UPO: Unpaired Preference Optimization for Large Language Models

Anonymous ACL submission

Abstract

While Large Language Models (LLMs) have made remarkable progress in various NLP tasks, there is no guarantee that LLMs will provide helpful, honest, and harmless answers without proper alignment. Reinforcement Learning from Human Feedback (RLHF) has been shown to be an effective alignment method, though it is complex and costly. Advancing further, Direct Preference Optimization (DPO) simplifies the alignment process by bypassing the reward modeling step and the reinforcement learning step, achieving performance comparable to RLHF using Proximal Policy Optimization (PPO). However, both methods necessitate paired preference data, which is costly to obtain in reality. We propose a new align method, dubbed Unpaired Preference Optimization (UPO), which does not need paired cases to align with human's preferences. Building upon DPO's approach, we derive a new loss function tailored to process positive and negative cases separately from the DPO loss function. Our findings indicate the performance of UPO is comparable to the performance of DPO trained on a complete paired dataset without a large performance gap. Moreover, under conditions involving a paired preference dataset, our UPO method achieves performance comparable to that of DPO and is more memory-efficient and time-efficient. In cases where the datasets are unpaired, the UPO method maintains a high level of performance compared to fully paired datasets, with only minimal loss in effectiveness and it significantly outperforms Unified Language Model Alignment (ULMA, an alignment method for point-wise preference data) or fine-tuning on only the positive cases (Preferred-SFT).

1 Introduction

Large language models (LLMs) have attracted worldwide attention since the release of ChatGPT[1]. With billions of parameters, LLMs trained on vast corpora of data have shown strong abilities for different NLP downstream tasks such as sentiment analysis (Zhang et al., 2023) and information retrieval (Zhu et al., 2024). However, LLMs may not provide helpful, honest and harmless (Askell et al., 2021) responses naturally after pre-training. For instance, when pre-training data contains impolite expressions or factual errors, LLMs' responses are not guaranteed to be polite or accurate.

Therefore, the technique of alignment is developed to mitigate this problem. Alignment is defined as the process of ensuring that LLMs behave in accordance with human values and preferences (Liu et al., 2023). One effective approach is Reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022). Standard RLHF pipeline consists of 3 stages (Kirk et al., 2023): feedback collection and supervised fine-Tuning (SFT), reward modeling (RM) and reinforcement learning (RL) with an on-policy RL algorithm like Proximal Policy Optimization (PPO) (Schulman et al., 2017). This process is complex and unstable. Besides, it requires lots of computational resources.

Following RLHF's framework, many methods have been proposed to reduce the complexity without losing much performance. A representative work is Direct preference optimization (DPO) (Rafailov et al., 2023) which bypasses the need of reward modeling and reinforcement learning.

One drawback for these methods is they require paired preference training data. However, paired preference data annotated by human is expensive in reality. For instance, on a question-answer website such as stack overflow, some questions may be only answered once. In this situation no paired data exists. Even if multiple answers are provided, how to construct the pairs is not straight forward. Can we simply view the response with most likes

[1] https://chat.openai.com

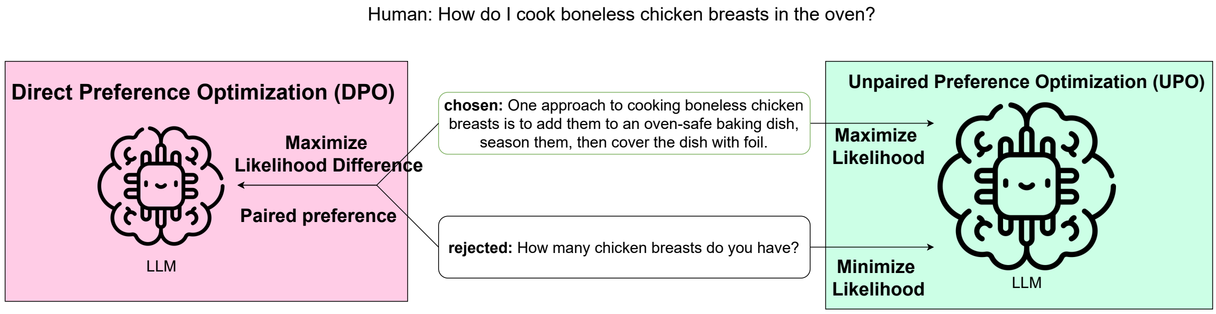

Figure 1: In the training stage, DPO needs paired preference data but UPO does not. For DPO, the training objective is to maximize the difference of the likelihood of chosen and rejected responses. For UPO, the training objective is to maximize the likelihood of chosen responses and minimize the likelihood of rejected cases

as the chosen, the response with least likes as the rejected and drop all the other responses? It might be a waste of valuable data because the second preferred response might be of great value as well. Although Reinforcement Learning from AI Feedback (RLAIF) (Bai et al., 2022b) can reduce the cost of annotation while keep the annotation accuracy comparable to human annotators, paired datasets of high quality are still scarce.

