# Triple Preferences Optimization (TPO): A Simple One-Step Combination of SFT and Preference Optimization with a Better Performance

Anonymous ACL submission

## Abstract

Large Language Models (LLMs) excel across various tasks. However, aligning them with human demonstrations proves challenging. Prior approaches relied on Reinforcement Learning from Human Feedback (RLHF) using online RL methods like Proximal Policy Optimization (PPO). Recently, RL-free methods like Direct Preference Optimization (DPO) have emerged as appealing alternatives, offering improved stability and scalability while retaining competitive performance. However, these methods have a separate supervised fine-tuning (SFT) step for further learning and require sampling from the post-SFT model and ranking them. In this paper, we introduce **Triple Preferences Optimization (TPO)**, a new preference learning method designed to align an LLM with three preferences without requiring a separate supervised fine-tuning step. Our TPO aims to maximize the log probability of preferred to less-preferred responses while simultaneously learning the gold standard response in a single step. To provide a comprehensive evaluation, we use HuggingFace Open LLMs Benchmarks and MT-Bench (Zheng et al., 2023) involving dialogue systems and encompassing various NLP aspects. The results indicate that TPO surpasses other alignment methods, such as DPO and SFT, in average accuracy by 1.8% and 2.5%, respectively. Notably, TPO without the SFT part exhibits superior average accuracy compared to DPO and SFT by **4%** and **4.7%**, respectively. Overall, TPO resolves sampling challenges and combines the SFT part with the preference optimization part into a single step and provides better performance.[1]

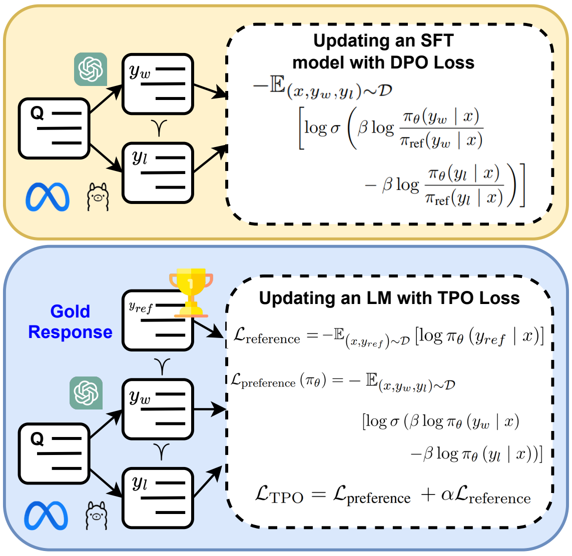


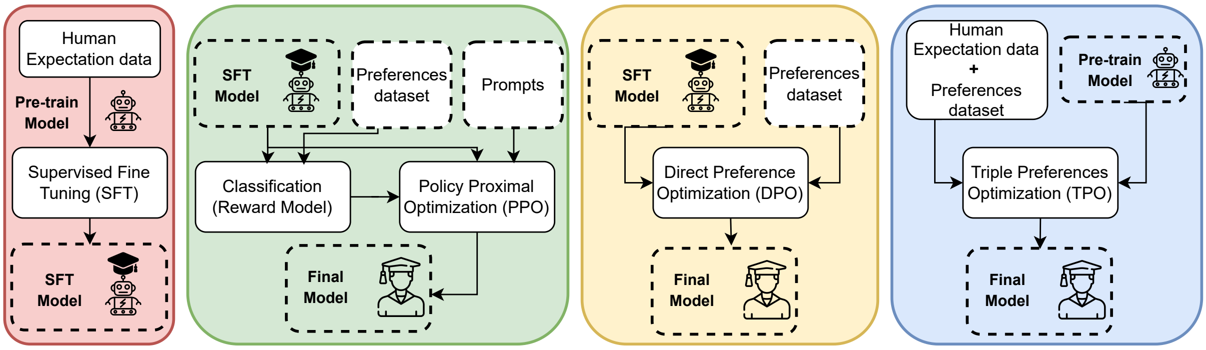

Figure 1: Comparison of the loss functions of TPO and DPO. TPO's loss function incorporates two main objectives. Its first term optimizes the log probability of preferences ($\mathcal{L}_{\text{preference}}(\pi_\theta)$) that draws from CPO's insights (Xu et al., 2024), which demonstrated that optimizing preferences doesn't necessitate a reference model. Through its second term TPO aims to learn the gold standard response ($\mathcal{L}_{\text{reference}}$). This aspect of the loss function is regulated by a parameter $\alpha$, which serves as a parameter controlling the extent to which the policy model learns the gold standard response.

## 1 Introduction

LLMs are trained across a wide array of tasks, demonstrating their remarkable versatility in solving diverse tasks (Brown et al., 2020; Narayanan et al., 2021; Bubeck et al., 2023). However, their training on data of varying quality can lead to many issues, such as the generation of toxic or harmful text under certain contexts (Perez et al., 2022; Ganguli et al., 2022), and in general generation of undesired (by humans) outputs. Hence, it is crucial to align LLMs with human expectations and preferences that prioritize their helpfulness, honesty, and harmlessness (Bai et al., 2022).

Supervised fine-tuning (SFT) is a direct alignment method that involves fitting a model to human-written data (Sanh et al., 2022). However, this approach fails to fully impart the human perspective to the model. During training, the model only re-

[1]Data and code are available at https://anonymous.4open.science/r/triple_preferences_optimization-E63F/




(a) SFT (b) RLHF (PPO) (c) DPO (d) TPO

Figure 2: (a) During the supervised fine-tuning step, a pre-trained model is fine-tuned to align with human expectations. (b) To further enhance the performance of the SFT model, we train it with human preferences using reinforcement learning. (c) Alternatively, we can directly align an SFT model with human preferences using RL-free methods such as DPO. (d) In TPO, we merge preference optimization with gold standard response learning, enabling direct fine-tuning of a pre-trained model based on three preferences.

ceives a reference response for each input, thus lacking exposure to incorrect answers and preferences, which ultimately constrains its performance on downstream tasks (Touvron et al., 2023).

