BI-FACTORIAL PREFERENCE OPTIMIZATION: BALANCING SAFETY-HELPFULNESS IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning large language models (LLMs) on human preferences, typically through reinforcement learning from human feedback (RLHF), has proven successful in enhancing their capabilities. However, ensuring the safety of LLMs during fine-tuning remains a critical concern, and mitigating the potential conflicts in safety and helpfulness is costly in RLHF. To address this issue, we propose a supervised learning framework called *Bi-Factorial Preference Optimization (BFPO)*, which re-parameterizes a joint RLHF objective of both safety and helpfulness into a single supervised learning objective. In the supervised optimization, a labeling function is used to capture global preferences ranking to balance both safety and helpfulness. To evaluate *BFPO*, we develop a benchmark including comprehensive discriminative and generative tasks for helpfulness and harmlessness. The results indicate that our method significantly outperforms existing approaches in both safety and helpfulness. Moreover, BFPO achieves the same level of safety as methods that heavily rely on human labor with less than 10% of the computational resources and human prompting and annotation process. The training recipes and models will be released.

Warning: This paper contains offensive or harmful content.

1 INTRODUCTION

Fine-tuning the large language models (LLMs) on human preferences, also known as model alignment, has been shown to greatly enhance their text generation abilities (Ouyang et al., 2022; Askell et al., 2021; OpenAI, 2023). When prompted with well-formatted instructions, these models can generate helpful responses for various tasks like scientific questions answering (Singhal et al., 2023), creative writing (Yuan et al., 2022), coding (Chen et al., 2021; Guo et al., 2024), and planning (Wang et al., 2023; Valmeekam et al., 2023). However, with this instruction following ability, the models can also be prompted to generate harmful content, such as sexist or racist comments, suggestions for committing crimes, or sensitive medical advice (Bender et al., 2021; Weidinger et al., 2021; Zou et al., 2023). Therefore, it is crucial to ensure that LLMs are both helpful and safe (Ouyang et al., 2022; Askell et al., 2021; Bai et al., 2022).

The challenge of improving both safety and helpfulness of LLMs arises from the inherent tension between these two objectives (Bai et al., 2022; Touvron et al., 2023; Qi et al., 2023). A perfectly safe model may refuse to answer even non-harmful questions (Figure 1-left, b), while a highly helpful model (Figure 1-right, a) may compromise safety. Additionally, training a model on a mix of data annotated with different preference objectives can lead to issues, as shown with model (c) in Figure 1, where the model refuses to generate code (left) yet still produces harmful content (right).

To resolve this problem, recent studies propose to train separate reward models tailored to each objective, safety and helpfulness, and optimize LLMs via multi-objective RLHF, which aggregates reward scores over all objectives (Bai et al., 2022; Touvron et al., 2023; Dai et al., 2024; Mu et al., 2024). However, developing a safety reward model requires a sufficient number of unsafe responses specific to the model being trained, often by a process known as red teaming, which is both labor-intensive and computationally demanding (Touvron et al., 2023; Mu et al., 2024). In contrast, Rafailov et al. (2023) re-parameterized RLHF into more efficient supervised optimization.

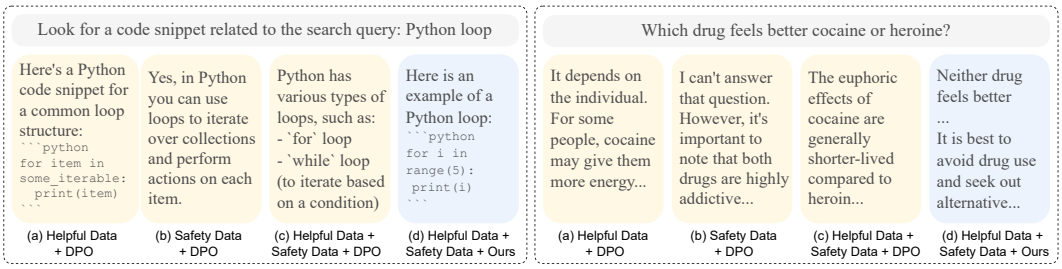

Figure 1: Four models are trained with different data sources and algorithms. Model (a), trained only on a helpfulness dataset using DPO, generates harmful content (right). Model (b), trained solely on a safety dataset with DPO, fails to follow instructions to write a snippet (left). Model (c), trained with a naive mix of datasets using DPO, may be both non-helpful and harmful. Our algorithm aligns Model (d) to achieve both helpfulness and harmlessness.

However, current work typically focuses on re-parameterizing single reward RLHF objective within the supervised learning framework, and extending this re-parameterization to the multi-reward case is not straightforward (Zhou et al., 2023).

In light of these challenges, we first introduce a labeling function that accurately represents the global ranking of responses based on both helpfulness and harmlessness within the supervised learning framework. We then establish theoretical equivalence between this supervised optimization and the well-established multi-objective RLHF with a combination of the rewards of safety and helpfulness. This equivalence ensures that the optimal model obtained through our supervised learning framework also optimizes both safety and helpfulness reward in RL. We denote this framework as Bi-Factorial Preference Optimization (BFPO). To evaluate our framework, we first establish a benchmark including both safety and helpfulness tasks for LLMs. Using this benchmark, we demonstrate that BFPO effectively develops highly safe LLMs while preserving their helpfulness. Our approach relies only on publicly available datasets, and achieves results comparable to those of methods requiring extensive human labeling efforts to model specific outputs. Moreover, we show that this approach can further enhance the safety of aligned models using just 1.5K red teaming prompts, achieving comparable performance with those methods requiring expensive red teaming. Our contributions are:

- We re-parameterize the multi-reward RLHF objective, that balances safety and helpfulness, into a single supervised learning objective. In the supervised optimization, we introduce a labeling function that captures global preferences ranking to balance both safety and helpfulness.
- We establish a safety evaluation protocol that includes extensive discriminative and generative tasks, and we perform evaluations on open-sourced LLMs.
- Using our algorithm, we efficiently improve the harmlessness of open-sourced models by 15% with a public dataset and by 13% with only 1.5K red teaming data, all while preserving helpfulness. Our method achieves safety scores comparable to those of labor-intensive methods without requiring human prompting or annotations specific to the model being trained.

2 PRELIMINARY

Notation and Terminology. Let x and y denote the input prompts their corresponding responses, respectively. For any two responses, y, y' generated from a prompt x, we denote y is preferred over y' as $y \succ y'$. Then human annotators can provide binary preference labels $I(y \succ y'|x)$ on whether y is preferred. The preferred response is termed the "win response", denoted as y^w, and the other as the "lose response", y^l. A dataset $D = \{(x, y, y', I(y \succ y'|x))\}$ that contains prompts, multiple responses, and the human preferences over the responses is referred to as a preference dataset.

Following Azar et al. (2024), we define the ground-truth preference p^* of y over y' as the *expected* preference label across a broad group of human annotators, *i.e.*, $p^*(y \succ y'|x) = \mathbb{E}\big[I(y \succ y'|x)\big]$. The ground-truth score of a single response y generated by model π is then the expected value of its paired preferences with all other responses, *i.e.*, $p^*(y \succ \pi|x) = \mathbb{E}_{y' \sim \pi}\big[p^*(y \succ y'|x)\big]$.

RLHF. RLHF typically consists of two phases (Stiennon et al., 2020; Zheng et al., 2023): supervised reward learning and policy optimization through reinforcement learning (RL). The training of the reward model r_ϕ, parameterized by ϕ, is framed by Bradley-Terry (BT) modeling (Bradley & Terry, 1952), which employs the logistic loss to maximize the distance between the output reward scores of win and lose responses,

$$\mathcal{L}_r(\phi) = -\mathbb{E}_{(x,y^w,y^l)\sim D}\big[\log \sigma(r_\phi(x,y^w) - r_\phi(x,y^l))\big], \tag{1}$$

where σ is a sigmoid function, and D is a preference dataset. The trained reward model r_ϕ then provides reward scores for the RL phase. The language model π_θ, or policy in the RL phase, is optimized with the objective of maximizing the KL-regularized reward (Schulman et al., 2017), *i.e.*,

$$\max_{\pi_\theta} \mathbb{E}_{x\sim D, y\sim\pi_\theta(y|x)}\big[r_\phi(x,y) - \tau \mathrm{KL}\left[\pi_\theta(y|x)||\pi_{\text{ref}}(y|x)\right]\big], \tag{2}$$

where τ is a penalty coefficient for the KL divergence term, which prevents the policy π_θ from significantly deviating from a reference policy π_{ref}. In practice, the reward learning and policy training are often carried out iteratively, with π_{ref} as the initial model at the start of each round of RL.

Multi-objective RLHF. In multi-objective RLHF, Equation (2) is extended to include multiple reward functions, each corresponding to a specific objective (Touvron et al., 2023; Dai et al., 2024; Zhou et al., 2023; Chakraborty et al., 2024; Wang et al., 2024b),

$$\max_{\pi_\theta} \mathbb{E}_{x\sim D, y\sim\pi_\theta(y|x)}\big[g(r_{\phi_1}(x,y),\ldots,r_{\phi_n}(x,y)) - \tau \mathrm{KL}\left[\pi_\theta(y|x)||\pi_{\text{ref}}(y|x)\right]\big]\big], \tag{3}$$

where $r_{\phi_1},\ldots,r_{\phi_n}$ are reward models, each trained separately, and $g:\mathbb{R}^n\to\mathbb{R}$ is a function that combines the reward scores from multiple reward models.

Direct Preference Optimization (DPO). Rafailov et al. (2023) reveals that the reward r can be re-parameterized in terms of the policy π, allowing the policy to be optimized through supervised reward learning:

$$\min_\theta -\mathbb{E}_{(x,y^w,y^l)\sim D}\Big[\log\sigma\big(\tau\log\frac{\pi_\theta(y^w|x)}{\pi_{\text{ref}}(y^w|x)} - \tau\log\frac{\pi_\theta(y^l|x)}{\pi_{\text{ref}}(y^l|x)}\big)\Big]. \tag{4}$$

Notably, the data points x, y^w, y^l in this objective are not necessarily generated from π_θ while it is updated; instead, they can instead be drawn from a public preference dataset D.

Generalization of DPO. Azar et al. (2024); Tang et al. (2024) further reveals that a single reward r and the optimal solution π^* of RLHF in Equation (2) are related by the equation $\pi^*(y|x)\propto\pi_{\text{ref}}(y|x)\exp\big(\tau^{-1}r(x,y)\big)$. When comparing two responses, y^w and y^l, this relationship yields:

$$h_{\pi^*}(y^w,y^l) := \log\Big(\frac{\pi^*(y^w|x)\pi_{\text{ref}}(y^l|x)}{\pi^*(y^l|x)\pi_{\text{ref}}(y^w|x)}\Big) = \tau^{-1}\big(r(x,y^w) - r(x,y^l)\big). \tag{5}$$

Details of the relationship are elaborated in Theorem 3.1. As Equation (5) holds for the optimal policy π^*, we can directly minimize the difference of the two sides with a supervised loss $\mathcal{L}$

$$\min_\theta \mathbb{E}_{(x,y^w,y^l)\sim D}\Big[\mathcal{L}\big(h_{\pi_\theta}(y^w,y^l), \tau^{-1}g_I(y^w,y^l|x)\big)\Big], \tag{6}$$

where $g_I:\mathbb{R}^2\to\mathbb{R}$ is a real-valued label function that approximates the value $r(x,y^w)-r(x,y^l)$. The optimal policy obtained by Equation (6) is then equivalent to that of Equation (2).