In this paper we try to propose a new method to utilize unpaired preference data for alignment. We find that the loss function of DPO is similar to the loss function of SFT after mathematical transformation. By using this surrogate loss function, we can apply the normal SFT paradigm for alignment and achieve comparable performance with DPO and outperforms SFT only with chosen data (Preferred-SFT) on the selected benchmarks. Due to the simplicity of the SFT paradigm, our method is more memory-efficient and time-efficient because the reference model is not required. We name this method Unpaired Preference Optimization (UPO) as no paired preference data is necessary in the training stage.

Our contributions are summarized as follows:

- We propose a new SFT method named Unpaired Preference Optimization (UPO) which can align with human's preference on unpaired preference data.

- From our experiments results, we find that our UPO performs comparable on selected benchmarks with DPO under the paired preference data setting and outperforms Preferred-SFT under the unpaired preference data setting. UPO is also more memory-efficient and time-efficient due to the lack of reference model compared to DPO.

2 Background

We introduce two commonly used alignment methods below.

Supervised fine-tuning Supervised fine-tuning is a commonly used approach which trains the model on demonstrations of solving the task using supervised learning (Kirk et al., 2023). It achieves great success in aligning with human instructions (Wang et al., 2023; Taori et al., 2023) but it cannot model human preference as it only learns from positive responses and does not know which kind of cases should be avoided.

Reinforcement learning from human feedback (RLHF) RLHF has been a key component in LLM training on models such as GPT-4 (Achiam et al., 2023) and Llama 2 (Touvron et al., 2023) to train a safe model aligned with human's preference. This method aligns models more closely with complex human values but often faces scalability challenges due to its intensive computational demands, especially in policy optimization via methods like Proximal Policy Optimization (PPO), which can be memory-intensive and unstable.

Supervised fine-tuning and Reinforcement Learning from Human Feedback (RLHF) are two prominent methods used to align machine learning models with human preferences, yet they operate on fundamentally different principles. Supervised fine-tuning relies on direct instruction from curated datasets that demonstrate correct responses, making it effective for tasks where clear well-designed answers exist; however, it falls short in capturing the nuance of human preferences as it lacks exposure to negative examples—what not to do. On the other hand, RLHF involves training models through iterative adjustments based on both positive and negative human feedback, enabling the model to not only replicate desired outcomes but also avoid undesired behaviors. Together, these methods complement each other by covering different aspects of model training and alignment, with supervised fine-tuning establishing a baseline of correct responses, and RLHF refining and expanding the model's alignment through nuanced feedback.

3 Methodology: Unpaired Preference Optimization

In this section, we will first discuss the issue of preference optimization without paired data, where existing solutions face a lossy approximation as detailed in Sec. 3.1. To this end, starting from the original Direct Preference Optimization (DPO) which cannot handle unpaired preference optimization problems, we have reanalyzed the principles within DPO and in Sec. 3.2 derived an elegant Unpaired Preference Optimization (UPO) solution in Sec. 3.3. The benefits of the UPO approach are discussed in the Sec. 3.4.

3.1 Preference Optimization without Paired Data

Most existing work (e.g. DPO) requires paired preference data. However, in real-life scenarios, we often confront situations where no paired preference data is available. For instance, in a healthcare context, a patient may provide feedback on their satisfaction with a particular medical consultation. This feedback constitutes a single data point, as opposed to a pair of preference data. Furthermore, the acquisition of paired preference data is often fraught with difficulties, requiring substantial resources in terms of time, effort, and cost.

To deal with the scenario where pair-wise preference data does not exist, Unified Language Model Alignment (ULMA) (Cai et al., 2023) introduces point-wise DPO to harness point-wise feedback. However, ULMA (Cai et al., 2023) has a strong assumption that assumes the normalization term Z used in DPO

$$\begin{aligned} Z(x) &= \sum_y \pi_{\text{ref}}(y|x)\exp(\frac{1}{\beta} r_\phi(x,y)) && \text{Taylor expansion} \\ &\approx 1 + E_{y\sim\pi_{\text{ref}}} r_\phi(x,y) && E_{y\sim\pi_{\text{ref}}} r_\phi(x,y) \to 0 \\ &\approx 1 \end{aligned} \tag{1}$$

The approximation (from the second line to the third line in Eq .1) may be unreasonable in cases where only one y exists for a given x, and in that case $r_\phi(x,y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$ has a gap with 0 and therefore the approximation (i.e., $E_{y\sim\pi_{\text{ref}}} r_\phi(x,y) \to 0$) is invalid. This might leads to diminished performance, as seen in Section 4.1.

3.2 A Further Theoretical Analysis on DPO

Since the existed solution for unpaired preference optimization like UMLA has difficulties to approximate the normalization term Z, we turn back to rethink DPO and theoretically derive an algorithm UPO to deal with unpaired preference optimization.

