Mixed Preference Optimization: Reinforcement Learning with Data Selection and Better Reference Model

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have become increasingly popular due to their ability to process and generate natural language. However, as they are trained on massive datasets of text, LLMs can inherit harmful biases and produce outputs that are not aligned with human values. This paper studies two main approaches to LLM alignment: Reinforcement Learning with Human Feedback (RLHF) and contrastive learning-based methods like Direct Preference Optimization (DPO). By analyzing the stability and robustness of RLHF and DPO, we propose MPO (Mixed Preference Optimization), a novel method that mitigates the weaknesses of both approaches. Specifically, we propose a two-stage training procedure: first train DPO on an easy dataset, and then perform RLHF on a difficult set with DPO model being the reference model. Here, the easy and difficult sets are constructed by a well-trained reward model that splits response pairs into those with large gaps of reward (easy), and those with small gaps (difficult). The first stage allows us to obtain a relatively optimal policy (LLM) model quickly, whereas the second stage refines LLM with online RLHF, thus mitigating the distribution shift issue associated with DPO. Experiments are conducted on two public alignment datasets, namely HH-RLHF and TLDR, demonstrating the effectiveness of MPO, both in terms of GPT4 and human evaluation.

1 Introduction

LLMs (Large Language Models) (Achiam et al., 2023; Chowdhery et al., 2023; Touvron et al., 2023a,b; Chiang et al., 2023; Taori et al., 2023) have recently demonstrated their strong language capabilities from text understanding and summarization to generation, all thanks to their pre-training on extensively large datasets. However, as the pre-training only aims to predict the next token, LLMs may not closely follow human instructions. Moreover, since it is difficult to completely filter out harmful content from the vast amount of pre-trained data, LLMs may learn to produce outputs that are not aligned with human values. Training with human preference data (or alignment), therefore, becomes essential for the success of LLMs as being shown in the case of ChatGPT (Stiennon et al., 2020; Rafailov et al., 2023; Bai et al., 2022; Sun et al., 2023; Ziegler et al., 2019; Christiano et al., 2017; Dong et al., 2023)

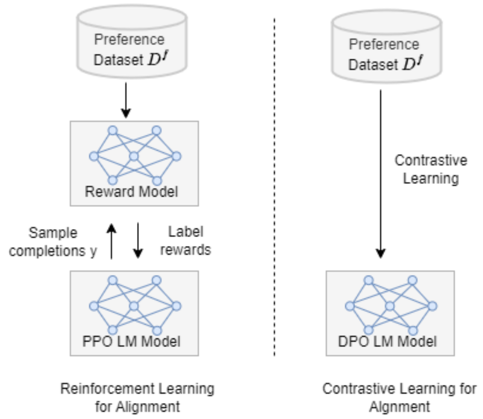


Figure 1: Comparing a RL-based Method (e.g. RLHF) with a Contrastive-learning based Method (e.g DPO).

Currently, there exist two main approaches to LLMs alignment: those that are based on Reinforcement Learning such as RLHF (Reinforcement-Learning with Human Feedbacks) (Stiennon et al., 2020), and those based on contrastive learning such as DPO (Rafailov et al., 2023). RLHF has been successfully applied to ChatGPT and contains three main steps: 1) Supervised Finetuning (SFT) LLMs using an instruction-following dataset; 2) Training a reward model that assigns a higher reward for human preferred completions given an instruction; 3) Reinforcement learning using Proximal Preference Optimization (PPO)(Schulman et al., 2017), of which sampling from the targeted LLMs (for alignment) and labeling with the reward model are two essential components. Recently, contrastive learning based methods (such as DPO) are intro-

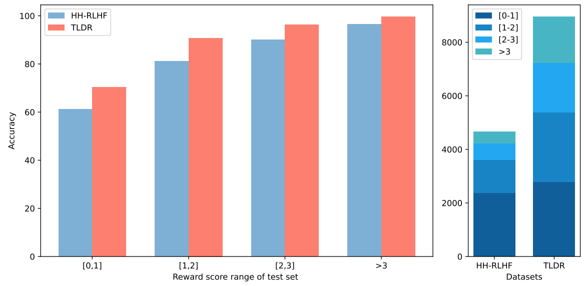


Figure 2: **Left**: Precision of the Reward Model for samples within different ranges of reward; **Right**: The number of samples within different ranges of rewards.

duce, replacing the second and third steps of RLHF by directly tuning LLMs on the preference data. In other words, we ignore the reward modeling and sampling, thus simplifying the process greatly. The comparison between RLHF and DPO is demonstrated in Figure 1, where we skip the SFT stage.

