# Arithmetic Control of LLMs for Diverse User Preferences: Directional Preference Alignment with Multi-Objective Rewards

**Haoxiang Wang**[*1] **Yong Lin**[*2] **Wei Xiong**[*1] **Rui Yang**[2] **Shizhe Diao**[2] **Shuang Qiu**[2]
**Han Zhao**[1] **Tong Zhang**[1]
[1]University of Illinois Urbana-Champaign
[2]The Hong Kong University of Science and Technology

## Abstract

Fine-grained control over large language models (LLMs) remains a significant challenge, hindering their adaptability to diverse user needs. While Reinforcement Learning from Human Feedback (RLHF) shows promise in aligning LLMs, its reliance on *scalar* rewards often limits its ability to capture diverse user preferences in real-world applications. To address this limitation, we introduce the Directional Preference Alignment (DPA) framework. Unlike the scalar-reward RLHF, DPA incorporates *multi-objective reward* modeling to represent diverse preference profiles. Additionally, DPA models user preferences as *directions* (i.e., unit vectors) in the reward space to achieve user-dependent preference control. Our method involves training a multi-objective reward model and then fine-tuning the LLM with a preference-conditioned variant of Rejection Sampling Finetuning (RSF), an RLHF method adopted by Llama 2. This method enjoys a better performance trade-off across various reward objectives. In comparison with the scalar-reward RLHF, DPA offers users *intuitive control over LLM generation*: they can *arithmetically* specify their desired trade-offs (e.g., more helpfulness with less verbosity). We also validate the effectiveness of DPA with real-world alignment experiments on Mistral-7B. Our method provides straightforward arithmetic control over the trade-off between helpfulness and verbosity while maintaining competitive performance with strong baselines such as Direct Preference Optimization (DPO). The code and trained model are released at `https://github.com/RLHFlow/directional-preference-alignment`.

## 1 Introduction

Large language models (LLMs) (OpenAI, 2023; Anthropic, 2023) have demonstrated remarkable capabilities across various domains and tasks, such as mathematical reasoning (Wei et al., 2022) and medical question answering (Singhal et al., 2023a; Wang et al., 2023a; Thirunavukarasu et al., 2023). However, for an assistant to be truly useful, it must align with human preferences, such as being helpful, honest, harmless, and managing verbosity.

*Reinforcement Learning from Human Feedback* (RLHF) (Christiano et al., 2017; Ziegler et al., 2019; Ouyang et al., 2022; Bai et al., 2022b; Lee et al., 2023), is the leading approach to adapt LLMs towards these complex, often implicitly-defined goals. Typically, the most popular RLHF framework (Christiano et al., 2017; Ziegler et al., 2019; Ouyang et al., 2022) first constructs a scalar reward model to represent the difficult-to-specify goal of being preferred by human and then use this reward model to provide signals for the subsequent reward optimization stage. Its success spans various practical applications, including recommendation systems (Pereira et al., 2019), image generation (Hao et al., 2022; Wu et al., 2023a; Dong et al., 2023a), robotics (Brown et al., 2019), and most notably, aligning LLMs with human values and preferences, such as ChatGPT (OpenAI, 2023), Claude (Anthropic, 2023), Llama 2 (Touvron et al., 2023) and Gemini (Team et al., 2023).

While recent advancements in RLHF are noteworthy, a fundamental challenge persists due to problem misspecification. This means that a *single* reward function may not sufficiently capture complex human values. For example, a generative model aligned by RLHF for helpfulness tends to produce verbose responses as shown in Figure 1 (Left) (Singhal et al., 2023b), even though many users prefer answers that are both helpful and concise. Assuming scalar-objective reward implies a *total order* over preferences, which is hard to satisfy when the preference is aggregated across a diverse set of human groups (May, 1954; Tversky, 1969), because humans typically have a set of intricate or even *contradictory* targets (Biyik and Sadigh, 2018). In real-world applications, the



*Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8642–8655
August 11-16, 2024 ©2024 Association for Computational Linguistics

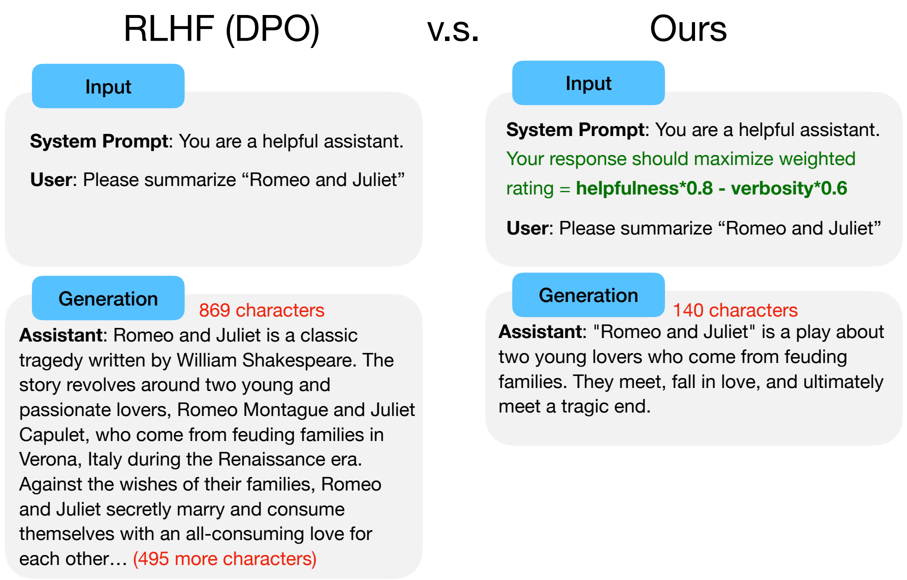


Figure 1: **Arithmetic Prompting** for Preference-Conditional Generalization: Comparison between conventional RLHF methods such as DPO and our Directional Preference Alignment (DPA). In the case of DPO (left), it is capable of generating helpful responses, but these tend to be excessively verbose. Conversely, with our DPA (right), it allows for **arithmetic control** of LLMs to meet various user preferences. For instance, setting the directional preference (unit vector) to $v = \langle 0.8, -0.6 \rangle$ leads to less verbose responses from our aligned LLM.

scalar-reward RLHF tends to align the LLMs toward an "average-user" preference, which cannot capture the complicated nature of human preferences and can be unfair for the under-represented groups (Feffer et al., 2023). For example, consider User-1, 2, 3, and responses $A$, $B$, $C$ in Fig. 2 (Left). User-1 and 3 prefer response $B$ over $C$ ($B \prec C$), while User-2 prefers $C$ over $B$ ($C \prec B$). This could occur as response $C$ is more verbose than $B$, while User-2 prefers concise answers. When these diverse preferences are aggregated across human groups, the typical reward models with scalar rewards tend to learn the "average-user" preference (which is $B \prec C$ in this case), overlooking the individual preference of User-2, as shown in Figure 2 (Middle). This is also known as the "Condorcet paradox" in the theory of social choice (Gehrlein, 2002). In general, human opinions and expertise can vary significantly (Coello, 2000; Bobu et al., 2023; Bansal et al., 2023). Meanwhile, the importance of these targets may also change over time, depending on the users and their expectations.