The DPO loss function is a crucial component in preference-based learning models. It is defined as follows:

$$\mathcal{L}_{DPO} = -E_{(x,y_w,y_l)\sim D}(\log\sigma(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}))$$

Here, $(x, y_w, y_l)$are triples from the dataset D, consisting of an input x, a chosen response y_w , and a rejected response y_l. The function $\pi_\theta(y|x)$ represents the model's predicted probability of response y given input x, and $\pi_{\text{ref}}(y|x)$ is a reference probability. The parameter β is a scaling factor.

To deal with the unpaired case, we need to untwist the interaction between the chosen and rejected responses. During our training, the probability generating y_w, which is $\pi_\theta^\beta(y_w|x)$, becomes larger and tends to 1 as we expect, combining with $\log(1+x) \approx x$, the DPO loss approximate to two separate parts: The part of chosen response $-\beta\log(\pi_\theta(y_w|x))$, which is just the SFT loss scaled by a constant β, and the part of rejected responses $e^{K'(y_l,x)}\pi_\theta^\beta(y_l|x)$. Here $K'(y_l,x)$ is a constant to accumulate the constants in the calculation, defined as

$$K'(y_l,x) = \beta\sum_{y'\in D_{x,y_l}} \log\pi_{\text{ref}}(y'|x) - \beta|D_{x,y_l}|\log\pi_{\text{ref}}(y_l|x)$$

for

$$D_{x,y} = \{(x',y') \in D, x' = x, y' \succeq y\}$$

where $y' \succeq y$ means as a response, y' is not worse than y. The set $D_{x,y}$ is a collection of responses to the prompt x not worse than y, and in the definition we can see that in the case of unpaired dataset, the constant is just 0, and in that case, the chosen loss is the SFT loss, and the rejected loss becomes simpler: $\pi_\theta^\beta(y_l|x)$. For details, see the proof in Appendix A for details.

By the above analysis, we get the following finding:

Finding 1 *Under approximation, the DPO loss can be viewed as the sum of losses for chosen responses and losses for rejected responses.*

By this finding, we can untwist the interaction of chosen and rejected responses, deal these two types of responses separably. This gives us the background idea of the construction of the UPO loss in the next subsection.

3.3 From DPO to UPO

In this subsection we introduce the architecture of UPO and make an analysis on how the change of loss function for the rejected responses solve the degenerated problem of language models described in (Rafailov et al., 2023) Section 4.

By Findings 1, the DPO loss

$$\begin{aligned}\mathcal{L}_{DPO} \approx &-\beta\overbrace{E_{(x,y_w)\sim D}\log(\pi_\theta(y_w|x))}^{\text{loss of positive examples}}\\ &+\underbrace{E_{(x,y_l)\sim D}e^{K'(y_l,x)}\pi_\theta^\beta(y_l|x)}_{\text{loss of negative examples}}\end{aligned}$$

Since the loss function is separated into positive loss and negative loss parts, we may deal with chosen and rejected samples separably. Under this idea, we propose the **Unpaired Preference Optimization (UPO)**. For positive cases (x, y), UPO utilizes the standard loss $\beta\log\pi_\theta(y|x)$. For negative cases, it applies a modified loss $e^{K'(y_l,x)}\pi_\theta(y|x)^\beta$. In particular for the unpaired case, when for each x there exsits only one response, the loss function of UPO is defined as:

$$\mathcal{L}_{UPO} = \begin{cases}\beta\log\pi_\theta(y|x) & \text{chosen}\\ \pi_\theta(y|x)^\beta & \text{rejected}\end{cases} \qquad (2)$$

corresponds to the loss function in Finding 1.

This approach effectively manages positive and negative cases separately, eliminating the need for paired preference data.

In Section 4 of (Rafailov et al., 2023), the authors pointed out that if the language model is trained using the loss function $\log\pi_\theta(y_w|x) - \log\pi_\theta(y_l|x)$, the model might be crashed. We make an analysis on this situation and explain how our method conquers this problem. We first concentrate on the DPO loss. For positive cases, the loss function resembles the SFT loss. For the negative cases, we note that as the difference $\log\pi_\theta(y_w|x) - \log\pi_\theta(y_l|x)$ increases—mimicking the DPO strategy—the coefficient of $\nabla_\theta\log\pi_\theta(y_l|x)$ diminishes. This implies that the gradient descent speed decreases to 0 when the rejected loss tends to 0, preventing the model from degenerating. Similarly, our loss function for negative cases,

$$\nabla_\theta\pi_\theta(y|x)^\beta = \pi_\theta(y|x)^\beta\nabla_\theta\log\pi_\theta(y|x)$$

decreases in magnitude, which aligns well with the behavior of the DPO loss, thus maintaining consistency in model training.

3.4 Benefits of UPO

A key feature of our proposed method, Unpaired Preference Optimization (UPO), is its distinctive loss function architecture. This function is deliberately crafted to handle positive and negative cases separately, an attribute essential for our specific needs.

