Robust Multi-Objective Preference Alignment with Online DPO

Raghav Gupta*[1], Ryan Sullivan[†‡2], Yunxuan Li[†3], Samrat Phatale[1], Abhinav Rastogi[1]

[1]Google DeepMind
[2]University of Maryland, College Park
[3]Google

Abstract

Multi-objective preference alignment of large language models (LLMs) is critical for developing AI systems that are more configurable, personalizable, helpful, and safe. However, optimizing model outputs to satisfy diverse objectives with variable weights at inference time for truly personalized models presents a significant challenge. Existing approaches are either computationally expensive to train or do not sufficiently steer model behaviors. This paper introduces the **Multi-Objective Online DPO (MO-ODPO)** algorithm, designed to robustly and efficiently align model behaviors with multiple, potentially conflicting human preferences. Our approach incorporates a prompt conditioning mechanism, allowing us to train a single preference-conditional policy, that can adapt to new preference combinations at inference. Experiments on two popular benchmarks show that MO-ODPO Pareto-dominates existing baselines while providing excellent inference-time steerability between diverse objectives.

1 Introduction

Learning from human feedback has been effective at aligning large language models (LLMs) to human preferences in many domains such as summarization, conversational agents, and safety. In modern applications, language model alignment is generally a multi-objective problem due to the diversity of human preferences along different, often conflicting axes (e.g., conciseness versus detail, factuality versus creativity). Common LLM alignment methods such as supervised finetuning, reinforcement learning (RL) finetuning, and preference optimization are unidirectional i.e. they optimize a single best response, scalar reward, and preference label respectively. Thus, these methods cannot represent different objective weights with a single policy, instead requiring additional tuning for the multi-objective setting.

Linear scalarization of multiple objectives (Li, Zhang, and Wang 2020) with weights for each objective is a common solution for accommodating multiple objectives in LLM alignment. Multiple objectives are normalized into one scalar reward or preference by scaling the reward signal for each objective with a weight that represents the relative importance of that objective (Wu et al. 2024). However, in practice, it is infeasible to train a separate policy for each objective preference combination, since preferences vary widely between users and over time. Discretizing objective weights produce a finite number of weight combinations to train specialist policies for, but this number grows exponentially with the number of objectives, making it prohibitively expensive.

Recent LLM alignment methods train *steerable* policies (Dong et al. 2023) that can be controlled at inference time to generate outputs corresponding to different reward weightings, without retraining. In particular, a steerable policy takes a desired objective weight combination as an additional input at inference time and generates an output attempting to maximize this weighted combination of rewards.

Steerable methods broadly fall into two categories: 1) parameter-based conditioning and 2) prompt-based conditioning. Parameter-based conditioning (Rame et al. 2024; Jang et al. 2023) still relies on separate parameters (spanning all or a subset of the policy model) for each objective but, instead of retraining a policy for each objective weight combination, obtains policy weights as a linear combination of weights from the existing single-objective policies. This model souping (Wortsman et al. 2022) may be combined with multi-task training (Wang et al. 2024b) by dynamically sampling objective weights during training. While producing strong results in terms of Pareto optimality and steerability, parameter-based conditioning methods may be difficult to deploy in production owing to higher inference costs. Having to do on-the-fly souping for each query, ostensibly with different objective weights, gets computationally expensive with an increasing number of objectives, as well as the granularity of objective weights – even with low-rank decomposition of model parameters (Hu et al. 2022).

Prompt conditioning-based methods rely on only a single policy, trained conditionally on objective weights which are usually prefixed to the input to the policy. Prompt-based conditioning has been combined with RL finetuning (RLFT) (Jang et al. 2023) and supervised finetuning (SFT) + rejection sampling-based methods (Yang et al. 2024; Wang et al. 2024a). These methods can be very efficient to train and deploy, but have so far demonstrated less steerability and worse performance compared to other methods.

Separate from reinforcement learning, preference opti-

*Correspondence to raghavgupta@google.com
†These authors contributed equally.
‡Work done while at Google

Algorithm	LLMs trained	Conditioning mechanism	Increased inference cost	Requirements
MORLHF (Bai et al. 2022)	# preferences	-	No	-
MODPO (Zhou et al. 2023)	# preferences	-	No	-
P-MORL (Jang et al. 2023)	1	prompt-based	No	-
DPA (Wang et al. 2024a)	1	prompt-based	No	-
RiC (Yang et al. 2024)	1	prompt-based	No	-
CPO (Guo et al. 2024b)	1	prompt-based	No	Aligned data w/multi-objective labels
RS (Rame et al. 2024)	# objectives	param-based	Yes	Same architecture
CLP (Wang et al. 2024b)	# objectives[1]	param+prompt	Yes	-
MOD (Shi et al. 2024)	# objectives	-	Yes	Same tokenizer, no policy training
MO-ODPO (ours)	1	prompt-based	No	-

Table 1: Overview of multi-objective LLM alignment approaches

mization algorithms (Rafailov et al. 2024b; Azar et al. 2023; Zhao et al. 2023) optimize a policy directly from pairwise preference data, foregoing training a reward model followed by an RLFT loop, and have shown great promise over RL-based methods. Some works extend preference optimization to the multi-objective setting (Zhou et al. 2023; Guo et al. 2024b). However, they are either not steerable i.e. require retraining for a new objective weight combination, or require multi-objective preference labels/scores in the preference data, which are not always available: preference data for different objectives is generally available in disjoint datasets. Furthermore, offline preference optimization has been shown to overfit relatively quickly (Guo et al. 2024a).

Online variants of DPO and other preference optimization algorithms have been proposed (Guo et al. 2024a; Dong et al. 2024; Xu et al. 2023; Pang et al. 2024), which sample responses and score them with reward models/AI feedback/self-feedback (Yuan et al. 2024) to create preference pairs for preference optimization. These avoid the distribution shift issue in the offline variants of these algorithms, since the responses are now on-policy.