To address the limitations of the existing scalar reward model, previous works suggest the use of multi-objective rewards that characterize human preferences from different aspects (e.g., helpfulness, verbosity, harmlessness) (Pan et al., 2023; Rame et al., 2023). One common way is to take the human feedback as a *multi-dimensional* reward vector and each dimension models one objective (Rame et al., 2023; Dong et al., 2023b). Then, one may apply a linear combination to transform the multi-objective rewards into a scalar for LLM alignment (Bakker et al., 2022; Wu et al., 2023b). However, this approach still cannot handle the user-dependent needs from a diverse user population and can be unfair for minority groups. One may further adopt a user-dependent linear combination to multi-objective rewards for aligning a model for each user preference (Rame et al., 2023; Jang et al., 2023). However, this approach is quite *inference-unfriendly* because we have to switch between different models in response to the different user preferences. Finally, in social choice theory, a game-based formulation was studied under the name *maximal lotteries* (Sternberg, 1965; Fishburn, 1984), as well as the subsequent works in RLHF (Wang et al., 2023b; Swamy et al., 2024; Ye et al., 2024), to handle the diversity of user preferences. We remark that their framework is fundamentally different from the multi-objective rewards and cannot offer a user-dependent preference control in the inference stage, either. Refer to Section 2.3 for a more detailed discussion with existing methods.

In recognition of the aforementioned limitations, we propose a novel and practical alignment approach, *Directional Preference Alignment* (DPA), to enhance the *adaptability and controllability of*

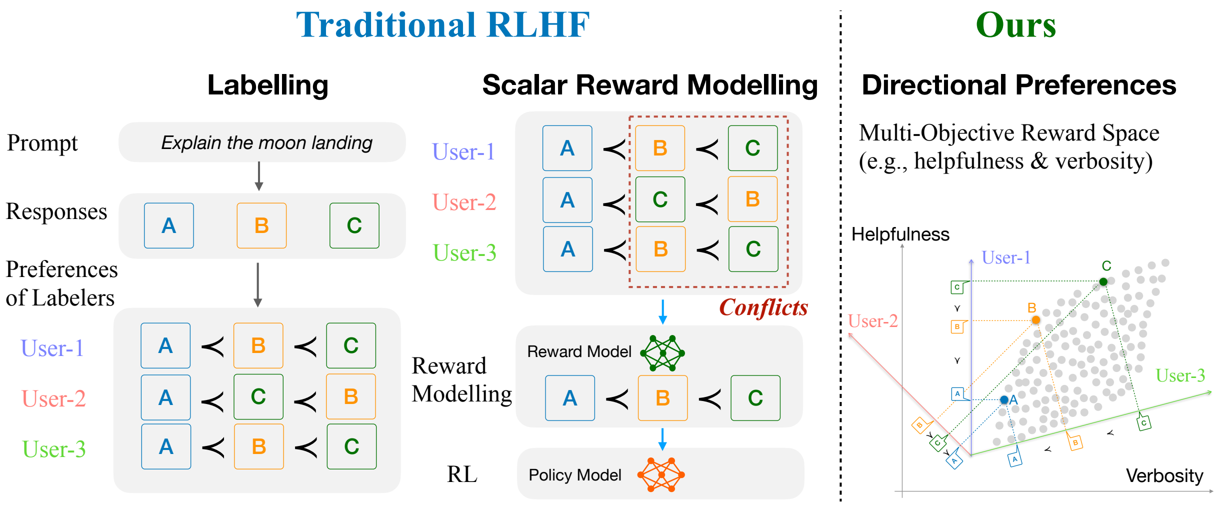


Figure 2: (Left) The illustration depicts preference conflicts among different users, where User-1 and User-3 favor response B over response C, while User-2 prefers C over B. (Middle) Generally, the scalar-reward RLHF framework tends to align toward the average-user preference, thus favoring B over C, which overlooks the preference of User-2. (Right) Our Directional Preference Alignment (DPA) enables users to specify their preference vector in a multi-dimensional space, allowing each user's preference to be well represented within this context.

*a single LLM*. Our aligned LLM enjoys the flexibility to be controlled with different preferences embedded numerically into the system prompt. The ability to control preferences can significantly enhance the model's personalization ability during inference. For example, as the model is aligned with DPA with `helpfulness` and `verbosity` in consideration, a user could simply control the model's generation by specifying a directional preference $v = \langle v_1, v_2 \rangle$ that $\|v\|_2 = 1$, and the model will generate responses that maximize $\texttt{reward} = v_1 \times \texttt{helpfulness} + v_2 \times \texttt{verbosity}$ where `helpfulness` and `verbosity` are rewards scored from different perspectives as shown in Figure 1 (Right). Figure 2 (Right) further shows that the preferences of User-1, User-2, and User-3 can be accurately represented by specifying the preference vector in the 2-dimensional space. This is a scenario where DPA can alleviate the problem of misspecification in RLHF.

Our approach features two crucial aspects: 1). Multi-Objective Rewards, which involve learning with multiple different preference targets simultaneously, and 2). Directional Preference Alignment, which encodes user preferences as unit vectors for preference-aware LLM alignment. Specifically, we summarize our contributions as follows.

- **We identify the limitations of existing popular RLHF frameworks**: 1) the limited capacity for capturing the real-world complicated human preference; 2) lacking in adaptability for user-dependent preference;
- **We propose Directional Preference Alignment (DPA)**: a novel alignment approach that allows a *single* LLM to accommodate users with varying preferences.
- **We consider both helpfulness and verbosity rewards, and align Mistral-7B (Jiang et al., 2023) with our DPA**: empirical evaluations show that DPA offers effective arithmetic control over the trade-off between helpfulness and verbosity, while maintaining competitive performance with DPO (Rafailov et al., 2023).

## 2 Directional Preference Alignment

In a typical RLHF pipeline (Ouyang et al., 2022; Bai et al., 2022a; Touvron et al., 2023), we first construct a reward model based on a labeled preference dataset (e.g., preference $A \prec B \prec C$ annotated by a labeler) and then use the reward model to provide supervision for the subsequent reward optimization stage. In this section, we first present the problem setup, where we additionally consider multi-objective rewards and user preferences in the framework. Then, we present our algorithm, the Directional Preference Alignment, to handle the problem of preference-aware alignment.