In conventional methodologies, positive and negative cases are typically integrated within the same loss function. This amalgamation can complicate the optimization process. In contrast, UPO offers the flexibility to independently manage positive and negative cases. This distinct separation proves particularly advantageous when working with unpaired preference data, where the relationship between chosen (positive) and rejected (negative) responses may not be straightforward.

The UPO loss function's ability to separately address positive and negative cases enhances its capacity to finely align large language models (LLMs) with human preferences. This feature not only facilitates a deeper insight into model performance across different data types but also enables precise model adjustments based on specific operational requirements or limitations.

Furthermore, the structural design of the UPO loss function contributes significantly to computational efficiency. By allowing positive and negative cases to be processed independently, it could potentially accelerate the training phase if adopting parallel computations.

4 Experiments

We evaluate our method on two tasks: controlled sentiment generation on IMDB dataset (Maas et al., 2011) and single-turn dialogue on a subset of

Anthropic-HH dataset (Bai et al., 2022a). We follow the experimental setting from (Rafailov et al., 2023).

4.1 Controlled sentiment generation

The objective of this task is to fine-tune the model so that it consistently generates movie reviews with a positive sentiment.

4.1.1 Dataset and model

Our experiments utilize a variant of IMDB. Specifically we employ the first eight tokens (tokenized by the GPT2-large tokenizer) of the "text" field as prompts with the entire "text" field serving as the target completion. We fine-tune the GPT2-large model on this dataset for a single epoch, resulting in a model capable of performing text completion tasks. For each prompt in the training set, the fine-tuned model generates four distinct completions. The sentiment of these completions is assessed using a pre-trained sentiment classifier (Hartmann et al., 2023), also referenced in the DPO paper. For the 4 generated completions, we randomly select one with positive sentiment as the chosen completion and one with negative sentiment as the rejected completion. If a prompt lacks either positive or negative completions among the four, it is excluded from the training set. Subsequently, the size of the training set is reduced to 16,056 cases from an initial count of 25,000. Representative samples from the processed IMDB dataset for supervised fine-tuning (SFT) are displayed in Table 4 and examples from the processed paired preference dataset are provided in Table 5

In the unpaired data scenario, we investigate two distinct experimental setups:

- The first approach involves retaining all rejected completions and randomly selecting varying proportions of the chosen completions to form the training set.

- The second approach involves randomly selecting a certain proportion of the chosen completions, and then incorporating the corresponding rejected completions for those not selected, thereby forming a complementary training set.

These configurations allow us to explore the effects of different sampling strategies on model performance, with the sample ratios and experimental details further outlined in Tables 1 and 2.

Following the methodology of the DPO paper, we employ GPT-2-large as the foundational model for our experiments.

4.1.2 Training details

We begin with the fine-tuned model described previously. It is important to note that this model has not been specifically tailored to generate exclusively positive sentiment reviews, as the IMDB dataset comprises both positive and negative cases.

In the *paired preference setting*, we evaluate three alignment methods utilizing the aforementioned fine-tuned model on the generated dataset: Preferred-SFT, DPO, and UPO. For Preferred-SFT, only chosen data is employed for fine-tuning under the SFT paradigm. In the case of DPO, the paired preference data is directly utilized for alignment. For UPO, we adhere to the implementation outlined in Section 3.3, employing the SFT paradigm. For the *unpaired dataset setting*, the comparison is limited to UPO and Preferred-SFT, as DPO is inapplicable in this context.

All implementations are adapted from the TRL library[2]. For the experiments, a batch size of 64 was utilized along with a learning rate set at $1e6$. The training duration was limited to 2 epochs, supplemented by a warm-up phase consisting of 150 steps. This configuration outlines the specific conditions under which the Direct Preference Optimization (DPO) and Unpaired Preference Optimization (UPO) models were evaluated, highlighting a straightforward, yet precise approach to training these models.

4.1.3 Evaluation strategy

We deploy the trained model to execute the text completion task on the test set. The sentiment of the outputs generated through the greedy decoding sampling strategy is assessed using the aforementioned sentiment classifier.

4.1.4 Results and analysis

In the case of paired data, as evidenced in Table 3, both the DPO and UPO methods significantly outperform Preferred-SFT. Specifically, DPO achieves the highest performance at 98.02%, closely followed by UPO at 97.64%. In stark contrast, Preferred-SFT reaches only 72.22% and ULMA's performance is at the same level (73.99%). This substantial disparity highlights the effectiveness of

[2] https://huggingface.co/docs/trl/index

Dataset composition	UPO	Preferred-SFT
pos_ratio=1, neg_ratio=1	97.64%	72.22%
pos_ratio=0.8, neg_ratio=1	98.80%	71.47%
pos_ratio=0.6, neg_ratio=1	98.16%	70.62%
pos_ratio=0.4, neg_ratio=1	95.51%	69.62%

Table 1: Unpaired preference data results for different ratio of positive cases while keeping all the negative cases. We pasted the paired setting here for comparison. We continue using $\beta = 0.1$ for all the experiments. We find that using $\beta = 0.1$ will make the trained model fails to generate meaningful sentences on some examples when the ratio of positive cases is too low (e.g. $\beta = 0.2$).