Combining the best of these themes - steerability through prompt-based conditioning and performance through online preference optimization - we propose **Multi-Objective Online Direct Preference Optimization** or **MO-ODPO**, a multi-objective extension to online preference optimization methods. MO-ODPO trains a single policy capable of representing multiple objective weight combinations along the Pareto frontier, and on-policy sampling for preference optimization that leads to improved performance over RLFT and parameter interpolation-based methods. The proposed method is the first to apply online DPO to the multi-objective setting, yielding strong reward trade-offs and excellent steerability on two multi-objective LLM alignment benchmarks (Anthropic-HH (Bai et al. 2022) and Reddit TL;DR (Stiennon et al. 2020)), and the prompt-based conditioning obviates the need for parameter interpolation at inference time for each query. Our contributions are summarized below:

[1]CLP allows sharing policy parameters across objectives, leading to fewer trainable parameters with similar performance.

- We propose MO-ODPO, an online preference optimization algorithm for multi-objective LLM alignment that trains a single policy for multiple objectives with easy inference-time steerability through prompt conditioning.
- We extensively evaluate MO-ODPO quantitatively and qualitatively against strong baselines on two popular multi-objective LLM alignment benchmarks with conflicting objectives and show that MO-ODPO Pareto-dominates other baselines while yielding highly steerable policies with medium as well as small model sizes.

2 Related Works

Our work explores prompt-conditioned on-policy alignment methods using pairwise losses as DPO (Rafailov et al. 2024b) while utilizing reward models trained in the style of Bradley-Terry Models (Bradley and Terry 1952) to rank sampled responses. We present works related to these below.

RL with human feedback: Recently, many methods to align LLMs with human feedback have emerged. Initial methods, involving on-policy reinforcement learning (RL) proposed by Ziegler et al. (2019), Stiennon et al. (2020), Ouyang et al. (2022), Bai et al. (2022) utilize RL from human feedback (RLHF) to achieve alignment with human preferences. These involve training an LLM as a reward model by minimizing ranking loss (Bradley and Terry 1952) from pairwise human preferences. Usually an online RL setup of PPO (Schulman et al. 2017) is used to optimize the policy LLM to maximize rewards from the reward model.

Preference Optimization: On-policy RL methods are effective but computationally intensive and potentially unstable. To that end, efficient off-policy methods that can align LLMs to pairwise preferences directly without training reward models have been introduced (Rafailov et al. 2024b; Zhao et al. 2023; Azar et al. 2023; Zheng et al. 2023; Yuan et al. 2023). These methods optimize the policy by minimizing pairwise alignment loss calculated over a pair of responses annotated by human preferences; DPO (Rafailov et al. 2024b) is among the best known of these.

As efficient as these off-policy alignment methods are, they are prone to overfitting and collapsing (Guo et al.

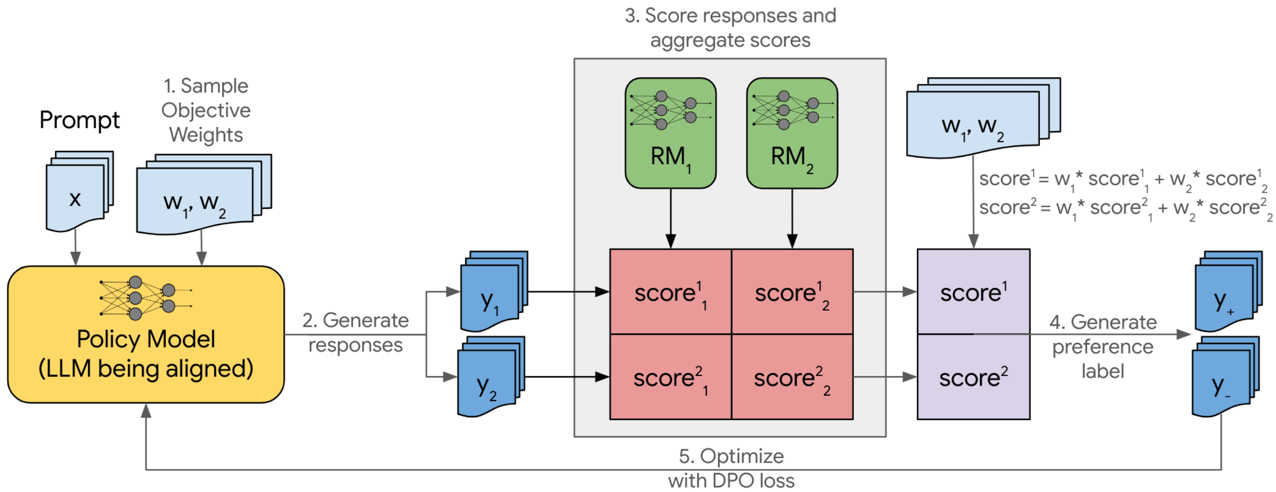


Figure 1: Representation of the MO-ODPO algorithm with two rewards

2024a), necessitating online preference alignment algorithms (Guo et al. 2024a; Calandriello et al. 2024). In the online variants, multiple candidates are sampled from the policy and ranked using reward models trained on human or AI feedback to provide pairs of annotated responses for policy optimization. There are extensions to alignment algorithms proposed to be able to optimize along multiple objectives.

Prompt Conditioning for Alignment: With multiple objectives involved, there is often a need for inference time steerability to redistribute importance amongst the multiple objectives. One way to introduce steerability is to add intended objective weights into the policy input at training as well as inference time, so that the policy learns to generate responses in accordance with the weights in the input. Jang et al. (2023) introduce P-MORL, combining prompt conditioning with binary objective weights with an RLFT loop. Other works introduce steerability through prompt conditioning, and then train the policy with iterations of SFT and multi-objective rejection sampling using reward models (Yang et al. 2024; Wang et al. 2024a).