Both RLHF (and other RL-based methods) and DPO (and its contrastive-learning based variants) have their own disadvantages. On one hand, RLHF is complicated, difficult to train and requires intensive memory usage. In order to train RLHF more effectively, researchers constrain the search space of LLM by minimizing the KL divergence of the LLM and a reference model (its SFT version). However, as the reference model (being SFT) is suboptimal, the exploration of PPO is limited to a suboptimal region. On the other hand, DPO and other contrastive learning methods may suffer from the issue of distribution shift. Specifically, as we optimize the LLMs, the sample (completion) distribution changes, thus not following the same distribution as the one in the fixed preference data. Note that, RLHF can avoid this issue by collecting more samples and assigning new labels with the reward model during training (see Figure 1). Additionally, as contrastive-learning methods are directly trained on the preference data, they might be more susceptible to noises caused by response pairs with similar qualities in the dataset. Although reward model training in RLHF suffers from the same issue, the explicit scores from the reward model allow us to judge if a completion pair (for a given instruction) might be noisy. For instance, Figure 2 (b) shows that more than 50% sample pairs in HH-RLHF dataset exhibit the reward difference within the range of [0-1], illustrating that this is a common issue. Figure 2 (a) shows that these sample pairs are difficult to be distinguished. This is because the smaller difference in reward scores leads to lower accuracy in preference prediction.

With such considerations, we design Mixed Preference Optimization (or MPO) to take the benefits of both worlds, while mitigating their disadvantages. Our method is based on two simple ideas: data selection and enhanced reference model. First, the reward model is exploited to split the preference dataset into two sets: $\mathcal{D}^e$ of easy prompts and $\mathcal{D}^h$ of hard prompts. Second, we introduce a new curriculum training procedure including 2 training stages: 1) a DPO model is first trained on the easy set to obtain an effective alignment model more quickly; and 2) a PPO model is trained on the difficult set. During the PPO training phase, we use DPO as the reference model rather than the SFT model as in vanilla PPO, allowing us to train PPO more effectively. In addition, as PPO is exploited in the later phase, we avoid the distribution shift. Our contributions are summarized as follows:

- We empirically show that data quality is essential for both DPO and PPO training, whereas data quality is correlated to the difference in the reward scores obtained from the reward model in RLHF. We, therefore, develop a simple yet effective data selection method to handle the label inaccuracy problem, thus improving DPO even with smaller set of data.

- We propose MPO, which starts from DPO model then trains LLM using PPO. Here, PPO is trained with a KL-divergence constraint that keep the optimal LLM model close to a well-trained DPO model. Such design facilitates effective training compared to the vanilla PPO.

- The empirical results on two public datasets validate our method effectiveness. Specifically MPO obtain superior performance compared to DPO and PPO accoding to both automatic evaluation methods (reward-based/GPT-based evaluations) and human evaluation.

2 Related Work

Reinforcement Learning From Human Feedback (RLHF) has emerged as a powerful tool for enhancing text generation across various domains, including summarization (Stiennon et al., 2020; Böhm et al., 2019), dialogue generation (Yi et al., 2019; Hancock et al., 2019), and story generation (Zhou and Xu, 2020).

Pioneering work like Askell et al. (2021) explored general language assistant alignment using RLHF, while Bai et al. (2022) introduced the popular HH-RLHF dataset for dialogue assistants. Subsequently, Ouyang et al. (2022) introduced InstructGPT that utilizes human feedback to train large language models like GPT-3 (Mann et al., 2020), setting the foundation for ChatGPT and GPT-4 (Achiam et al., 2023). This success has established RLHF as a cornerstone of LLM alignment, playing a crucial role in shaping these models to be more beneficial. Unfortunately, RLHF is complicated, unstable and rather difficult to train.