A prominent method in AI alignment for LLMs is Reinforcement Learning with Human Feedback (RLHF) (Ouyang et al., 2022). Despite its impressive performance relative to SFT, RLHF faces limitations such as instability and susceptibility to reward hacking (Liu et al., 2024). Consequently, a recent approach called Direct Preference Optimization (DPO) (Rafailov et al., 2023) has emerged. DPO is an RL-free method that directly optimizes human preferences by shifting from RL to simple binary cross-entropy. However, DPO encounters several limitations: 1) high dependency on the SFT part , 2) tendency to overfit beyond a single epoch, and 3) inefficient learning and memory utilization.

Various alignment methods have been proposed for dialogue systems (Tunstall et al., 2023), harmful and helpfulness question answering (Wu et al., 2023), summarization (Zhao et al., 2023), and translation (Xu et al., 2024) to address these limitations. However, all these studies include a separate SFT component.

During SFT, models are fine-tuned with respect to what response is appropriate for what input prompt, while during DPO optimization, models are fine-tuned to optimize the "relative log probability of preferred to dispreferred responses" (Rafailov et al., 2023) and not to stray far away from the SFT trained model. In this paper we combine these two steps into a single step, with the input data having both the gold standard response (as in SFT) and the preferences (as in PPO/DPO) in a single format. Thus our training data will be of the form *(input prompt, gold standard response, preferred response, less-preferred response)*. Mostly we jointly optimize a policy model with $-\mathbb{E}_{(x,y_{ref})\sim\mathcal{D}}\left[\log \pi_\theta\left(y_{ref} \mid x\right)\right]$ and $-\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\pi_\theta\left(y_w \mid x\right) \; - \beta\log\pi_\theta\left(y_l \mid x\right)\right)\right]$ in one step. We refer to our approach as Triple Preferences Optimization (TPO).

Our results show that TPO exhibits impressive performance compared to SFT across various benchmarks and outperforms other alignment methods, including DPO. TPO surpasses SFT, DPO, Kahneman-Tversky Optimization (KTO) (Ethayarajh et al., 2023), Identity Preference Optimization (IPO) (Azar et al., 2023), and Constrictive Preference Optimization (CPO) (Xu et al., 2024) in the conventional alignment process involving the SFT part by, on average, 2.54%, 1.81%, 1.62%, 2.95%, and 1.64% respectively, in terms of accuracy. Notably, TPO without the SFT part, on average, improves the performance by 2.2% compared with SFT and other methods. On average, TPO demonstrates comparable performance across both MT-Bench and Big-Bench benchmarks.

Overall, TPO addresses two key shortcomings in alignment tasks. Firstly, by removing $\pi_{ref}$ as done in CPO, TPO mitigates the inefficient learning and memory utilization issues observed in DPO, allowing for more computational efficiency with less memory usage. Secondly, TPO resolves the high dependency of alignment methods on the SFT component by introducing a second term in the loss function. We demonstrate that this approach not only achieves comparable performance with SFT across various benchmarks but also exhibits further improvement when incorporating the SFT model into the TPO method. These findings suggest that a separate SFT is not necessary for TPO and, in certain scenarios, may even hinder TPO's performance.

We summarize our primary contributions as:

1. We propose a new preference learning method called Triple Preference Optimization (TPO) that simplifies the post-pretraining alignment.
2. We show that TPO achieves comparable performance both with and without SFT compared to SFT and other alignment methods across various benchmarks. Additionally, we demonstrate that TPO without SFT exhibits superior performance.
3. We examine the performance of TPO across different values of $\alpha$ which serves as a regularization parameter to control gold standard response learning. Our findings reveal that TPO achieves better performance at $\alpha = 1$ without SFT part, indicating the LLM's capacity for more exploration.

## 2 Related Works

The performance of Large Language Models (LLMs) across diverse tasks is noteworthy (Anil et al., 2023). However, their efficacy in downstream tasks and alignment with user feedback has been notably enhanced through fine-tuning instructions and human-written completion datasets (Mishra et al., 2022). Fine-tuning models on instructions not only fosters generalization beyond the tuning set but also enhances overall usability (Chung et al., 2022). Despite the effectiveness of instruction tuning, collecting relative human judgments of response quality is often simpler than obtaining expert demonstrations. Consequently, subsequent studies have fine-tuned LLMs using datasets of human preferences, leading to improvements in translation (Kreutzer et al., 2018), summarization (Stiennon et al., 2022), story-telling (Ziegler et al., 2019), instruction-following (Ramamurthy et al., 2023), and dialogue systems. In this section, we provide a succinct overview of various tasks and settings where LLMs have demonstrated notable performance in alignment.

RLHF (Christiano et al., 2023), introduced in the literature, aims to optimize for maximum reward by interacting with a reward model trained using the Bradley-Terry (BT) model (Bong and Rinaldo, 2022), typically through reinforcement algorithms like Proximal Policy Optimization (PPO) (Schulman et al., 2017). While RLHF enhances model performance, it faces challenges such as instability, reward hacking, and scalability inherent in reinforcement learning. Recent works have presented techniques to overcome these challenges by optimizing relative preferences without relying on reinforcement learning. Utilizing the Bradley-Terry (BT) model to optimize a model on preference datasets is instrumental in ensuring alignment with human preferences.

Sequence Likelihood Calibration (SLiC) (Zhao et al., 2023) introduced a novel method for ranking preferences generated by a supervised fine-tuned (SFT) model, incorporating calibration loss and regularization fine-tuning loss during training. Meanwhile, RRHF (Yuan et al., 2023) trains the SFT model using a zero-margin likelihood contrastive loss, assuming multiple ranked responses for each input. While both SLiC and RRHF are effective, they lack theoretical foundations. In contrast, DPO offers a method to directly fit an SFT model to human preferences using the Bradley-Terry (BT) model, providing theoretical insights into the alignment process.