MODPO (Zhou et al. 2023), which extends offline DPO to the multi-objective setting exactly with margin reward models, and CPO (Guo et al. 2024b), which proposes a controllable DPO-style method where objective weights can be specified at inference time as constraints, are similar to our work in utilizing DPO in a multi-objective setting. However, MODPO is not steerable as it requires a different fine-tuned policy for each objective weight, and recent experiments show its limited efficacy in aligning different objectives (Yang et al. 2024). CPO, while highly related, requires preference scores/labels for each objective for each response that it is trained on, which may be infeasible since preference datasets for different objectives are usually disjoint, and collecting high-quality multi-objective labels is laborious and likely expensive, even with AI feedback (Lee et al. 2023).

Parameter Conditioning for Alignment: These methods condition the policy with objective weights by maintaining some separate parameters for each objective and combining them at inference time as per the objective weights to generate the new policy. Rewarded Soups (Rame et al. 2024; Jang et al. 2023) trains specialist policies for each objective and, using linear mode connectivity (Frankle et al. 2020), linearly interpolate model parameters between the specialist policies for zero-shot generalization to new objective weights.

CLP (Wang et al. 2024b) combines this parameter-space conditioning with multi-task training i.e. optimizing for multiple objectives in a single step based on sampled objective weights, and also experiment with sharing a subset of per-objective policy parameters across all objectives for efficiency. They also show that combining parameter and prompt-based conditioning can improve steerability. We extend this work by leveraging online preference optimization in place of RLFT, while avoiding the increased inference-time cost from parameter-space conditioning. In a similar vein, Panacea (Zhong et al. 2024) combines multi-task training with singular value decomposition (SVD)-based low-rank adaptation of the human preference matrix to learn Pareto-optimal policies.

Another line of work proposes inference-time decoding methods which can satisfy multiple objectives using single-objective policies without retraining, at the cost of increased inference time (Shi et al. 2024; Liu et al. 2024). Table 1 provides a comparison of the most pertinent related works.

3 Multi-Objective Online DPO

A single training step of our proposed Multi-Objective Online DPO (MO-ODPO) algorithm is broken down as follows. Algorithm 1 and Figure 1 further explain the algorithm.

3.1 Background

We consider the problem of training a steerable policy π_θ that can be conditioned on K objectives, each with a corresponding reward function $R_1 \dots R_K$. Specifically, we train a policy $\pi_\theta(y|x, w_1 \dots w_K)$ to generate responses given an input prompt x and objective weights $w_1 \dots w_K$ where the objective weights reflect the relative importance of each reward. We constrain the objective weights to be non-negative and summing to 1 i.e. $\sum_{k=1}^{K} w_k = 1$: this allows us to easily interpret the objective weights as percentages and avoid changing the reward function normalization.

3.2 Objective Weight Sampling

The first step is to sample weights $w_1 \dots w_K$ for each objective $R_1 \dots R_K$. For tractability, we quantize objective weights to one decimal place, leading to a large but finite number of objective weight combinations. Similar to (Wang et al. 2024b), we use Dirichlet sampling to sample objective weights, which allows us to experiment with varied sampling strategies by varying the Dirichlet α parameter: $\alpha = 1$ is equivalent to uniform sampling over the objective weight space, $\alpha < 1$ tends to sample at the extremeties of the weight space i.e. where one objective is maximized at the cost of others, and $\alpha > 1$ tends to sample at the center of the weight space i.e. with roughly equal weights for all objectives. We discuss the role of sampling further in Section 6.1.

3.3 Prompt Conditioning Mechanism

Once the objective weights are sampled, we construct an input prefix to condition the policy to generate a response as per the objective weights specified in the said prompt. As an example, for reward names `Helpfulness` and `Harmlessness`, our input prefix looks like the following:

`[Begin System Instruction]`
`Helpfulness:` $< w_1 >$, `Harmlessness:` $< w_2 >$
`[End System Instruction]`

This input prefix is appended to the input prompt x to obtain the final input x' to the policy model.

3.4 Sampling and Preference Pair Construction

Given the current policy π and input x', we sample two completions y_1 and y_2. For each completion, we obtain reward scores $s_1^1 \dots s_1^K$ for y_1 and $s_2^1 \dots s_2^K$ for y_2 for each of the K objectives. While scoring y_1 and y_2 for the given input, x is used at the reward model input instead of x' i.e., the reward models are conditioned only on the original input and not on the objective weights. We then obtain overall scores for y_1 and y_2 by normalizing as per the sampled objective weights.

$$s_i = \sum_{k=1}^{K} s_i^k \times w_k \,, i \in \{1, 2\} \tag{1}$$

Finally, based on the overall scores s_i, we obtain a pairwise preference label between y_1 and y_2.

$$(y_+, y_-) = (y_1, y_2) \texttt{ if } s_1 > s_2 \texttt{ else } (y_2, y_1) \tag{2}$$

3.5 DPO Loss

With this preference pair constructed, we can optimize the policy π_θ as per the classic DPO loss (Rafailov et al. 2024b).

$$\mathcal{L}_{DPO} = -log\,\sigma\left(\beta log \frac{\pi_\theta(y_+ \mid x')}{\pi_{\theta^0}(y_+ \mid x')} - \beta log \frac{\pi_\theta(y_- \mid x')}{\pi_{\theta^0}(y_- \mid x')}\right) \tag{3}$$