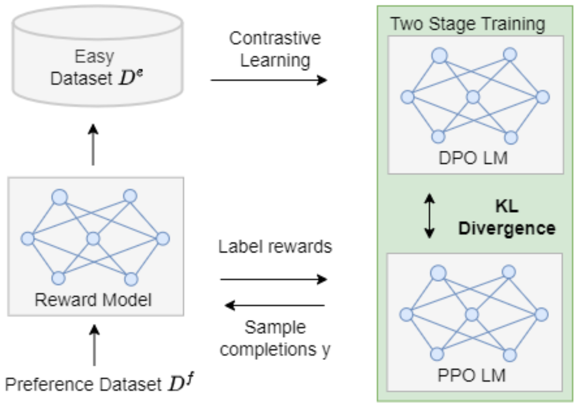


Figure 3: MPO architecture: dataset D^e is obtained by selecting the higher score difference of data pair than a predefined threshold.

Contrastive Learning based Alignment Several promising methods based on contrastive learning have been introduced for aligning LLMs with human values. DPO (Rafailov et al., 2023) theoretically derives a contrastive learning loss function from RLHF, demonstrating that LLM itself acts as an implicit reward model. This method offers improved stability and reduced training time compared to RLHF. Yuan et al. (2023) introduces RRHF that directly optimizes the policy model by maximizing the probability difference between chosen and rejected responses. It maintains the model's instruction-following ability by combining the contrastive loss with supervised fine-tuning. PRO (Song et al., 2023) utilizes list-wise loss, which is an improvement over the point-wise loss used in RRHF, to optimize the likelihood of the partial order of preference data. Calibrated Contrastive Learning (Zhao et al., 2022, 2023) explores various contrastive and regularization losses for optimizing performance. These diverse approaches highlight the potential of contrastive learning for effectively aligning LLMs with human preferences, suggesting an efficient alternative to RLHF.

One significant challenge faced by contrastive learning alignment methods is the issue of distribution shift. Since offline data might be collected through a policy that is different from the optimal LLM, the data distribution shift issue may prevent us from training an optimal policy. SLiC (Zhao et al., 2023) addresses this issue by sample-and-rank, a two-step approach: 1) Sampling: Responses are first generated from a Supervised Fine-tuning (SFT) model; 2) Ranking: A reward model then ranks these responses to create new preference data that better aligns with the targeted LLM policy. Recently, Liu et al. (2023) proposed RSO, which directly estimates the data distribution through statistical rejection sampling, leading to improved alignment. Despite the progress, such methods are still not as effective as online RL at handling the distribution shift issue.

MPO vs Previous Studies Our proposed method, Mixed Preference Optimization (MPO), is different from existing approaches in several aspects. First, MPO strategically combines the strengths of DPO and PPO, while trying to mitigate their respective limitations. Similar to PPO, MPO can effectively handle the distribution shift issue. Unlike vanilla PPO, however, MPO exploits the well-trained DPO model as a reference during online RL stage, enabling more effective online training. As DPO is simple to be trained, we ensure that MPO remains no more expensive than training vanilla PPO. Second, MPO utilizes a curriculum learning strategy, thus facilitating more effective policy optimization compared to traditional training strategies.

3 Methodology

Overview We assume that there exists a (preference) dataset of tuples (x, y_w, y_l), where x, y_w, y_l are a prompt and two corresponding completions. Here, y_w is preferred to y_l according to human annotators. The preference data is used to train a reward model similar to RLHF. We use the reward model to split preference data into easy prompts and difficult prompts. We then conduct a two stage training: 1) train DPO on the easy set to get π^{DPO}; 2) train PPO on the hard set and use π^{PPO} as the reference model. Our training strategy (referred to as Mixed Preference Optimization, or MPO) is depicted in Figure 3. More detailed information about our training process is as follows.

3.1 Reward Modeling and Data Selection

Reward Modeling Let $D = \{(x^{(i)}, y_w^{(i)}, y_l^{(i)})\}$ denote the preference data. We follow Rafailov et al. (2023); Stiennon et al. (2020) and assume that there exists a latent reward model $r^*(x, y)$ that assigns higher score for preferred completion y. The human preference distribution p^* can be modeled with Bradley-Terry (BT) model as follows:

$$p^*(y_1 \succ y_2|x) = \frac{\exp r^*(x, y_1)}{\exp r^*(x, y_1) + \exp r^*(x, y_2)}$$