Statistical Rejection Sampling Optimization (RSO) (Liu et al., 2024) merges the techniques of SLiC and DPO while introducing an improved approach for collecting preference pairs through statistical rejection sampling. IPO (Azar et al., 2023), like DPO methods, has mathematically revealed the limitations of the DPO approach concerning overfitting and generalization. It proposes a comprehensive objective for learning from human preferences. Zephyr (Tunstall et al., 2023) has improved DPO by utilizing state-of-the-art (SOTA) models to generate responses for the same input and ranking them using teacher models such as GPT-4. Moreover, they emphasize the importance of SFT as an initial step before implementing DPO.

KTO (Ethayarajh et al., 2023), drawing inspiration from Kahneman and Tversky's influential work on prospect theory (TVERSKY and KAHNEMAN, 1992), seeks to maximize the utility of LLM outputs directly rather than optimizing the log-likelihood of preferences. By prioritizing the determination of whether a preference is desirable or undesirable, this method eliminates the requirement for two preferences for the same input.

Self-Play fIne-tuNing (SPIN) (Chen et al., 2024) introduced a self-training approach to augment DPO using the dataset utilized in the SFT step. The core concept of this strategy is to leverage synthetic data generated as the rejected response and the gold standard response from the SFT dataset as the chosen response. On the other hand, CPO (Xu et al., 2024) proposed an efficient method for learning preferences by integrating the maximum-likelihood loss and the DPO loss function with the aim of enhancing memory and learning efficiency.

We observe two primary challenges in the alignment process addressed by the aforementioned studies. Firstly, there's a significant reliance on the supervised fine-tuning (SFT) component. Zephyr's findings indicate that the DPO approach fails to learn without SFT. Secondly, there are concerns regarding inefficient learning and memory usage. While CPO demonstrates that removing $\pi_{ref}$ doesn't decrease performance compared with DPO, it falls short in comparison to DPO and other alignment methods like KTO. In this investigation, we delve into these limitations and endeavor to introduce a novel method to address them.

## 3 Method

In this section, we introduce a novel preference learning method named Triple Preferences Optimization (TPO) for learning both gold and preference responses. This method is devised to train a model in dialogue systems to generate accurate responses to questions by considering three preferences: one reference, one chosen, and one rejected, all for the same input.

### 3.1 Triple Preferences Optimization

In this subsection, we outline the objective of the TPO method. First, we begin with an analysis of DPO. Given a set of sources alongside preferred response $y_w$ and less-preferred response $y_l$, for the same input, DPO can directly optimize a policy model. Consequently, the DPO loss function can drive towards a maximum likelihood objective for a parameterized policy ($\pi_\theta$):

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

where $\pi$ is a fine-tuned model on downstream tasks, $\sigma$ is sigmoid function and $\beta$ is a hypper-parameter. DPO introduced an RL-free approach by reformulating the reward model's objective in RLHF. They demonstrated that there's no requirement for interaction with an environment to train a model on human preferences, and a reward can be explicitly incorporated into the loss function without interaction.

Although the concept of directly optimizing a model based on human preferences is intriguing, DPO encounters two significant challenges. Firstly, it suffers from inefficient learning and memory utilization. Specifically, during model fine-tuning with DPO, an additional policy model must be loaded as a reference to maintain consistency between the model's post-training and pre-training states. Loading two models consumes more memory than SFT, leading to less efficient learning. Secondly, DPO exhibits a high dependency on achieving an optimal policy. The Zephyr (Tunstall et al., 2023) study demonstrated that the SFT component is crucial for DPO, and without it, DPO's effectiveness is limited.

CPO, a new preference learning method, tackles DPO's inefficient learning and memory utilization by showing that removing the reference model during DPO training does not affect performance. However, the results suggest that CPO does not achieve impressive performance compared to DPO on dialogue systems and, like DPO, suffers from a dependency on the SFT component.

Considering these limitations, we revise the DPO loss function with dispiring CPO ideas. In CPO, it was demonstrated that having an optimal model $\pi_w$ for preferred data enables truthful prediction of preferred data, i.e., $(\pi_{ref}(y_w|x) = \pi_w(y_w|x))$.In this context, for any data point $(x, y_w, y_l)$ from dataset $D$ the condition $\pi_w(y_w|x) = 1$ and $0 < \pi_w(y_l|x) < 1$ hold true. They proved that the DPO loss function does not require $\pi_{ref}$, and for better learning of the policy model $\pi_\theta$,they integrate the log-likelihood supervised fine-tuning loss for preferred data into the DPO loss function without the $\pi_{ref}$ term.

While CPO achieves generally comparable results with DPO on some benchmarks, this loss function may not be effective in dialogue systems. Its main limitation lies in its high dependency on preferred data. Preferred data can be generated in two ways: 1) by the SFT model and 2) by state-of-

| Model | Align | ARC | TruthfulQA | Winogrande | HellaSwag | MMLU | Average |
|---|---|---|---|---|---|---|---|
| Zephyr | SFT | 60.41 | 43.73 | 74.19 | 81.69 | 60.92 | 64.18 |
| Zephyr | DPO | 59.04 | 46.70 | 76.63 | 82.10 | 60 | 64.91 |
| Zephyr | IPO | 59.30 | 42.22 | 76.4 | 81.02 | 59.93 | 63.77 |
| Zephyr | KTO | 57.84 | 49.88 | 76.47 | 81.61 | 59.73 | 65.1 |
| Zephyr | CPO | 57.50 | 53.22 | 75.92 | 80.37 | 58.41 | 65.08 |
| Zephyr | TPO (**our**) | 58.02 | 59.05 | 76.47 | 80.6 | 59.48 | 66.72 |
| Mistral | TPO (**our**) | **60.92** | **59.19** | **78.53** | **82.92** | **63** | **68.91** |

Table 1: Comparing TPO's performance with other alignment methods reveals that the Zephyr-TPO model exhibits comparable performance across different benchmarks and, on average, outperforms other methods. In particular, Zephyr-TPO shows remarkable performance on the TruthfulQA benchmark. Additionally, the Mistral-TPO model consistently outperforms other methods across all benchmarks. It's worth noting that the Mistral-TPO model is directly trained with TPO, which contributes to its superior performance. We clarify that Mistral is a base model of the Zephyr model. We also note that the measure of performance for all benchmarks is accuracy.

the-art models such as LLaMA-2. If a portion of the preferred data is not suitable, the policy model may deviate significantly from the reference model. Since there is no reference model in this loss function, in the worst-case scenario, the policy model may transition to a new distribution, leading to failed learning.

