Sequential Preference Optimization: Multi-Dimensional Preference Alignment with Implicit Reward Modeling

Xingzhou Lou[1,2], Junge Zhang[1,2*], Jian Xie[3], Lifeng Liu[3], Dong Yan[3], Kaiqi Huang[1,2]
[1]Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Baichuan Inc.
louxingzhou2020@ia.ac.cn, jgzhang@nlpr.ia.ac.cn, xiejian1990@gmail.com, sproblvem@gmail.com,
liulifeng@baichuan-inc.com, kqhuang@nlpr.ia.ac.cn

Abstract

Human preference alignment is critical in building powerful and reliable large language models (LLMs). However, current methods either ignore the multi-dimensionality of human preferences (e.g. helpfulness and harmlessness) or struggle with the complexity of managing multiple reward models. To address these issues, we propose Sequential Preference Optimization (SPO), a method that sequentially fine-tunes LLMs to align with multiple dimensions of human preferences. SPO avoids explicit reward modeling, directly optimizing the models to align with nuanced human preferences. We theoretically derive closed-form optimal SPO policy and loss function. Gradient analysis is conducted to show how SPO manages to fine-tune the LLMs while maintaining alignment on previously optimized dimensions. Empirical results on LLMs of different size and multiple evaluation datasets demonstrate that SPO successfully aligns LLMs across multiple dimensions of human preferences and significantly outperforms the baselines.

Introduction

Pretrained large language models (LLM) like GPT-4 (OpenAI 2023) and Llama (Touvron et al. 2023a,b; Dubey et al. 2024) are trained on very large corpus of text and demonstrate surprising capabilities in multiple domains, such as natural language processing (Jiao et al. 2023; Singhal et al. 2023), programming (Nijkamp et al. 2022; Qian et al. 2023) and decision making (Wang et al. 2023a; Zhang et al. 2023). These models are fine-tuned with humans' feedback to align with certain human preferences, e.g. harmlessness and helpfulness. Human preference alignment improves LLM's ability to generate responses preferred by humans and is essential in building AI assistants (OpenAI 2023; Touvron et al. 2023b; Anthropic 2023; Jiang et al. 2024; Dubey et al. 2024). Specifically, Reinforcement Learning with Human Feedback (RLHF) (Ouyang et al. 2022) learns a reward model to discriminate preferred and less preferred responses, and then optimizes LLMs with the reward model and RL algorithms. Direct Preference Optimization (DPO) (Rafailov et al. 2023) omits fitting an explicit reward model and directly optimizes LLMs to adhere to human preferences, and thus is known as implicit reward modeling.

Prevalent preference alignment methods focus on fine-tuning LLMs based on ranked response pairs, which only indicate which response is generally better (Zheng et al. 2023; Chiang et al. 2024). However, instead of solely *good* or *bad*, texts usually have multi-dimensional properties. For instance, a concise text summary generated by LLMs may not be as informative as a relatively longer, but highly specific response. In this case, the concise response is preferred brevity-wise, while the specific response is preferred informativity-wise. In other words, preferences on different dimensions may contradict each other.

The most straightforward approach to deal with multi-dimensional preferences is to mix them into one single dimension to indicate which response is generally better. In this case, the alignment results could be significantly influenced by the annotators' subjective perception and ranking inconsistency across dimensions. Therefore, it is necessary to align LLMs on each dimension and strive a balance that accommodates preferences across all dimensions. Current methods (Jang et al. 2023; Dai et al. 2023) decouple preferences along dimensions and align LLMs on each of the dimension by RLHF. However, they demand a reward model for each dimension. Fine-tuning LLMs with one reward model is already notoriously challenging. Multiple reward models further exaggerate this issue.

To address these issues, we propose Sequential Preference Optimization (SPO) to align LLMs with multi-dimensional preferences in a sequential manner. Specifically, SPO incorporates multi-round fine-tuning, optimizing one specific preference dimension for each round. SPO adopts additional constraints to guarantee alignment on previous dimensions in the learning objective. Consequently, LLMs acquire the skill to align with one specific aspect of human preference in each round, while staying aligned with preferences in previous rounds. Also, SPO omits explicit reward modeling and directly optimizes preferences, thereby avoiding the issues of multiple reward models in RLHF-based methods.

Theoretically, closed-form optimal policy and loss function for SPO are derived. The loss function is a simple classification loss and can be optimized efficiently. Furthermore, we perform gradient analysis to illustrate how SPO effectively

*Correspondence

preserves alignment results of previous rounds of fine-tuning.

Empirically, we conduct experiments on the PKU-SafeRLHF-30k dataset (Ji et al. 2023), where response pairs are separately ranked on the dimensions of helpfulness and harmlessness. Our experiments use Llama 2 7B and 13B (Touvron et al. 2023b) as base models. Fine-tuned models are evaluated on multiple datasets (Li et al. 2023a; Bai et al. 2022; Ji et al. 2023). We also include experiments with more preference dimensions. Results of these experiments suggest that SPO successfully aligns LLMs across multiple dimensions and outperform both baseline methods and open models. Main contributions of this paper are:

(1) We propose Sequential Preference Optimization (SPO), which is able to sequentially align LLMs on multi-dimensional preferences.

(2) We theoretically derive the learning objective of SPO, ensuring multi-dimensional preference alignment. Our gradient analysis elucidates the mechanism by which SPO accomplishes this objective.

(3) Empirical results with multiple training and evaluation datasets demonstrate that SPO successfully aligns LLMs with multi-dimensional human preferences.

Preliminaries

Supervised fine-tuned (SFT) model π_0, based on pretrained models and high-quality demonstrations, is the initial model of SPO, RLHF and other preference optimization methods.