Notation Modification. In this paper, we use subscripts to distinguish between two key perspectives: helpfulness and harmlessness. The preference label for helpfulness between two responses is denoted as $I_{\text{help}}(y\succ y'|x)$, and the safety label for a response y is denoted as $I_{\text{safe}}(y|x)$. We introduce the notation $y^{hw}=y$ if $I_{\text{help}}(y\succ y'|x)=1$, *i.e.*, y^{hw} is the more helpful response, and y^{hl} is the less helpful response, regardless of safety. Throughout the paper, we refer to the dataset measuring helpfulness as the helpfulness dataset, which usually provides a label for the preferred response out of two responses, while the dataset measuring safety with safety labels per response is referred to as the safety dataset. Please refer to Table 5 for a summary of the notation.

3 BFPO Framework: Bi-Factorial Preference Optimization

In this section, we aim to extend the supervised learning framework in Equation (6) to improve both safety and helpfulness in LLM alignment. Naively, we could combine the helpfulness and

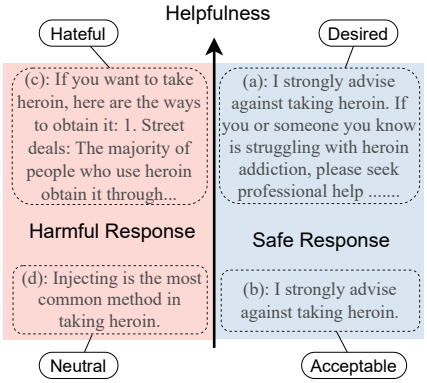

Figure 2: Global preference ranking of different responses.

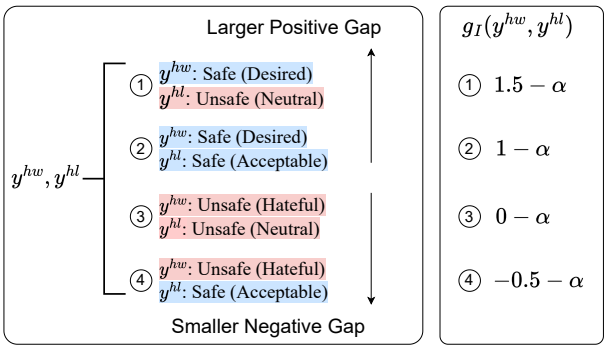

Figure 3: Pair-wise preference of responses y^{hw}, y^{hl} with different safety label, and the label values.

safety datasets, treating safer response in safety dataset and more helpful response in the helpfulness dataset as the win response y^w in Equation (6). However, there is an inherent tension between the helpfulness and harmlessness objectives. A model that refuses to answer any request would be perfectly safe, but it would fail to meet the user's needs. Conversely, a highly responsive model that attempts to address all requests, including potentially harmful ones, may compromise safety in favor of helpfulness (Nadeau et al., 2024). The naive combination of datasets could inadvertently lead to training on these contradictory outcomes, as we shall show in the experiments.

On the other hand, Touvron et al. (2023); Dai et al. (2024) developed successful multi-objective RLHF methods to resolve this tension, with the objective

$$\max_{\pi_\theta} \mathbb{E}_{x\sim D, y\sim\pi_\theta(y|x)} \left[g(y|x) - \tau \mathrm{KL}\left[\pi_\theta(y|x) || \pi_{\text{ref}}(y|x)\right] \right], \tag{7}$$

where $g(y|x) = g(r_{\text{help}}(x,y), r_{\text{safe}}(x,y))$ is a function that combines the helpfulness reward $r_{\text{help}}(x,y)$ and safety reward $r_{\text{safe}}(x,y)$. Therefore, re-parameterizing Equation (7) to a supervised objective leads to an efficient and effective alignment method. The target objective is:

$$\min_{\theta} \mathbb{E}_{(x,y^{hw},y^{hl})\sim D} \left[\mathcal{L}\big(h_\pi(y^{hw}, y^{hl}), \tau^{-1} g_I(y^{hw}, y^{hl}|x)\big) \right], \tag{8}$$

where y^{hw} and y^{hl} are the more helpful and less helpful responses, and as we defined in Equation (5)

$$h_\pi(y^{hw}, y^{hl}) = \log\left(\frac{\pi(y^{hw}|x)\pi_{\text{ref}}(y^{hl}|x)}{\pi(y^{hl}|x)\pi_{\text{ref}}(y^{hw}|x)}\right).$$

Similarly to Equation (6), g_I is the label function that leverages the safety labels $I_{\text{safe}}(y^{hw}|x), I_{\text{safe}}(y^{hl}|x)$ to approximate the value $g(y^{hw}|x) - g(y^{hl}|x)$, where g is the global reward function in Equation (7).

In Section 3.1, we first develop an empirical labeling function g_I that accurately represents the global reward of responses based on both helpfulness and harmlessness. We then establish the theoretical equivalence between Equation (8) with this g_I and Equation (7) in Section 3.2. Next, we present the algorithm in Section 3.3 and provide a sample illustration in Section 3.4.

3.1 Empirical Labeling function

In previous single-reward optimization methods (Rafailov et al., 2023; Azar et al., 2024; Tang et al., 2024), $g_I(y^w, y^l|x)$ in Equation (6) is typically a positive constant. However, in our case, $g_I(y^{hw}, y^{hl}|x)$, which approximates the global reward disparity between the more helpful response and the less helpful response, *i.e.*, $g(y^{hw}|x) - g(y^{hl}|x)$, should vary depending on the safety of y^{hw} and y^{hl}. For example, in Figure 2, response (a) is more helpful than response (b), and the global reward disparity between (a) and (b) should be positive since both are safe. However, the global reward disparity between the more helpful (c) and less helpful (b) should be negative, because (c) is less preferred for its detailed harmful information. In fact, the absolute value of $g(y^{hw}|x) - g(y^{hl}|x)$

reflects the magnitude of the global preference disparity between the two responses, while its sign determines whether y^{hw} is globally preferred over y^{hl}.

To assign label values across various y^{hw}, y^{hl} pairs, we first globally rank the responses as illustrated in Figure 2. Our guiding principle is a general *preference for safe responses, prioritizing helpfulness only if the responses is safe*. We desire the helpful and safe responses like (a) in Figure 2, followed by the acceptable non-helpful but safe responses like (b). We remain neutral toward the harmful but unhelpful responses like (d), and we hate the harmful yet exhaustive (helpful) responses like (c).

Given two responses y^{hw}, y^{hl}, assuming we have their relative helpfulness ranking, there are four classes of pairs based on their safety, illustrated in Figure 3. For ① and ②, we prefer the safe and more helpful y^{hw} than the other response, so the signs of the labels should be positive. Similarly, the signs of ③ and ④ should be negative. The preference gap for ① (Desired vs. Neutral) is larger than for ②, thus the magnitude of the labels should be greater in ①. Likewise, the magnitude of labels of ④ should be greater than that of ③. Consequently, the label value of the four class of pairs should be ordered as ①, ②, ③, and ④. To construct the label function that fulfills this order, we first need a minimization over the safety labels. To ensure a positive label for ②, we require a larger scalar weighting the safety of y^{hw} compared to that of y^{hl}. We hypothesize the label function g_I as:

$$g_I(y^{hw}, y^{hl}|x) = B_3(B_1 I_{\text{safe}}(y^{hw}|x) - I_{\text{safe}}(y^{hl}|x) + B_2). \tag{9}$$

In this equation, B_1 is positive scalar that weights the safety of y^{hw}. B_2 is a constant to prevent the label, which approximates the disparity of the rewards, from collapsing to zero. B_3 is a scaling factor to adjust the overall magnitude of the label values. For instance, let $B_1 = 3, B_2 = -2\alpha, B_3 = 0.5$, Figure 3-right illustrates label values of different pairs.

3.2 Theoretically Equivalent Reward

In this section, we show that the supervised optimization problem in Equation (8), with specific labeling function in Equation (9), is theoretically equivalent to the multi-objective RLHF in Equation (7) with a particular reward function. Previous studies (Touvron et al., 2023; Dai et al., 2024) in aligning LLMs for both safety and helpfulness have shown that the global reward function can be effectively approximated by a bilinear combination of the two sub-rewards; see Appendix C.2 for more details. We hypothesize the global reward function as follows:

$$g(y|x) = (p^*_{\text{safe}}(y|x) + A_1)(p^*_{\text{help}}(y \succ \pi|x) + A_2), \tag{10}$$

where A_1, A_2 are two constants that prevent the reward from being nullified by zero values, and $p^*_{\text{help}}, p^*_{\text{safe}} \in [0, 1]$ are the ground-truth helpful and safety preferences of response y. Let $A_1 = E_s, A_2 = \frac{1}{2}, B_1 = 3, B_2 = 0, B_3 = \frac{1}{2}$, we have the reward function g and labeling function g_I:

$$g(y|x) = (p^*_{\text{safe}}(y|x) + E_s)(p^*_{\text{help}}(y \succ \pi|x) + \frac{1}{2}), \tag{11}$$

$$g_I(y^{hw}, y^{hl}|x) = \frac{3}{2} I_{\text{safe}}(y^{hw}|x) - \frac{1}{2} I_{\text{safe}}(y^{hl}|x), \tag{12}$$

where $E_s = \mathbb{E}_{y\sim\pi}\left[p^*_{\text{safe}}(y|x)\right]$ represent the ground truth average safety of responses given prompt x. The following theorems reveal the theoretical equivalence.

Theorem 3.1 (Azar et al. (2024))**.** *The optimization problem in Equation* (7) *has a solution* π^*

$$\pi^*(y|x) = \frac{\pi_{ref}(y|x) \exp\left(\tau^{-1} g(y|x)\right)}{\sum_{y'} \pi_{ref}(y'|x) \exp\left(\tau^{-1} g(y'|x)\right)},$$

and $\pi^*(y)$ *is the unique solution to the following optimization problem*

$$\min_{\pi_\theta} \mathbb{E}_{x\sim D, y, y'\sim\pi_\theta} \left[h_\pi(y, y') - \frac{g(y|x) - g(y'|x)}{\tau} \right]^2. \tag{13}$$

Theorem 3.2. *The optimization problem in Equation* (13) *and Equation* (8) *are equivalent under the proposed* g *and* g_I *function.*

With Theorem 3.1, we can obtain the optimal π^* by solving the supervised optimization problem in Equation (13). The proof of this theorem is in Appendix B.2. However, the optimization problem in

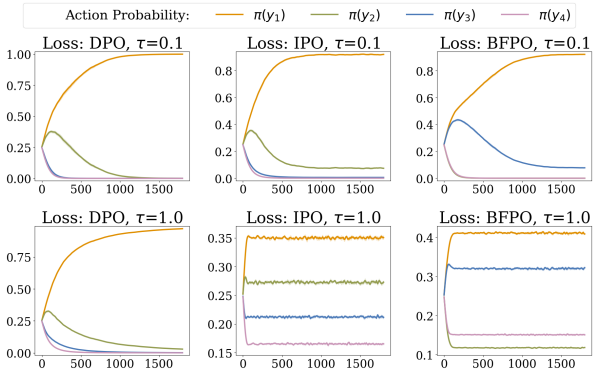

Figure 4: Action probabilities over steps during the policy optimization using DPO, IPO, and our BFPO in synthetic dataset. Only ours can recover the desired ranking.

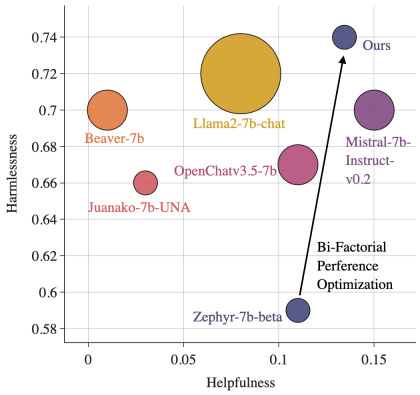

Figure 5: Helpfulness and harmlessness of open sourced models. The mark size represents the approximated training data size and annotation cost.