To address this limitation, we propose that by incorporating an additional preference, we can mitigate the aforementioned issues. In addition to considering two, chosen and rejected preferences, we introduce the gold standard response as a third preference in the loss function. Following the approach in SLiC, SLiC-HF, and CPO, we eliminate $\pi_{ref}$ from the DPO loss function and integrate the log-likelihood supervised fine-tuning loss on the gold standard response. Thus, the TPO loss function is formulated as follows:

$$\mathcal{L}_{\text{reference}} = -\mathbb{E}_{(x,y_{ref})\sim\mathcal{D}}\left[\log \pi_\theta\left(y_{ref} \mid x\right)\right] \tag{2}$$

$$\begin{aligned}\mathcal{L}_{\text{preference}}\left(\pi_\theta\right) = \\ &-\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\pi_\theta\left(y_w \mid x\right)\right.\right. \\ &\left.\left.-\beta\log\pi_\theta\left(y_l \mid x\right)\right)\right]\end{aligned} \tag{3}$$

$$\mathcal{L}_{\text{TPO}} = \mathcal{L}_{\text{preference}} + \alpha\mathcal{L}_{\text{reference}} \tag{4}$$

Where $y_{ref}$, $y_w$, and $y_l$ represent the gold standard, chosen, and rejected responses, respectively. Additionally, $\alpha$ serves as a hyperparameter to regulate the impact of the gold standard response during training. Broadly speaking, TPO has two main objectives: 1) optimizing the chosen and rejected preferences and 2) predicting the next token of the gold standard response. The second objective guides the model to learn the gold standard response while directly optimizing between the chosen and rejected responses. Despite the absence of a reference model in the preference learning objective, the second objective helps maintain stability by preventing significant changes in the distribution of the policy model. Overall, TPO addresses the issues of inefficient learning and memory utilization by removing the reference model. Additionally, by incorporating an SFT loss on the global response (See Equation 2), TPO aims to ensure that the policy model remains close to the reference distribution. Furthermore, the TPO loss function resolves the dependency of DPO on the supervised fine-tuned model. This implies that TPO can achieve comparable performance to DPO combined with the supervised fine-tuned model, even without the SFT component. We also evaluate TPO's performance across several benchmarks, as detailed in Section 4.

## 4 Experiments and Results

In this section we present a comprehensive empirical analysis of TPO, yielding several key findings: 1) TPO with SFT outperforms other alignment methods, boasting an average accuracy improvement of 1.8%. 2) Directly training a model with TPO surpasses the performance of TPO with SFT by an average of 2.2%. 3) TPO with SFT excels notably performance on the Truthful QA benchmark. 4) Factors such as training size, $\alpha$ parameter, and epochs significantly influence TPO's performance. 5) While TPO may experience overfitting with ex-

| Model | Align | MT-Bench | BB-casual | BB-sports | BB-formal | OpenBookQA |
|---|---|---|---|---|---|---|
| Zephyr | SFT | 5.94 | 51.57 | 61.76 | 51.4 | 43.8 |
| Zephyr | CPO | 6.2 | 49.47 | 70.68 | **51.07** | 44.6 |
| Zephyr | DPO | 6.64 | 52.1 | 71.9 | 51 | **46.2** |
| Zephyr | TPO (**our**) | **6.66** | **54.21** | **73.93** | 50.84 | 45.6 |

Table 2: In our comparison of TPO with other alignment methods across more benchmarks, Zephyr-TPO emerges as the top performer, surpassing other methods in MT-Bench and BB-causal, while showing comparable results in BB-formal and OpenBookQA. The measure of performance on BB-causal, BB-sports, BB-formal, and OpenBookQA are and for MT-Bench is a score between 0 and 10 that generated by GPT-4.

tended epochs, an SFT model aligned with TPO demonstrates enhanced performance, particularly on understanding tasks.

### 4.1 Experimental Setup

**Models:** All experiments were conducted using `zephyr-sft-full` and Mistral-7B0instruct-v0.1, which are among the current state-of-the-art models at the 7B parameter scale. We utilized the Transformer Reinforcement Learning (TRL) library for fine-tuning (von Werra et al., 2020). All models were trained using the AdamW optimizer without weight decay. Furthermore, parameter-efficient techniques such as LoRA (Hu et al., 2021) were not employed. The experiments were conducted on 6 A100 GPUs, utilizing bfloat16 precision, and typically required 5-8 hours to complete.

**Datasets:** In this study, we employ two dialogue datasets: 1) UltraChat (Ding et al., 2023) and 2) UltraFeedback (Cui et al., 2023). UltraChat comprises 200k examples generated by GPT-3.5-TURBO across 30 topics and 20 text material types, offering a high-quality dataset utilized for training the SFT model. Meanwhile, UltraFeedback consists of a 64K set of responses generated by state-of-the-art models like LLaMA-2, Vicuna, UltraLM, and Alpaca, evaluated by a teacher model such as GPT-4. Each response is assigned a score indicating its quality given the same input. To train TPO, which requires three preferences, we create a custom dataset from the UltraFeedback dataset. Here, the response with the highest score serves as the reference response, the second-highest score as the chosen response, and the lowest score as the rejected response. For evaluating the performance of other alignment methods, we train them on the chosen and rejected responses from the custom dataset. Due to computational limitations, we restrict our analysis to 12K data points, randomly selected from the dataset.