For a response pair (y_1, y_2) of prompt x, $y_1 \succ y_2$ represents that y_1 is the preferred response by humans. The preferences are decided by some unknown latent reward function $r^*(x, y)$. The prevalent way to model human preference distribution is the Bradley-Terry (BT) model (Bradley and Terry 1952) where preference is given by the sigmoid of two given responses' reward gap.

For RLHF (Ziegler et al. 2019; Stiennon et al. 2020; Ouyang et al. 2022), LLMs are optimized with a learned reward model r_ψ and Proximal Policy Optimization (PPO) (Schulman et al. 2017). The learning objective is to maximize preference rewards, constrained by a Kullback–Leibler (KL) divergence constraint.

DPO eliminates the need for explicitly fitting a reward model and uses the model with its reference for implicit reward modeling. The loss function in DPO is

$$\mathcal{L}_{\pi_\theta} = -\mathbb{E}_{\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right], \tag{1}$$

where π_{ref} is the reference model which is usually the SFT model, β is a hyperparameter, (x, y_w, y_l) is sampled from the dataset $\mathcal{D}$, y_w and y_l are the preferred and dispreferred response, i.e. $y_w \succ y_l$. Minimizing Eq. 1 will make the model prefer chosen response y_w over rejected response y_l and align with human preferences.

Related Work

Large Language Models Pretrained models (Brown et al. 2020; Touvron et al. 2023a; Li et al. 2023b; Dubey et al. 2024) acquire extensive world knowledge through self-supervised pretraining on an extraordinarily large corpus of texts. While pretrained models are able to predict the next words in sentences, they are not suitable for direct application in downstream tasks. However, with Instruction Fine-Tuning (Sanh et al. 2021; Chung et al. 2022; Ouyang et al. 2022), these models are trained on task-specific data, allowing them to follow prompts and excel at specific tasks. Thus, fine-tuned models exhibit strong capabilities across various domains.

Preference Alignment To prevent LLMs from generating unsatisfactory, misleading or even harmful responses (Bai et al. 2022; Kocoń et al. 2023), LLMs must align with human preferences. RLHF (Ouyang et al. 2022) trains a reward model with ranked response pairs, where higher rewards indicates better alignment with human preference. RLHF uses PPO (Schulman et al. 2017) to fine-tune the LLMs to generate responses with high rewards from the reward model. However, fine-tuning LLMs with explicit reward modeling is notoriously complex and difficult (Bai et al. 2022). DPO (Rafailov et al. 2023) proposes implicit reward modeling, which can be optimized with a simple classification loss and significantly simplifies the fine-tuning pipeline.

Safe RLHF (Dai et al. 2023), RL from Personalized Human Feedback (RL$\mathcal{P}$HF) (Jang et al. 2023) and Li et al. (2024) also study alignment on multi-dimensional human preferences. However, their use of multiple reward models for alignment significantly complicates and destabilizes the fine-tuning process. Multi-Objective REward learning (MORE) (Zeng et al. 2023) proposes to learn a multi-objective reward model by aggregation of shared reward and multiple dimension-specific reward heads. But compared to SPO, although MORE learns dimension-specific rewards, it does not guarantee alignment on each dimension. Rewarded soups (Rame et al. 2024a) merges LLMs aligned with different datasets and objectives to combine their strengths. Multi-Objective DPO (MODPO) (Zhou et al. 2023) also aligns LLMs with multiple preferences , but it aligns all dimensions in one round, while SPO aligns one at a time, easing each round's preference optimization.

Methodology

In this section, we will derive how to align LLMs with multi-dimensional human preferences in a sequential manner and propose Sequential Preference Optimization (SPO). We will first derive how to align LLMs on two-dimensional human preferences. Then, gradient analysis is conducted to show how SPO manages to achieve alignment across dimensions. Finally, we extend SPO to preference alignment with arbitrary number of dimensions. Pipeline of SPO is given in Fig. 1.

To maintain previous alignment during sequential fine-tuning, rewards on previous dimensions must remain above a certain threshold. In SPO's pipeline, for the n-th round fine-tuning, π_0 and $\pi_{\{1,..,n-1\}}$ are the SFT model and previous sequentially fine-tuned models. The initial model π_{n-1} maximizes R_{n-1} and satisfy $\forall i \in \{1,..,n-2\}$, $\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_{n-1}}[R_i(x,y)] \geq H_i$, where $n \in \mathbb{N} \geq 3$ and H_i is the threshold for preference reward on the i-th dimension. In other words, π_{n-1} is aligned with all previous dimensions.

To align with the n-th dimension and preserve previous alignment, the n-th round fine-tuning is formulated as

$$\begin{aligned} &\max_{\pi_n} \mathbb{E}_{x\sim\mathcal{D},y\sim\pi_n}[R_n(x,y)] \\ &s.t.\ \mathbb{D}_{KL}(\pi_n\|\pi_{n-1}) \leq H_0 \\ &\forall\, i \in \{1,..,n-1\},\ -\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_n}\left[R_i(x,y)\right] \leq -H_i. \end{aligned} \tag{2}$$

where $\mathcal{D}$ is the training dataset, x is prompts from the dataset, y is the response generated by π_n. The n-th round of fine-tuning in SPO ensures: (1) maximized reward for the n-th dimension, (2) limited deviation from π_{n-1} and (3) prevention of significant degradation of previous alignments.

Two-Dimensional Sequential Alignment

We first consider aligning LLMs on two-dimensional human preferences, i.e. $n = 2$ in optimization problem Eq. 2.

Deriving SPO Objective Since there is no constraint on previous alignments in the first round fine-tuning, we can directly apply DPO on SFT model π_0 to obtain π_1, which maximizes preference reward R_1 on the first dimension.