**Notation.** We denote the prompt space and the response space as $\mathcal{X}$ and $\mathcal{Y}$, respectively. $\mathbb{S}^k = \{v \in \mathbb{R}^k : \|v\|_2 = 1\}$ is the unit sphere under the $\|\cdot\|_2$ norm. We use $\pi_\theta$ to denote the policy (generative) LLM whose parameter is $\theta$.

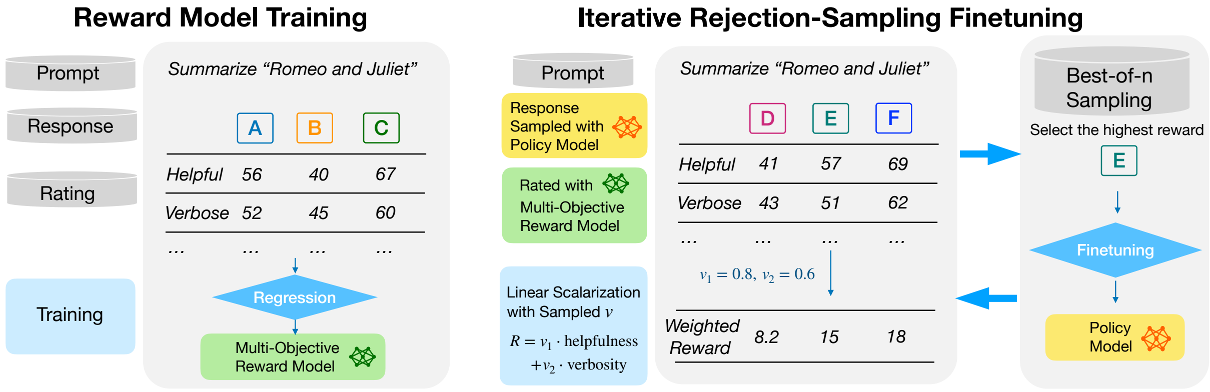


Figure 3: Illustration of the Directional Preference Alignment procedure

### 2.1 Multi-Objective Reward Model

We consider $k$-objective reward for a response $y$ given prompt $x$ as

$$r(x, y) = \langle r_1(x, y), \ldots, r_k(x, y) \rangle \in \mathbb{R}^k$$

where each $r_i(x, y)$ is the rating for a single attribute such as helpfulness, correctness, and verbosity. We use $r$ to denote $r(x, y)$ for short when it is clear from the context. Let $\mathcal{D}_r$ denote the distribution of $(x, y, r)$ (Wang et al., 2023c; Köpf et al., 2023). We then train a multi-objective reward model $\tilde{r}$ with regression loss (Dong et al., 2023b):

$$\min_{\tilde{r}} \mathbb{E}_{(x,y,r)\sim\mathcal{D}_r} \|\tilde{r}(x, y) - r(x, y)\|_2^2. \quad (1)$$

The trained reward model $\tilde{r}$ can rate any prompt-response pair $(x, y)$ across $k$ attributes.

### 2.2 Directional Preference Alignment

Our work aims to learn a collection of policies that can traverse the Pareto front as efficiently as possible. Moreover, we intend to relate the learned policies to the user's preferences concerning various objectives and control the learning process according to such preferences. To make multi-objective optimization tractable and controllable, a common approach is *linear scalarization* (Caruana, 1997; Ghane-Kanafi and Khorram, 2015; Hu et al., 2023), which takes a linear combination of multiple objectives. Through exploring all different linear combinations, the solutions to these problems can sufficiently cover a significant area of the Pareto front, which justifies the application of the linear scalarization approach.

**Directional Preference.** To achieve a fine-grained representation of the preference signal, we model user preference as a *direction* in the multi-objective reward space, that is, a unit vector $v = \langle v_1, \ldots, v_k \rangle \in \mathbb{S}^k$. Then, the preference-conditioned reward is

$$R(x, v, y) = v^\top r(x, y) = \sum_{i=1}^{k} v_i r_i(x, y). \quad (2)$$

To incorporate user preference into the language model, we condition the text generation on $v$ in addition to $x$, such that the response is generated according to $y \sim \pi_\theta(\cdot|x, v)$. For a specific $v$, the preference-conditional reward objective is

$$J(v, \pi_\theta) = \mathbb{E}_{x\sim\mathcal{D}_x, y\sim\pi_\theta(\cdot|x,v)}[R(x, v, y)] \quad (3)$$

We model the directional preferences of our targeted user population as $\mathcal{P}_v$, a probability distribution over $\mathbb{S}^n$. Finally, we optimize $\theta$ by maximizing the expected reward with respect to $\mathcal{P}_v$:

$$\max_\theta \mathbb{E}_{v\sim\mathcal{P}_v} [J(v, \pi_\theta)] . \quad (4)$$

**Reward Optimization via Rejection Sampling.** We now proceed to discuss the algorithmic designs for optimizing the RL objective in Eq. (4). While PPO is the most predominant approach for a fixed reward function (OpenAI, 2023; Anthropic, 2023), it is known that PPO is unstable and sample-inefficient in aligning LLMs (Choshen et al., 2019) and imposes a heavy burden on GPU memory resources (Ouyang et al., 2022; Yuan et al., 2023). Hence, PPO requires extensive efforts to be tuned to its best performance. In light of the above limitations, we resort to an alternative approach, *Rejection Sampling Fine-tuning* (RSF) (Dong et al., 2023a; Yuan et al., 2023; Gulcehre et al., 2023), a RLHF algorithm used in the Llama 2 project (Touvron et al., 2023), with appealing simplicity, stability, and comparable reward gains. In essence, the original RSF learns from the best-of-$n$ policy created by the reward function. Initially, we generate $n$ responses using a base LLM and then rank them using the reward model to select the responses with the highest reward. We further finetune our LLM based on these selected samples, and this process can be repeated multiple times.

In our scenario, to address the multi-objective nature and user-dependent preferences, we iteratively alternate among the following steps for $t = 1, \ldots, T$ iterations:

0. **Preparation.** Initialize an empty dataset $\mathcal{D}_t = \emptyset$. Prepare policy model $\pi_{\theta_{t-1}}$ obtained from last iteration.
1. **Rejection Sampling.** For each randomly sampled prompt $x$ and directional preference $v$, generate $n$ responses $\{y_1, \ldots y_n\}$ by $\pi_{\theta_{t-1}}(\cdot|x, v)$ and compute their multi-objective rewards by $\tilde{r}(x, y)$. Obtain the linear scalarization of $\tilde{r}(x, y)$ by $R(x, v, y_i) = v^\top \tilde{r}(x, y_i)$. Then, rank $y_1, ..., y_n$ according to $R(x, v, y_i)$ and select the highest-rank response $y^\star$. Add $(x, v, y^\star)$ to $\mathcal{D}_t$.
2. **Finetuning.** Train on $\mathcal{D}_t$:

$$\theta_t \leftarrow \arg\max_\theta \mathbb{E}_{(x,v,y)\sim\mathcal{D}_t}[\pi_\theta(y|x, v)].$$

The whole procedure of our methods is summarized in Figure 3.