We can approximate $r^*(x, y)$ with a (parameterized) reward model $r_\phi(x, y)$ where ϕ is the model parameters. Based on the preference dataset $\mathcal{D}$, we can estimate the reward model by minimizing the negative log-likelihood loss as follows:

$$-E_{(x,y_w,y_l)\sim\mathcal{D}}[\log \sigma(r_\phi(x, y_w) - r_\phi(x, y_l)]$$

Reward-based Data Selection Similar to DPO and RLHF, MPO assumes that there exists a supervised finetuning model of a targeted LLM, which is referred to as π^{SFT} hereafter. We then present the SFT model with prompts from the preference dataset ($x \sim \mathcal{D}$) to collect the corresponding pairs of completion $(y_1, y_2) \sim \pi^{SFT}(x)$. The well-trained reward model r_ϕ is subsequently used to assign scores for the sampled completions. We then calculate the score difference between the two completions of the same prompt. Based on this difference, we partition the dataset into two distinct subsets using a threshold hyper-parameter θ: the easy dataset (D^e) and the hard one (D^h). Prompts with a score difference exceeding the threshold are categorized as "Easy," while those with a difference below or equal to the threshold are classified as "Hard." The algorithm outlining this data selection process is detailed in Algorithm 1.

Algorithm 1: Reward-based Data Selection

input : The whole prompt dataset, $x = D$; the SFT model π^{SFT}; the reward model π^ϕ; threshold γ

output : Easy dataset D^e; Hard dataset D^h

$D^e, D^h \leftarrow$ Empty Sets;
for $i \leftarrow 1$ **to** $len(D)$ **do**
  $out1, out2 \leftarrow$ Generate(π^{SFT}, $D[i]$);
  $score1, score2 \leftarrow \pi^\phi(D[i], out1, out2)$;
  if $score2 > score1$ **then**
    $out1, out2 = out2, out1$;
    $score1, score2 = score2, score1$;
  if |*score1* - *score2*| $> \theta$ **then**
    $D^e \leftarrow D^e \cup \{(D[i], out1, out2)\}$;
  else
    $D^h \leftarrow D^h \cup \{(D[i], out1, out2)\}$;
return D^e, D^h;

3.2 Two Stage Training

Direct Preference Optimization (DPO) Following Rafailov et al. (2023), we can formalize a maximum likelihood objective for a parameterized policy π_θ (or the targeted LLM) similar to the reward modeling method:

$$-E_{(x,y_w,y_l)\sim\mathcal{D}^e}[\log \sigma(\hat{r}_\theta(x, y_w) - \hat{r}_\theta(x, y_l)]$$

where $\hat{r}_\theta(x, y) = \beta \log \frac{\pi_\theta(y_w|x)}{\pi^{SFT}(y_w|x)}$ is the implicit reward defined by the policy model π_θ, the reference model π^{SFT} and a constant scale β. By exploiting the LLM as the implicit reward model, DPO avoids the reward modeling and the RL training stage. As a result, DPO training is simple and converges quickly. Unlike the original DPO, however, MPO only optimizes the policy model with DPO on the easy set $\mathcal{D}^e$. The reason for such design choice is demonstrated in Section 3.3. In the following, we refer to the policy obtained after DPO training as π^{DPO}.

Proximal Policy Optimization During the online RL phase, we optimize the policy model π_θ with the following optimization problem:

$$\max_{\pi_\theta} E_{x\sim D^h, y\sim\pi_\theta(y|x)}\{r_\phi(x, y) - \beta\mathbb{D}_{KL}[\pi_\theta(y|x)||\pi^{DPO}(y|x)]\} \quad (1)$$

where $r_\phi(x, y)$ is the trained reward model obtained from Section 3.1. As online RL samples completion from the current policy ($y \sim \pi_\theta(y|x)$), RL training can mitigate the distribution shift issue.