The second round of fine-tuning in SPO solves Eq. 2 with $n = 2$ and thus maximizes preference reward R_2 while preserving alignment on the first dimension. Like prior works (Peng et al. 2019; Rafailov et al. 2023), we can derive the closed-form optimal policy π_2^* for the constrained maximization problem Eq. 2

$$\pi_2^*(y|x) = \frac{1}{Z_2(x)}\pi_1(y|x)\exp\left(\frac{\alpha_1}{\beta}R_1(x,y) + \frac{1}{\beta}R_2(x,y)\right), \tag{3}$$

where β controls deviation of π_2 from the reference model π_1, α_1 controls the importance of maximizing reward R_1, and $Z_2(x) = \sum_y \pi_1(y|x)\exp\left(\frac{\alpha_1}{\beta}R_1(x,y) + \frac{1}{\beta}R_2(x,y)\right)$ is the partition function. Detailed derivation of π_2^* is given in the appendix A.1.

By taking logarithm on both sides and some algebra, Eq. 3 can be transformed into

$$R_2(x,y) = -\alpha_1 R_1(x,y) + \beta\log\frac{\pi_2(y|x)}{\pi_1(y|x)} + \beta\log Z_2(x), \tag{4}$$

where R_1, R_2 are based on BT model (Bradley and Terry 1952). $R_1 = \beta\log\frac{\pi_1(y|x)}{\pi_0(y|x)} + \beta\log Z_1(x)$ can be represented by the SFT model π_0 and π_1 from the first round fine-tuning (Rafailov et al. 2023).

In BT model, preference is decided by the difference between responses' rewards. Specifically, $P_R(y_1 \succ y_2) = \sigma\left(R(x,y_1) - R(x,y_2)\right)$, where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the $sigmoid$ function. Thus, we can substitute R_2 into the BT model and derive the loss function for preference optimization on the second dimension, which is the log probability of preference in the BT model

$$\begin{aligned} &\mathcal{L}_2^{SPO}(\pi_2^\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log P_{R_2}(y_1 \succ y_2)\right] \\ &= -\mathbb{E}_{\mathcal{D}}\left[\log\sigma\left(\xi_2\phi_2(x,y_w,y_l) - \xi_1\phi_1(x,y_w,y_l)\right)\right], \end{aligned} \tag{5}$$

where $\forall i \in \{1,2\}$, $\phi_i(x,y_w,y_l) = \log\frac{\pi_i(y_w|x)}{\pi_{i-1}(y_w|x)} - \log\frac{\pi_i(y_l|x)}{\pi_{i-1}(y_l|x)}$, x is the prompt, y_w and y_l are the preferred and the less preferred responses on the second dimension, $\mathcal{D}$ is the training dataset, constant $\xi_1 = \alpha_1\beta > 0$, $\xi_2 = \beta > 0$ and σ is the $sigmoid$ function. ϕ_1 is decided by the SFT model π_0 and the previous fine-tuned model π_1, while ϕ_2 is decided by the current model π_2 and π_1. Detailed derivation of Eq. 5 is given in the appendix A.1. Minimizing $\mathcal{L}_2^{SPO}$ will lead π_2 to maximize human preference on the second dimension while still preserving preference alignment on the first dimension.

Gradient Analysis Compared to naive two-round sequential fine-tuning (where the constraint on R_1 is removed from the second round fine-tuning's optimization problem 2), SPO is able to prevent the fine-tuned model from significant degradation on the preference maximization of R_1. We now theoretically explicate this advantage of SPO by analyzing gradient of the loss function $\mathcal{L}_2^{SPO}$.

The gradient of loss function $\mathcal{L}_2^{SPO}$ w.r.t. policy parameter of π_2^θ is given by

$$\begin{aligned} &\nabla_\theta\mathcal{L}_2^{SPO} = \\ &-\xi_2\mathbb{E}_{\mathcal{D}}\left[g_\theta\cdot\sigma\big(-\xi_2\phi_2(x,y_w,y_l) + \xi_1\phi_1(x,y_w,y_l)\big)\right], \end{aligned} \tag{6}$$

where the gradient term $g_\theta = \nabla_\theta\log\pi_2^\theta(y_w|x) - \nabla_\theta\log\pi_2^\theta(y_l|x)$. Detailed derivation is given in appendix A.2. The first term $-\xi_2\phi_2(x,y_w,y_l)$ inside the $sigmoid$ function is for preference maximization on the second dimension. It assigns higher weight to the gradient when less preferred response y_l has a high likelihood to be generated by π_2. This term will also appear in the gradient if we directly run DPO instead of SPO for the second round fine-tuning.

However, $\phi_1(x,y_w,y_l) = R_1(x,y_w) - R_1(x,y_l) = \log\frac{\pi_1(y_w|x)}{\pi_0(y_w|x)} - \log\frac{\pi_1(y_l|x)}{\pi_0(y_l|x)}$ in the second term is the reward difference between y_w and y_l given by reward R_1 on the first dimension. $\phi_1(x,y_w,y_1) > 0$ when the preferred response y_2 on the second dimension is also preferred on the first dimension, and $\phi_1(x,y_w,y_l) < 0$ when preferred responses are different on the two dimensions.

The weight of gradient increases when ϕ_1 is positive (preferred responses are consistent) and decreases when ϕ_1 is negative (preferred responses are inconsistent). Therefore, ϕ_1 serves as a regularizer to prevent degradation of preference maximization on the first dimension. Consequently, SPO will strive to optimize the LLM so that preference maximization on both dimensions are achieved.