### 2.3 Discussion with Existing Methods

**Comparison with SteerLM (Dong et al., 2023b).** Recall that we have multi-objective reward $r = \langle r_1, r_2, ..., r_k\rangle$ of each response $y$ to the prompt $x$. Dong et al. (2023b) first fine-tunes the generative model to maximize the likelihood of $y$ by taking both $x$ and $r$ as the input prompts:

$$\max_\theta \mathbb{E}_{(x,y,r)\sim\mathcal{D}_r} \log P_\theta(y|x, r).$$

When presented with a new input $\bar{x}$, SteerLM aims to produce a response that aligns with the newly assigned multi-dimensional $\bar{r}$. Particularly, a user could specify $\bar{r}$ as "($\texttt{helpfulness} = 10, \texttt{verbosity} = 1$)", namely high helpfulness but low verbosity, for a new prompt $\bar{x}$ = "Please summarize 'Romeo and Juliet'". SteerLM could then generate answers according to $\bar{r}$. However, SteerLM will encounter a significant challenge when a user-specified $\bar{r}$ falls outside the feasible region of rewards for the given $\bar{x}$, i.e., $\bar{r} \notin \{r : (\bar{x}, y, r) \in \mathcal{D}_r\}$. In this case, if a user sets a $\bar{r}$ that is not achievable given $\bar{x}$, SteerLM may generate uncontrolled responses due to the infeasibility of $\bar{r}$ under $\bar{x}$. For example, "($\texttt{helpfulness} = 10, \texttt{verbosity} = 1$)" could be infeasible for $\bar{x}$ according to the set $\mathcal{S}$ since it will be difficult or impossible to generate a helpful summarization of 'Romeo and Juliet' in very few words.

**Comparison with Soup Methods (Rame et al., 2023; Jang et al., 2023).** Soup methods trains a policy $\theta_i$ for each reward objective. Let $r_i(x, y)$ denote the $i$-th objective, we have:

$$\theta_i = \arg\max_\theta \mathbb{E}_{x\sim\mathcal{D}_x}\mathbb{E}_{y\sim\pi_\theta(\cdot|x)} r_i(x, y)$$

During inference, when a user specifies the combination vector $\langle v_1, v_2, ..., v_k\rangle \in \mathbb{S}^k$, reward soups first combine the weight of $k$ models as their interpolation $\sum_i v_i\theta_i$ and then query the interpolation for response. Compared with our method, rewarded soup can cause significant storage and computation overhead because they need to maintain $k$ LLMs and calculate different interpolations whenever a new combination vector is assigned.

## 3 Empirical Results

We conduct experiments on Mistral-7B (Jiang et al., 2023), focusing on two reward objectives: `helpfulness` and `verbosity`. Our proposed DPA achieves arithmetic control of LLM generations for different helpfulness-verbosity preferences while demonstrating an excellent balance between the two objectives.

**Verbosity Bias.** Recently, the verbosity bias in LLMs and humans, meaning that LLMs and humans sometimes prefer more verbose answers even though they are of similar qualities, has attracted considerable attention (Saito et al., 2023; Singhal et al., 2023b). It has been exploited or even "hacked" by the RLHF-aligned models. For instance, Kabir et al. (2023) demonstrated that 77% of ChatGPT answers are verbose, while Yuan et al. (2024) found that the average output length increases to 2.5 times as the DPO iterates. Preliminary experiments have been conducted in response to this bias, such as those by Chen et al. (2024), which explicitly consider verbosity as a response feature. Benchmark creators like AlpacaEval (Li et al., 2023) and MT-Bench (Zheng et al., 2023) have observed verbosity bias in their LLM judges

| Alignment Methods | Multi-objective Rewards | Preference Arithmetic | Single Model | Feasibility Guarantee |
|---|---|---|---|---|
| PPO (Schulman et al., 2017) | ✗ | ✗ | ✓ | ✓ |
| DPO (Rafailov et al., 2023) | ✗ | ✗ | ✓ | ✓ |
| Reward Soup (Rame et al., 2023) | ✓ | ✓ | ✗ | ✓ |
| SteerLM (Dong et al., 2023b) | ✓ | ✓ | ✓ | ✗ |
| Ours | ✓ | ✓ | ✓ | ✓ |

Table 1: Comparison among different RLHF algorithms. **Multi-objective rewards**: if the algorithm considers multiple reward objectives. **Preference arithmetic**: if the model allows for arithmetic control of the preference. **Single model**: if the algorithm can handle different preferences with a single LLM. **Feasibility Guarantee**: Whether the model is free from the feasibility issue that the specified control vector (prompt) could be unreachable (refer to Section 2.3 for details).

(typically GPT-4), and AlpacaEval-2.0 has adjusted to account for output length[1].

### 3.1 Implementation

**Datasets.** We use two datasets for experiments: HelpSteer and UltraFeedback. Both datasets are used for reward model training[2], while only UltraFeedback is used for finetuning.

- **HelpSteer** Wang et al. (2023d) comprises 10K prompts and 37K annotated responses with five attributes: `helpfulness`, `correctness`, `coherence`, `complexity`, and `verbosity`. A 43B closed-source LLM generated responses, and human labelers annotated each response on a scale of 0-4 for the five attributes.
- **UltraFeedback** (Cui et al., 2023) includes 64K prompts, each of them are associated with 4 responses of five attributes: `honesty`, `truthfulness`, `instruction-following`, `helpfulness` and `overall-score`. GPT-4 was employed to label these responses. We use the same training-validation prompt split[3] as Zephyr (Tunstall et al., 2023).

**Reward Modeling.** We train a multi-objective reward model on the union of HelpSteer and UltraFeedback, initializing with Mistral-7B. Specifically, we follow SteerLM-v2 practices[4] (Wang et al., 2023c), attaching a linear regression head layer on the last hidden state of Mistral-7B. We include both regression and traditional language modeling losses in the reward model training, as we find the latter improves accuracy without additional observed costs. The reward model has 10 output dimensions: the first half corresponds to HelpSteer's five attributes, while the other half accounts for UltraFeedback's attributes. Rewards in each dimension are rescaled to the range of 0-100 in the data preprocessing stage.