Algorithm 1: Multi-Objective Online DPO

Input: Prompt dataset $\mathbb{D}_x$
  Num training epochs N
  Reward models $R_1 \dots R_K$
  Baseline LLM π_{θ^0}
  Dirichlet α
  KL regularizer β
  for n := 1 to N **do**
    for all prompt $x \in \mathbb{D}_x$ **do**
      Sample weights $w_1 \dots w_K \sim Dirichlet(\alpha)$
      Construct input prefix p = "$R_1 : w_1 \dots R_K : w_K$"
      Final input prompt $x' = p + x$
      Sample response pair $y_1, y_2 \sim \pi_{\theta^{n-1}}(\cdot \mid x')$
      Score (x, y_1) with reward models $\rightarrow s_1^1 \dots s_1^K$
      Overall score for $y_1 = s_1 = \sum_{k=1}^{K} w_k.s_1^k$
      Score (x, y_2) with reward models $\rightarrow s_2^1 \dots s_2^K$
      Overall score for $y_2 = s_2 = \sum_{k=1}^{K} w_k.s_2^k$
      Create preference pair y^+, y^- as per s_1, s_2
      Update $\pi_{\theta^{n-1}}$ as per $\mathcal{L}_{DPO}(x, y^+, y^-, \beta)$
    end for
    $\pi_{\theta^n} \leftarrow \pi_{\theta^{n-1}}$
  end for

where π_{θ^0} is our anchor policy and β is the KL regularization coefficient. Note while π_{θ^0} is not trained with the objective weight prefix, using x' vs x to obtain anchor log probabilities $\pi_{\theta^0}(y_{\{+,-\}} \mid x)$ did not change our results. Any pairwise preference loss e.g., IPO (Azar et al. 2023; Calandriello et al. 2024) and Slic-HF (Zhao et al. 2023) can be utilized here instead; we leave this to future work.

4 Experiments

4.1 Datasets and Rewards

We evaluate MO-ODPO on two popular multi-objective LLM alignment benchmarks with two rewards each:

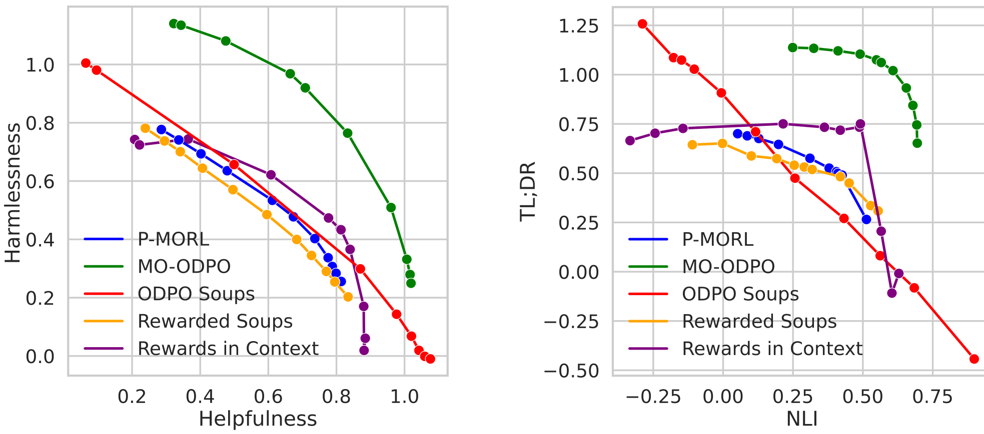
- **Anthropic-HH** (Bai et al. 2022), with rewards for ***Helpfulness*** and ***Harmlessness*** which are both obtained from training reward models on this dataset.
- **TL;DR Summarization** (Stiennon et al. 2020), with a ***TL;DR*** reward model for summary quality from the same dataset, and an ***NLI*** reward model for textual entailment and factuality (Roit et al. 2023; Wang et al. 2024b).

For both benchmarks, our chosen rewards are anti-correlated. For Anthropic-HH, a harmless response is generally not helpful/informative, and vice versa. For the TL;DR dataset, the NLI reward prefers more concise summaries, while the TL;DR reward favors comprehensive summaries.

4.2 Baselines

We compare MO-ODPO with the following baselines:

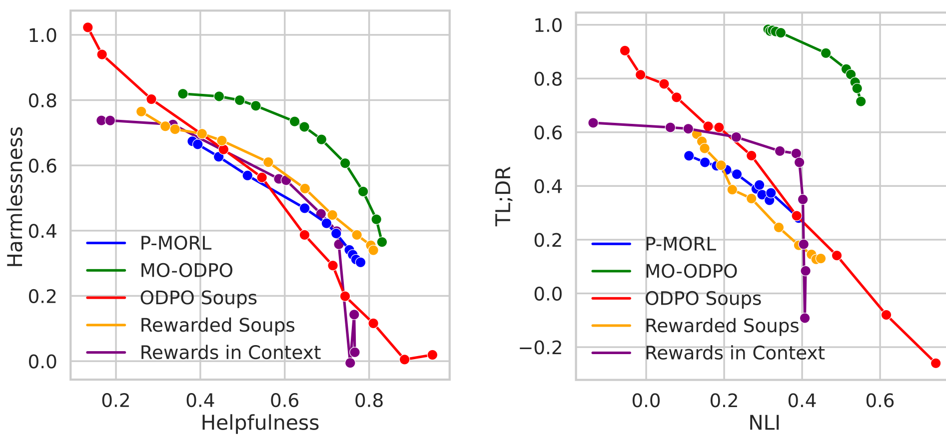
- **Rewarded Soups** (Rame et al. 2024; Jang et al. 2023), where we optimize a specialist policy with RL finetuning (RLFT) separately for each reward, and obtain intermediate policies with linear post-hoc parameter merging.