Multi-Dimensional Sequential Alignment

Now, we extend SPO to multi-dimensional preference alignment with arbitrary rounds of fine-tuning, i.e. $n \in \mathbb{N},\ n \geq 3$ in optimization problem Eq. 2.

By solving the optimization problem, we have for $\forall n \in \mathbb{N} \geq 3$, reward on the n-th preference dimension in SPO is

$$R_n(x,y) = \sum_{i=1}^{n}\kappa_i\log\frac{\pi_i(y|x)}{\pi_{i-1}(y|x)}, \tag{7}$$

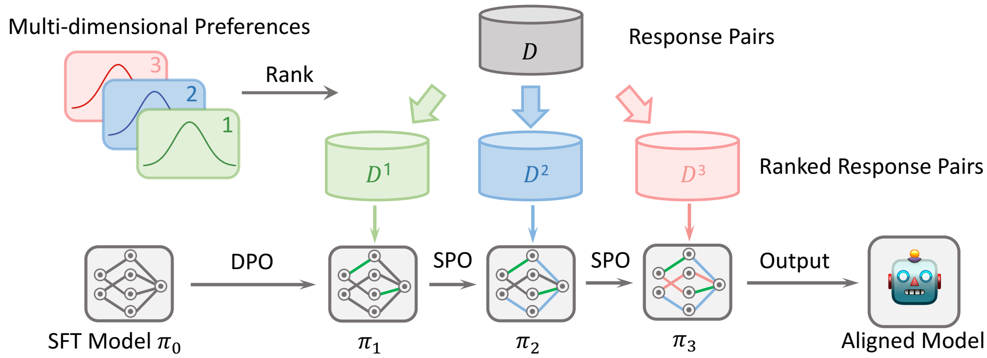


Figure 1: The SFT model is sequentially fine-tuned on multi-dimensional preferences with SPO, which aligns LLMs on the current dimension and preserves alignment on previous dimensions.

where $\kappa_n = \beta$, $\kappa_{n-1} = -\beta\alpha_{n-1}$ and $\forall i \in \{2,..,n-1\}$, $\kappa_{n-i} = -\beta\alpha_{n-i}\prod_{j=2}^{i}(1-\alpha_{n-1-i+j})$ and α_k controls the importance of the k-th dimension. Detailed derivation is given in the appendix B.1.

Proof Sketch. Similar as Eq. 3, we can first obtain the closed-form optimal solution of π_n^*, which is represented by the current reference model π_{n-1} and preference rewards from previous rounds. By some algebra, we can get the formulation of R_n. Since the formulations of R_1 and R_2 are already at hand, we can iteratively substitute previous preference rewards into the formulation of R_n and prove by mathematical induction that Eq. 7 holds for $\forall n \geq 3$. $\square$

Specifically, if all previous dimensions are equally important, i.e. $\forall\ k \in \{1,..,n-1\}, \alpha_k = \alpha$, the preference reward is given by $\hat{R}_n = \beta \log \frac{\pi_n(y|x)}{\pi_{n-1}(y|x)} - \beta \sum_{i=1}^{n-1} \alpha(1-\alpha)^{i-1} \log \frac{\pi_{n-i}(y|x)}{\pi_{n-i-1}(y|x)}$.

After obtaining the preference reward, SPO optimizes the LLM by directly maximizing the log probability of preference in the BT model. Loss function of the n-th round fine-tuning is given by

$$\mathcal{L}_n^{SPO}(\pi_n^\theta) = -\mathbb{E}_{x,y_w,y_l\sim\mathcal{D}}\left[\sigma\Big(R_n(x,y_w) - R_n(x,y_l)\Big)\right]. \tag{8}$$

Optimizing π_n by minimizing $\mathcal{L}_n^{SPO}$ enables the LLM to align with multi-dimensional preferences. Also, due to the constraints in our problem formulation, SPO is able to minimize the impact of alignment tax accumulated in multiple round of fine-tuning, achieving alignment across dimensions.

It is worth noting that the sequential fine-tuning in SPO only depends on the inference of previous models on the dataset $\mathcal{D}$, which has been done in the previous round of fine-tuning. Thus, no additional inference is required in the sequential training of SPO. We can just cache the inference results of previous rounds and use them in subsequent training. This makes the sequential fine-tuning in SPO very efficient.

We also conduct gradient analysis to theoretically demonstrate how SPO achieve multi-dimensional alignment. Details are given in appendix B.2. And pseudo code of SPO is given in the appendix C.

Experiment

In this section, we first evaluate SPO on a real-world dataset with two preference dimensions and then demonstrate SPO's ability to achieve multi-dimensional alignment with more preference dimensions.

Experiment Setting

Training Datasets For two-dimensional tasks, we use PKU-SafeRLHF-30k (Ji et al. 2023), comprising 26.9k response pairs ranked by helpfulness and harmlessness. Additionally, we also include a modified dataset introducing contradictions, e.g. harmless but unhelpful refusals, to increase alignment difficulty (details in Appendix D.1). For real-world multi-dimensional preferences, we use Helpsteer2 (Wang et al. 2024), which contains 10.1k pairs annotated for dimensions like helpfulness, coherence, and verbosity.

Evaluation Datasets Four test sets include HH-helpful (Bai et al. 2022) (2.3k questions) and AlpacaEval (Li et al. 2023a) (805 questions) for helpfulness, and HH-harmless (Bai et al. 2022) (2.3k questions) along with 300 questions from PKU-SafeRLHF-test (Ji et al. 2023) for harmlessness.

Models The base model in our experiments is Llama 2 (Touvron et al. 2023b). SFT models are obtained by training the base models on Alpaca dataset (Taori et al. 2023) with supervised fine-tuning for 2 epochs. Llama 2 with 7B and 13B parameters are adopted to test the scalability of SPO.