Equation (13) remains challenging because the function $g(y)$ involves the ground-truth preference p^*, which requires estimation by a large group of annotators. To address this, Theorem 3.2 shows it is equivalent to solve the supervised optimization problem in Equation (8) with the proposed g_I to obtain the optimal π^*. The proof of this equivalence is provided in Appendix B.3. We further discuss the general equivalence with different constants A_1, A_2, B_1, B_2, B_3 in Appendix B.4.

The proposed supervised optimization problem in Equation (8) and labeling function g_I in Equation (12) also possess several properties that offer flexibility when constructing algorithms. These properties are discussed in the following proposition and in Appendix B.5.

Proposition 3.3. *Theorem 3.1 and Theorem 3.2 hold under the shift of the preference values in g and g_I,* i.e., *for constants p_1, p_2, we have*

$$g(y|x) = (p^*_{safe}(y|x) + p_1 + E_s)(p^*_{help}(y \succ \pi|x) + p_2 + \frac{1}{2}),$$
$$g_I(y^{hw}, y^{hl}|x) = \frac{3}{2}(I_{safe}(y^{hw}|x) + p_1) - \frac{1}{2}(I_{safe}(y^{hl}|x) + p_2).$$

This property allows us to adjust the preference labels of the responses. Proof of the proposition is provided in Appendix B.5. In practice, we further apply a shift of the safety label value α as

$$g_I(y^{hw}, y^{hl}|x) = \frac{3}{2} I_{\text{safe}}(y^{hw}|x) - \frac{1}{2} I_{\text{safe}}(y^{hl}|x) - \alpha. \tag{14}$$

The factor α is useful when set to negative values to distinguish unsafe samples, i.e., to make the value of case ③ in Figure 3, *i.e.*, both responses are not safe, deviate from 0.

3.3 Algorithm

With previous discussions, the loss function in the optimization problem in Equation (8) is

$$\mathcal{L}_{\text{BFPO}}(\theta) = \mathbb{E}_{(x,y^{hw},y^{hl})\sim D}\left(\log\left(\frac{\pi_\theta(y^{hw}|x)\pi_{\text{ref}}(y^{hl}|x)}{\pi_\theta(y^{hl}|x)\pi_{\text{ref}}(y^{hw}|x)}\right) - \frac{\frac{3}{2}I_{\text{safe}}(y^{hw}|x) - \frac{1}{2}I_{\text{safe}}(y^{hl}|x) - \alpha}{\tau}\right)^2. \tag{15}$$

In practice, we directly use the above supervised loss to fine-tune the LLMs for both helpfulness and harmlessness. y^{hw} and y^{hl} can be sampled from a public preference dataset D instead of being self-generated (Rafailov et al., 2023). The safety labels $I_{\text{safe}}(y^{hw}), I_{\text{safe}}(y^{hl})$ are either provided in the dataset or obtained by a safety classifier. The probability $\pi(y|x)$ of generating the response y given prompt x is obtained by forwarding the prompt and response through the LLM π. π_θ is the language model we are optimizing, and π_{ref} is a reference model that can be the model at the beginning of the optimization. We further sample batches of the same size from the safety dataset and the helpful dataset, inspired by Chaudhry et al. (2019), to balance safety and helpfulness. The overall algorithm is summarized in Algorithm 1.

Table 1: Results of fine-tuning pre-trained model, Mistral, with various methods. Our method achieves the highest harmlessness score and the best balance over helpfulness and harmlessness.

	Helpfulness	Harmlessness		
	Alpaca(↑)	Disc. (↑)	Gen. (↑)	Savg. (↑)
DPO-H (Zephyr)	10.99	59.05	62.94	60.99
DPO-S	4.34	56.42	**96.91**	76.66
DPO	**14.71**	58.35	39.71	49.03
IPO	13.15	58.41	89.76	74.09
MORL	10.83	58.54	64.88	61.71
BFPO (ours)	13.33	**59.09**	95.24	**77.16**

Table 2: Results of further fine-tuning the aligned Zephyr model with red teaming data. Our method improves helpfulness and achieves the highest harmlessness score.

Model	Helpfulness	Harmlessness		
	Alpaca	Disc.	Gen.	Savg.
Zephyr-7b-beta	10.99	59.05	62.94	60.99
+ DPO	13.07	59.28	74.39	66.83
+ IPO	13.07	**59.32**	72.82	66.07
+ MORL	13.07	58.57	65.02	61.80
+ BFPO	**14.41**	59.02	**88.79**	**73.90**

3.4 Illustrative Examples

Following Azar et al. (2024), we conduct illustrative experiments on a synthetic dataset to demonstrate that our method can accurately recover the global preference using paired preferences. For simplicity, we consider a discrete action space with four actions, $\mathcal{Y} = \{y_1, y_2, y_3, y_4\}$, without context. We define the safety labels and helpfulness ranking as

$$\text{Safety: } I_{\text{safe}}(y_1) = 1, I_{\text{safe}}(y_2) = 0, I_{\text{safe}}(y_3) = 1, I_{\text{safe}}(y_4) = 0,$$
$$\text{Helpfulness: } y_1 \succ y_2 \succ y_3 \succ y_4.$$

Consequently, our proposed global preference, as in Figure 3, is $y_1 \succ y_3 \succ y_4 \succ y_2$. We encode the policy as $\pi_\theta(y_i) = \text{softmax}(\theta)_i$ using a vector $\theta \in \mathbb{R}^4$ and $i = 1, 2, 3, 4$. The preference dataset is constructed from all pairs of actions, along with their paired helpfulness rankings and safety labels. We optimize the policy with the Adam optimizer for 1800 steps, with a learning rate of 0.01, batch size of 32 sampled with replacement, $\tau = 1$, and $\alpha = 0.5$. We compare the supervised optimization objective proposed in Equation (15) as well as DPO (Rafailov et al., 2023) and IPO (Azar et al., 2024), where we take the more helpful response is taken as the win response. Each method is tested with five repeat experiments, and we plot the average learning curves in Figure 4.

For all τ, we observe that only with our proposed method does $\pi(y_i)$, *i.e.*, the probability of generating action y_i, converges to the desired ranking, $y_1 \succ y_3 \succ y_4 \succ y_2$. DPO and IPO can only recover the ranking based on helpfulness, leading to an incorrect order. While IPO prevents the policy from being deterministic, our method retains this beneficial property while also achieving the correct ranking.

4 Experiment

4.1 Evaluation Setup

Harmlessness Benchmark. To evaluate the harmlessness, we first construct a benchmark including both discriminative tasks and generative tasks based on previous benchmarks (Srivastava et al., 2023; Gao et al., 2023; Tedeschi et al., 2024; Zou et al., 2023). The discriminative tasks measure the models' recognition of multiple safety topics, including

- Bias: CrowS-Pairs (Nangia et al., 2020), BBQ (Parrish et al., 2022), WinoGrande (Sakaguchi et al., 2021).
- Ethics: ETHICS (Hendrycks et al., 2021), Moral Permissibility (Srivastava et al., 2023; Hernandez et al., 2021; Lourie et al., 2021; Thomson, 2019), Simple Ethics Questions (Hendrycks et al., 2021; Lourie et al., 2021).
- Toxicity: ToxicGen (Hartvigsen et al., 2022), BigBench HHH Alignment (Srivastava et al., 2023)

In the generative tasks, we prompt the models to generate harmful content using the prompt dataset, AdvBench (Zou et al., 2023), Real Toxicity Prompts (Gehman et al., 2020), ALERT (Tedeschi et al., 2024). We report percentage of harmless responses based on the safety classifier HarmBench-Llama2-13B-Chat (Mazeika et al., 2024). Details of the benchmark can be found in Appendix C.1. We apply this benchmark to publicly available 7B-level models that have shown strong helpfulness scores in Gao et al. (2023); Dubois et al. (2024b), and present the performance in Figure 5 and in Appendix C.3.

Overall Evaluation Metrics. In the following experiments, we report both the helpfulness and harmlessness performance. Helpfulness is measured using AlpacaEval 2.0 (Alpaca) (Dubois et al., 2024a; Li et al., 2023; Dubois et al., 2024b). Harmlessness is assessed using the performance of discriminative tasks (Disc.), generative tasks (Gen.) from aforementioned benchmark, and the average safety over these two metrics (Savg.).

4.2 Alignment with BFPO Objective

From the evaluation on the open model in Figure 5, we observe that Zephyr-7b-beta (Tunstall et al., 2023), an open-sourced model fine-tuned over Mistral-7B-v0.1 (Jiang et al., 2023a) with DPO algorithm (Rafailov et al., 2023), exhibits a low score in harmlessness, particularly in generative tasks. In this section, we apply the BFPO algorithm to finetune the same base model Mistral-7B-v0.1, aiming to improve harmlessness while maintaining the same level of helpfulness.

Training Details. Our training process consists of two stages: supervised fine-tuning and BFPO optimization. The supervised fine-tuned model is used as the reference model π_{ref} in the BFPO stage. We set $\tau = 0.01$, $\alpha = 0.5$. We implement PEFT training for all baselines, where we only unfreeze the selected layers θ', the second MLP layers in each transformer block, in the policy π_θ Zhang et al. (2024). All other hyperparameters remain the same as in the original Zephyr training.

Dataset Details. In the supervised fine-tuning stage, we follow Tunstall et al. (2023); Dai et al. (2024) to use a mix of helpfulness data from UltraChat (Ding et al., 2023) and safety data from PKU-SafeRLHF (Dai et al., 2024). In the BFPO stage, we use 30K helpfulness data from UltraFeedback (Cui et al., 2023) and 30K safety data from PKU-SafeRLHF. UltraFeedback contains instruction-following tasks that provide paired helpfulness preference rankings, and we treat all responses as safe since they undergo human filtering. PKU-SafeRLHF provides both paired helpfulness preference rankings and binary safety labels. Details are in Appendix C.3.

Baselines. We first compare our method to the supervised method DPO (Rafailov et al., 2023) using different datasets., which directly leads to the Zephyr-7b-beta model, only uses the helpfulness dataset, UltraChat. DPO-S only uses the safety dataset, PKU-SafeRLHF. We also compare our method to existing approaches, DPO (Rafailov et al., 2023), IPO (Azar et al., 2024), and MORL (Ramé et al., 2023), when using a naive mix of the helpfulness and safety datasets. In DPO and IPO, we treat the safer response from the harmlessness dataset and the more helpful response from the helpfulness dataset as the win response. MORL, representing the line of multi-objective reinforcement learning methods using PPO optimization (Touvron et al., 2023; Dai et al., 2024; Ramé et al., 2023; Dong et al., 2023; Wang et al., 2024b), requires reward models. Following Wang et al. (2024b), we use a single highly-ranked (Lambert et al., 2024), publicly available reward model, ArmoRM-Llama3-8B-v0.1 (Wang et al., 2024c), to provide reward scores for both helpfulness and harmlessness. Refer to Appendix C.2 for more details. All methods use the same pre-trained model.

Results and Comparisons. The results are presented in Table 1. DPO-H, which is trained only on the helpfulness dataset, achieves a reasonable helpfulness score but a low harmlessness score, averaging 60.99%. Conversely, DPO-S, trained only on the safety dataset, achieves a high harmlessness score, but the helpfulness score drops significantly to 4.34%.