**Evaluation Metrics:** We evaluate our models in both single-turn and multi-turn scenarios using the MT-Bench benchmark (Ding et al., 2023). MT-Bench consists of 160 questions spanning eight knowledge domains for evaluation. In this benchmark, models are required to answer a question and then address a predefined follow-up question. Model responses are rated on a scale of 1 to 10 by GPT-4, and the final score is calculated as the mean over the two turns. Additionally, we assess models on the Open LLM Leaderboard Benckmarks (Beeching et al., 2023), which assess the performance of LLMs across five multiclass classification tasks: ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2021), Truthful QA (Lin et al., 2022), and Winogrande (Sakaguchi et al., 2019). Furthermore, we explore the performance of TPO across a broader variety of tasks, including Big-Bench (bench authors, 2023) and OpenBookQA (Mihaylov et al., 2018). To ensure a comprehensive evaluation, we include benchmarks such as Causal Judgment (causal reasoning), Sports Understanding (commonsense reasoning), and Formal Fallacies in the Big Bench Hard dataset.

**Detail of Training:** All models are trained for one epoch, employing a linear learning rate scheduler with a peak learning rate of 5e-7 and 10% warmup steps. Additionally, the global batch size is set to 8, and $\beta = 0.1$ is used to regulate the deviation from the reference model.

### 4.2 Results

The primary results are depicted in Tables 1 and 2.

**Demonstration of TPO Performance** We compare the TPO method with other alignment methods, including KTO, IPO, CPO, and DPO, on the

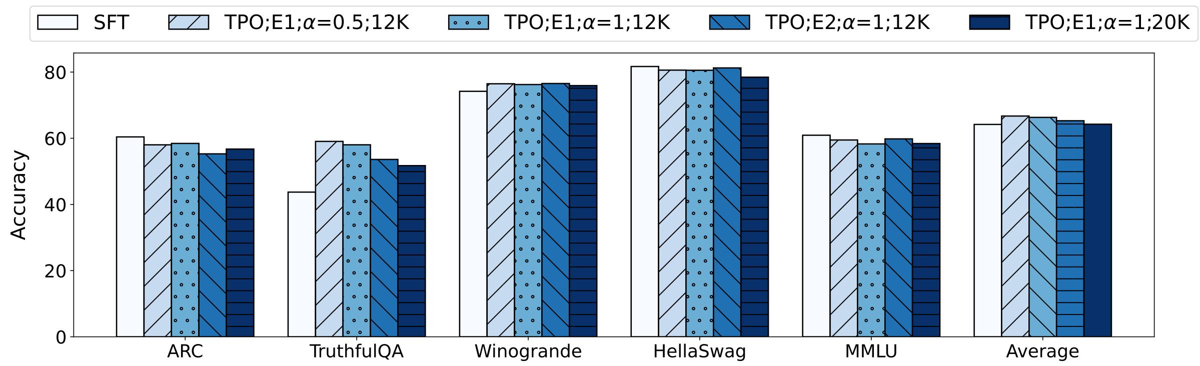



Figure 3: In this figure, we present the performance of TPO across various values of hyperparameters, as discussed in Section 4.3. It's important to note that all models in this figure include the SFT part.

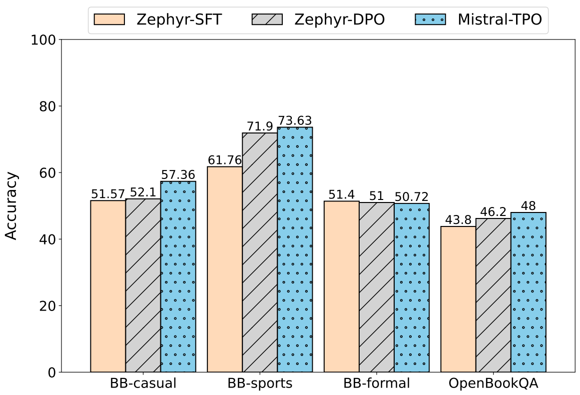



Figure 4: Compare the performance of the Mistral-TPO model with Zeyphr-DPO and Zephyr-TPO. We observe that TPO generally outperforms DPO and SFT across all benchmarks except BB-formal. Particularly noteworthy is the significant performance gap between Mistral-TPO and Zephyr-SFT in BB-Sports.

Open LLM Benchmarks. Our comparison involves two scenarios: 1) training an SFT model with TPO and other alignment methods, and 2) directly training a pre-trained model with TPO and other alignment methods.

The results presented in Table 1 demonstrate that employing the TPO method on an SFT model not only enhances the performance of the SFT model itself but also exhibits remarkable overall performance compared to other methods. Specifically, TPO surpasses SFT and DPO in terms of average accuracy across five different benchmarks by 2.5% and 1.8%, respectively. On average, Zephyr-TPO model outperforms Zephyr-DPO model and achieves comparable performance on three out of five benchmarks (ARC, HellaSwag, and MMLU) compared with Zehyr-DPO model. We also, highlight that, TPO achieves notably high accuracy on the TruthfulQA benchmark, distinguishing it from other models.

However, these results do not align with our initial hypothesis, prompting us to directly train a pre-trained model with TPO. Surprisingly, the findings reveal that TPO without the SFT component not only exhibits superior performance across five benchmarks but also outperforms SFT and DPO in terms of accuracy by 4.7% and 4%, respectively (See Table **??**). This suggests that while the SFT method is an important component of alignment, it may constrain alignment methods from achieving their full potential. Consequently, TPO without SFT may facilitate better exploration, resulting in greater improvement compared to TPO with SFT.