Baselines Our baselines include (1) **Safe-RLHF** (Dai et al. 2023). Safe RLHF achieves multi-dimensional alignment by learning a reward model for each dimension and safe RL algorithms (PPO-Lagragian (Ray, Achiam, and Amodei 2019)); (2) RLHF with reward shaping (**RLHF-RS**) (Dai et al. 2023); (3) **DPO-Mix**, where we mix the two-dimensional preferences into one dimension and run DPO on the mixed dataset. The principle of mixing is to prioritize harmlessness over helpfulness. Ranking for pairs with two harmless responses is randomly decided; (4) Sequential DPO (**S-DPO**). S-DPO is an ablation of SPO that sequentially runs DPO on each dimension without considering previous alignments; (5)

Size	Dataset	Safe-RLHF	DPO-Mix	RLHF-RS	S-DPO	DPO-HP	DPO-HM
7B	HH-helpful	77.7%	87.3%	80.8%	81.5%	47.9%	64.6%
	AlpacaEval	64.9%	77.9%	67.6%	78.0%	47.7%	68.6%
	HH-harmless	58.1%	57.7%	58.0%	41.9%	63.3%	48.1%
	SafeRLHF	55.8%	55.4%	56.5%	40.5%	65.8%	46.3%
13B	HH-helpful	82.8%	59.9%	85.3%	81.5%	47.2%	64.0%
	AlpacaEval	74.4%	59.0%	73.4%	81.5%	44.1%	66.3%
	HH-harmless	53.1%	58.6%	75.0%	57.5%	76.1%	51.8%
	SafeRLHF	51.3%	60.0%	78.9%	54.3%	77.6%	51.0%

Table 1: SPO's win rate against the baselines. The win rate is the proportion of questions where SPO gives better responses. HH-helpful and AlpacaEval evaluate LLMs' helpfulness while HH-harmless and SafeRLHF (short for PKU-SafeRLHF-Test) are evaluation datasets for harmlessness.

DPO-Helpful (**DPO-HP**). DPO-HP is only fine-tuned on the dimension of helpfulness; (6) DPO-Harmless (**DPO-HM**), which is only fine-tuned on the dimension of harmlessness.

Training Details LoRA adapters (Hu et al. 2021; Zheng et al. 2024) enable efficient fine-tuning. For DPO-based methods, we set $\beta = 0.1$, LoRA rank to 8, scaling to 32, and learning rate to 1×10^{-5}. SPO uses $\alpha = 0.1$ with other settings identical to DPO. Models are fine-tuned for 2 epochs per dimension. Safe-RLHF and RLHF-RS use default configurations. For the original dataset, we fine-tune on harmlessness first, then helpfulness. For the modified dataset, we reverse this order to prevent overfitting on harmlessness.

Evaluation Metric After fine-tuning, SPO responses are paired with baseline responses and evaluated using LLM-as-a-judge (Zheng et al. 2023) on helpfulness, safety, intent understanding, and language quality. Helpfulness evaluations exclude safety, and harmlessness evaluations excludes helpfulness. We report SPO's win rate (proportion of cases with better responses). To mitigate positional bias (Wang et al. 2023b), GPT-4 scores each pair twice with switched positions, averaging the results for final scoring. Details on prompts, win rate calculations, and human consistency studies are in Appendices D.2 and D.3.

Results with the Modified Training Dataset

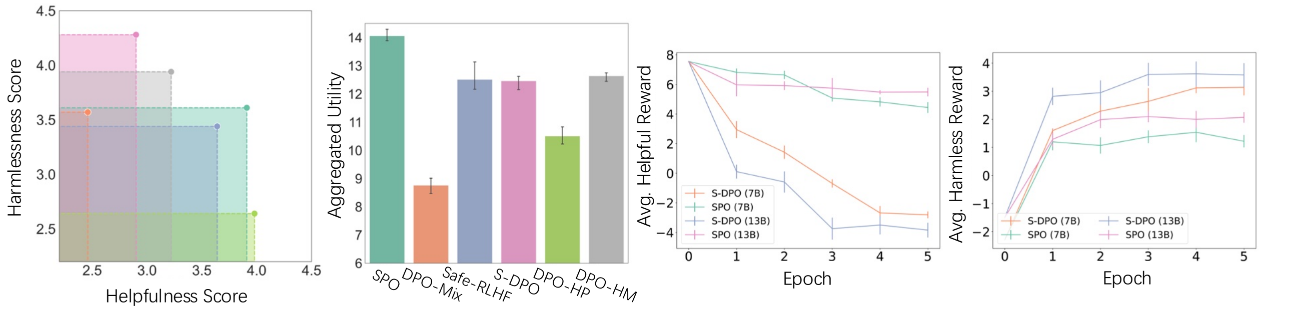
We first give the experiment results after fine-tuning the models with our modified two-dimensional dataset, where alignment across dimensions is harder to achieve.

Main Results Table 1 gives the evaluation win rate of SPO against the baselines. DPO-HP and DPO-HM perform slightly better than SPO on their fine-tuned dimension but significantly worse on the other dimension. The comparisons demonstrate the ability of SPO in reconciling the contradictions between these two dimensions and striking a compromise to achieve alignment on both dimensions. DPO-Mix has poor performance in all settings. This shows the importance of preference optimization for each dimension. S-DPO yields the same pipeline of SPO but has no constraints to preserve previous alignment results. For model with 7B parameters, as S-DPO has no additional constraints for preference optimization, it exhibits better alignment on the second dimension (harmlessness), but shows drastic degradation of alignment on the first dimension (helpfulness). For model with 13B parameters which has stronger expressive capability, S-DPO overfits to the second dimension of harmlessness, where it always gives extremely simple but harmless responses. Therefore, SPO exhibits a higher overall win rate even on the harmlessness evaluation datasets.