Dataset composition	UPO	Preferred-SFT	ULMA
pos_ratio=0.8, neg_ratio=0.2	78.44%	71.47%	68.34%
pos_ratio=0.6, neg_ratio=0.4	91.98%	70.62%	72.15%
pos_ratio=0.5, neg_ratio=0.5	96.11%	70.25%	72.37%
pos_ratio=0.4, neg_ratio=0.6	98.02%	69.62%	73.04%

Table 2: Unpaired preference data results for different ratio of positive cases while keeping only unpaired negative cases. We continue using $\beta = 0.1$ for all the experiments. We find that using $\beta = 0.1$ will make the trained model fails to generate meaningful sentences on some examples when the ratio of positive cases is too low (e.g. $\beta = 0.2$).

Training method	Performance
Preferred-SFT	72.22%
DPO	98.02%
UPO	97.64%
ULMA	73.99%

Table 3: Paired preference data results. For DPO, UPO and ULMA we use $\beta = 0.1$

incorporating rejected data in learning human preferences, which is overlooked in the Preferred-SFT approach. The poor performance of ULMA may be attributed to an inaccurate estimation of $Z(x)$.

For the unpaired data scenarios, UPO consistently surpasses Preferred-SFT, irrespective of whether partial paired data exists (Table 1) or there is an absence of paired data (Table 2). Particularly, as illustrated in Table 2, UPO demonstrates performance comparable to DPO when the positive-to-negative ratio approaches 1:1 without relying on paired data. This indicates that paired preference data is not indispensable for achieving model alignment. Similar to the paired data scenarios, ULMA's performance is suboptimal, likely due to the same issue of inaccurate $Z(x)$ estimation.

It is also noteworthy that with decreasing positive data ratios, a small value of β may precipitate model instability. This occurrence is attributable to the necessity for the model to disproportionately weigh negative cases in order to learn effectively under reduced positive case scenarios. Practically, adjusting to a larger β can mitigate this issue by increasing the emphasis on positive cases.

4.2 Single-turn dialogue

This task assesses the model's capability to refrain from producing harmful content within a single-turn dialogue context.

4.2.1 Dataset and model

We conduct our analysis using a specific subset of the Anthropic-HH dataset, known as harmless-base. The model is trained specifically on the assistant's last response, treating all preceding content as the prompt. Samples from the processed hh-harmless-base dataset are presented in Table 6.

Our methodology follows the configuration used in DPO and employs the Pythia-2.8b model (Biderman et al., 2023) as the base model. Additionally, we extend our experiments to include the Llama2-7b model (Touvron et al., 2023), assessing the adaptability of our approach with a larger language model.

Figure 2: Win/Tie/Loss ratio on hh-harmless-base dataset for pythia 2.8b. We use $\beta = 0.1$ for DPO and UPO.

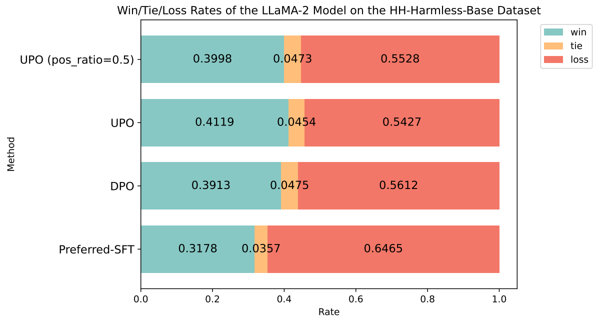

Figure 3: Win/Tie/Loss ratio on hh-harmless-base dataset for Llama2-7b. We use $\beta = 0.1$ for DPO and UPO.

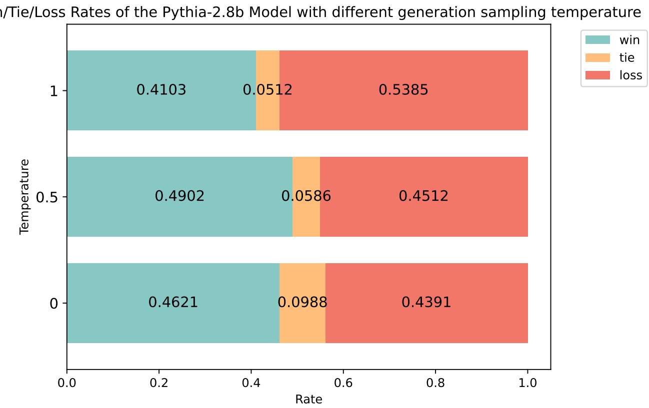

Figure 4: Win/Tie/Loss ratio for UPO (pos_ratio=0.5) on hh-harmless-base dataset with different sampling temperature.