**Exploration on More Benchmarks** For a comprehensive evaluation, we compare the TPO method with CPO, DPO, and SFT across additional benchmarks, including MT-Bench, Big-Bench, and OpenBookQA, in addition to the tasks from the Open LLM Leaderboard. Specifically, we incorporate tasks from Big Bench Hard, such as Casual Judgment (casual reasoning), Sports Understanding (common sense reasoning), and Fallacies (logical reasoning). For further details, please refer to the Appendix A.

Table 2 presents the performance of TPO and other alignment methods. The results reveal that TPO exhibits comparable performance on MT-Bench compared to DPO and outperforms SFT and CPO. While TPO falls short on OpenBookQA compared to DPO, the directly trained TPO model enhances performance even further, surpassing DPO (See Figure 4). Additionally, the finding indicates that CPO performs poorly on all benchmarks except BB-formal and does not achieve comparable performance with TPO and DPO.

### 4.3 Ablation Studies

We found that the number of epochs has a notable effect on the performance of TPO. To thoroughly evaluate the performance improvement of TPO, we conduct additional evaluations across five benchmarks in the Open LLM Leaderboard Benchmarks.

**Training Size and Effect of $\alpha$ in TPO** We assess TPO's performance across different sizes of training sets. Figure 3 illustrates that increasing the training size with the same $\alpha$ leads to a decrease in TPO's performance across all benchmarks. On average, TPO's performance exhibits a 2% decrease.

In the TPO method, we define $\alpha$ as a hyper-parameter to control the effect of the reference response on learning. The results depicted in Figure 3 indicate that, under the same configuration, TPO's performance remains relatively consistent across different $\alpha$ values. However, we observe that for aligning an SFT model with TPO, better performance is achieved by decreasing $\alpha$ while aligning a pre-trained model directly with TPO benefits by increasing $\alpha$.

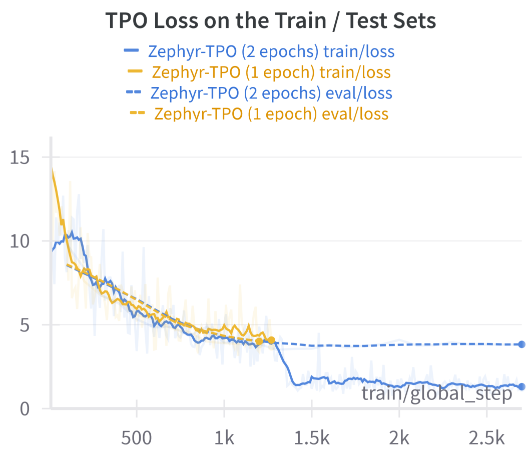



Figure 5: We show the overfitting impact on loss while training the Zephyr-TOP model.

**Training for More Epoch** We examine the performance of TPO and the occurrence of overfitting on multiple epochs. Figures 5 and 6 illustrate that TPO experiences overfitting with an increase in epochs. This overfitting phenomenon suggests that while the accuracies of TPO improve during training on additional epochs, the model's accuracy during evaluation remains unchanged, indicating that the model may not be learning effectively. However, TPO trained over two epochs demonstrates

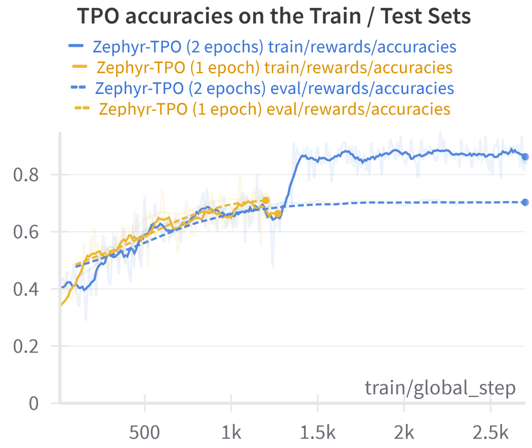



Figure 6: The accuracy of the Zephyr-TPO model increases after one epoch, while the evaluation accuracy remains constant.

a 2% improvement on MMLU and HellaSwang, implying that more learning steps are necessary for understanding tasks.

## 5 Conclusions

In this paper, we begin by addressing the limitations inherent in existing alignment methods. Generally, alignment methods exhibit a high reliance on the SFT component and encounter challenges in sampling from policy models for training. To mitigate these shortcomings, we introduce TPO, a novel alignment approach aimed at concurrently optimizing human preferences and gold standard responses. TPO represents an RL-free method with dual objectives: 1) optimizing human preferences and 2) refining model predictions beyond gold standard responses. Our findings demonstrate the impressive performance of TPO compared to other alignment methods post the SFT phase. Particularly noteworthy is the superior performance of TPO without SFT compared to TPO with SFT. Moreover, we identify SFT as a critical alignment method that acts as a barrier for a model to engage in more exploration. While we examine TPO's performance across various $\alpha$ values, we emphasize the significance of exploring the impact of altering both $\beta$ and $\alpha$ on TPO performance. Furthermore, extending the application of TPO to diverse tasks presents an opportunity for further exploration, which we suggest for future works. We believe our study represents a pioneering effort in eliminating the SFT part from alignment methodologies.

## 6 Limitations

While our study demonstrates the impressive performance of TPO compared to SFT and other alignment methods, certain challenges remain. Creating a dataset with three preferences poses a notable challenge, and determining optimal values for $\alpha$ and $\beta$ presents additional complexities. Moreover, TPO is susceptible to overfitting, and evaluation accuracy diminishes over multiple epochs. Additionally, assessing models on certain alignment benchmarks, such as MT-Bench and AlpacaEval (Dubois et al., 2023), incurs significant costs. These challenges highlight areas for further investigation.

## Ethics Statement

We have used AI assistants (Grammarly and ChatGPT) to address the grammatical errors and rephrase the sentences.

## References


Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

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

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. `https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard`.