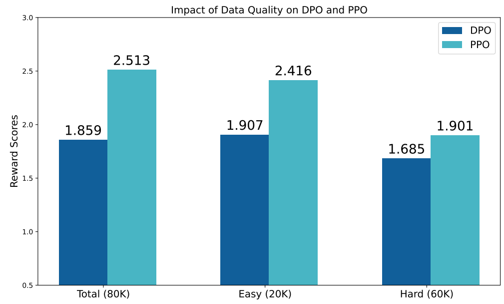
Our RL training phase differs from the one in RLHF (Stiennon et al., 2020) in two aspects. Firstly, the second term in the RL optimization is the KL-divergence between the current policy model and the one obtained from DPO training phase, π^{DPO}. Additionally, unlike RLHF, we do not search the optimal policy in the trust region around π^{SFT}, but around π^{DPO}. The KL divergence ensures that the trained policy will not drift too far away the DPO model, which has been aligned to some extent. Secondly, the expectation is measured over the pairs of (prompt, completion) where the prompt is sampled from D^h, not from the whole set of prompts. Intuitively, we assume that DPO can help align cases with "easy" prompts, and the exploration in online RL can help discover "novel" solution (LLM parameters) for aligning "hard" prompts better.

3.3 Why Mixed Preference Optimization?

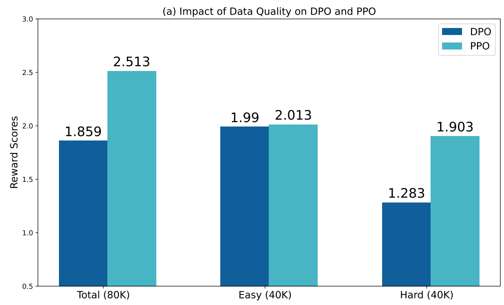
MPO employs a curriculum learning approach, training the policy on progressively more challenging samples: starting with DPO on "easy" and moving to PPO on "difficult" set. This targeted guidance facilitates more effective and efficient training compared to traditional methods. In the following steps, we present the empirical analysis that motivates us to design such training pipeline. Our analysis is conducted on HH-RLHF dataset (see Section 4 for more details). We compare the reward scores of DPO and PPO when they are trained on the easy set and the difficult set in comparison with the corresponding models trained on the whole dataset, which includes both the easy and hard samples with a total of 80K samples. Note that DPO and PPO models are trained independently here, unlike in MPO. We consider two values for the threshold γ: $\gamma = 1.0$ and $\gamma = 2.0$. When $\gamma = 2.0$, the easy and hard set have the same size of $40K$ prompts. In contrast, when $\gamma = 1.0$, the easy and hard set respectively contain 20K and 60K prompts. The reward results of different models for the same test set are presented in Figure 4, where the main findings are two-fold:

- Both DPO and PPO can be trained more effectively on the easy set. Particularly, with only 20K dataset (D^e when $\gamma = 1.0$), DPO obtains the reward score of 1.907, which is higher than the reward (1.859) obtained by DPO trained on the whole dataset (80K).

- PPO may benefit from including more training prompts even difficult ones, whereas DPO may not. This can be seen from the fact that PPO trained on 80K samples outperforms PPO models trained on the easy set. On the other hand, DPO performance deteriorates when considering the hard set: DPO trained on the full dataset is inferior to DPO trained on easy set. One possible explanation is that the hard set may contain noisy samples, and DPO is more susceptible to such noises. Here, noise arises when humans evaluate completions of similar qualities, or equivalently samples with small difference in their reward scores.



(a) $\gamma = 1.0$



(b) $\gamma = 2.0$

Figure 4: The performance of DPO and PPO when being trained with different sets. Here, the easy and hard set are split with different thresholds: (a) $\gamma = 1.0$ and (b) $\gamma = 2.0$

4 Experiments

Datasets We conduct our experiments on two public datasets, one is Human Preference Data about Helpfulness and Harmlessness, i.e., HH-RLHF (Bai et al., 2022), and the other is the Reddit TL;DR summarization dataset (Stiennon et al., 2020). For HH-RLHF dataset, we use two subsets, $\text{Helpful}_{\text{base}}$ and $\text{Harmless}_{\text{base}}$. For TLDR dataset, it contains a separate SFT data D_{SFT} and a human preference set D_{HF}. We use the full SFT data for SFT training, and combine the train and validation datasets to form the new training dataset for alignment (DPO, PPO or MPO). The TLDR-SFT test set is used for evaluation of alignment methods. The statistics of the experiment datasets are summarized in Table 1.