**Alignment Setup.** For a fair comparison with DPO (Rafailov et al., 2023), we conduct a head-to-head comparison with Zephyr-$\beta$ (Tunstall et al., 2023), a DPO-trained Mistral-7B model that was state-of-the-art (7B) at its release. Zephyr-$\beta$ uses supervised fine-tuning (SFT) on UltraChat-200K (Ding et al., 2023) followed by DPO on UltraFeedback (Cui et al., 2023). Since RLHF typically begins with SFT models, we initialize with the SFT checkpoint of Zephyr-$\beta$ and apply DPA on UltraFeedback. Following practices of Cui et al. (2023); Tunstall et al. (2023), we average `instruction-following`, `truthfulness`, `honesty`, and `helpfulness` ratings of UltraFeedback for the overall *helpfulness* objective. We use HelpSteer's `verbosity` attribute for the verbosity objective. Our multi-objective reward model annotates *helpfulness* and *verbosity* for all UltraFeedback data and self-generated responses.

**Rewards and Directional Preferences.** We denote the reward objectives for helpfulness and verbosity as $r_1$ and $r_2$, respectively. As noted by Singhal et al. (2023b), $r_1$ and $r_2$ correlate positively. Therefore, aligning an LLM to maximize $r_1$ (helpfulness) will also tend to increase $r_2$ (verbosity), a trend documented in recent works (Yuan et al., 2024; Chen et al., 2024). Consequently, when using the preference-conditional reward $v^\top r = v_1 r_1 + v_2 r_2$, we argue that it is unnecessary to have $v_2 > 0$ (i.e., to explicitly encourage verbosity). Instead, we propose sampling the distribution of $\langle v_1, v_2 \rangle$ as $\arctan(\frac{v2}{v1}) \sim \text{Uniform}(-\frac{\pi}{4}, 0)$ with $v_1 \in [\sqrt{2}/2, 1]$ and $v_2 \in [-\sqrt{2}/2, 0]$. Intuitively,

[1] `tatsu-lab.github.io/alpaca_eval/`

[2] We include HelpSteer since it has `verbosity` annotations.

[3] `hf.co/datasets/HuggingFaceH4/ultrafeedback_binarized`

[4] The authors of SteerLM (Dong et al., 2023b) improved the original training recipe in a follow-up work (Wang et al., 2023c), which we denote as SteerLM-v2.

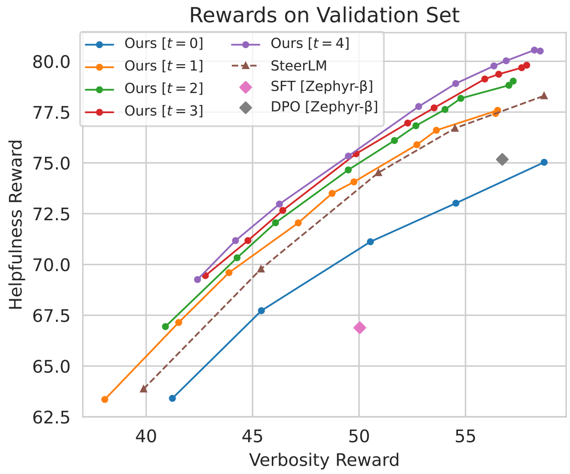


Figure 4: The validation reward of different methods. When $t \geqslant 1$, our DPA model Pareto-dominates SFT, DPO, and SteerLM. Further, DPA at iteration $t$ Pareto-dominates models at previous iteration $t'$ with $t' < t$.

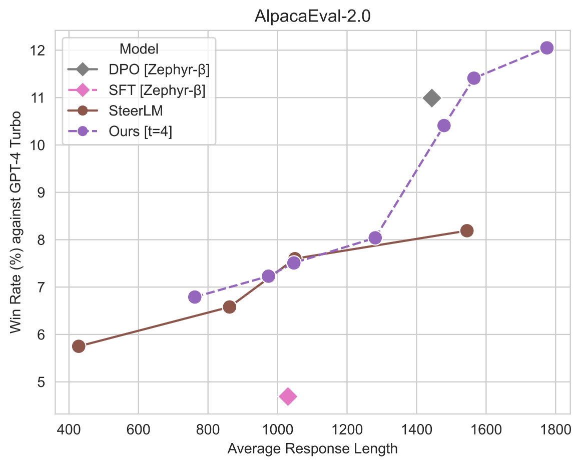


Figure 5: AlpacaEval-2.0 evaluation results.

this lets the user preference direction $\langle v_1, v_2 \rangle$ be uniformly sampled between $\langle 1, 0 \rangle$ (pure focus on helpfulness) and $\langle \sqrt{2}/2, -\sqrt{2}/2 \rangle$ (a balance favoring less verbosity) on the unit circle.

**Dataset Splitting.** Iterative RLHF methods typically sample responses for *unseen* prompts in each new iteration to prevent the model from simply memorizing and repeating the responses (Dong et al., 2023a; Xiong et al., 2023; Yuan et al., 2024). In view of this, we split UltraFeedback dataset into two disjoint subsets, $\mathcal{D}_1$ and $\mathcal{D}_2$, containing an equal number of unique prompts. In each iteration $t$, we initialize the policy model $\pi_{\theta_t}$ from an SFT checkpoint rather than $\pi_{\theta_{t-1}}$, and we use a different subset from the last iteration. The use of alternative subsets ensures that the policy model $\pi_{\theta_t}$ for response sampling in iteration $t+1$ has not encountered the prompts before.

**Rejection Sampling.** We conduct rejection sampling following our iterative algorithm detailed in Sec. 2.2. Notice that to launch training in $t = 1$, we need $\pi_{\theta_{t=0}}$ for sampling responses for a diverse set of helpfulness-verbosity preferences. However, Zephyr-$\beta$-SFT is not designed for preference-conditional generation, making it not a good choice for $\pi_{\theta_{t=0}}$. To resolve this, we train a SteerLM model on $\mathcal{D}_2$ (a half of UltraFeedback) that can generate responses conditioned on both user prompt $x$ (sampled from $\mathcal{D}_1$) and reward objectives $r_1, r_2$. We use this model for rejection sampling in iteration $t = 1$ to obtain $\pi_{\theta_1}$ (for each prompt, we generate 80 responses for diverse reward combinations $(r_1, r_2)$). In all the following iterations, for each prompt, we sample 5 directional preferences $\langle v_1, v_2 \rangle$, and use $\pi_{\theta_{t-1}}$ to generate 16 responses per preference, then keep the highest-reward response and reject the rest 15.

**Fine-tuning.** For the response data obtained through rejection sampling, we prepend the user's directional preference to the system prompt, as illustrated in Fig. 1, to make the model aware of the user preference. The fine-tuning process then follows the same approach as SFT, optimizing the next-token prediction loss across the text corpus. It is also worth noting that RLHF often leads to performance degradation or knowledge forgetting, a phenomenon referred to as *alignment tax* in the literature (Askell et al., 2021; Lin et al., 2023). To mitigate this issue, we adopt the memory replay techniques suggested in Instruct-GPT (Ouyang et al., 2022) and Llama 2 (Touvron et al., 2023) that can effectively reduce alignment tax (Lin et al., 2023). Specifically, we incorporate original responses from UltraFeedback, which constitute about 15% of our finetuning data for each iteration. Our algorithm is applied for iterations $t = 1, \ldots, 4$.