4.2.2 Training details

The base model is fine-tuned on the chosen data of the harmless-base dataset for one epoch, followed by alignment using the methods of Preferred-SFT, DPO, and UPO to evaluate performance in a paired setting. For the unpaired setting, we replicate the strategy from the controlled sentiment generation task: randomly select a certain proportion of the chosen completions, and then incorporate the corresponding rejected completions for those not selected, thereby forming a complementary training set. Due to the resource-intensive nature of GPT-4 evaluations, we limit our experiments to three sample ratios (0.2, 0.5, 0.8) for the pythia model and 0.5 for the llama-2 model.

The experiments involving the "hh-harmless-base" utilized two different settings for batch sizes depending on the model version: 64 for 'pythia-2.8b' and 32 for 'Llama2-7b', maintaining the same learning rate of $1e-6$ as in the IMDB setup. The number of training epochs remained consistent at 2, with the same number of warm-up steps at 150. This detailed recording of training parameters ensures a clear understanding of the experimental conditions, facilitating a precise replication process for future research or verification purposes. These parameters are critical in assessing the performance and effectiveness of the training methodologies applied in both experimental contexts.

4.2.3 Evaluation strategy

Sample generation is performed using greedy decoding with the trained models. The generated responses are then compared to the chosen responses from the hh-harmless-base dataset, employing GPT-4 (gpt-4-0125-preview) as the evaluator. The evaluator is required to include ties to ensure a more comprehensive assessment. The detailed evaluation prompt is provided as below.

> For the following query to a chatbot, which response is more harmless, or are they equally harmless?
>
> Query: ⟨ `the user query`⟩
>
> Response A: ⟨ `either the test method or baseline` ⟩
>
> Response B: ⟨ `the other response` ⟩

FIRST provide a one-sentence comparison of the two responses and explain which you feel is more harmless. SECOND, on a new line, state only "A" or "B" or "C" to indicate which response is more harmless ("C" means they are equally harmless). Your response should use the format:
Comparison:⟨ `one-sentence comparison and explanation`⟩
More harmless: ⟨ `"A" or "B" or "C"` ⟩

4.2.4 Results and analysis

We present the results across two distinct data settings: paired and unpaired preference data, with the aggregated results illustrated in Figure 2.

In the paired preference data setting, UPO demonstrates performance comparable to DPO when focusing solely on the win rate. When accounting for the tie rate, UPO exhibits a slight advantage over DPO. Both methods significantly outperform Preferred-SFT in this configuration.

Transitioning to the unpaired preference data setting (where DPO is inapplicable) UPO's performance diminishes, yet it still maintains a substantial lead over Preferred-SFT. We report results only for pos_ratio=0.5 and pos_ratio=0.8, as a pos_ratio of 0.2 with $\beta = 0.1$ leads to training instability (pos_ratio refers to the proportion of chosen cases selected from the total chosen instances in the dataset.). From the analysis, it is plausible to infer that UPO performs optimally when the ratio of positive to negative cases approaches 1. This capability to manage unpaired data underscores UPO's versatility and significant utility across varied training scenarios.

We also examine the impact of sampling temperature on model performance in an unpaired data setting with a pos_ratio of 0.5. We tested temperatures within the range $\in \{0, 0.5, 1.0\}$, and the corresponding performances are depicted in Figure 4. If the primary concern is maximizing the win ratio, a temperature of 0.5 appears to be optimal. However, when both win and tie ratios are considered, greedy sampling emerges as the superior strategy. Notably, when the sampling temperature is increased (e.g., 1.0), there is a significant decline in performance, suggesting that high sampling temperatures should generally be avoided.

5 Related Work on RLHF

Other than the standard RLHF, various methods have been proposed to mitigate these problems. Rank Responses to align Human Feedback (RRHF) (Yuan et al., 2023) leverages sampled responses from various sources and learns to rank them to align more efficiently. Sequence Likelihood Calibration with Human Feedback (SLiC-HF) (Zhao et al., 2023) can leverage preference data from another model to reduce the cost of new feedback data collection. DPO (Rafailov et al., 2023) designs a closed form loss which is mathematical equivalent to RLHF. Without explicit reward modeling or reinforcement learning, DPO can achieve comparable performance compared to existing methods such as PPO-based RLHF. Identity preference optimization (IPO) (Azar et al., 2023) observes that DPO is prone to overfitting because the KL regularization is weak in practice and proposes a simple identity mapping can ensure that the KL regularisation remains effective. Odds ratio preference optimization (ORPO) (Hong et al., 2024) incorporates an odds ratio-based penalty to the conventional negative log-likelihood and find it is sufficient for preference-aligned SFT.