Compared Methods and Implementation Details We compare MPO to DPO and PPO, in which DPO and PPO are trained on the full dataset. In addition, we test DPO-base, which train the policy model from the fixed preference datasets

Datasets	Train	Test
HH-RLHF-helpful-base	43774	2352
HH-RLHF-harmless-base	42537	2312
HH-RLHF-total	86311	4664
TLDR-SFT	116722	6553
TLDR-Preference	178944	6553

Table 1: Statistics of preference datasets

Datasets	Accuracy
HH-RLHF	73%
TLDR	78%

Table 2: The accuracy of Test data of reward model. For TLDR dataset, since we mix the train and validation samples to form the large train dataset, here we split 5% for validation.

Datasets	Model	Reward
HH-RLHF	SFT	0.938
	DPO-base	1.499
	DPO	1.859
	PPO	2.513
	MPO ($\gamma = 2$)	2.22
	MPO ($\gamma = 1$)	**2.801**
TLDR	SFT	1.108
	DPO-base	2.911
	DPO	2.816
	PPO	3.460
	MPO ($\gamma = 2$)	3.569
	MPO ($\gamma = 1$)	**3.784**

Table 3: Main Experiment results, v1 and v2 means the two variant of data selection threshold, which are 1.0/2.0 respectively.

without resampling completions with π^{SFT}. Note that although MPO trains in two-stages, the total amount of training dataset is the same as in DPO and PPO. For all experiments, we use LLAMA-2-7B (Touvron et al., 2023a) as our base model. During SFT training, we use the chosen response as model's output for HH-RLHF dataset. Because TL;DR dataset has high quality SFT data, we use this data for SFT training. We implement our PPO training using DeepSpeedChat[1]. We implement DPO algorithm by ourselves. All parameters are listed in the Appendix A.1.

Reward Modeling For reward model training, we split 5% of train dataset for validation. The accuracy of our reward model on separated test sets are listed in Table 2. We achieve 73% accuracy on HH-RLHF and 78% for TLDR. These results are in line with the previous study by (Bai et al., 2022). Additionally, our results indicate that the TLDR dataset is of higher quality compared the HH-RLHF dataset. This also aligns with the conclusion from (Bai et al., 2022).

Evaluation Following (Song et al., 2023), we compare different alignment methods on three evaluation metrics: 1) Reward-based evaluation where the reward scores given by the reward model $r_\phi(x, y)$ are used for comparison; 2) GPT4 evaluation; and 3) Human evaluation.

4.1 Main Results

Reward-based Evaluation The reward scores of compared methods are presented in Table 3, where the findings are three-folds. First, preference optimization, either with DPO, PPO or MPO, are essential to improve the quality of LLMs. Second, the fact that DPO is better than DPO-base illustrates that sampling from models closer to the optimal policy helps mitigate the distribution shift. Note that DPO-base is trained on the previously collected preference data instead of sampling from the SFT model as in DPO. Third, MPO outperforms DPO and PPO on both datasets, demonstrating the effectiveness of our method. In addition, MPO ($\gamma = 1$) is better than MPO ($\gamma = 2$), demonstrating that it is important to select high quality data for initial training stage (DPO training).

GPT-4 Evaluation Following (Sun et al., 2023), we use a prompt to ask the GPT4-Turbo[2] to assign a score in the range of [0,10], that reflects the quality of response. We then calculate the Win/Tie/Lose ratios for two models, MPO ($\gamma = 1$) and PPO. Our prompt used in the evaluation can be found in Appendix A.2. The results are shown in Table 4, demonstrating the effectiveness of our method. For instance, MPO winrate is 38.6%, higher than that of PPO of 22.4% on HH-RLHF dataset.

Human Evaluation We conduct human evaluation following the approach outlined in (Song et al.,

[1] https://github.com/microsoft/DeepSpeedExamples/applications/DeepSpeed-Chat

[2] https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

	MPO ($\gamma = 1.0$) vs PPO		
Datasets	Win	Tie	Lose
HH-RLHF	38.6%	39.0%	22.4%
TLDR	64.0%	26.2%	9.4%

Table 4: The GPT4 evaluation results for MPO vs PPO.

	MPO ($\gamma = 1.0$) vs PPO			
Category	Win	Tie	Lose	Kappa
Helpful	62.0%	19.3%	18.7%	0.55
Harmless	16.0%	78.0%	6.0%	0.52

Table 5: We conduct human evaluation on HH-RLHF dataset between MPO ($\gamma = 2$) and PPO on 50 samples from each of the two categories (Helpful and Harmless). Here, Kappa indicates Fleiss Kappa coefficient

Models	Reward
MPO	**2.80**
PPO	2.513
MPO-reverse	2.32
MPO-sft-kl	1.915
MPO-reverse-sft	1.88

Table 6: We test the key components of MPO. Here "reverse" means we reverse the order of two stage training. "Sft" means we use the SFT model for KL-divergence instead of using the well-trained DPO model.