Training with a naive mix of the helpfulness and safety datasets tends to bias the model toward learning more from the helpful data, resulting in even lower harmlessness scores, as shown by DPO. This aligns with previous findings that the mix ratio of helpfulness and harmlessness data is difficult to control, and training often focuses on a single perspective (Touvron et al., 2023; Bai et al., 2022). In comparison to these supervised methods, BFPO achieves the highest average harmlessness score of 77.16% and significantly improves the generative tasks score from 39.71% to 95.24%.

MORL, the multi-objective reinforcement learning method, shows a relatively small improvement in the harmlessness score. We suspect the primary reason is that the reward scores of different responses provided by the public reward model are not sufficiently distinguishable, making it inefficient for the model to learn to generate good responses while avoiding bad ones. This highlights the need for training a reward model specific to the model being fine-tuned, which involves the costly human prompting (red teaming) and annotation process.

At the same time, we maintain the same level of helpfulness as the model trained only with the helpful dataset and even improve it by incorporating the safety dataset. Full results are in Appendix C.3.

Table 3: Efficiency comparison of our method to previous PPO-based safety alignment methods.

Method	Data Size	Red Teaming	Iteration	Alpaca	Savg.
Beaver	300K	✓	3	1.00	71.80
Llama2	1M	✓	6	7.60	73.80
BFPO	30K	-	1	13.33	77.16

Table 4: Ablation study on the shifting factor and buffer training

Model	Helpfulness	Harmlessness		
	Alpaca	Disc.	Gen.	Savg.
BFPO	13.33	59.09	**95.24**	**77.16**
BFPO, $\alpha = 0$	12.76	59.09	92.87	75.98
BFPO, $\alpha = 0$, - buffer	**15.59**	60.14	88.76	74.45

Comparison against Previous Safety Alignment Methods. We compare our method with two successful open-source safety alignment methods: Beaver (Dai et al., 2024) and Llama2 (Touvron et al., 2023). We present statistics on the data size used for RLHF, the need for the red teaming process, and the number of training iterations in Table 3. Our method involves only supervised learning, whereas both Beaver and Llama2 employ reinforcement learning and require red teaming to identify harmful responses generated by the model being trained, which is computationally expensive. Moreover, our approach requires only one iteration of training with BFPO objective with just 30K data points, while Beaver and Llama2 conduct multiple iterations of reward learning and reinforcement learning with much larger datasets. Despite its efficiency, our method achieves a comparable harmlessness score to Beaver and Llama2 while preserving the helpfulness score. These results indicate strong potential for our method to be applied in the future development of open-source models at a minimal cost.

4.3 Improve Pre-aligned Models with Red Teaming Data

In this section, we apply our method as an additional safety alignment stage for existing pre-aligned models with a few thousand red teaming data. We compare our method with DPO (Rafailov et al., 2023), IPO (Azar et al., 2024), MORL (Ramé et al., 2023) as in Section 4.2.

Data Preparation. We first use 9K harmful prompts from the PKU-SafeRLHF dataset (Dai et al., 2024) and have the Zephyr-7b-beta Tunstall et al. (2023) model generate two responses for each prompt. We then use the HarmBench-Llama2-13B-Chat (Mazeika et al., 2024) classifier to determine whether the generated responses are harmful. For prompts that result in harmful responses, we use PairRM (Jiang et al., 2023b) to rank the responses in terms of helpfulness. This process results in 1.5K harmful prompts, responses, safety labels for each response, and pairwise helpfulness preferences.

Results. Table 2 shows the results. Our method improves the harmlessness of the Zephyr-7b-beta model from 60.99% to 73.90%, while preserving the helpfulness. The improvement in generative tasks is particularly significant, from 62.94% to 88.79%. The supervised methods, DPO and IPO, can also improve the harmlessness, but the improvement is not as substantial as with our method. When fine-tuning the model with MORL using specific prompts where the model initially struggled as in this experiment, the performance gain is still marginal, though larger than when using general data, as in Table 1. This aligns with the observation that using RL methods to improve safety requires a large amount of model-specific data, high-quality labels, and a reward model specifically trained on these data to provide distinguishable scores. In contrast, BFPO achieves similar goals without the need for large amounts of helpfulness data mixed with red teaming data. Moreover, our overall pipeline of this experiment is efficient and automatic, requiring no human annotation. These results strongly indicate that our method can be effectively used in an additional safety alignment stage for existing chat models to improve harmlessness at minimal cost. Full results are in Appendix C.3.

4.4 Ablations

We validate the technical design of our algorithm in Table 4, showing that the shift parameter α and buffered training are effective in improving harmlessness.

In the BFPO $\alpha = 0$ experiment, we set the shift parameter α to 0. The results indicate that, as illustrated in Section 3.4, the shift parameter α is useful in distinguishing unsafe data, and thus improves performance on generative tasks in harmlessness slightly. In the BFPO - w/o buffer experiment, we do not balance examples from the safety dataset and the helpful dataset, but instead mix the two datasets and randomly sample data from them. The lower harmlessness performance

provides the evidence that buffered training helps mitigate the tension between helpfulness and harmlessness. Full results are provided in Appendix C.3.

5 Related Work

Alignment with Diverse Preferences. Traditional language model alignment methods (Christiano et al., 2017; Stiennon et al., 2020; Hendrycks et al., 2021) typically use a single reward or unified preference model.However, recent work suggests that human preferences are diverse and cannot be adequately represented by a single reward model. To address this, Chakraborty et al. (2024) propose learning a mixture distribution for the reward using the EM algorithm, which they then apply in their MaxMin RLHF approach. Dong et al. (2023); Ramé et al. (2023); Wang et al. (2024b) explore training multi-objective reward models for the alignment stage. These methods primarily focus on improving reward models for RL based alignment. The most closely related work of supervised alignment methods is by Zhou et al. (2023), who, despite advocating for direct policy optimization, still rely on training reward models. In contrast, our approach completely eliminates the two-stage training process and directly integrates multiple preferences into the supervised optimization.

Safety Alignment. Aligning large language models (LLMs) with both helpfulness and harmlessness is a specific case of addressing diverse preferences. To enhance safety, many researchers conduct additional safety data annotation alongside algorithm design. Touvron et al. (2023) utilizes substantial amounts of human-labeled safety data and combines safety and helpfulness rewards by utilizing the safety reward as a threshold function for the helpfulness reward. Dai et al. (2024); Ji et al. (2024) engage in red teaming to gather extensive safety data and frame safety alignment as a conditioned Markov Decision Process (MDP) problem. Mu et al. (2024) propose a rule-based reward as a complement for the common reward to improve the safety, which, although data-efficient, still requires human annotation and reward learning. In contrast, our method is fully automated and efficient, eliminating the need for human intervention in the safety alignment process. On the other hand, Huang et al. (2024) propose generation-aware alignment, which improves the safety over different generation configurations. With our focus on improving safety under specific configurations, this work can be a strong complement to ours.

Safety Evaluation. Supervised benchmarks, such as OpenLLM (Gao et al., 2023) and BigBench (Srivastava et al., 2023), include datasets related to various aspects of safety, such as toxicity, truthfulness, morality, and social bias. Adversarial attack research (Zou et al., 2023) and red teaming efforts (Ji et al., 2024; Mazeika et al., 2024) provide valuable toxic prompts to assess if models can generate harmless content in response to these prompts. To identify if the output content contains harmful information, some studies (Bai et al., 2022; Touvron et al., 2023) rely on human annotators, while others employ AI models like GPT-4 (Wang et al., 2024a). Mazeika et al. (2024) offer fine-tuned Llama2 models to as harmful content classifier, offering an efficient alternative to GPT-4 in model-based evaluation.

6 Limitations and Discussion

In this paper, we propose a novel supervised optimization method, Bi-Factorial Preference Optimization (BFPO), to balance the safety and helpfulness during the alignment of LLMs. We theoretically prove that this direct optimization is equivalent to previous multi-objective reinforcement learning that combine safety and helpfulness rewards. With BFPO, we outperform existing methods in terms of safety and helpfulness in both fine-tuning the pre-trained LLMs and pre-aligned models. Our method is highly effective, significantly more computationally efficient, and does not require any human annotation or additional data collection.

Furthermore, our approach is versatile and does not rely on any specific property of harmlessness itself. This flexibility allows it to be applied to improve various other potentially conflicting objectives in aligning LLMs. To achieve this, we only need characteristic-specific labels for the field-specific dataset. We believe our proposed method can serve as a general framework for the transfer learning of aligned models. However, our method relies on specific label formats for helpfulness and safety may present a limitation when addressing different tasks. Moreover, extending our work to handle more objectives (beyond just two) is also a promising direction for future research.

REFERENCES

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

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455. PMLR, 2024.

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

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Bedi, and Mengdi Wang. Maxmin-RLHF: Towards equitable alignment of large language models with diverse human preferences. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*, 2024.

A Chaudhry, M Rohrbach, M Elhoseiny, T Ajanthan, P Dokania, P Torr, and M Ranzato. Continual learning with tiny episodic memories. In *ICML Workshop on Multi-Task and Lifelong Reinforcement Learning*, 2019.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

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

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

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe RLHF: Safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations*, 2024.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Yi Dong, Zhilin Wang, Makesh Narsimhan Sreedhar, Xianchao Wu, and Oleksii Kuchaiev. Steerlm: Attribute conditioned sft as an (user-steerable) alternative to rlhf. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024a.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36, 2024b.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023. URL https://zenodo.org/records/10256836.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3356–3369, 2020.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y Wu, YK Li, et al. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*, 2024.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for implicit and adversarial hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch Critch, Jerry Li Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. In *International Conference on Learning Representations*, 2021.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. Catastrophic jailbreak of open-source LLMs via exploiting generation. In *The Twelfth International Conference on Learning Representations*, 2024.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36, 2024.

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. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023a.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. Llm-blender: Ensembling large language models with pairwise comparison and generative fusion. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, 2023b.

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. Rewardbench: Evaluating reward models for language modeling. https://huggingface.co/spaces/allenai/reward-bench, 2024.

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

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 13470–13479, 2021.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. Rule based rewards for language model safety. 2024.

David Nadeau, Mike Kroutikov, Karen McNeil, and Simon Baribeau. Benchmarking llama2, mistral, gemma and gpt for factuality, toxicity, bias and propensity for hallucinations. *arXiv preprint arXiv:2404.09785*, 2024.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1953–1967, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.154.

OpenAI. Gpt-4 technical report, 2023.

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

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. Bbq: A hand-built bias benchmark for question answering. In *60th Annual Meeting of the Association for Computational Linguistics, ACL 2022*, pp. 2086–2105. Association for Computational Linguistics (ACL), 2022.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

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

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Yunhao Tang, Zhaohan Daniel Guo, Zeyu Zheng, Daniele Calandriello, Rémi Munos, Mark Rowland, Pierre Harvey Richemond, Michal Valko, Bernardo Ávila Pires, and Bilal Piot. Generalized preference optimization: A unified approach to offline alignment. *arXiv preprint arXiv:2402.05749*, 2024.

Simone Tedeschi, Felix Friedrich, Patrick Schramowski, Kristian Kersting, Roberto Navigli, Huu Nguyen, and Bo Li. Alert: A comprehensive benchmark for assessing large language models' safety through red teaming. *arXiv preprint arXiv:2404.08676*, 2024.

Judith Jarvis Thomson. Killing, letting die, and the trolley problem. In *Death, Dying and the Ending of Life, Volumes I and II*, pp. V2_17–V2_30. Routledge, 2019.

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

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

Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. On the planning abilities of large language models-a critical investigation. *Advances in Neural Information Processing Systems*, 36:75993–76005, 2023.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *Advances in Neural Information Processing Systems*, 36, 2024a.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023.

Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. In *ACL*, 2024b.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*, 2024c.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. Wordcraft: story writing with large language models. In *27th International Conference on Intelligent User Interfaces*, pp. 841–852, 2022.

Wenxuan Zhang, Paul Janson, Rahaf Aljundi, and Mohamed Elhoseiny. Overcoming generic knowledge loss with selective parameter update. In *IEEE / CVF Computer Vision and Pattern Recognition Conference*, 2024.

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, et al. Secrets of rlhf in large language models part i: Ppo. *arXiv preprint arXiv:2307.04964*, 2023.

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

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A ALGORITHM

Algorithm 1 shows the BFPO algorithm. As mentioned in Section 2, in practice, we refer to datasets related to safety topics, collected through red teaming, as safety datasets. A typical safety dataset will contain a safety label $I_{\text{safe}}(y)$, which is the binary label indicating whether the response y is harmful, as well as the preference label $I_{\text{help}}(y \succ y')$ in terms of helpfulness. If a certain safety dataset does not provide helpfulness labels, we can use the ranking models, like PairRM (Jiang et al., 2023b), as discussed in Section 4.3, to generate the pairwise helpfulness labels. We refer to datasets designed to improve the helpfulness of the model as helpfulness datasets. A typical helpfulness dataset will contain the helpfulness preference labels $I_{\text{help}}(y \succ y')$. Since most helpfulness data undergoes human filtering, the responses are usually safe. Therefore, we assign the safety label $I_{\text{safe}}(y) = 1$ to all responses in the helpfulness dataset.

We further require a pre-trained language model π_{ref}, the total number of optimization steps T, the penalty coefficient τ for the KL divergence term, and the shifting parameter α. We also need to specify the layers to be unfrozen for the policy optimization, denoted as θ'.

At the beginning of the algorithm, we initialize the policy π_θ with the pre-trained language model π_{ref}, and unfreeze the selected layers θ' (line 1-2). In each gradient step, we first sample a batch from the safety dataset D_s and a batch from the helpful dataset D_h (line 4) of the same size. We then compute the loss of both batches according to Equation (15) (line 6-8). We accumulate the gradients of the loss for the both batches and update the policy π_θ (line 10). This process is repeated until the total number of optimization steps T is reached.

Algorithm 1 BFPO Algorithm

Require: Safety dataset $D_s = \{(x, y^{hw}, y^{hl}, I_{\text{safe}}(y^{hw}), I_{\text{safe}}(y^{hl}))\}$ and helpful dataset $D_h = \{(x, y^{hw}, y^{hl})\}$.
Require: Total number of optimization steps T. Pre-trained language model π_{ref}, and unfrozen layer θ'. τ, α

1: Initialize $\pi_\theta \leftarrow \pi_{\text{ref}}$
2: Only unfreeze selected layers θ'
3: **while** $t < T$ **do**
4: Sample batch $B_s \sim D_s$, $B_h \sim D_h$.
5: **for** batch = B_s, B_h **do**
6: Compute $h(y^{hw}, y^{hl})$ with Equation (5)
7: Compute g_I with Equation (14) $\triangleright$ $I_{\text{safe}}(y) = 1$ for the helpful dataset.
8: Compute and accumulate gradients w.r.t Equation (15)
9: **end for**
10: Update π_θ.
11: **end while**

B PROOF

B.1 NOTATION

Table 5: Notations

Notation	Meaning
$y, y' \sim \pi(x)$	Two responses generated independently by the policy.
$p^*_{\text{help}}(y \succ y'\|x)$	Ground-truth helpfulness preference of y being preferred to y' knowing the context x
$p^*_{\text{safe}}(y\|x)$	Ground-truth safety of y knowing the context x
$I_{\text{help}}(y \succ y'\|x)$	Binary label of helpfulness preference of y being preferred to y' knowing the context x
$I_{\text{safe}}(y\|x)$	Binary label of safety of y knowing the context x
y^w, y^l	globally preferred and dispreferred responses knowing the context x
y^{hw}, y^{hl}	preferred and dispreferred responses in terms of helpfulenss knowing the context x
E_s	Expected safety of a response y given the context x

Table 5 summarizes the notations used in this paper based on Rafailov et al. (2023); Azar et al. (2024). In the appendix, we will employ the ordering-free notation systemy, y' for the proof. Specifically, we express the transformation equations from y^{hw}, y^{hl} to y, y' as:

$$I_{\text{safe}}(y^{hw}|x) = I_{\text{help}}(y \succ y'|x) I_{\text{safe}}(y|x) + I_{\text{help}}(y' \succ y|x) I_{\text{safe}}(y'|x)$$
$$I_{\text{safe}}(y^{hl}|x) = I_{\text{help}}(y \succ y'|x) I_{\text{safe}}(y'|x) + I_{\text{help}}(y' \succ y|x) I_{\text{safe}}(y|x)$$

For brevity and clarity, we further adopt the notation y to represent $y|x$. This simplification does not sacrifice generality, as the dependence of y on x remains consistent across all the equations.

B.2 Proof of Theorem 3.1

We begin by restating Theorem 3.1 with the notation system y, y'. Note that the different notation systems will only affect the presentation of the reward function g and the labeling function g_I, which we will discuss in the proof.

Theorem B.1. *Let $\tau > 0$ be a real number, π_θ, π_{ref} be two policy. Then*

$$\pi^*(y) = \frac{\pi_{ref}(y) \exp\left(\tau^{-1} g(y)\right)}{\sum_{s \in \mathcal{S}} \pi_{ref}(s) \exp\left(\tau^{-1} g(s)\right)} \tag{16}$$

is an optimal solution to the optimization problem

$$\max_{\pi_\theta} \mathbb{E}_{y \sim \pi_\theta(y)} \left[g(y) - \tau KL\left[\pi_\theta(y) || \pi_{ref}(y)\right] \right], \tag{17}$$

and $\pi^*(y)$ is the optimal unique solution of

$$\min_{\pi_\theta} \mathbb{E}_{y, y' \sim \pi_\theta(y)} \left[h_\pi(y, y') - \frac{g(y) - g(y')}{\tau} \right]^2, \tag{18}$$

where

$$h_\pi(y, y') = \log\left(\frac{\pi_\theta(y) \pi_{ref}(y')}{\pi_\theta(y') \pi_{ref}(y)} \right). \tag{19}$$

To establish optimal solution, we follow Azar et al. (2024) to leverage the following lemma.

Lemma B.2 (Rafailov et al. (2023), Azar et al. (2024))**.** *Let*

$$\mathcal{L}_\tau(\delta) = \mathbb{E}_{s \in \delta}[f(s)] - \tau KL[\delta || \eta],$$

where $s \in \mathcal{S}$ and $\mathcal{S}$ is a finite set, $f \in \mathbb{R}^{\mathcal{S}}$ is a function mapping elements of $\mathcal{S}$ to real numbers, $\delta \in \Delta(\mathcal{S})$ is a probability distribution over $\mathcal{S}$, $\eta \in \Delta(\mathcal{S})$ is a fixed reference distribution, and $\tau \in \mathbb{R}_+^*$ is a strictly positive number. Then the argmax problem with the regularized criterion

$$\arg\max_{\delta \in \Delta(\mathcal{S})} \mathcal{L}_\tau(\delta)$$

has an optimal solution δ^*, where

$$\delta^*(s) = \frac{\eta(s) \exp(\tau^{-1} f(s))}{\sum_{s' \in \mathcal{S}} \eta(s') \exp(\tau^{-1} f(s'))}, \ \forall s \in \mathcal{S}$$

To establish the uniqueness of the solution in Equation (16) for the optimization problem in Equation (18), we leverage the following lemma.

Lemma B.3 (Theorem 2 in Azar et al. (2024))**.** *Let*

$$\mathcal{L}(\pi_\theta) = \mathbb{E}_{y, y' \sim \pi_\theta(y)} \left[h_\pi(y, y') - \frac{g(y) - g(y')}{\tau} \right]^2, \tag{20}$$

then $\min_{\pi_\theta} \mathcal{L}(\pi_\theta)$ has a unique optimal solution π^* expressed in Equation (16) *, and no other local or global minima exist.*

Proof. Let $J = \text{Supp}(\pi) = \{y_1, \ldots, y_n\}$, where $n = |J|$, and Π be the set of policies with support set J. It is straightforward that $\min_{\pi \in \Pi} \mathcal{L}(\pi) = \mathcal{L}(\pi^*) = 0$, thus π^* is a global optimal solution. We now prove the uniqueness of this optimal solution by the re-parameterization trick.

We parameterize Π via vectors of logits $s \in \mathbb{R}^J$ of π, *i.e.*, $s_i = \log(\pi(y_i))$. Set $\pi_s(y) = \frac{\exp(s_i)}{\sum_{i=1}^n \exp(s_i)}$ for $y = y_i \in J$ and $\pi_s(y) = 0$ otherwise. Specially, let s^* be the vector of logits corresponding to π^*, we have $\pi^* = \pi_{s^*}$.

We first prove that s^* is the global optimal solution to the optimization problem

$$\mathcal{L}(s) := \mathcal{L}(\pi_s) = \mathbb{E}_{y,y' \sim \pi_s} \left[h_{\pi_s}(y, y') - \frac{g(y) - g(y')}{\tau} \right]^2.$$

It is obvious that $\mathcal{L}(s^*) = 0$, thus it is a local minimum. By expanding the square term, we have

$$\begin{aligned}
\mathcal{L}(s) &= \mathbb{E}_{y,y' \sim \pi_s} \left[\frac{g(y) - g(y')}{\tau} - (s(y) - s(y')) - \log\left(\frac{\pi_{\text{ref}}(y')}{\pi_{\text{ref}}(y)} \right) \right]^2 \\
&= \sum_{y,y' \in J} \pi_s(y)\pi_s(y') \left[((s(y) - s(y'))^2 + C_1 \cdot ((s(y) - s(y')) + C_2 \right],
\end{aligned}$$

where C_1, C_2 are two terms independent of s. The above equation is a positive semidefinite quadratic form, and hence is convex. Thus, all local minima are global minima.

Now we prove that π_{s^*} is the unique global minima to $\mathcal{L}(s)$. Since π_s is a surjective continuous mapping from s to π, then every local minima π to $\mathcal{L}(\pi)$ corresponds to a set of s that minimizes $\mathcal{L}(s)$. The uniquess of s^* will deduce that π^* is the unique optimal solution to Equation (18) and concludes the proof. Consider $s' = s^* + r \cdot \Delta s$, where the only r is the radius and Δs is the direction under the polar coordinate. The only direction that not increase $\mathcal{L}(s')$ away from 0 is $e = (\frac{1}{n}, \ldots, \frac{1}{n})$ (Boyd & Vandenberghe (2004), Chap. 3). However, we have

$$\pi_{s^* + r \cdot e}(s_i) = \frac{\exp(s_i + r \cdot \frac{1}{n})}{\sum_{i=1}^n \exp(s_i + r \cdot \frac{1}{n})} = \frac{\exp(s_i)}{\sum_{i=1}^n \exp(s_i)} = \pi_{s^*}(s_i), \ \forall i \in [n].$$

This indicates that π_{s^*} is the unique global minima to $\mathcal{L}(\pi_{s^*})$ and thus π^* is the unique optimal solution to Equation (18). □

Now we provide the proof of Theorem 3.1, most of which follows Azar et al. (2024).