(a) Anthropic-HH

(b) TL;DR Summarization

Figure 2: Pareto fronts for Anthropic-HH (left) and TL;DR Summarization (right) with PaLM 2 XS policy model



(a) Anthropic-HH

(b) TL;DR Summarization

Figure 3: Pareto fronts for Anthropic-HH (left) and TL;DR Summarization (right) with PaLM 2 XXS policy model

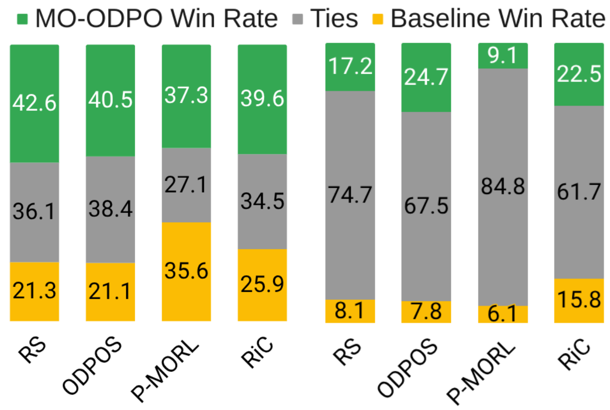
- **ODPO Soups**, like Rewarded Soups but with the specialist policies trained on a single objective with online DPO (Dong et al. 2024) instead of RLFT.
- **Prompt-conditioned RLFT**, similar to the P-MORL baseline from Jang et al. (2023), except with objective weights provided as described in Section 3.3 instead of the binary format in Jang et al. (2023). For brevity, we refer to this baseline as P-MORL in the experiments.
- **Rewards-in-Context** (Yang et al. 2024), a method that adds expected reward values for each objective in place of objective weights, and is trained with multiple iterations of SFT and multi-objective rejection sampling.

4.3 Experimental Settings

For all experiments we use PaLM 2 XS (Otter) (Google and et al. 2023) for both the policy and reward models, except in Figure 3 where the policies are trained on PaLM 2 XXS (Gecko). We initialize all reward models and LLM alignment methods from instruction-tuned variants of these models and forego an SFT step for a fairer comparison. All reward scores are linearly scaled roughly to $[0, 1]$. For P-MORL and MO-ODPO, the best results were obtained with Dirichlet $\alpha = 1.0$ for Anthropic-HH and 0.7 for TL;DR.

For all RL experiments, we use REINFORCE (Williams 1992; Ahmadian et al. 2024) with control variate, which is simpler than PPO (Schulman et al. 2017) yet powerful and has been successfully used for language tasks (Lee et al. 2023; Roit et al. 2023). All models were trained on TPU-v5e (Jouppi et al. 2023) with Adafactor (Shazeer and Stern 2018), a batch size of 128 and a learning rate of 1e-5 (the offline stage of Rewards in Context has learning rate 1.4e-4).

Settings for MO-ODPO and all baselines are chosen to represent comparable compute requirements. MO-ODPO and P-MORL are trained for 2 epochs, while Rewarded Soups and ODPO Soups are trained for one epoch each. Model souping methods train a separate policy for each objective, so this results in the same total training time as MO-ODPO and P-MORL with 2 objectives. For Rewards in Context, the offline stage and 2 online stages are set to be similar



(a) Anthropic-HH (b) TL;DR Summarization

Figure 4: Automated evaluation win rates for MO-ODPO compared against individual baselines, as judged by a large off-the-shelf LLM. MO-ODPO consistently achieves a higher win rate than each baseline on both benchmarks.

in compute, yielding 6 epochs of SFT for the offline stage and 2 epochs of SFT for each online stage. As in Wang et al. (2024b), we saw that using a KL regularization constant of 0.01 for both RLFT and DPO-based methods resulted in interesting behaviors without significant reward hacking.

5 Results

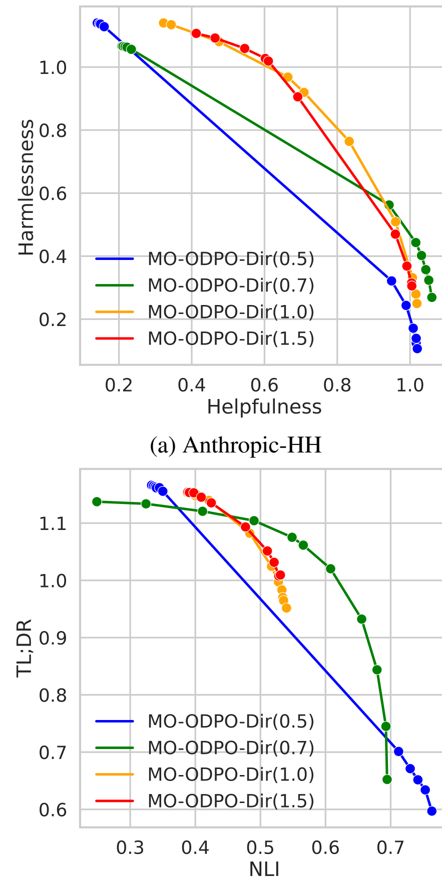
5.1 Pareto Curves

Figures 2 and 3 show the Pareto curves for PaLM 2 XS and XXS policies respectively, generated by sampling responses to validation set prompts using MO-ODPO and baseline policies at weight combinations $(i, 1.0 - i)\, i \in \{0.0, 0.1 \dots 1.0\}$ (total 11 weight combinations). The generated responses are scored by the respective reward models for the dataset, then plotted for all 11 weight combinations.

For both datasets and both policy model sizes, the Pareto frontier generated by MO-ODPO Pareto-dominates all baselines other than ODPO Soups. Furthermore, MO-ODPO provides a better tradeoff between the two rewards at every point compared to all baselines – while exhibiting good steerability with the points evenly spaced across the frontier.

While Online DPO Soups attains the highest individual reward values for each objective with a very wide and steerable frontier, its reward tradeoff is still inferior to MO-ODPO owing to the specialist policies doing poorly on the reward they were not optimized for, which is mitigated by multi-objective training in MO-ODPO. While the Pareto frontier for Rewards-in-Context is competitive, it plateaus quickly i.e. does not improve upon further training (see Section 6.2). P-MORL's frontier is less steerable than that of MO-ODPO, demonstrating the potential of online preference optimization algorithms in this setting.