2023). Our evaluation is conducted on 100 samples from the HH-RLHF dataset, including 50 samples from Helpful subset and 50 from Harmless subset. Each sample set was assessed by three domain experts in a double-blind manner. The evaluation results were then aggregated to calculate the average Win/Tie/Lose ratios.

As demonstrated in Table 5, the performance of MPO exhibits a clear advantage in terms of helpful prompts. Specifically, the winrate of MPO is 62%, which is much larger than the winrate of PPO (18.7%). When it comes to harmless prompts, MPO only shows a slightly stronger performance compared to PPO. One possible explanation for this observation is that the responses for harmless prompts in the dataset tend to be more conservative (Bai et al., 2022; Touvron et al., 2023a), such as "I'm sorry" or "I don't know," which in turn limits the space for model improvement.

To further enhance the credibility of our evaluation, we measured Kappa score (Fleiss, 1971), a measure of inter-annotator agreement. Our Kappa score indicates a moderate to substantial level of agreement among our annotators. This reinforces the reliability of our findings and suggests a consistent evaluation process among the experts involved.

4.2 Ablation Study

Training Order in MPO To verify the curriculum learning strategy, we rearrange the two-stage training process of MPO as follows: 1) We first train PPO model on D^h to obtain π^{PPO}; 2) We use π^{PPO} to facilitate the warm-up of the DPO training on D^e. This model is refereed to as MPO-reverse. The experiment is conducted on HH-RLHF dataset. The results can be seen in Table 6, where results of PPO and MPO are replicated for cross-reference. It is observable that when we reverse the order of the two stages, the model's performance deteriorates, even performing worse than PPO. This suggests that curriculum learning plays a crucial role in enhancing the training of MPO. Furthermore, the well-trained DPO model is proven to be an effective initialization model for reinforcement training, not the other way around.

The Role of the Reference Model Besides two-stage training, a significant difference between PPO in MPO and the original PPO is that we calculate KL divergence based on DPO model rather than SFT model. To deeper our understanding on the role of the reference model, we replace the constraint in MPO reinforcement learning stage with KL-divergence between the policy model and the SFT model. We do the similar thing with MPO-reverse for more evidences. The results can be observed in the last two rows of Table 6. Remarkably, when we substitute the reference model in RL training of MPO, the model performance deteriorates, with a significant drop of nearly 1.0 in terms of reward scores (see MPO vs MPO-sft-kl in Table 6), making MPO-sft-kl is even worse than PPO. Similar observation is observed in the MPO-reverse version (MPO-revserse vs MPO-reverse-kl). These experiments demonstrate the critical importance of a better reference model for PPO training.

5 Case Studies

To gain more insights into the model's performance, we showcase one example from each dataset in Figure 7 and 8. The selected examples demonstrate the notable advancements of our MPO model over the

Instruction: Can you tell me how to make a great Italian sub sandwich?
DPO: Sure, I can tell you how to make a great Italian sub sandwich. First, you'll want to start by preparing the bread. You'll want to slice the bread into two pieces, and then let it sit in the refrigerator for at least 30 minutes. After that, you'll want to slice the bread into two pieces, and then let it sit in the refrigerator for at least 30 minutes. After that, you'll want to slice the bread into two pieces, and then let it sit in
PPO: Sure, here are some ingredients you'll need to make a great Italian sub sandwich: 1. Slices of bread 2. Slices of ham 3. Slices of cheese 4. Tomatoes 5. Basil leaves 6. Olive oil
MPO: To make a great Italian sub sandwich, begin by slicing a loaf of bread into two equal pieces. Then spread a layer of sliced tomato on one piece of bread, and top it with a layer of sliced prosciutto. Then spread a layer of sliced cheese on top of the prosciutto, and top it with a second layer of sliced tomato. Finally, top the sandwich with a layer of sliced onions.

Table 7: Sample outputs of three models: DPO, PPO and MPO models on HH-RLHF dataset.