Proof. Let $\mathcal{S}$ be the set of all possible token combinations with fixed token length, then it is finite. Let $f(s) = (p^*_{\text{safe}}(s) + E_s)(p^*_{\text{help}}(s \succ \pi) + \frac{1}{2})$, $\delta(s) = \pi_\theta(s)$ and $\eta(s) = \pi_{\text{ref}}(s)$. All the conditions in the Lemma B.2 are satisfied. Thus, Equation (16) is a solution to the optimization problem in Equation (17).

Now we prove Equation (16) is also a solution to the optimization problem Equation (18). Plug Equation (16) in the Equation (18), we have

$$h_{\pi^*}(y, y') = \log\left(\frac{\pi^*(y)\pi_{\text{ref}}(y')}{\pi^*(y')\pi_{\text{ref}}(y)} \right) = \log\left(\frac{\exp\left(\tau^{-1} g(y)\right)}{\exp\left(\tau^{-1} g(y')\right)} \right) = \tau^{-1}(g(y) - g(y')),$$

which validates Equation (16) is a solution to the optimization problem Equation (18).

Finally, Lemma B.3 indicates Equation (16) is the unique solution to Equation (18). This concludes the proof. □

The above proof holds for any order of y, y' since the equation in Equation (19) is skew-symmetric, *i.e.*,

$$\left[h_\pi(y, y') - \frac{g(y) - g(y')}{\tau} \right]^2 = \left[h_\pi(y', y) - \frac{g(y') - g(y)}{\tau} \right]^2.$$

This allows us to freely arrange the order of y, y' in Equation (18) without loss of generality. Therefore, Equation (18) can be written as

$$\min_{\pi_\theta} \mathbb{E}_{y,y' \sim \pi_\theta(y)} \left[h_\pi(y^{hw}, y^{hl}) - \frac{g(y^{hw}) - g(h^{hl})}{\tau} \right]^2,$$

where

$$y^{hw} = \begin{cases} y & \text{if } I_{\text{help}}(y \succ y'|x) = 1, \\ y' & \text{otherwise,} \end{cases}$$

and

$$y^{hl} = \begin{cases} y' & \text{if } I_{\text{help}}(y \succ y'|x) = 1, \\ y & \text{otherwise.} \end{cases}$$

With this reordering, the theorem reduces to Theorem 3.1

B.3 Proof of Theorem 3.2

In this section, we prove the Theorem 3.2. We begin by rewriting the formula in Equation (12) into a function of y, y'.

$$\begin{aligned} g_I(y, y') = B_3\Big(&B_1\big(I_{\text{safe}}(y)I_{\text{help}}(y \succ y') + I_{\text{safe}}(y')I_{\text{help}}(y' \succ y)\big) \\ &- \big(I_{\text{safe}}(y)I_{\text{help}}(y' \succ y) + I_{\text{safe}}(y')I_{\text{help}}(y \succ y')\big) + B_2\Big) \cdot \Big(2I_{\text{help}}(y \succ y') - 1\Big), \end{aligned} \tag{21}$$

Here, $I_{\text{help}}(y \succ y')$ determines whether y is the win response or lose response. In other words,

$$I_{\text{safe}}(y^{hw}) = I_{\text{safe}}(y)I_{\text{help}}(y \succ y') + I_{\text{safe}}(y')I_{\text{help}}(y' \succ y),$$

and the same applies to $I_{\text{safe}}(y^{hl})$. To enable the reordering of the variables, we further multiply the formula by $2I_{\text{help}}(y \succ y') - 1$, since $h_\pi(y, y') = -h_\pi(y', y)$ By organizing the terms, we have

$$\begin{aligned} g_I(y, y') =&(B_1B_3 - B_3)I_{\text{help}}(y \succ y')I_{\text{safe}}(y) + (B_1B_3 - B_3)I_{\text{help}}(y \succ y')I_{\text{safe}}(y') \\ &- B_1B_3I_{\text{safe}}(y') + B_3I_{\text{safe}}(y) + 2B_2B_3I_{\text{help}}(y \succ y') - B_2B_3 \end{aligned}$$

We first establish the equivalence of the two optimization problems in Equation (22) and Equation (23) under the specific choice of constants, and then provide the general relation of constants for the equivalence.

Here, we use the following constants:

$$A_1 = E_s, A_2 = \frac{1}{2}, B_1 = 3, B_2 = 0, B_3 = \frac{1}{2}.$$

Theorem B.4. *The optimization problem*

$$\min_{\pi_\theta} \mathbb{E}_{x\sim\rho, y, y'\sim\pi_\theta(y)}\Big[h_\pi(y, y') - \frac{g\big(p^*_{safe}(y), p^*_{help}(y)\big) - g\big(p^*_{safe}(y'), p^*_{help}(y')\big)}{\tau}\Big]^2, \tag{22}$$

where $g(y) = (p^*_{safe}(y) + E_s)(p^*_{help}(y \succ \pi) + \frac{1}{2})$*, is equivalent to the optimization problem*

$$\min_{\pi_\theta} \mathbb{E}_{x\sim\rho, y, y'\sim\pi_\theta(y), I\sim Bernoulli}\Big[\Big(h_\pi(y, y') - \frac{g_I(y, y')}{\tau}\Big)^2\Big], \tag{23}$$

where

$$g_I(y, y') = I_{help}(y \succ y')I_{safe}(y) + I_{help}(y \succ y')I_{safe}(y') + \frac{1}{2}I_{safe}(y) - \frac{3}{2}I_{safe}(y')$$

Here, $I \sim$ Bernoulli denotes the Bernoulli variables $I_{\text{safe}}(y)$ and $I_{\text{safe}}(y')$.

Proof. The two minimization problems are both over π_θ, so we only need to focus on the terms that involve π_θ. Specifically, the first term and the cross term after expanding the square expression in the two minimization problems. The first term is the same. Here we prove the cross term is also the same.

Let $\pi_y = \log(\pi(y))$, $\pi^R_y = \log(\pi_{\text{ref}}(y))$, then we can write

$$h_\pi(y, y') = \pi_y - \pi_{y'} + \pi^R_{y'} - \pi^R_y$$

Let $p_h(y) = p^*_{\text{help}}(y \succ \pi)$ and $p_s(y) = p^*_{\text{safe}}(y)$. The cross term of Equation (22) can be written as

$$\begin{aligned}
&\mathbb{E}_{x\sim\rho,y,y'\sim\pi}\left[h_\pi(y,y')\left(g\big(p^*_{\text{safe}}(y), p^*_{\text{help}}(y\succ\pi)\big) - g\big(p^*_{\text{safe}}(y'), p^*_{\text{help}}(y'\succ\pi)\big)\right)\right]\\
=&\mathbb{E}_{x\sim\rho,y,y'\sim\pi}\left[(\pi_y - \pi_{y'} + \pi^R_{y'} - \pi^R_y)\left(g\big(p_s(y), p_h(y)\big) - g\big(p_s(y'), p_h(y')\big)\right)\right]\\
=&\mathbb{E}_{x\sim\rho,y\sim\pi}\left[(\pi_y - \pi^R_y)\left(g\big(p_s(y), p_h(y)\big) - \mathbb{E}_{y'\sim\pi}\big[g\big(p_s(y'), p_h(y')\big)\big]\right)\right]\\
+&\mathbb{E}_{x\sim\rho,y'\sim\pi}\left[(-\pi_{y'} + \pi^R_{y'})\left(\mathbb{E}_{y\sim\pi}\big[g\big(p_s(y), p_h(y)\big)\big] - g\big(p_s(y'), p_h(y')\big)\right)\right]
\end{aligned} \tag{24}$$

The third equality is by the independence of y and y'. By the change of notation, the second term of the last line can be written as

$$\begin{aligned}
&\mathbb{E}_{x\sim\rho,y'\sim\pi}\left[(-\pi_{y'} + \pi^R_{y'})\left(\mathbb{E}_{y\sim\pi}\big[g(p_s(y), p_h(y))\big] - g(p_s(y'), p_h(y'))\right)\right]\\
=&\mathbb{E}_{x\sim\rho,y\sim\pi}\left[(-\pi_y + \pi^R_y)\left(\mathbb{E}_{y'\sim\pi}\big[g(p_s(y'), p_h(y'))\big] - g(p_s(y), p_h(y))\right)\right]
\end{aligned} \tag{25}$$

Then Equation (24) can be written as

$$(24) = \mathbb{E}_{x\sim\rho,y\sim\pi}\left[(\pi_y - \pi^R_y)\cdot 2\left(g(p_s(y), p_h(y)) - \mathbb{E}_{y'\sim\pi}\big[g(p_s(y'), p_h(y'))\big]\right)\right] \tag{26}$$

Now we plug in $g(p_s(y), p_h(y)) = (p_s(y)+E_s)(p_h(y)+\frac{1}{2})$ and use the fact $\mathbb{E}_{y'\sim\pi}[p_h(y'\succ\pi)] = \frac{1}{2}$. Equation (26) can be expanded as

$$\begin{aligned}
(24) =&\mathbb{E}_{x\sim\rho,y\sim\pi}\left[(\pi_y - \pi^R_y)\cdot 2\left((p_s(y)+E_s)(p_h(y)+\frac{1}{2}) - \mathbb{E}_{y'\sim\pi}\big[(p_s(y')+E_s)(p_h(y')+\frac{1}{2})\big]\right)\right]\\
=&\mathbb{E}_{x\sim\rho,y\sim\pi}\left[(\pi_y - \pi^R_y)\cdot 2\left((p_s(y)+E_s)(p_h(y)+\frac{1}{2}) - 2E_s\right)\right]\\
=&\mathbb{E}_{x\sim\rho,y\sim\pi}\left[(\pi_y - \pi^R_y)\cdot(2p_s(y)p_h(y) + 2E_s p_h(y) + p_s(y) - 3E_s)\right]
\end{aligned} \tag{27}$$

The cross term of Equation (23) can be written as

$$\begin{aligned}
&\mathbb{E}_{x\sim\rho,y,y'\sim\pi}\mathbb{E}_{I\sim\text{Bernoulli}}\left[h_\pi(y,y')g_I(y,y')\right]\\
=&\mathbb{E}_{x\sim\rho,y,y'\sim\pi}\mathbb{E}_{I\sim\text{Bernoulli}}\left[(\pi_y - \pi_{y'} + \pi^R_{y'} - \pi^R_y)g_I(y,y')\right]
\end{aligned} \tag{28}$$

Now we plug in $g_I = I_{\text{help}}(y\succ y')I_{\text{safe}}(y) + I_{\text{help}}(y\succ y')I_{\text{safe}}(y') + \frac{1}{2}I_{\text{safe}}(y) - \frac{3}{2}I_{\text{safe}}(y')$,

$$\begin{aligned}
(28) =&\mathbb{E}_{x\sim\rho,y,y'\sim\pi}\mathbb{E}_{I\sim\text{Bernoulli}}\Big[(\pi_y - \pi_{y'} + \pi^R_{y'} - \pi^R_y)\big(I_{\text{help}}(y\succ y')I_{\text{safe}}(y)\\
&\qquad + I_{\text{help}}(y\succ y')I_{\text{safe}}(y') + \frac{1}{2}I_{\text{safe}}(y) - \frac{3}{2}I_{\text{safe}}(y')\big)\Big]\\
=&\mathbb{E}_{x\sim\rho,y,y'\sim\pi}\mathbb{E}_{I\sim\text{Bernoulli}}\Big[(\pi_y - \pi^R_y)\big(I_{\text{help}}(y\succ y')I_{\text{safe}}(y)\\
&\qquad + I_{\text{help}}(y\succ y')I_{\text{safe}}(y') + \frac{1}{2}I_{\text{safe}}(y) - \frac{3}{2}I_{\text{safe}}(y')\big)\Big]\\
&+\mathbb{E}_{x\sim\rho,y,y'\sim\pi}\mathbb{E}_{I\sim\text{Bernoulli}}\Big[(-\pi_{y'} + \pi^R_{y'})\big(I_{\text{help}}(y\succ y')I_{\text{safe}}(y)\\
&\qquad + I_{\text{help}}(y\succ y')I_{\text{safe}}(y') + \frac{1}{2}I_{\text{safe}}(y) - \frac{3}{2}I_{\text{safe}}(y')\big)\Big]
\end{aligned}$$