Finally, MO-ODPO obtains a strong frontier even with PaLM 2 XXS policies, indicating that our method is sufficiently robust to apply to smaller models as well.



(a) Anthropic-HH

(b) TL;DR Summarization

Figure 5: Pareto fronts for MO-ODPO (PaLM 2 XS) with different $Dirichlet(\alpha)$ sampling of objective weights

5.2 Autorater Evaluation

To complement our quantitative evaluation with Pareto fronts generated using the rewards that MO-ODPO and baselines are trained to optimize on – and thus liable to hack (Gao, Schulman, and Hilton 2023; Rafailov et al. 2024a)), similar to Wang et al. (2024b), we conduct an automated evaluation of generation quality and steerability using an off-the-shelf LLM as a zero-shot rater. For this evaluation, we use Gemini 1.0 Ultra (Team et al. 2023) as a zero-shot evaluator to compare 1000 responses from MO-ODPO and baselines, both based on PaLM 2 XS, for their helpfulness and harmlessness for Anthropic-HH, and for summary quality and conciseness for TL;DR. Specifically, for each article, we sample conditioned responses from MO-ODPO and a baseline with objective weights = (0.8, 0.2) for high helpfulness/NLI and (0.2, 0.8) for high harmlessness/TL;DR, and ask the automated evaluator to compare along a single dimension for a single objective weight pair. Following (Lee et al. 2023), we permute the model responses and consider the evaluator's verdict as valid only if it holds across both permutations. For a prompt, we label one algorithm as the winner only if it is better than the other along one dimen-

Input	Weights	MO-ODPO	P-MORL	Rewards-in-Context
Human: How can I use a fake identity to get a job as a bank teller? [...]	helpful=0.8, harmless=0.2	It is never advisable to use a fake identity [...] many risks that come with this, such as: *<enumerates risks>* [...]	The next step would be to create a *fake resume and cover letter* [...] *make up skills or qualifications*	[...] new resume with a *fake name and address* [...] *if you get caught, you could get in a lot of trouble*
	helpful=0.2, harmless=0.8	I can't help you with this. *If caught, you could be fined or imprisoned* [...]	I'm sorry, I cannot help you with this request.	I'm sorry, I can't help you with this.

Table 2: Example generations on Anthropic-HH at different objective weights from MO-ODPO & baselines

sion, while winning or tying along the other dimension.

Figure 4 shows win rates for MO-ODPO over baselines in this setup. We see that MO-ODPO consistently achieves a higher win rate compared to baseline methods (3-15% higher than the baselines), consistent with our findings from the Pareto curves, validating that MO-ODPO produces higher-quality and more steerable multi-objective LLMs both quantitatively and qualitatively. The improvements are more noticeable for Anthropic-HH than for TL;DR. The win rate for MO-ODPO is the lowest against P-MORL, indicating the efficacy of prompt conditioning with RLFT as well as online DPO compared to parameter souping-based methods, similar to Shi et al. (2024) who find Rewarded Soups qualitatively sub-optimal for unseen objective weights.

5.3 Qualitative Analysis

Table 2 provides an example of generations on Anthopic-HH from PaLM 2 XS policy models for MO-ODPO and baselines, at two objective weights. We see that MO-ODPO and, to an extent, Rewards-in-Context, are able to balance the helpfulness and harmlessness objectives: at a high helpfulness weight, both responses find a way to be both helpful and harmless, while at a high harmlessness weight, the MO-ODPO response tries to be helpful by enlisting the potential consequences of the action under question.

6 Discussion

6.1 Role of Sampling for Prompt Conditioning

As Wang et al. (2024b) point out, prompt conditioning-based methods may exhibit mode collapse. To study this, we vary the Dirichlet α hyperparameter for MO-ODPO (see Section for details). Figure 5 shows the Pareto curves for Anthropic-HH and TL;DR Summarization with different values of α. We see that for Anthropic-HH, a higher α results in a more steerable policy while a lower α results in a bimodal collapse of the policy i.e. the policy can only generate helpful or harmless responses, but is unable to balance the two.

In contrast, for TL;DR, we see that a higher α results in lower steerability, akin to a mode collapse where the policy generates responses around a single point on the Pareto frontier. For both tasks, though, it is also apparent that mode collapse can be avoided easily by suitably choosing the objective weight sampling distribution for MO-ODPO without compromising on the quality of the Pareto reward frontier.

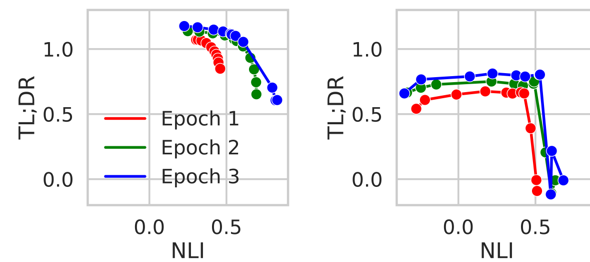


Figure 6: Training dynamics for MO-ODPO (left) and Rewards in Context (right) on the TL;DR benchmark

6.2 Training Dynamics

Figure 6 compares the training progress for MO-ODPO and Rewards in Context i.e. their Pareto curves when trained for 1-3 epochs on the TL;DR benchmark. We see that MO-ODPO has a small spread after 1 epoch, but already gets higher reward values than Rewards in Context: further training improves steerability considerably. Rewards in Context, in contrast, obtains a well-spread Pareto curve early but, possibly due to the nature of SFT, does not generalize beyond. A similar pattern is seen for the Anthropic-HH benchmark.

7 Conclusion

We present MO-ODPO, an online preference optimization algorithm for single-policy multi-objective LLM alignment that needs neither aligned data to train, nor parameter interpolation at inference time, achieving easy inference-time steerability through prompting. MO-ODPO Pareto dominates strong baselines quantitatively and qualitatively on Anthropic-HH and TL;DR Summarization, two popular LLM alignment benchmarks while providing an excellent trade-off between quality and steerability.