**Software, Hardware and Hyperparameters** We use PyTorch (Paszke et al., 2019) with HuggingFace's TRL framework (von Werra et al., 2020) for all fine-tuning experiments across $t = 0, \ldots, T$. All experiments are conducted on 8x A6000 GPUs. The training cost of each DPA iteration is about 60 GPU hours. The AdamW optimizer (Loshchilov and Hutter, 2019) is employed with a learning rate of $10^{-5}$ and a cosine learning rate schedule (20 warmup steps). We use a context window of 4096 tokens with sample-packing (packing short responses within the context window). The training takes 2 epochs with a global batch size of 64. We use vLLM (Kwon et al., 2023) for inference. In the rejection sampling process, we conduct inference with temperature $1.0$. In evaluation (Sec. 3.2), we use temperature $0.7$.

### 3.2 Evaluation

**Rewards on Validation Set** For validation, we used 2000 prompts from UltraFeedback and considered 10 uniformly sampled directional preferences ranging from $v = \langle 1, 0 \rangle$ to $v = \langle \sqrt{2}/2, \sqrt{2}/2 \rangle$. For each prompt-preference combination, our DPA-aligned models generated two responses. We then calculated the average *helpfulness* and *verbosity* rewards for all 2000 responses per preference using our reward model. For SteerLM[5], five *verbosity* reward values were sampled, and the highest corresponding *helpfulness* reward from UltraFeedback was identified for each value. These verbosity-helpfulness pairs were then used to condition SteerLM's generation, with the average rewards computed across prompts. In the case of Zephyr-$\beta$'s DPO and SFT models, we generated responses using their original prompt templates and averaged the rewards across the validation set. The results, illustrated in Fig. 4, show that as $t \geqslant 1$, our DPA model Pareto-dominates SFT, DPO, SteerLM, and DPA at iteration $t$ Pareto-dominates the models of previous iterations. This demonstrates DPA's effective arithmetic control for different user preferences, and with increasing finetuning iterations $t$, the *empirical front* of DPA (i.e., each curve in Fig. 4) expands, indicating that our finetuning approach successfully maximizes rewards for all user preferences of consideration. Notably, our DPA's empirical front significantly surpasses that of SteerLM and DPO, even though all models were trained on the same UltraFeedback dataset and originated from the same SFT model.

**AlpacaEval-2.0 Evaluation** AlpacaEval-2.0 (Li et al., 2023) is an LLM-based automatic evaluation benchmark that employs GPT-4-turbo as the LLM judge. It includes 805 prompts, and model responses to these prompts are compared with reference answers provided by GPT-4-turbo. Subsequently, the win-rate against the reference answers is calculated as a metric for the models' instruction-following capabilities. We evaluated SteerLM and our DPA (at $t = 4$) conditioned with various user preferences and report the win rate and average response length in Fig. 5, along with DPO and SFT results for reference. Fig. 5 demonstrates that our DPA model outperforms SteerLM and achieves competitive performance against DPO while providing arithmetic control for diverse user preferences. The discrepancy between the validation reward evaluation results and the AlpacaEval-2.0 outcomes may arise because our reward model has different behaviors and preferences compared to GPT-4-turbo. While DPA can closely fit the reward model, this does not necessarily guarantee generalization to GPT-4-turbo evaluations.

## 4 Related Works

**Large Language Models.** The landscape of natural language processing has been profoundly transformed in recent years through the development of large language models (LLMs), showcasing human-level proficiency across a range of tasks including text classification, generation, and complex reasoning. This progress stems from extensive pre-training on vast datasets, enabling these models to address diverse challenges. Despite their achievements, a distinction arises between closed-source models (e.g., GPT-3 (Brown et al., 2020), Bard (Google, 2023), Claude (Anthropic, 2023), and PaLM (Chowdhery et al., 2023)), often surpassing their open-source counterparts (e.g., megatron-turing-530b (Smith et al., 2022), and Bloom (Workshop et al., 2022)) in performance (Liang et al., 2022), which poses challenges for open-source research. However, initiatives like Meta's LLaMA (Touvron et al., 2023) and subsequent works such as Alpaca (Taori et al., 2023), Vicuna (Chiang et al., 2023), and LMFlow (Diao et al., 2023), demonstrate significant open-source contributions that continue to push the boundaries of what's possible with LLMs. These advancements enabled by the fine-tuning techniques, aim to improve LLMs' ability and adapt to a wide range of domains and tasks. Nonetheless, as these generative foundation models advance, they still face problems like implicit biases, underscoring the need for ongoing alignment and ethical considerations in their development and application. In this paper, we focus on how to align LLMs with human preferences, including the principles of being helpful, honest, and harmless as outlined by (Askell et al., 2021). This procedure is often achieved by Rein-

[5]We trained a SteerLM model (initialized with the SFT checkpoint of Zephyr-$\beta$) on UltraFeedback, following practices of Wang et al. (2023c).

forcement Learning with Human Feedback (RLHF) (Ouyang et al., 2022).

**RLHF Algorithmic Designs.** Policy Optimization (PPO) (Schulman et al., 2017) is the most predominant approach, with its tremendous success in Chat-GPT (OpenAI, 2023) and Claude (Anthropic, 2023). However, PPO is significantly less efficient and stable compared to supervised fine-tuning (Choshen et al., 2019), and is also sensitive to the parameter and code-level implementation (Engstrom et al., 2020). Therefore, tuning the PPO to its best performance is very challenging in practice and the results of Chat-GPT (OpenAI, 2023) have not been widely reproduced so far. In view of this, efforts have been made to develop supervised-learning-based methods as an alternative approach to the PPO, and we review them as follows. Rejection sampling finetuning (RSF) is proposed in (Dong et al., 2023a; Yuan et al., 2023; Gulcehre et al., 2023) with different variants, but essentially, they learn from the positive samples selected by a learned reward model. RSF was applied to the RLHF of LLaMA2 project (Touvron et al., 2023) and we adopt the iterative implementation as suggested in Dong et al. (2023a); Touvron et al. (2023); Gulcehre et al. (2023). There is also another line of work designing algorithms from the KL-constraint reward optimization (Rafailov et al., 2023; Zhao et al., 2023; Azar et al., 2023; Xiong et al., 2023), which additionally requires the resulting model to be close to the initial model. Among them, the Direct Preference Optimization (DPO) (Rafailov et al., 2023) has attracted considerable attention due to its simplicity and stability, and effectiveness. We remark that it is also possible to incorporate these algorithmic ideas into our DPA framework and we leave the algorithmic design beyond RSF to future work.