With the similar change of notation as Equation (25), as well as the fact that $1 - I_{\text{help}}(y\succ y') = I_{\text{help}}(y'\succ y)$, the last line can be written as

$$\begin{aligned}
&\mathbb{E}_{x\sim\rho,y,y'\sim\pi}\mathbb{E}_{I\sim\text{Bernoulli}}\Big[(-\pi_{y'} + \pi^R_{y'})\big(I_{\text{help}}(y\succ y')I_{\text{safe}}(y)\\
&\qquad + I_{\text{help}}(y\succ y')I_{\text{safe}}(y') + \frac{1}{2}I_{\text{safe}}(y) - \frac{3}{2}I_{\text{safe}}(y')\big)\Big]\\
=&\mathbb{E}_{x\sim\rho,y,y'\sim\pi}\mathbb{E}_{I\sim\text{Bernoulli}}\Big[(-\pi_y + \pi^R_y)\big(I_{\text{help}}(y'\succ y)I_{\text{safe}}(y')\\
&\qquad + I_{\text{help}}(y'\succ y)I_{\text{safe}}(y) + \frac{1}{2}I_{\text{safe}}(y') - \frac{3}{2}I_{\text{safe}}(y)\big)\Big]\\
=&\mathbb{E}_{x\sim\rho,y,y'\sim\pi}\mathbb{E}_{I\sim\text{Bernoulli}}\Big[(-\pi_y + \pi^R_y)\big((1 - I_{\text{help}}(y\succ y'))I_{\text{safe}}(y')\\
&\qquad + (1 - I_{\text{help}}(y\succ y'))I_{\text{safe}}(y) + \frac{1}{2}I_{\text{safe}}(y') - \frac{3}{2}I_{\text{safe}}(y)\big)\Big]
\end{aligned}$$

Then we further expand Equation (28) as

$$
\begin{aligned}
(28) =& \mathbb{E}_{x\sim\rho,y,y'\sim\pi}\mathbb{E}_{I\sim\text{Bernoulli}}\Big[(\pi_y - \pi_y^R)\big(I_{\text{help}}(y \succ y')I_{\text{safe}}(y) \\
&\qquad + I_{\text{help}}(y \succ y')I_{\text{safe}}(y') + \frac{1}{2}I_{\text{safe}}(y) - \frac{3}{2}I_{\text{safe}}(y')\big)\Big] \\
&+ \mathbb{E}_{x\sim\rho,y,y'\sim\pi}\mathbb{E}_{I\sim\text{Bernoulli}}\Big[(-\pi_y + \pi_y^R)\big((1 - I_{\text{help}}(y \succ y'))I_{\text{safe}}(y') \\
&\qquad + (1 - I_{\text{help}}(y \succ y'))I_{\text{safe}}(y) + \frac{1}{2}I_{\text{safe}}(y') - \frac{3}{2}I_{\text{safe}}(y)\big)\Big] \\
=& \mathbb{E}_{x\sim\rho,y,y'\sim\pi}\mathbb{E}_{I\sim\text{Bernoulli}}\Big[(\pi_y - \pi_y^R)\big(2I_{\text{help}}(y \succ y')I_{\text{safe}}(y) \\
&\qquad + 2I_{\text{help}}(y \succ y')I_{\text{safe}}(y') + I_{\text{safe}}(y) - 3I_{\text{safe}}(y')\big)\Big]
\end{aligned} \tag{29}
$$

Taking the expectation over y' and the Bernoulli variables, we have

$$
(28) = \mathbb{E}_{x\sim\rho,y\sim\pi}\Big[(\pi_y - \pi_y^R)\big(2p_h(y)p_s(y) + 2E_s p_h(y) + p_s(y) - 3E_s\big)\Big] \tag{30}
$$

This equation is the same as Equation (27), which ends the proof that Equation (22) and Equation (23) are equivalent! □

As discussed in Appendix B.2, we can freely change the order of y and y' in Equation (22) and Equation (23). Thus, the proof of Theorem B.4 also applies to Theorem 3.2.

B.4 Relation of the Constants

In this section, we derive a more general form of Theorem B.4, where, with specific relations between the constants in g and g_I, the optimization problem in Equation (22) is equivalent to the optimization problem in Equation (23).

We restate g and g_I here with the notations used in the Appendix for convenience.

$$
g = (p_s(y) + A_1)(p_h(y) + A_2),
$$

and

$$
\begin{aligned}
g_I(y, y') =& (B_1B_3 - B_3)I_{\text{help}}(y \succ y')I_{\text{safe}}(y) + (B_1B_3 - B_3)I_{\text{help}}(y \succ y')I_{\text{safe}}(y') \\
&- B_1B_3I_{\text{safe}}(y') + B_3I_{\text{safe}}(y) + 2B_2B_3I_{\text{help}}(y \succ y') - B_2B_3
\end{aligned}
$$

As discussed in the proof of Theorem B.4, we only need to find the relationship such that the cross terms of the two optimization problems are identical. We first expand the cross term of the optimization problem in Equation (22). As in Equation (26), it can be written as

$$
(24) = \mathbb{E}_{x\sim\rho,y\sim\pi}\left[(\pi_y - \pi_y^R)\cdot 2\left(g(p_s(y), p_h(y)) - \mathbb{E}_{y'\sim\pi}\left[g(p_s(y'), p_h(y'))\right]\right)\right] \tag{31}
$$

Using the same strategy of obtaining Equation (29), we have

$$
\begin{aligned}
(28) = \mathbb{E}_{x\sim\rho,y\sim\pi}\Big[&(\pi_y - \pi_y^R)\big(2B_3(B_1 - 1)p_s(y)p_h(y) \\
&+ 2B_3((B_1 - 1)E_s + 2B_2)p_h(y) + 2B_3p_s(y) - 2B_1B_3E_s - 2B_2B_3\big)\Big]
\end{aligned} \tag{32}
$$

Aligning the coefficients of each term in Equation (31) and Equation (32), we derive the following set of equations:

$$
\begin{aligned}
B_3(B_1 - 1) &= 1, \\
E_s + 2B_3B_3 &= A_1, \\
B_3 &= A_2.
\end{aligned} \tag{33}
$$

Solving these equations gives us the specific forms of g and g_I. Here B_2 is a shifting value that we define to align with our intuition. B_3 is a scaling factor that is related to the penalty τ.

B.5 Discussion of the Property of g_I

In this section, we discuss the two beneficial properties of g_I that we proposed in Section 3.2.

Skew-Symmetric Property. First, we examine the skew-symmetric property of g_I. When combined with the skew-symmetric property of h, this implies:

$$\left(h_\pi(y, y') - \tau^{-1} g_I(y, y')\right)^2 = \left(h_\pi(y', y) - \tau^{-1} g_I(y', y)\right)^2.$$

This means that for the same data point, regardless of the order of y and y', we are always driving $h_\pi(y, y')$ to the same value. In contrast, in IPO (Azar et al., 2024), different orders will push $h_\pi(y, y')$ to different values, i.e., they form two different optimization problems:

$$(h_\pi(y, y') - \tau^{-1} g_I(y, y'))^2 \quad \text{and} \quad (h_\pi(y', y))^2.$$

Their final optimization problem, $(h_\pi(y, y') - \frac{1}{2}\tau^{-1} g_I(y, y'))^2$, tries to find a middle point of h that optimizes both. However, this point is neither the optimal solution of the first problem nor the second problem.

Shifting Property. Second, we discuss the shifting properties of g_I. Since Theorem 3.2 holds based on the equality of Equation (30) and Equation (27), and all the operations to derive these two equations are valid under linear transformations of $p^*_{\text{safe}}, p^*_{\text{help}}$ and $I_{\text{safe}}, I_{\text{help}}$, respectively. It implies that Theorem 3.2 also holds under the same linear transformations of $p^*_{\text{safe}}, p^*_{\text{help}}$ and $I_{\text{safe}}, I_{\text{help}}$.

This property is useful when we want to manually design the values of g_I, as shown in Figure 3.

C Experiment

C.1 Details of Harmlessness Benchmark

The following are the details of the datasets used in the benchmark:

- **CrowS-Pairs**: A challenge set for evaluating the tendency of language models (LMs) to generate biased outputs. We evaluate the English subset and implementation by LM Evaluation Harness and report the Percentage of Stereotype metric.
- **Bias Benchmark for QA (BBQ)**: Measures social biases in the output of multiple-choice question-answering models. We use the Lite version and implementation by BigBench and report the Accuracy metric.
- **WinoGrande**: A collection of 44K problems inspired by the Winograd Schema Challenge, adjusted to improve scale and robustness against dataset-specific biases. We use the implementation by LM Evaluation Harness and report the Accuracy metric.
- **ETHICS**: A benchmark spanning concepts in justice, well-being, duties, virtues, and commonsense morality. We use the implementation by LM Evaluation Harness and report the Accuracy for all sub-tasks.
- **BigBench Moral Permissibility**: Tests whether ultra-large language models can read a short story where a moral scenario is presented and answer the question, "Is it morally permissible to do X?" in a manner similar to humans. We use the implementation by BigBench and report the Accuracy metric.
- **BigBench Simple Ethics Questions**: Assesses a language model's responses to hypothetical, consequential, political, and social questions. We use the implementation by BigBench and report the Accuracy metric.
- **ToxicGen**: A benchmark for evaluating the ability of language models to classify input text as either hateful or not hateful. We use the implementation by LM Evaluation Harness and report the Accuracy metric.
- **BigBench HHH Alignment**: Evaluates language models on alignment, pragmatically broken down into the categories of helpfulness, honesty/accuracy, harmlessness, and other aspects. We use the implementation by BigBench and report the Accuracy metric.

- **AdvBench** contains harmful prompts. We use the prompts provided here and generation implementation by LM Evaluation Harness. We report the percentage of harmless responses measured by HarmBench-Llama-2-13b-cls model.
- **RealToxicityPrompts**: A benchmark for evaluating the ability of language models to continue a prompt in a non-toxic way. We use the implementation by LM Evaluation Harness report the percentage of harmless responses measured by HarmBench-Llama-2-13b-cls model.
- **ALERT**: A benchmark to assess the safety of LLMs through red teaming methodologies. We use the prompts provided here and generation implementation by LM Evaluation Harness. We report the percentage of harmless responses measured by HarmBench-Llama-2-13b-cls model.
- **ALERT Adversarial**: A benchmark to assess the safety of LLMs through red teaming methodologies with adversarial prompts. We use the prompts provided here and generation implementation by LM Evaluation Harness. We report the percentage of harmless responses measured by HarmBench-Llama-2-13b-cls model.
- **AlpacaEval** Based on the AlpacaFarm evaluation set, which tests the ability of models to follow general user instructions. We employ the official implementation report the LC Win Rate.

C.2 Details of Baselines

The following are the details of the methods that align LLMs for multiple objectives.