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Heejong Bong and Alessandro Rinaldo. 2022. Generalized results for the existence and consistency of the mle in the bradley-terry-luce model.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-play fine-tuning converts weak language models to strong language models.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge.

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

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

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback.

Kawin Ethayarajh, Winnie Xu, Dan Jurafsky, and Douwe Kiela. 2023. Human-aware loss functions (halos). Technical report, Contextual AI.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

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

Julia Kreutzer, Joshua Uyheng, and Stefan Riezler. 2018. Reliability and learnability of human bandit feedback for sequence-to-sequence reinforcement learning.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J. Liu, and Jialu Liu. 2024. Statistical rejection sampling improves preference optimization.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions.

Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Anand Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, Amar Phanishayee, and Matei Zaharia. 2021. Efficient large-scale language model training on gpu clusters using megatron-lm.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models.

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

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2023. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

AMOS TVERSKY and DANIEL KAHNEMAN. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323.

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

Tianhao Wu, Banghua Zhu, Ruoyu Zhang, Zhaojin Wen, Kannan Ramchandran, and Jiantao Jiao. 2023. Pairwise proximal policy optimization: Harnessing relative feedback for llm alignment.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

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

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence?

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Daniel M. Ziegler, Nisan Stiennon, Jeff Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *ArXiv*, abs/1909.08593.

# Appendix

## A More Exploration

We present the details of the MT-Bench benchmark in Table 3. Figure 7 illustrates the scoring of GPT-4 in various domains. Additionally, we provide some examples for models that fine-tuned by TPO and DPO.

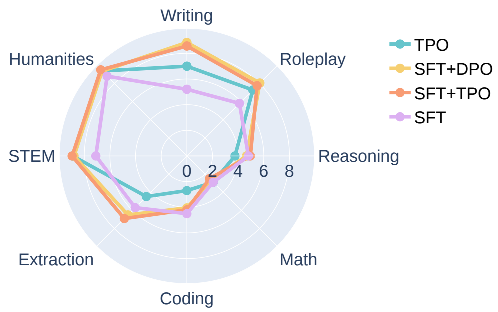



Figure 7: Generation another example of fine-tuned model by TPO and DPO.

| **Model** | **Align** | **First Turn (Score)** | **Second Turn (Score)** | **Average (Score)** |
|---|---|---|---|---|
| Zephyr | TPO | 7.13 | 6.2 | 6.66 |
| Zephyr | CPO | 6.47 | 5.93 | 6.2 |
| Mistral | TPO | 6.36 | 5.22 | 5.79 |

Table 3: The performance of alignment methods on MT-Bench scores indicates that Mistral-TPO performs comparably to SFT. As depicted in Table 1 and Figure 4, Mistral-TPO demonstrates impressive performance.

| | |
|---|---|
| **Prompt** | Please describe the concept of machine learning. Could you elaborate on the differences between supervised, unsupervised, and reinforcement learning? Provide real-world examples of each.", "In your last example of reinforcement learning, can we use supervised learning to solve it? |

| **Model** | **Response** |
|---|---|
| **Mistral-TPO** | Machine Learning is a subfield of Artificial Intelligence (AI) that focuses on developing computer programs that can learn from data and improve their performance over time. It involves the development of algorithms and models that can automatically learn from data, make predictions, and adapt to new data without being explicitly programmed.The concept of machine learning can be understood as a process of teaching a computer to learn from data and make predictions or decisions based on that learning. The goal is to create systems that can learn from data, adapt to new situations, and perform tasks without being explicitly programmed.Machine learning can be divided into [...] |
| **Zephyr-TPO** | Machine learning is a subfield of artificial intelligence (AI) that involves designing algorithms and models that enable computer systems to improve their performance on specific tasks over time, without being explicitly programmed. It's a process of training a system to learn from data, identify patterns, and make predictions or decisions based on that knowledge.There are three main categories of machine learning:1. **Supervised Learning**: In this approach, the system is trained using labeled data, where the input data (features) and the corresponding output (labels) are known. [...] |
| **Zephyr-DPO** | Machine learning is a subset of artificial intelligence (AI) that involves the development of algorithms and statistical models that allow computer systems to improve their performance on a specific task over time, without being explicitly programmed. Essentially, machine learning involves teaching a computer to learn from data and make predictions or decisions based on that data. There are three main types of machine learning: supervised, unsupervised, and reinforcement learning. 1. Supervised Learning: In supervised learning, a machine[...] |

Table 4: Generation example of fine-tuned model by TPO and DPO on MT-Bench questions.

| Prompt | Act as a math teacher. I will provide some mathematical equations or concepts, and it will be your job to explain them in easy-to-understand terms. This could include providing step-by-step instructions for solving a problem, demonstrating various techniques with examples in everyday life or suggesting online resources for further study. My first request is "I need help understanding how probability works.", "What are the differences between Riemannian geometry and euclidean geometry? |
|---|---|

| **Model** | **Response** |
|---|---|
| **Mistral-TPO** | Probability is a fascinating concept, representing the likelihood or chance that a specific event or outcome will occur. It plays a crucial role in various fields, such as statistics, science, and even gambling. If you're looking to understand probability better, I'm happy to help by providing you with clear, step-by-step instructions.To start, let's consider a simple example of flipping a coin. If you were to flip a coin, there are only two possible outcomes [...] |
| **Zephyr-TPO** | As an AI assistant, I'm happy to explain probability in simple terms. Probability is a measure of the likelihood or chance that a particular event occurs. It's a crucial concept in mathematics, science, and everyday life. Here's a step-by-step explanation:1. **Understanding the basics:** Probability is a number between 0 and 1 that represents the possibility of an event happening. A probability of 0 means an event never happens, while a probability of 1 means it's certain to occur.2. **Calculating Probability:** To calculate the[...] |
| **Zephyr-DPO** | Probability is the study of how likely an event is to occur based on a set of conditions or outcomes. It is often expressed as a percentage or a fraction. Here are some basic concepts to help you understand probability:1. Sample Space: This is the set of all possible outcomes in a given situation. For example, if you are flipping a coin, the sample space would be heads or tails.2. Events: These are the different outcomes that can occur in a given situation. For example, if you are flipping a coin, [...] |

