi-stage knowledge injection advanced medical large language model, 2023.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest x-ray radiology report generation. 2023. URL https://doi.org/10.1016/j.patter.2023.100802.

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

Zijian Zhou, Miaojing Shi, Meng Wei, Oluwatosin Alabi, Zijie Yue, and Tom Vercauteren. Large model driven radiology report generation with clinical quality reinforcement learning, 2024b.

Appendix A. GPT-4 Prompt for Removing References to Prior Exams.

We dynamically created few-shot prompts for each report, depending on which common "prior exam keywords" appeared in it. We divide reports into sentences by splitting on this substring: ". ". All prompts began with the following prefix:

```
Instructions: Return only a json object for this radiology report, with
a key-value pair for every line.
Each line starts with a numerical id. The key will be the id. The value
will be another JSON object.
Inside the value object, set the ``prior_cat'' attribute to say whether
this line makes a comparison to a prior exam.
``prior_cat'' must take one of three possible values: ``none'',
``partial'', ``all''.
1.) If the sentence has some clear information about the current exam,
set ``prior_cat'' as ``partial'' and then add a ``partial_rewrite''
attribute to the value JSON object. For the value of ``partial_rewrite'',
rewrite the sentence to keep only that information, without any
comparison to a prior report.
2.) If there is no comparison, set ``prior_cat'' as ``none''. Do not
rewrite the sentence.
3.) In the rare case that the sentence has absolutely no clear
information about the current exam (e.g. does not mention that any
abnormality is present or absent), set ``prior_cat'' as ``all''. Do not
rewrite the sentence.
Here is an example:
Report: [0] No acute pulmonary process. [1] No significant changes from
last exam.
JSON: {``0'':{``prior_cat'':``none''}, ``1'':{``prior_cat'':``all''}}
Examples of individual lines and their value objects:
Heart is enlarged. -> {``prior_cat'':``none''}
Heart size is stable. -> {``prior_cat'':``all''}
In comparison with the study on ___, heart has enlarged, while lungs
remain clear. -> {``prior_cat'': ``partial'', ``partial_rewrite'':
``Heart is enlarged, while lungs are clear''.}
```

Additionally, we identified 37 other terms that commonly appear in references to prior exams and, in collaboration with radiologists, provided 1-2 examples showing how to handle that keyword in different circumstances. For instance, here are the examples for the keyword "stable":

```
``stable'': ``Heart size and mediastinum are stable -> {``prior_cat'':
``all''}
The cardiomediastinal silhouettes are stable reflective of a tortuous
thoracic aorta. -> {``prior_cat'':``partial'', ``partial_rewrite'':
``The cardiomediastinal silhouettes are reflective of a tortuous
thoracic aorta.''}''
```

For every term that appeared in a report, we appended the corresponding examples to the prompt.

Appendix B. Estimating Number of Lines Referencing Prior Exams.

To estimate how many lines in each report reference a prior exam, we split each report into sentences. Using case-insensitive matching, we then count how many sentences contain one of the following 43 substrings, which commonly appear in references to prior exams:

> 'more', 'regress', 'advanc', 'less', 'fewer', 'constant', 'unchanged', 'prior', 'new', 'stable', 'progressed', 'interval', 'previous', 'further', 'again', 'since', 'increase', 'improve', 'remain', 'worse', 'persist', 'remov', 'similar', 'cleared', 'earlier', 'existing', 'decrease', 'reduc', 'recurren', 'redemonstrat', 'resol', 'still', 'has enlarged', 'lower', 'larger', 'extubated', 'smaller', 'higher', 'continue', 'compar', 'change','develop', 'before'

To validate our automatic metric, we use the results from our expert evaluation, which provides annotations for 165 AI-generated reports (33 studies x 5 generations per study). We find that the automatic metric and the radiologist count find the same number of lines with references to prior exams for 135 out of 165 AI-generated reports. For the remaining 30 reports, our automatic metric tends to overcount the number of references. Compared to radiologists, the automatic metric finds more references to prior exams in 22 reports; across all reports, it finds an average of .76 references to prior exams per report, compared to .63 from radiologists. The mean absolute difference between the automatic metric and radiologist counts is .23.