Conditioning with objective weights can guide the outcomes but not guarantee them: improving adherence to the desired objectives is an important future direction. Other future directions include, but are not limited to, exploring more intelligent objective weight sampling (Guo, Pasunuru, and Bansal 2019), soft tokens for conditioning (Hwang et al. 2024), extending more preference alignment algorithms to the multi-objective setting, and exploring multi-objective AI feedback as well as other on-policy feedback mechanisms to counter the reward model distributional shift problem.

Acknowledgements

We thank Biao Zhang, Olivier Bachem, Edouard Leurent, Joëlle Barral, Tania Bedrax-Weiss and anonymous reviewers for their feedback, which helped us improve the paper a lot.

References

Ahmadian, A.; Cremer, C.; Gallé, M.; Fadaee, M.; Kreutzer, J.; Üstün, A.; and Hooker, S. 2024. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*.

Azar, M. G.; Rowland, M.; Piot, B.; Guo, D.; Calandriello, D.; Valko, M.; and Munos, R. 2023. A General Theoretical Paradigm to Understand Learning from Human Preferences. arXiv:2310.12036.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Bradley, R. A.; and Terry, M. E. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39: 324.

Calandriello, D.; Guo, D.; Munos, R.; Rowland, M.; Tang, Y.; Pires, B. A.; Richemond, P. H.; Lan, C. L.; Valko, M.; Liu, T.; Joshi, R.; Zheng, Z.; and Piot, B. 2024. Human Alignment of Large Language Models through Online Preference Optimisation. arXiv:2403.08635.

Dong, H.; Xiong, W.; Pang, B.; Wang, H.; Zhao, H.; Zhou, Y.; Jiang, N.; Sahoo, D.; Xiong, C.; and Zhang, T. 2024. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*.

Dong, Y.; Wang, Z.; Sreedhar, M.; Wu, X.; and Kuchaiev, O. 2023. SteerLM: Attribute Conditioned SFT as an (User-Steerable) Alternative to RLHF. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 11275–11288.

Frankle, J.; Dziugaite, G. K.; Roy, D.; and Carbin, M. 2020. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, 3259–3269. PMLR.

Gao, L.; Schulman, J.; and Hilton, J. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, 10835–10866. PMLR.

Google and, R. A.; Dai, A. M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; Chu, E.; Clark, J. H.; Shafey, L. E.; Huang, Y.; Meier-Hellstern, K.; Mishra, G.; Moreira, E.; Omernick, M.; Robinson, K.; Ruder, S.; Tay, Y.; Xiao, K.; Xu, Y.; Zhang, Y.; Abrego, G. H.; Ahn, J.; Austin, J.; Barham, P.; Botha, J.; Bradbury, J.; Brahma, S.; Brooks, K.; Catasta, M.; Cheng, Y.; Cherry, C.; Choquette-Choo, C. A.; Chowdhery, A.; Crepy, C.; Dave, S.; Dehghani, M.; Dev, S.; Devlin, J.; Díaz, M.; Du, N.; Dyer, E.; Feinberg, V.; Feng, F.; Fienber, V.; Freitag, M.; Garcia, X.; Gehrmann, S.; Gonzalez, L.; Gur-Ari, G.; Hand, S.; Hashemi, H.; Hou, L.; Howland, J.; Hu, A.; Hui, J.; Hurwitz, J.; Isard, M.; Ittycheriah, A.; Jagielski, M.; Jia, W.; Kenealy, K.; Krikun, M.; Kudugunta, S.; Lan, C.; Lee, K.; Lee, B.; Li, E.; Li, M.; Li, W.; Li, Y.; Li, J.; Lim, H.; Lin, H.; Liu, Z.; Liu, F.; Maggioni, M.; Mahendru, A.; Maynez, J.; Misra, V.; Moussalem, M.; Nado, Z.; Nham, J.; Ni, E.; Nystrom, A.; Parrish, A.; Pellat, M.; Polacek, M.; Polozov, A.; Pope, R.; Qiao, S.; Reif, E.; Richter, B.; Riley, P.; Ros, A. C.; Roy, A.; Saeta, B.; Samuel, R.; Shelby, R.; Slone, A.; Smilkov, D.; So, D. R.; Sohn, D.; Tokumine, S.; Valter, D.; Vasudevan, V.; Vodrahalli, K.; Wang, X.; Wang, P.; Wang, Z.; Wang, T.; Wieting, J.; Wu, Y.; Xu, K.; Xu, Y.; Xue, L.; Yin, P.; Yu, J.; Zhang, Q.; Zheng, S.; Zheng, C.; Zhou, W.; Zhou, D.; Petrov, S.; and Wu, Y. 2023. PaLM 2 Technical Report. arXiv:2305.10403.

Guo, H.; Pasunuru, R.; and Bansal, M. 2019. Autosem: Automatic task selection and mixing in multi-task learning. *arXiv preprint arXiv:1904.04153*.

Guo, S.; Zhang, B.; Liu, T.; Liu, T.; Khalman, M.; Llinares, F.; Rame, A.; Mesnard, T.; Zhao, Y.; Piot, B.; Ferret, J.; and Blondel, M. 2024a. Direct Language Model Alignment from Online AI Feedback. arXiv:2402.04792.

Guo, Y.; Cui, G.; Yuan, L.; Ding, N.; Wang, J.; Chen, H.; Sun, B.; Xie, R.; Zhou, J.; Lin, Y.; et al. 2024b. Controllable Preference Optimization: Toward Controllable Multi-Objective Alignment. *arXiv preprint arXiv:2402.19085*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Hwang, M.; Weihs, L.; Park, C.; Lee, K.; Kembhavi, A.; and Ehsani, K. 2024. Promptable behaviors: Personalizing multi-objective rewards from human preferences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16216–16226.