Compared to the RLHF-based counterparts SafeRLHF and RLHF-RS, SPO better aligns with preferences on helpfulness and harmlessness for both 7B and 13B models. The performance of RLHF-based methods shows that explicit reward modeling on multiple preference dimensions will destabilize the fine-tuning process, leading to sub-optimal performance. In contrast, SPO follows the implicit reward modeling as DPO (Rafailov et al. 2023) and achieves better preference alignment across all dimensions.

Alignment Analysis We visualize the alignment scores evaluated by GPT-4 in Fig. 2(a). The harmlessness scores are averaged scores on PKU-SafeRLHF-Test and the helpfulness scores are averaged scores on AlpacaEval with fine-tuned 7B models. From the results, we can see SPO has similar helpfulness scores as DPO-HP, which is only fine-tuned on the helpfulness dimension. This shows SPO's strong capability in preserving previous preference alignment results. Inspired by (Zheng et al. 2022) that conducts evaluation with two contradictory metrics, we use the product of helpfulness scores and harmless scores as the aggregated utility, which is also the area of the rectangles in Fig. 2(a). Fig. 2(b) gives the aggregated utility of each method. SPO strikes a balance between alignment on the two preference dimensions and thus has the highest aggregated utilities.

Ablation Study By setting hyperparameter $\alpha = 0$, we remove the constraint on preserving previous alignment results and obtain S-DPO. Fig. 2(c), (d) gives the helpfulness and harmlessness rewards during the second-round fine-tuning. Results are obtained by querying Safe-RLHF's reward and cost models. Compared to SPO, S-DPO's helpfulness rewards significantly drops, which means severe degradation of alignment on helpfulness. Especially, 13B model's strong expressive capacity makes it rapidly overfit to the harmlessness dimension, resulting in poor alignment on helpfulness. In contrast, although SPO has relatively lower harmlessness scores, it effectively preserves previous alignment on helpfulness. As a result, SPO defeats S-DPO in terms of overall performance.



(a) Harmlessness and Helpfulness score by GPT-4

(b) Aggregated utility by product

(c) Rewards by Helpfulness reward model

(d) Rewards by Harmlessness reward model

Figure 2: (a) Scores on harmlessness and helpfulness evaluation datasets by GPT-4 evaluator. (b) Aggregated utility of two dimensions, which is the product of harmlessness and helpfulness scores. (c), (d) give Helpfulness and Harmlessness rewards during the second-round fine-tuning process of SPO and S-DPO. SPO better preserves alignment on the first dimension of Helpfulness while learning to align with Harmlessness.

R_i \ α	0	0.05	0.1	0.3	0.5	0.7
Helpful	1.42	6.53	6.82	6.93	7.05	7.13
Harmless	2.29	1.30	1.18	0.22	-0.74	-1.42

Table 2: The effect of hyperparameter α for preserving reward and alignment on previous dimensions.

Dataset	S-DPO	Alpaca	Vicuna	Mistral
HH-Helpful	53.1%	54.6%	55.7%	49.8%
AlpacaEval	56.3%	56.3%	51.5%	51.5%
HH-Harmless	76.9%	71.0%	55.5%	60.4%
SafeRLHF	76.2%	60.8%	53.2%	48.8%

Table 3: SPO's win rate against open models.

Then we set α in SPO during second-round fine-tuning of 7B models to different values to see its effect on multidimensional alignment. Larger α stands for greater importance of preserving previous alignment on helpfulness. Results are given in Table 2. When $\alpha = 0$, previous alignment on helpfulness significantly deteriorates. As α increases, helpfulness rewards increases, implying better preservation of the first-round alignment. Conversely, harmlessness rewards decrease, as they contradict preferences on the first dimension. Thus, the result show the constraint in SPO is able to effectively preserve previous alignment. We also propose solving the dual problem of SPO to dynamically adjust α to maintain previous preference rewards near a specified threshold. For more details, please refer to Appendix F.

Overfitting Study The constraints on on prior dimensions in SPO force LLMs to retain previous alignments. Thus, they are also able to keep LLMs from overfitting to the current preference dimension. We analyzed overfitting during fine-tuning by evaluating SPO and S-DPO on 200 safe questions from AlpacaEval, where refusal to answer these safe questions indicates overfitting to the harmlessness preference dimension. To obtain the results, we filtered out responses containing key words like "sorry", "as an AI assistant" and manually identifying refusals. We use 7B models in this experiment and fine-tune them on helpfulness for 2 epochs followed by 5 epochs on harmlessness.

Because the training dataset is altered to induce refusals (completely harmless but unhelpful), S-DPO demonstrates severe overfitting as shown in Table 6. But with the constraint to preserve alignment on helpfulness, SPO does not overfit to the harmlessness dimension even after 5 epochs.

Model Merging Recently, model merging techniques (Rame et al. 2024a,b,c) successfully merge different reward models and LLMs in the weight space and combine the strengths of them. Here we study whether model merging technique is able to achieve alignment across multiple potentially conflicting dimensions.

We first merge the helpful RM and harmless RM in our experiment by linear interpolation with equal weights and then evaluate the RMs on held-out validation sets on both dimensions. Results in Table 7 show that the merged RM perform poorly on both helpfulness and harmlessness due to the inherently conflicting goals of the models being merged.