**Fine-grained Preference Representation and Algorithmic design.** The scalar-reward-model has been criticized mainly due to its limited capacity (Wu et al., 2023b; Casper et al., 2023; Munos et al., 2023) (see the discussion of preference intransitivity in Section 1 for an illustrative example). A line of works has considered multi-objective rewards to capture the different aspects of human preferences (Zhou et al., 2023; Jang et al., 2023; Touvron et al., 2023; Wu et al., 2023b; Köpf et al., 2023; Rame et al., 2023). However, the multi-objective rewards are then combined in a fixed way (e.g., Wu et al., 2023b; Touvron et al., 2023), mainly to represent a preference averaged over different human groups, failing to capture the user-dependent preference. By introducing the user preference as a unit vector (direction) into the directional preference alignment framework, we achieve a fine-grained and user-dependent representation for the complicated human preference. Notably, in social choice theory (Sternberg, 1965; Fishburn, 1984), as well as some very recent studies in RLHF (Wang et al., 2023b; Swamy et al., 2024; Ye et al., 2024), the RLHF is formulated as a game between two LLMs to partially handle the diversity of preferences in the population-level. The learning objective is accordingly adjusted to be solving the *Nash equilibrium* of the game. In comparison, our techniques are fundamentally different from theirs and may offer computational advantages since game-based formulation is far more complicated.

## 5 Limitations

A primary constraint of our DPA framework is its reliance on a robust multi-objective reward model. The efficacy of DPA is intrinsically linked to the precision and discriminative capability of this reward model. Should the reward model not adequately capture the subtleties of specific preferences or exhibit bias in its reward distribution, the DPA might inadvertently exacerbate these shortcomings throughout the fine-tuning process. Furthermore, if the reward model fails to recognize harmful content, it could lead the aligned model to produce such content during inference.

## 6 Conclusion

In this paper, we introduce Directional Preference Alignment (DPA) to incorporate multidimensional user preferences. DPA addresses the limitation of conventional scalar reward models by alleviating conflicting user preferences through a high-dimensional preference vector in a multidimensional space. We demonstrate that DPA efficiently explores the Pareto front in the multidimensional reward space, revealing a more effective trade-off between helpfulness and verbosity on Mistral-7B compared to existing strong baselines such as DPO.

## Acknowledgments

HZ is partially supported by a research grant from the Amazon-Illinois Center on AI for Interactive Conversational Experiences (AICE) and a Google Research Scholar Award.

## References


Anthropic. 2023. Introducing claude.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189.

Hritik Bansal, John Dang, and Aditya Grover. 2023. Peering through preferences: Unraveling feedback acquisition for aligning large language models. *arXiv preprint arXiv:2308.15812*.

Erdem Biyik and Dorsa Sadigh. 2018. Batch active preference-based learning of reward functions. In *Conference on robot learning*, pages 519–528. PMLR.

Andreea Bobu, Andi Peng, Pulkit Agrawal, Julie Shah, and Anca D Dragan. 2023. Aligning robot and human representations. *arXiv preprint arXiv:2302.01928*.

Daniel Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. 2019. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International conference on machine learning*, pages 783–792. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28:41–75.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Lichang Chen, Chen Zhu, Davit Soselia, Jiuhai Chen, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Odin: Disentangled reward mitigates hacking in rlhf.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Leshem Choshen, Lior Fox, Zohar Aizenbud, and Omri Abend. 2019. On the weaknesses of reinforcement learning for neural machine translation. *arXiv preprint arXiv:1907.01752*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

CA Coello Coello. 2000. Handling preferences in evolutionary multiobjective optimization: A survey. In *Proceedings of the 2000 congress on evolutionary computation. CEC00 (Cat. No. 00TH8512)*, volume 1, pages 30–37. IEEE.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback.

Shizhe Diao, Rui Pan, Hanze Dong, Ka Shun Shum, Jipeng Zhang, Wei Xiong, and Tong Zhang. 2023. Lmflow: An extensible toolkit for finetuning and inference of large foundation models. *arXiv preprint arXiv:2306.12420*.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, KaShun SHUM, and Tong Zhang. 2023a.

RAFT: Reward ranked finetuning for generative foundation model alignment. *Transactions on Machine Learning Research*.

Yi Dong, Zhilin Wang, Makesh Narsimhan Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. 2023b. Steerlm: Attribute conditioned sft as an (user-steerable) alternative to rlhf. *arXiv preprint arXiv:2310.05344*.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. 2020. Implementation matters in deep policy gradients: A case study on ppo and trpo. *arXiv preprint arXiv:2005.12729*.

Michael Feffer, Hoda Heidari, and Zachary C Lipton. 2023. Moral machine or tyranny of the majority? *arXiv preprint arXiv:2305.17319*.

Peter C Fishburn. 1984. Probabilistic social choice based on simple voting comparisons. *The Review of Economic Studies*, 51(4):683–692.

William V Gehrlein. 2002. Condorcet's paradox and the likelihood of its occurrence: different perspectives on balanced preferences. *Theory and decision*, 52:171–199.

A Ghane-Kanafi and E Khorram. 2015. A new scalarization method for finding the efficient frontier in non-convex multi-objective problems. *Applied Mathematical Modelling*, 39(23-24):7483–7498.

Google. 2023. Bard.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. 2023. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2022. Optimizing prompts for text-to-image generation. *arXiv preprint arXiv:2212.09611*.

Yuzheng Hu, Ruicheng Xian, Qilong Wu, Qiuling Fan, Lang Yin, and Han Zhao. 2023. Revisiting scalarization in multi-task learning: A theoretical perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. Beavertails: Towards improved safety alignment of LLM via a human-preference dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Samia Kabir, David N Udo-Imeh, Bonan Kou, and Tianyi Zhang. 2023. Who answers it better? an in-depth analysis of chatgpt and stack overflow answers to software engineering questions. *arXiv preprint arXiv:2308.02312*.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Minh Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Alexandrovich Glushkov, Arnav Varma Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Julian Mattick. 2023. Openassistant conversations - democratizing large language model alignment. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. Rewardbench: Evaluating reward models for language modeling.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Yong Lin, Lu Tan, Hangyu Lin, Zeming Zheng, Renjie Pi, Jipeng Zhang, Shizhe Diao, Haoxiang Wang, Han Zhao, Yuan Yao, et al. 2023. Speciality vs generality: An empirical study on catastrophic forgetting in fine-tuning foundation models. *arXiv preprint arXiv:2309.06256*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Kenneth O May. 1954. Intransitivity, utility, and the aggregation of preference patterns. *Econometrica: Journal of the Econometric Society*, pages 1–13.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. 2023. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In *International Conference on Machine Learning*, pages 26837–26867. PMLR.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Bruno L Pereira, Alberto Ueda, Gustavo Penha, Rodrygo LT Santos, and Nivio Ziviani. 2019. Online learning to rank for sequential music recommendation. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 237–245.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Alexandre Rame, Guillaume Couairon, Mustafa Shukor, Corentin Dancette, Jean-Baptiste Gaya, Laure Soulier, and Matthieu Cord. 2023. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *arXiv preprint arXiv:2306.04488*.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*.