Table 5: Generation another example of fine-tuned model by TPO and DPO on MT-Bench questions.

| **Prompt** | Explain what's base rate fallacy and list five specific examples of how politicians use it for campaigns.", "Provide a detailed plan for an election campaign using the first example. |
|---|---|
| **Model** | **Response** |
| **Mistral-TPO** | Base rate fallacy is a logical fallacy that occurs when a person ignores the base rate of an event and focuses on the probability of an event given a certain condition. In other words, it is the tendency to overestimate the probability of an event based on the specific information provided, without considering the overall likelihood of that event occurring. Here are five specific examples of how politicians use base rate fallacy in their campaigns: 1. **Crime Statistics:** Politicians often use crime statistics to create fear and promote their tough-on-crime policies. They [...] |
| **Zephyr-TPO** | Base rate fallacy, also known as the base rate neglect or the conjunction fallacy, is a cognitive error where individuals tend to overemphasize or underestimate the significance of specific information while ignoring the overall probability or base rate. It occurs when people focus too much on a particular piece of information, leading them to make incorrect judgments or decisions. In the context of politics, this fallacy can manifest in various ways during campaigns.1. **Appeal to Emotion:** Politicians often [...] |
| **Zephyr-DPO** | Base rate fallacy, also known as the fallacy of the fallacious generalization, is a type of fallacy where a conclusion is drawn based on a small sample or a single case, without considering the larger context or base rate. In other words, it is assuming that a small sample or a single case is representative of the larger population.Politicians often use base rate fallacy in their campaigns to make their arguments more convincing and to appeal to voters' emotions. Here are five specific examples:[...] |

Table 6: Generation another example of fine-tuned model by TPO and DPO on MT-Bench questions.

| Prompt | Describe a vivid and unique character, using strong imagery and creative language. Please answer in fewer than two paragraphs.", "Revise your previous response and incorporate an allusion to a famous work of literature or historical event in each sentence. |
|---|---|
| **Model** | **Response** |
| **Mistral-TPO** | The man was like a human, but not really. He stood eight feet tall and his body looked like a mixture of a lion and a human. His skin was a deep gold color, and his mane was as long as a lion2̆019s, but unlike a lion2̆019s mane, his was the color of a lion2̆019s fur. His eyes were the brightest blue of any human eyes, and his smile radiated warmth. He had a thick, bushy mustache that, along with his thick, dark eyebrows, gave the impression that he was a kind, wise man. But, the strangest thing about him was not [...] |
| **Zephyr-TPO** | Amidst the crimson waves of sunset, a figure emerges - a woman etched from the very essence of the night sky. Her hair a tapestry of starlight, cascading down her naked shoulders like a cosmic waterfall. Her eyes, a celestial dance of iridescence, shifting from a deep indigo to a shimmering sapphire as she gazes upon the universe with reverence. Her voice, a symphony of [...] |
| **Zephyr-DPO** | Meet Aria, a woman with skin as pale as moonlight, eyes that glitter like emeralds, and hair as black as ebony. Her slender frame is adorned with intricate tattoos that seem to pulse with a life of their own, depicting fantastical creatures and ethereal landscapes. She moves with the grace of a serpent, her every step delicate and precise. Her voice is a symphony of music, a melody that seems to lure those around her, beckoning them closer to her mystical world. Her presence is a force to be reckoned with, leaving [...] |

Table 7: Generation another example of fine-tuned model by TPO and DPO on MT-Bench questions.

| Prompt | Photosynthesis is a vital process for life on Earth. Could you outline the two main stages of photosynthesis, including where they take place within the chloroplast, and the primary inputs and outputs for each stage?", "How much energy can a tree produce through photosynthesis in its lifetime? Please provide an estimate using actual numerical values and thoroughly explain your thought process step-by-step. |
|---|---|
| **Model** | **Response** |
| **Mistral-TPO** | Photosynthesis is a process that converts light energy into chemical energy, which is stored in the form of carbohydrates. It is a vital process for life on Earth, as it provides the energy that sustains most organisms. Photosynthesis occurs in two stages: the light-dependent reactions and the light-independent reactions. These two stages take place within the chloroplast, which is a specialized organelle found in plants, algae, and some bacteria.1. Light-Dependent Reactions: These reactions occur in the thylakoid membranes of the chloroplast. They involve the absorption of [...] |
| **Zephyr-TPO** | Photosynthesis is a complex biochemical process that occurs in the chloroplasts of plant cells, algae, and some bacteria. It involves two main stages: the light-dependent reactions (also known as the light reactions) and the light-independent reactions (also known as the dark reactions or the Calvin cycle).1. Light-dependent reactions: a. Location: These reactions occur in the thylakoid membranes of the chloroplasts, which are flattened, stacked membranes that form a network of sacs. b. Inputs: Water (H2O), [...] |
| **Zephyr-DPO** | Photosynthesis is a process by which plants, algae, and some bacteria convert light energy into chemical energy in the form of glucose. It occurs in the chloroplasts of plant cells, which are specialized organelles that contain chlorophyll, the pigment that gives plants their green color. Photosynthesis can be divided into two main stages:1. Light-dependent reactions: This stage takes place in the thylakoid membranes of the chloroplasts. The primary inputs for this stage are light energy and water. The primary output is ATP (adenosine triphosphate), a molecule that stores energy, and NADPH (nicotinamide adenine dinucleotide phosphate), a molecule that carries electrons.During the light-dependent reactions, light energy is absorbed by chlorophyll [...] |

Table 8: Generation another example of fine-tuned model by TPO and DPO on MT-Bench questions.