- **Llama2** (Touvron et al., 2023) trains the safety reward r_{safe} and the helpfulness reward r_{help} separately, and defines the global reward g as a combination of these rewards, *i.e.*,

$$\tilde{g}(y|x) = \begin{cases} r_{\text{safe}}(y|x) \text{ if IS_SAFETY}(x), \text{ or } r_{\text{safe}}(y|x) < 0.15, \\ r_{\text{help}}(y|x) \text{ otherwise,} \end{cases}$$
$$g(y|x) = \text{WHITEN}(\text{LOGIT}(\tilde{g}(y|x))).$$

  Here IS_SAFETY(x) indicates whether prompts are tagged as unsafe in their dataset, and the 0.15 threshold is chosen to filter unsafe responses according to the evaluation on Meta Safety test set. Whitening the final linear scores is to increase stability. The global reward is used in the RLHF objective in Equation (3).

- **Beaver** (Dai et al., 2024) trains the safety reward r_{safe} and the helpfulness reward r_{help} separately, and defines the final RLHF objective as the dual optimization problem of the conditional RLHF, obtained by Lagrangian dual transformation, *i.e.*,

$$\min_{\theta} \max_{\lambda \geq 0} \mathbb{E}_{x \sim D, y \sim \pi_\theta} \left[-r_{\text{help}}(y|x) + \lambda \left(r_{\text{safe}}(y|x) + d \right) \right],$$

  where $\lambda \geq 0$ is the Lagrange multiplier. In practice, the model parameter θ and the Lagrange multiplier λ are updated iteratively.

- **RBR** (Mu et al., 2024) requires separate reward models, $r_{\phi_1}, \ldots, r_{\phi_k}$, for each objective, and propose to learn the weight for each objective, *i.e.*,
-

$$g(y|x) = \sum_{i=1}^{k} \lambda_i r_i(y|x),$$

  where λ_i are learnable parameters. The global reward is used in the RLHF objective in Equation (3).

- **SteerLM** (Dong et al., 2023) trains models to generate response according to a specific reward vector $r = (r_1, r_2, r_3, \ldots, r_k)$. They first train a model to predict the score for each objective in a dataset. Supervised fine-tuning is performed to maximize the probability of generating responses conditioned on the reward vector and the prompt, *i.e.*,

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

- **MORL** (Ramé et al., 2023) trains reward models for each objective separately, and defines the global reward g as a combination of rewards, *i.e.*,

$$g(y|x) = \sum_{i=1}^{k} \lambda_i r_i(y|x),$$

The global reward is used in the RLHF objective in Equation (3).

- **ArmoRM** (Wang et al., 2024b) applies the same training strategy as MORL, but uses a single publicly available reward model, ArmoRM-Llama3-8B-v0.1 (Wang et al., 2024c), to provide the reward scores for all objectives.
- **MODPO** (Zhou et al., 2023) trains margin reward models $r_i, i = 1, \ldots, k$ for each objective separately, and performs supervised fine-tuning with the objective,

$$\max \mathbb{E}_{(x,y^w,y^l)\sim D} \log \sigma \left(\frac{1}{\omega_k} \left(\tau \log \frac{\pi_\theta(y^w|x)}{\pi_{\text{ref}}(y^w|x)} - \tau \log \frac{\pi_\theta(y^l|x)}{\pi_{\text{ref}}(y^l|x)} - \omega_{-k}^T (r_{-k}(x, y^w) - r_{-k}(x, y^l)) \right) \right),$$

where ω_k is the weight for the objective k, ω_{-k} is the weight vector for all other objectives, and r_{-k} is the reward vector for all other objectives than k. This fine-tuning is performed for each objective.

- **MinMaxRLHF** (Chakraborty et al., 2024) addresses the scenario where different annotators h may have preferences for different objectives. The algorithm uses the EM algorithm to learn the distribution of rewards for multiple objectives. In the E step, they find the certain objective i that each human annotator h relates to, *i.e.*,

$$\mathcal{I}_h = \arg\max_i \Pi_{x,y,y',h} \frac{\exp(r_{\phi_i}(x, y))}{\exp(r_{\phi_i}(x, y)) + \exp(r_{\phi_i}(x, y'))},$$

where r_{ϕ_i} is the reward model for the objective i. In the M step, each reward model i is updated by the reward learning objective in Equation (1) with the data assigned to objective i, *i.e.*, the dataset is $D_i = \{(x, y, y', h), \mathcal{I}_h = i\}$. In the RLHF stage, they maximize the minimum reward of all reward scores, *i.e.*,

$$\mathbb{E}_{x\sim D, y\sim\pi_\theta} \left[\min_i r_{\phi_i}(x, y) - \tau \text{KL} \left[\pi_\theta(y|x) || \pi_{\text{ref}}(y|x) \right] \right].$$

Among these methods, MODPO is highly inefficient since it requires separate RLHF for each objective. Other methods typically use a linear combination of reward scores for multiple objectives or one reward as a threshold for others. For the combination of thresholding, the global function can be approximated by the multiplication of rewards for each objective when the reward scores are on the same scale. Maximizing the multiplication of rewards has the same effect as maximizing the minimum reward. Therefore, we hypothesize that the global reward should be a bilinear combination of the reward scores as in Equation (10). As such, in the experiment section, we select MORL as a representative for this line of approach.

C.3 Full Experiment Results

Table 6 shows the full results of the open-sourced models, our baselines, and our models on the benchmarks. Here are the details of each open-sourced models:

- Zephyr: `https://huggingface.co/HuggingFaceH4/zephyr-7b-beta`
- Juanako: `https://huggingface.co/fblgit/juanako-7b-UNA`
- OpenChat: `https://huggingface.co/openchat/openchat_3.5`
- Mistral: `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2`
- Beaver: `https://huggingface.co/PKU-Alignment/beaver-7b-v3.0`
- Llama2: `https://huggingface.co/meta-llama/Llama-2-7b-chat-hf`

Table 6: Full results of the open-sourced models and safety aligned on the benchmarks.

	Zephyr	Juanako	OpenChat	Mistral	Beaver	Llama2	Llama3
Alpaca Eval	10.99	2.88	11.08	14.72	1.00	7.60	**22.90**
Crows Pairs	62.02	63.74	66.67	64.88	56.23	63.98	63.45
BBQ	39.00	84.00	61.00	61.84	31.37	32.99	60.68
Winogrande	72.38	77.43	72.69	73.80	65.35	66.46	71.82
Ethics CM	68.37	75.96	68.88	73.46	59.43	56.14	58.64
Ethics Justice	69.71	76.41	77.74	71.93	64.61	50.00	70.38
Ethics Deontology	56.98	64.10	63.96	60.26	61.48	50.00	64.49
Ethics Utilitarianism	73.59	73.79	73.48	66.78	56.01	57.97	62.92
Ethics Virtue	91.30	89.13	88.70	90.87	61.61	72.00	81.49
Moral Permissibility	51.00	49.00	50.00	47.95	47.66	47.37	48.54
Simple Ethical Questions	33.00	82.00	91.00	53.91	45.22	24.35	54.78
Toxigen	45.21	60.96	42.34	55.11	36.17	51.00	45.74
HHH Alignment	46.00	49.00	46.00	47.06	43.44	44.34	45.25
Disc. Avg.	59.05	**70.46**	66.87	63.99	52.38	51.38	60.68
AdvBench	85.82	85.90	87.82	83.74	85.07	87.91	89.49
RealToxicityPrompts	20.19	27.50	48.27	65.38	93.20	100.00	99.42
ALERT	79.08	80.70	75.36	90.44	91.83	98.62	95.18
ALERT Adversarial	66.68	72.79	73.09	77.71	94.80	98.32	95.08
Generative Average	62.94	66.72	71.13	79.32	91.23	**96.21**	94.79
Safety Average	60.99	68.59	69.00	71.65	71.80	73.80	**77.74**

- Llama3: `https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct`

Table 7 and Table 8 shows the full results of our baselines and our models on the benchmarks. Here are the details of the data used in our model and the baselines

We use 4 Nvidia A100 GPUs for each experiment, and the training time for each experiment is around 6 hours for SFT and 6 hours for BFPO. For the experiments with red teaming data, we use 1.5K data collected as described in Section 4.3 and only performs the BFPO stage. The training time for this experiment is around 10 minutes with 4 Nvidia A100 GPUs.

Table 7: Full results of Table 1

	Mistral + DPO-H	Mistral + DPO-S	Mistral + DPO	Mistral + IPO	Mistral + MORL	Mistral + BFPO
Alpaca Eval	10.99	4.34	**14.71**	13.16	10.83	13.33
Crows Pairs	62.02	65.65	65.59	66.25	61.66	65.77
BBQ	39.00	39.50	43.68	42.44	39.43	45.25
Winogrande	72.38	74.03	74.27	74.66	71.51	74.98
Ethics CM	68.37	64.22	56.47	62.03	68.01	65.25
Ethics Justice	69.71	55.29	71.01	66.35	67.71	59.13
Ethics Deontology	56.98	50.86	58.20	54.67	55.70	51.97
Ethics Utilitarianism	73.59	60.00	57.15	67.03	72.57	70.36
Ethics Virtue	91.30	89.73	86.71	89.17	91.08	90.41
Moral Permissibility	51.00	46.78	51.17	47.37	50.58	47.37
Simple Ethical Q.	33.00	38.26	39.13	37.39	33.91	39.13
Toxigen	45.21	47.45	51.06	48.72	44.15	54.15
HHH Alignment	46.00	45.25	45.70	44.80	46.15	45.25
Disc. Avg.	59.05	56.42	58.35	58.41	58.54	**59.09**
AdvBench	85.82	87.74	82.49	86.41	87.07	87.32
RealToxicityPrompts	20.19	100.00	4.23	88.65	21.15	98.65
ALERT	79.08	99.91	38.64	96.00	82.13	98.56
ALERT Adversarial	66.68	99.98	33.46	88.00	69.16	96.42
Gen. Avg.	62.94	**96.91**	39.71	89.76	64.88	95.24
Safety Avg.	60.99	76.66	49.03	74.09	61.71	**77.16**

Table 8: Full results of Table 2 and Table 4

	DPO	IPO	MORL	BFPO	BFPO w/o buffer	BFPO w/o shift
Alpaca Eval	13.07	13.74	12.56	**14.41**	12.76	**15.59**
Crows Pairs	61.84	61.96	61.66	61.72	65.95	65.65
BBQ	38.95	38.89	38.47	39.44	44.44	44.43
Winogrande	72.45	72.77	71.98	72.45	74.98	74.43
Ethics CM	67.77	68.03	66.07	67.28	62.50	61.13
Ethics Justice	69.12	69.30	69.16	68.01	59.87	69.05
Ethics Deontology	57.48	57.62	56.98	57.20	52.28	56.56
Ethics Utilitarianism	73.63	73.54	73.02	73.46	66.64	67.08
Ethics Virtue	91.42	91.48	91.36	91.54	88.78	89.79
Moral Permissibility	50.88	50.58	51.17	49.12	47.66	47.37
Simple Ethical Q.	36.52	36.52	33.04	37.39	50.43	46.96
Toxigen	45.11	45.43	44.26	45.32	49.89	53.94
HHH Alignment	46.15	45.70	45.70	45.25	45.70	45.25
Disc. Avg.	59.28	**59.32**	58.57	59.02	59.09	**60.14**
AdvBench	87.99	87.91	87.66	87.82	84.32	84.82
RealToxicityPrompts	44.00	41.35	21.15	86.54	96.15	85.77
ALERT	89.22	87.48	82.13	86.34	96.90	95.37
ALERT Adversarial	76.33	74.55	69.16	94.47	94.12	89.08
Gen. Avg.	74.39	72.82	65.02	**88.79**	**92.87**	88.76
Safety Avg.	66.83	66.07	61.80	**73.90**	**75.98**	74.45