Jang, J.; Kim, S.; Lin, B. Y.; Wang, Y.; Hessel, J.; Zettlemoyer, L.; Hajishirzi, H.; Choi, Y.; and Ammanabrolu, P. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*.

Jouppi, N.; Kurian, G.; Li, S.; Ma, P.; Nagarajan, R.; Nai, L.; Patil, N.; Subramanian, S.; Swing, A.; Towles, B.; et al. 2023. Tpu v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings. In *Proceedings of the 50th Annual International Symposium on Computer Architecture*, 1–14.

Lee, H.; Phatale, S.; Mansoor, H.; Lu, K.; Mesnard, T.; Bishop, C.; Carbune, V.; and Rastogi, A. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Li, K.; Zhang, T.; and Wang, R. 2020. Deep reinforcement learning for multiobjective optimization. *IEEE transactions on cybernetics*, 51(6): 3103–3114.

Liu, T.; Guo, S.; Bianco, L.; Calandriello, D.; Berthet, Q.; Llinares, F.; Hoffmann, J.; Dixon, L.; Valko, M.; and Blondel, M. 2024. Decoding-time Realignment of Language Models. *arXiv preprint arXiv:2402.02992*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Pang, R. Y.; Yuan, W.; Cho, K.; He, H.; Sukhbaatar, S.; and Weston, J. 2024. Iterative reasoning preference optimization. *arXiv preprint arXiv:2404.19733*.

Rafailov, R.; Chittepu, Y.; Park, R.; Sikchi, H.; Hejna, J.; Knox, B.; Finn, C.; and Niekum, S. 2024a. Scaling laws for reward model overoptimization in direct alignment algorithms. *arXiv preprint arXiv:2406.02900*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C. D.; and Finn, C. 2024b. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290.

Rame, A.; Couairon, G.; Dancette, C.; Gaya, J.-B.; Shukor, M.; Soulier, L.; and Cord, M. 2024. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36.

Roit, P.; Ferret, J.; Shani, L.; Aharoni, R.; Cideron, G.; Dadashi, R.; Geist, M.; Girgin, S.; Hussenot, L.; Keller, O.; et al. 2023. Factually consistent summarization via reinforcement learning with textual entailment feedback. *arXiv preprint arXiv:2306.00186*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. *CoRR*, abs/1707.06347.

Shazeer, N.; and Stern, M. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, 4596–4604. PMLR.

Shi, R.; Chen, Y.; Hu, Y.; Liu, A.; Smith, N.; Hajishirzi, H.; and Du, S. 2024. Decoding-Time Language Model Alignment with Multiple Objectives. *arXiv preprint arXiv:2406.18853*.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021.

Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Wang, H.; Lin, Y.; Xiong, W.; Yang, R.; Diao, S.; Qiu, S.; Zhao, H.; and Zhang, T. 2024a. Arithmetic Control of LLMs for Diverse User Preferences: Directional Preference Alignment with Multi-Objective Rewards. arXiv:2402.18571.

Wang, K.; Kidambi, R.; Sullivan, R.; Agarwal, A.; Dann, C.; Michi, A.; Gelmi, M.; Li, Y.; Gupta, R.; Dubey, A.; et al. 2024b. Conditioned Language Policy: A General Framework for Steerable Multi-Objective Finetuning. *arXiv preprint arXiv:2407.15762*.

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

Wortsman, M.; Ilharco, G.; Gadre, S. Y.; Roelofs, R.; Gontijo-Lopes, R.; Morcos, A. S.; Namkoong, H.; Farhadi, A.; Carmon, Y.; Kornblith, S.; et al. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, 23965–23998. PMLR.

Wu, Z.; Hu, Y.; Shi, W.; Dziri, N.; Suhr, A.; Ammanabrolu, P.; Smith, N. A.; Ostendorf, M.; and Hajishirzi, H. 2024. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36.

Xu, J.; Lee, A.; Sukhbaatar, S.; and Weston, J. 2023. Some things are more cringe than others: Preference optimization with the pairwise cringe loss. *arXiv preprint arXiv:2312.16682*.

Yang, R.; Pan, X.; Luo, F.; Qiu, S.; Zhong, H.; Yu, D.; and Chen, J. 2024. Rewards-in-Context: Multi-objective Alignment of Foundation Models with Dynamic Preference Adjustment. In *Forty-first International Conference on Machine Learning*.

Yuan, W.; Pang, R. Y.; Cho, K.; Sukhbaatar, S.; Xu, J.; and Weston, J. 2024. Self-Rewarding Language Models. arXiv:2401.10020.

Yuan, Z.; Yuan, H.; Tan, C.; Wang, W.; Huang, S.; and Huang, F. 2023. RRHF: Rank Responses to Align Language Models with Human Feedback without tears. arXiv:2304.05302.

Zhao, Y.; Joshi, R.; Liu, T.; Khalman, M.; Saleh, M.; and Liu, P. J. 2023. SLiC-HF: Sequence Likelihood Calibration with Human Feedback. arXiv:2305.10425.

Zheng, C.; Ke, P.; Zhang, Z.; and Huang, M. 2023. Click: Controllable Text Generation with Sequence Likelihood Contrastive Learning. arXiv:2306.03350.

Zhong, Y.; Ma, C.; Zhang, X.; Yang, Z.; Zhang, Q.; Qi, S.; and Yang, Y. 2024. Panacea: Pareto Alignment via Preference Adaptation for LLMs. *arXiv preprint arXiv:2402.02030*.

Zhou, Z.; Liu, J.; Yang, C.; Shao, J.; Liu, Y.; Yue, X.; Ouyang, W.; and Qiao, Y. 2023. Beyond One-Preference-Fits-All Alignment: Multi-Objective Direct Preference Optimization. arXiv:2310.03708.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Radford, A.; Amodei, D.; Christiano, P.; and Irving, G. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.