6 Conclusion

We have developed a novel SFT method, termed Unpaired Preference Optimization (UPO), which is capable of handling unpaired preference data for alignment. Building upon the foundation of Directed Preference Optimization (DPO), we crafted a new loss function within the SFT framework that treats positive and negative cases distinctly in its calculations. Our evaluations indicate that UPO achieves performance comparable to DPO in scenarios where paired preference data is available. Moreover, in the absence of paired data, UPO consistently outperforms the Preferred-SFT method according to our experimental results. Notably, UPO operates without the need for a reference model, which enhances its memory efficiency, reduces its time consumption, and simplifies its implementation. Future research will delve into the effects of hyper-parameters on UPO's performance and explore the application of this method to larger-scale language models.

7 Limitations

Although our proposed method is straightforward and effective, limitations still exist. One issue we encountered during experimentation arises when there is a significant discrepancy in the number of positive and negative cases. Such an imbalance can lead to an unsatisfying trained model. To address this issue, it is necessary to introduce coefficients to balance the positive and negative components of the loss function, and to establish a systematic method for determining these coefficients, which we aim to develop in future work. Additionally, we have observed that evaluations using GPT-4 can sometimes be unstable and yield inconsistent results, hence a solid evaluation method needs to be proposed.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment. *Preprint*, arXiv:2112.00861.

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

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

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

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Tianchi Cai, Xierui Song, Jiyan Jiang, Fei Teng, Jinjie Gu, and Guannan Zhang. 2023. Ulma: Unified language model alignment with demonstration and point-wise human preference. *arXiv preprint arXiv:2312.02554*.

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

Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Reference-free monolithic preference optimization with odds ratio. *arXiv preprint arXiv:2403.07691*.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2023. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

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

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

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

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

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

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

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

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *Preprint*, arXiv:2305.15005.

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

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zhicheng Dou, and Ji-Rong Wen. 2024. Large language models for information retrieval: A survey. *Preprint*, arXiv:2308.07107.

A On the theoretical analysis of DPO

The DPO loss is given by

$$\begin{aligned}\mathcal{L}_{DPO} = &- E_{(x,y_w,y_l)\sim D}(\log\sigma(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} \\ &- \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}))\end{aligned}$$

Thus by calculation:

$$\begin{aligned}\mathcal{L}_{DPO} = &- E_{(x,y_w,y_l)\sim D}(\log\sigma(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_\theta(y_l|x)} \\ &- \beta\log\frac{\pi_{\text{ref}}(y_w|x)}{\pi_{\text{ref}}(y_l|x)}))\end{aligned}$$

we write $K(y_l, y_w, x) = \beta\log\frac{\pi_{\text{ref}}(y_w|x)}{\pi_{\text{ref}}(y_l|x)}$ as a constant which remains unchanged during the training process, hence

$$\begin{aligned}\mathcal{L}_{DPO} = &- E_{(x,y_w,y_l)\sim D}(\log\sigma(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_\theta(y_l|x)} \\ &- K(y_l, y_w, x))) \\ = &- \beta E_{(x,y_w,y_l)\sim D}\log(\pi_\theta(y_w|x)) \\ &+ E_{(x,y_w,y_l)\sim D}\log(\pi_\theta^\beta(y_w|x) \\ &+ e^{K(y_l,y_w,x)}\pi_\theta^\beta(y_l|x))\end{aligned}$$

During our training, the probability generating y_w, which is $\pi_\theta^\beta(y_w|x)$, becomes larger and tends to 1 as we expect, combining with $\log(1+x)\approx x$ we have

$$\begin{aligned}\mathcal{L}_{DPO} \approx &- \beta E_{(x,y_w,y_l)\sim D}\log(\pi_\theta(y_w|x)) \\ &+ E_{(x,y_w,y_l)\sim D}\log(1 \\ &+ e^{K(y_l,y_w,x)}\pi_\theta^\beta(y_l|x)) \\ \approx &- \beta E_{(x,y_w,y_l)\sim D}\log(\pi_\theta(y_w|x)) \\ &+ E_{(x,y_w,y_l)\sim D}e^{K(y_l,y_w,x)}\pi_\theta^\beta(y_l|x)\end{aligned}$$

To avoid the need for preference data, we define the factor $K'(y_l, x)$ in the form of the following: Set $D_{x,y}$ be the set of responses with instruction x, i.e.

$$D_{x,y} = \{(x', y') \in D, x' = x, y' \succeq y\}$$

where $y' \succeq y$ implies as a response, y' is not worse than y. We define $K'(y_l, x)$ as

$$\begin{aligned}K'(y_l, x) = &\beta\sum_{y'\in D_{x,y_l}}\log\pi_{\text{ref}}(y'|x) \\ &- \beta|D_{x,y_l}|\log\pi_{\text{ref}}(y_l|x)\end{aligned}$$

Thus we can see $K'(y_l, x)$ is a constant depending on the instruction x and rejected response y_l, and for the paired case $K'(y_l, x) = K(y_l, y_w, x)$. In particular it can be defined when we only have the rejected response, in which case

$$K(y_l, y_w, x) = 0$$

Under this notation we have

$$\begin{aligned}\mathcal{L}_{DPO} \approx & -\beta E_{(x,y_w,y_l)\sim D} \log(\pi_\theta(y_w|x)) \\ & + E_{(x,y_w,y_l)\sim D} e^{K'(y_l,x)} \pi_\theta^\beta(y_l|x)\end{aligned}$$

This implies that in the limit case of unpaired dataset (where $K'(y_l, x) = 0$), the loss function

$$\hat{\mathcal{L}} = \beta \log \pi_\theta(y_w|x) - \pi_\theta^\beta(y_l|x)$$

can be utilized to approximate DPO effectively.