Appendix C. Effect of GPT-4 Annotation on References to Prior Exams.

When using GPT-4 to remove references to prior exams, we find that GPT-4 reduces but does not fully eliminate these unwanted references. It most commonly applies the label "none", indicating that a line has no reference to a prior. The second most common label is "partial". It rarely uses "all", indicating that most lines have some useful information that can be preserved even when removing references to prior exams.

Dataset	None	Partial	All	Total
Training	58178	37591	11468	107237
Validation	5077	1174	373	6624
Test	7426	3024	998	11448

Table 4: GPT-4 label frequencies for lines in the training, validation and test sets, followed by the total number of lines in each dataset before GPT-4 processing.

To assess GPT-4's performance on this annotation task, we use the automatic metric from Appendix B to estimate the number of lines that refer to prior exams in our datasets, before and after GPT-4 processing. This analysis includes the 19,806 studies in our training set, 915 studies in our validation set, and 1,383 studies in our test set.

We find that lines labeled "partial" or "all" almost always refer to a prior exam, as desired. However, a substantial portion of lines labeled "none" also appear to refer to a

18

prior exam, particularly in the training set. Additionally, lines labeled "partial" sometimes continue to refer to prior exams, even after being rewritten. Based on manual review, we find that, when a line refers to a prior exam multiple times, GPT-4 sometimes removes some but not all of the references. For example, "Similar to findings as compared to prior chest radiograph, perhaps increased opacity along the left lateral mid to lower lung" is rewritten as "Findings show increased opacity along the left lateral mid to lower lung"; the term "increased" still refers to a prior exam. We therefore find that GPT-4 provides noisy annotations, removing over half of all references to prior exams but not eliminating them entirely.

Dataset	None	Partial	All	Total
Training (Original)	10583	36627	11302	58512
Training (Processed)	10583	12586	0	**23169**
Validation (Original)	496	1083	362	1941
Validation (Processed)	496	291	0	**787**
Test (Original)	915	2854	980	4749
Test (Processed)	915	802	0	**1717**

Table 5: The first three columns show the estimated number of lines from each label category that refer to prior exams, before and after GPT-4 processing. The final column shows the total number of lines that refer to prior exams, before and after GPT-4 processing.

Appendix D. Lexical Metrics for Clinical Accuracy.

When measuring lexical similarity using general-purpose, non-clinical metrics, the pretrained model outperforms all DPO methods on both datasets. The difference in BLEU-1 scores is not significant on the processed dataset for two of our DPO models ($\gamma = 0$ or 1). Otherwise, the differences are significant.

We note that all BLEU scores increased when shifting from the original to the processed dataset. This finding can be explained by the fact that our original MIMIC test set refers to prior exams even more frequently than our pretrained model does, so the processed dataset represents a distribution of reports that more closely matches our models. We hypothesize that removing even more references to prior exams from the test set would lead to further increases in the BLEU scores of our DPO models.

Experiment	Bleu-1 (Original)	Bleu-2 (Original)	Bleu-1 (Processed)	Bleu-2 (Processed)
pretrained	0.2394 (0.2283, 0.2509)	0.1745 (0.1656, 0.1842)	0.2532 (0.2417, 0.2650)	0.1816 (0.1726, 0.1915)
sft	0.3058 (0.2967, 0.3152)	0.2195 (0.2119, 0.2277)	0.3305 (0.3215, 0.3399)	0.2353 (0.2274, 0.2434)
k1	0.2033 (0.1944, 0.2121)	0.1439 (0.1369, 0.1510)	0.2388 (0.2293, 0.2482)	0.1685 (0.1610, 0.1761)
k.5	0.1869 (0.1784, 0.1955)	0.1302 (0.1237, 0.1368)	0.2207 (0.2113, 0.2299)	0.1535 (0.1463, 0.1607)
k0	0.1998 (0.1907, 0.2094)	0.1437 (0.1365, 0.1513)	0.2316 (0.2217, 0.2417)	0.1656 (0.1578, 0.1738)

Table 6: Comparison of Fine-Tuning Methods, based on lexical similarity. For both metrics, higher scores indicate better performance.