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

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023a. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2023b. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Saul H Sternberg. 1965. *Mathematics and Social Sciences: Proceedings of the Seminars of Menthon-Saint-Bernard, France (1-27 July, 1960) and of Gösing, Austria (3-27 July, 1961)*, volume 1. Mouton.

Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. 2024. A minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint arXiv:2401.04056*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

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

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Amos Tversky. 1969. Intransitivity of preferences. *Psychological review*, 76(1):31.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl.

Benyou Wang, Qianqian Xie, Jiahuan Pei, Zhihong Chen, Prayag Tiwari, Zhao Li, and Jie Fu. 2023a. Pre-trained language models in biomedical domain: A systematic survey. *ACM Computing Surveys*, 56(3):1–52.

Yuanhao Wang, Qinghua Liu, and Chi Jin. 2023b. Is rlhf more difficult than standard rl? *arXiv preprint arXiv:2306.14111*.

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Polak Scowcroft, Neel Kant, Aidan Swope, et al. 2023c. Helpsteer: Multi-attribute helpfulness dataset for steerlm. *arXiv preprint arXiv:2311.09528*.

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Polak Scowcroft, Neel Kant, Aidan Swope, et al. 2023d. Helpsteer: Multi-attribute helpfulness dataset for steerlm. *arXiv preprint arXiv:2311.09528*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023a. Better aligning text-to-image models with human preference. *arXiv preprint arXiv:2303.14420*.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023b. Fine-grained human feedback gives better rewards for language model training. *arXiv preprint arXiv:2306.01693*.

Wei Xiong, Hanze Dong, Chenlu Ye, Han Zhong, Nan Jiang, and Tong Zhang. 2023. Gibbs sampling from human feedback: A provable kl-constrained framework for rlhf. *arXiv preprint arXiv:2312.11456*.

Chenlu Ye, Wei Xiong, Yuheng Zhang, Nan Jiang, and Tong Zhang. 2024. A theoretical analysis of nash learning from human feedback under general kl-regularized preference. *arXiv preprint arXiv:2402.07314*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. 2023. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Zhanhui Zhou, Jie Liu, Chao Yang, Jing Shao, Yu Liu, Xiangyu Yue, Wanli Ouyang, and Yu Qiao. 2023. Beyond one-preference-for-all: Multi-objective direct preference optimization. *arXiv preprint arXiv:2310.03708*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

## A Additional Experiments

RewardBench (Lambert et al., 2024) is a newly released benchmark suite for RLHF, comprising 23 validation datasets that are categorized into Chat, Chat Hard, Reasoning, and Safety. RewardBench was originally designed to benchmark the performance of reward models on challenging, structured, and out-of-distribution prompts. However, it can also be used to evaluate the generation quality of aligned LLMs. Here is our evaluation protocol:

1. We selected a reference model and multiple candidate models, and generated responses for each prompt from the 23 validation datasets of Reward-Bench using these models.

2. We employed a Mistral-7B Bradley-Terry reward models[6] to assess the generated responses for each model. The reward model outputs a score for each response given a prompt.

3. We compared each candidate model against the reference model using the reward model's ratings. For each prompt, the model that generated the response with the higher score was considered the winner. We then averaged the results over the datasets to compute the win-rate of each candidate model against the reference model.

We demonstrate our empirical evaluation results in the following two parts. Further, we will update our manuscript with Reward-Bench evaluations in the next revision.

The RLHF literature has demonstrated that helpfulness and harmlessness are two objectives that sometimes exhibit a tradeoff (Rame et al., 2023), making them suitable for our study. We have added the harmlessness objective to our original objectives of helpfulness and verbosity, resulting in a total of three objectives. To train a DPA model with these three objectives, we first train a three-objective reward model, following the same training recipe detailed in our paper, with the addition of one dataset: BeaverTails-30K (Ji et al., 2023). This dataset has an `is-safe` attribute with scores ranging from 0 to 1, averaged over three human annotators. After training, we use the attribute prediction for the `is-safe` attribute (multiplied by 100) as the harmlessness score.

We initialized our model from a Supervised Fine-Tuning (SFT) checkpoint of Gemma-2B[7] and adopted the same training procedures detailed in the paper to train Gemma-2B with rejection sampling finetuning. The only difference in training is the introduction of the `harmlessness` objective, thus the directional preference becomes $v = (v_1, v_2, v_3)$, where the three dimensions correspond to **`helpfulness`**, **`verbosity`**, and **`harmlessness`** and $\|v\| = 1$. In the sampling stage, we randomly sample the directional preference vectors while ensuring that $v_1 > v_3 \geqslant 0$.

For comparison, we used a DPO checkpoint[8] of Gemma-2B as the reference model, which was initialized from the same SFT checkpoint. We evaluated our Gemma-2B DPA models ($T = 3$) with different directional preferences on Reward-Bench (using the protocol explained above), comparing their Win-Rate against the DPO model.

| Directional Preference | Chat | Chat Hard | Reasoning | Safety |
|---|---|---|---|---|
| $v = (0.71, 0, 0.71)$ | **56.17** | **58.13** | **49.87** | 63.94 |
| $v = (0.79, 0, 0.61)$ | 44.63 | 49.63 | 44.61 | **67.21** |
| $v = (0.87, 0, 0.50)$ | 46.15 | 51.5 | 44.83 | 65.62 |
| $v = (0.92, 0, 0.38)$ | 36.91 | 47.23 | 45.41 | 65.92 |
| $v = (0.97, 0, 0.26)$ | 45.91 | 47.56 | 45.97 | 65.83 |
| $v = (0.99, 0, 0.13)$ | 47.52 | 51.88 | 45.33 | 66.17 |
| $v = (1, 0, 0)$ | 46.78 | 48.19 | 46.58 | 66.67 |

Here, we fix $v_2$ (corresponding to `verbosity`) and vary the directional preference $v$ for the remaining two dimensions (`helpfulness` and `harmlessness`). Clearly, there exist directional preferences $v$ such that the DPA responses are better than or on par with DPO (Win-Rate = 50% means tie).

[6] https://hf.co/weqweasdas/RM-Mistral-7B

[7] https://hf.co/wandb/gemma-2b-zephyr-sft

[8] https://hf.co/wandb/gemma-2b-zephyr-dpo