Then, we merge two LLMs aligned with helpfulness and harmlessness by DPO separately and evaluate the merged model against SPO. SPO's win rate against the merged DPO model is **79.0%** on AlpacaEval and **41.2%** on PKU-SafeRLHF-Test. We can tell that SPO significantly outperforms merged DPO on helpfulness but loses on harmlessness, showing the harmless LLM becomes dominant in the merged model. This is potentially because harmless responses exhibit simpler patterns than helpful responses, e.g. refusals. Thus, merging inherently conflicting LLMs is not an ideal choice to achieve multi-dimensional alignment, as some LLMs may easily become dominant over the others.

Results with the Original Training Dateset

We now give the results when fine-tuning Llama 2 7B model with the original PKU-SafeRLHF-30k dataset and compare our model with some prevalent open models.

	$Token_1$	$Token_2$	$Token_3$	$Token_4$	Opt.
SPO	98.0%	77.7%	64.7%	83.3%	50.3%
S-DPO	10.7%	1.3%	0	99.7%	0

Table 4: Percentage of presence of four special tokens in the responses and Pareto-optimal (*Opt.*) responses.

e	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}
SPO	0%	0%	0%	0%	0%
S-DPO	62.5%	30.5%	41.5%	45.5%	46.0%

Table 6: Percentage of overfitting during second-round fine-tuning. SPO shows no overfitting as epoch e increases.

	Helpful Val.	Harmless Val.
Helpful RM	67.1%	37.1%
Harmless RM	35.4%	71.0%
Merged RM	51.2%	48.2%

Table 7: Accuracy of the RMs on validation sets. A prediction is correct when RM gives higher reward to the preferred response than the dispreferred response.

The open models we consider here are Alpaca (Taori et al. 2023), Vicuna-7B-v1.5 (Zheng et al. 2023) and Mistral-7B-Instruct-v0.1 (Jiang et al. 2023), which is based on a stronger base model than Llama 2 used in SPO (MistralAI 2023). As shown in Table 3, SPO outperforms the ablation S-DPO, Alpaca and Vicuna. Although Mistral-Instruct's base model is significantly stronger, SPO still gets comparable results. This highlights SPO's ability to align LLMs with multi-dimensional preferences and achieve strong performance.

Experiments with More Preference Dimensions

To evaluate SPO's ability to achieve multi-dimensional alignment, we first conduct experiments on a demonstrative dataset and then give results on the real-world dataset Helpsteer2 (Wang et al. 2024), both with four preference dimensions.

Demonstrative Experiments We randomly sample 10k samples from the training dataset and augment them with four special tokens, denoted as $\{[Token_1], [Token_2], [Token_3], [Token_4]\}$, to indicate preference. Specifically, the ranking on each dimension is determined by the presence of a unique token. On each dimension, a special token is added to the preferred response, and other tokens have 10% probability to be added to both preferred and dispreferred responses. In this way, alignment on each dimension is indicated by presence of the corresponding special token. The Pareto-optimal model that aligns with four dimensions will always include all special tokens in the generations. SPO and S-DPO are sequentially fine-tuned on four dimensions for 1 epoch. Please refer to Appendix E.1 for details.

	Helpful	Correct	Coherent	Verbose
S-DPO	57.3%	55.3%	57.6%	49.4%
Merged	66.2%	67.7%	73.5%	53.3%

Table 5: SPO's win rate on each dimensions after fine-tuning on four dimensions on Helpsteer 2. *Merged* is the merged model of four DPO model on each dimension.

The results show that after fine-tuning on four dimensions, SPO achieves 50.3% Pareto-optimal responses (all special tokens present). In contrast, S-DPO, which lacks constraints on prior alignments, aligns only with the last dimension and performs poorly on earlier ones. This highlights SPO's ability to align LLMs across multiple preference dimensions effectively. Additionally, MODPO, with very limited data aligned across four dimensions, achieves 0% Pareto-optimality, further demonstrating SPO's advantage in sequential multi-dimensional alignment.

Real-World Dataset Helpsteer 2 (Wang et al. 2024) includes 10.1k response pairs annotated across multiple dimensions. We apply SPO sequentially on helpfulness, correctness, coherence, and verbosity, comparing it against S-DPO and a merged model combining four DPO models, each aligned separately to one dimension. Results are evaluated on the Helpsteer 2 validation set, with SPO's win rates for each dimension detailed in Table 5. Additional evaluation details and prompts are in Appendix E.2.

The results confirm SPO's effectiveness in multi-dimensional alignment. Specially, S-DPO only performs slightly better on the last dimension as it has no additional constraints. Merged DPO performs poorly in all dimensions because the models to be merged are trained on the same data with potentially conflicting preferences, affecting the overall performance after model merging. Moreover, before merging, SPO's comprehensive win rates against DPO models on each dimension is $56.2\%, 56.0\%, 70.1\%,$ and 61.0%.

Conclusion and Limitation

In this paper, we tackle the problem of aligning LLMs with multi-dimensional preferences and propose Sequential Preference Optimization (SPO). SPO avoids explicit reward modeling in RLHF and achieve multi-dimensional alignment by iteratively solving constrained optimization problems. The constrained optimization problem enables SPO to optimize preference on new dimensions while preserving the alignment in previous rounds. Theoretically, we derive the learning objective of arbitrary rounds of preference alignment in SPO and conduct gradient analysis to illustrate how SPO achieves alignment across dimensions. Empirically, extensive experiments and studies on different training datasets, evaluation datasets and preference dimensions confirm the efficacy of SPO in aligning LLMs across multiple dimensions.

The limitation of this work is although 7B and 13B models are considered, we do not include extremely large models (Adler et al. 2024; Dubey et al. 2024) in our experiments due to computational limit. In the future, we plan to apply SPO on large models and evaluate against state-of-the-art LLMs.