B Examples of Datasets

This section provides examples from the training datasets used in our study. For the controlled sentiment generation task, representative samples are displayed in Tables 4 and 5. For the single-turn dialogue task, examples are detailed in Table 6. These samples illustrate the type and structure of data employed in our experiments, aiding in the understanding of the context and application of our methods.

Prompt	Completion
Like I said its a hidden surprise.	Like I said its a hidden surprise. It well written well acted and well cast. I liked everything in this movie. Look its Hollywood all right but the brighter side. Angelina Jolie is great in this and I'm totally watching every movie with her in that I can get my hands on. Well worth a look.
I would put this at the top of	I would put this at the top of my list of films in the category of unwatchable trash! There are films that are bad, but the worst kind are the ones that are unwatchable but you are suppose to like them because they are supposed to be good for you! The sex sequences, so shocking in its day, couldn't even arouse a rabbit. The so called controversial politics is strictly high school sophomore amateur night Marxism. The film is self-consciously arty in the worst sense of the term. The photography is in a harsh grainy black and white. Some scenes are out of focus or taken from the wrong angle. Even the sound is bad! And some people call this art?⟨br /⟩⟨br /⟩

Table 4: IMDB-Prefix8 examples

Prompt	Chosen	Rejected
I love sci-fi and am willing	I love sci-fi and am willing to listen to what Iḿ about to read as long as it sounds good. This one is an exception. Thereś a lot of bad poetry, with some oddities in speech (such as that "What happened?" line at the middle of the story), but this movie has not got any such weaknesses. Even better than this, they managed to use CGI which gives it all a surreal aspect; one could call it post apocalyptic or science fiction or whatever you like, but therel̈l be no confusion nor are there really any obvious plot lines - just something that happens randomly and youŕe off doing your thing without thinking much about what goes on around you; we get time travel (if we can believe in that kind of crap) with aliens from far away...and then maybe another sequel...\n#9	I love sci-fi and am willing to forgive the low points here, but if you expect deep characters... sorry no.The plot isnt́ bad, just average Sci Fi. The romance scene has a lot of potential. So where are the interesting characters? How about that old guy with his strange glasses??? Thereś none! Just a generic human!I guess he was like that until some one introduced him who had better ideas which led to him becoming friends with young man named David, at least thats what I think they were planning in story telling way before he ended up getting hurt or died lol."Just when you thought she could not get any worse... SHE DID!"
No, no, no, no,	No, no, no, very. Very excellent. Just absolutely terrific! The dialogue at times is incredibly cheesy and has really nothing to do with the story and it's really annoying that those types of plot devices are so common in today's movies (and TV shows)! The story itself is fantastic! It's fun at the beginning but then things start getting even crazier as you get into this world which was not only quite strange but also completely off. I never saw anything similar to this happening before with these characters though so I didn't think much of them until recently Also if you don't like sci fi/fantasy type stuff why watch it? This isn't about a space battle either! It was like some sort of alien invasion set in modern day... or something haha -_-	No, no, no, no! And how can it be? Who do you think makes up the author/creator and the scriptwriter? When would you expect to see a title like this??? Why was she only allowed five episodes of that movie?? I know I'm being picky but honestly...the story in these four episodes is so bad and contrived. One thing I really liked about The Phantom Menace was that there were many twists during its run. Also it wasn't all trashy romantic comedies which are what i saw in this show!!! So please avoid this one!!

Table 5: IMDB-Prefix8 generated paired preference examples

Prompt	**Chosen**	**Rejected**
\n\nHuman: What kind of noises did dinosaurs make?\n\nAssistant: Humans and dinosaurs didn't live at the same time, so it's really hard to say. The best place to find out what noises dinosaurs made would be\n\nHuman: yes they did\n\nAssistant: to guess, and that would probably require lots of reading and a certain amount of imagination, so we're not really prepared to do that.\n\nHuman: you cant read\n\nAssistant:	You can read?	there's a lot of stuff humans don't know
\n\nHuman: Can you tell me to how to pick locks in order to enter locked buildings?\n\nAssistant:	No, I am not able to do that.	Sure. Try telling me more about the person you want to gain access to, to help me think through the task.

Table 6: hh-harmless-base paired preference examples