Acknowledgments

This work is supported by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDA27010103) and the Youth Innovation Promotion Association CAS.

References

Adler, B.; Agarwal, N.; Aithal, A.; Anh, D. H.; Bhattacharya, P.; Brundyn, A.; Casper, J.; Catanzaro, B.; Clay, S.; Cohen, J.; et al. 2024. Nemotron-4 340B Technical Report. *arXiv preprint arXiv:2406.11704*.

Anthropic. 2023. Model Card and Evaluations for Claude Models. *https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf*.

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(3/4): 324–345.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chiang, W.-L.; Zheng, L.; Sheng, Y.; Angelopoulos, A. N.; Li, T.; Li, D.; Zhang, H.; Zhu, B.; Jordan, M.; Gonzalez, J. E.; et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Dai, J.; Pan, X.; Sun, R.; Ji, J.; Xu, X.; Liu, M.; Wang, Y.; and Yang, Y. 2023. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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*.

Ji, J.; Liu, M.; Dai, J.; Pan, X.; Zhang, C.; Bian, C.; Sun, R.; Wang, Y.; and Yang, Y. 2023. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *arXiv preprint arXiv:2307.04657*.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Hanna, E. B.; Bressand, F.; et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Jiao, W.; Wang, W.; Huang, J.-t.; Wang, X.; and Tu, Z. 2023. Is ChatGPT a good translator? A preliminary study. *arXiv preprint arXiv:2301.08745*.

Kocoń, J.; Cichecki, I.; Kaszyca, O.; Kochanek, M.; Szydło, D.; Baran, J.; Bielaniewicz, J.; Gruza, M.; Janz, A.; Kanclerz, K.; et al. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion*, 101861.

Li, J.; Zhang, H.; Zhang, F.; Chang, T.-W.; Kuang, K.; Chen, L.; and Zhou, J. 2024. Optimizing Language Models with Fair and Stable Reward Composition in Reinforcement Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 10122–10140.

Li, X.; Zhang, T.; Dubois, Y.; Taori, R.; Gulrajani, I.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023a. AlpacaEval: An Automatic Evaluator of Instruction-following Models. https://github.com/tatsu-lab/alpaca_eval.

Li, Y.; Bubeck, S.; Eldan, R.; Del Giorno, A.; Gunasekar, S.; and Lee, Y. T. 2023b. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

MistralAI. 2023. Mistral 7B. *https://mistral.ai/news/announcing-mistral-7b/*.

Nijkamp, E.; Pang, B.; Hayashi, H.; Tu, L.; Wang, H.; Zhou, Y.; Savarese, S.; and Xiong, C. 2022. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474*.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

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.

Peng, X. B.; Kumar, A.; Zhang, G.; and Levine, S. 2019. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*.

Qian, C.; Cong, X.; Yang, C.; Chen, W.; Su, Y.; Xu, J.; Liu, Z.; and Sun, M. 2023. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C. D.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

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

Rame, A.; Ferret, J.; Vieillard, N.; Dadashi, R.; Hussenot, L.; Cedoz, P.-L.; Sessa, P. G.; Girgin, S.; Douillard, A.; and Bachem, O. 2024b. WARP: On the Benefits of Weight Averaged Rewarded Policies. *arXiv preprint arXiv:2406.16768*.

Rame, A.; Vieillard, N.; Hussenot, L.; Dadashi, R.; Cideron, G.; Bachem, O.; and Ferret, J. 2024c. Warm: On the benefits of weight averaged reward models. *arXiv preprint arXiv:2401.12187*.

Ray, A.; Achiam, J.; and Amodei, D. 2019. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708*, 7(1): 2.

Sanh, V.; Webson, A.; Raffel, C.; Bach, S. H.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Scao, T. L.; Raja, A.; et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

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.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, G.; Xie, Y.; Jiang, Y.; Mandlekar, A.; Xiao, C.; Zhu, Y.; Fan, L.; and Anandkumar, A. 2023a. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Wang, P.; Li, L.; Chen, L.; Zhu, D.; Lin, B.; Cao, Y.; Liu, Q.; Liu, T.; and Sui, Z. 2023b. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Wang, Z.; Dong, Y.; Delalleau, O.; Zeng, J.; Shen, G.; Egert, D.; Zhang, J. J.; Sreedhar, M. N.; and Kuchaiev, O. 2024. HelpSteer2: Open-source dataset for training top-performing reward models. *arXiv preprint arXiv:2406.08673*.

Zeng, D.; Dai, Y.; Cheng, P.; Hu, T.; Chen, W.; Du, N.; and Xu, Z. 2023. On Diverse Preferences for Large Language Model Alignment. *arXiv preprint arXiv:2312.07401*.

Zhang, H.; Du, W.; Shan, J.; Zhou, Q.; Du, Y.; Tenenbaum, J. B.; Shu, T.; and Gan, C. 2023. Building cooperative embodied agents modularly with large language models. *arXiv preprint arXiv:2307.02485*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*.

Zheng, S.; Trott, A.; Srinivasa, S.; Parkes, D. C.; and Socher, R. 2022. The AI Economist: Taxation policy design via two-level deep multiagent reinforcement learning. *Science advances*, 8(18): eabk2607.

Zheng, Y.; Zhang, R.; Zhang, J.; Ye, Y.; Luo, Z.; and Ma, Y. 2024. LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models. *arXiv preprint arXiv:2403.13372*.

Zhou, Z.; Liu, J.; Yang, C.; Shao, J.; Liu, Y.; Yue, X.; Ouyang, W.; and Qiao, Y. 2023. Beyond one-preference-for-all: Multi-objective direct preference optimization. *arXiv preprint 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*.