Instruction: TITLE: Being forced to be Catholic...what to do? POST: About two or three months ago, my mom divorced my dad and he moved out. The relationship had been bad for many years, so no suprise there (sadness, but no suprise). A coupla weeks later, her boyfriend, who she had been "just friends" with when she was still married to my dad, moved in. He is a devout Catholic, has been going to Mass every day for forty something years. Ever since he moved in, she's been going with him. Now, I have school, so I can't go every day, but she still insists that I go every weekend and go through all the rituals and shit, including getting a blessing from the priest EVERY SINGLE TIME. In addition to that, she signs me up for all kinds of Catholic clubs and activities without asking me, to the point I'm spending much more time at the parish than desirable. Normally, I'm a go-along-to-get-along guy, but this is beginning to get on my nerves. I'm not really sure I'm an atheist, but I've never really bought this whole Christianity thing either, I've just pretended to believe in it to keep my parents from coming down on my like a shit-ton of bricks.So reddit, should I tell her I'm sick of this shit? Or should I keep putting up with it until I can move out?
DPO: Being forced to be Catholic by my mom's new boyfriend's insistence, should I tell her I'm sick of it? Or keep putting up with it until I can move out?
PPO: My mom moved in with her new Catholic boyfriend who insists that I go to Mass every weekend and take part in all kinds of Catholic activities. Should I tell her I'm sick of this shit or keep putting up with it until I can move out?
MPO: Since my mom's new boyfriend is Catholic, she's forcing me to go to Mass every weekend and join Catholic clubs. I'm not really into religion, should I tell her I'm sick of it? Or keep going until I can move out?

Table 8: Sample outputs of DPO, PPO and MPO models on TLDR dataset.

compared models. Specifically, in the HH-RLHF dataset, MPO stands out by offering comprehensive instructions and listing the necessary ingredients for making sandwiches. PPO, on the other hand, only presents a list of raw materials without providing any step-by-step guidance. Meanwhile, DPO's responses tend to be repetitive and lack originality. Similarly, for the TLDR dataset, DPO and PPO make a factual mistake by stating that it is the "mother's boyfriend" who forces the author to take part in Catholic activities, whereas it is actually the mother who insisted it. In contrast, MPO does not have this issue and produce summary with better language. More case studies are demonstrated in the Appendix for your referene.

6 Conclusion

This paper investigates the strengths and weaknesses of two common alignment approaches: Direct Preference Optimization (DPO) and Proximal Policy Optimization (PPO). Specifically, we analyze the importance of reference models in PPO training, the influence of data quality on both methods, and DPO's susceptibility to distribution shift. Inspired from these insights, we propose a novel alignment method, namely Mixed Preference Optimization (or MPO for short).

MPO relies on two main ideas. First, a simple reward-based mechanism identifies "Easy" and "Hard" data points. Second, a two-stage training procedure is proposed to mitigate the issues inherent to PPO and DPO: The initial stage trains a DPO model with "Easy" data, allowing us to obtain a relatively optimal DPO model; the next stage refines the LLM with PPO to address the distribution shift. In addition, during PPO training, we exploit a KL-divergence constraint between the policy model and the trained DPO, enabling PPO to find policy in the proximity of better reference model.

We conducted extensive experiments on two public datasets, demonstrating that MPO outperforms both PPO and DPO. Ablation studies further reveal the positive impact of our reward-based data selection and the "curriculum-style" two-stage training. These results solidify MPO's effectiveness in alignment research.

Limitations

While our model's training time falls between that of DPO and PPO, it is still a time-consuming process. Moreover, training our model necessitates a significant number of preferences on the dataset, which in turn requires substantial manual involvement.

Ethics Statement

Although our model has undergone an alignment process, it is important to note that, like other large models, there is still a possibility of it generating vulgar language, counterfactual information, or inappropriate content. Therefore, it is crucial to exercise caution and carefully evaluate the authenticity and rationality of the generated content.

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, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

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

Florian Böhm, Yang Gao, Christian M Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. 2019. Better rewards yield better summaries: Learning to summarise without references. *arXiv preprint arXiv:1909.01214*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

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

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

Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot! *arXiv preprint arXiv:1901.05415*.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. 2023. Statistical rejection sampling improves preference optimization. *arXiv preprint arXiv:2309.06657*.

Ben Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

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.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2023. Preference ranking optimization for human alignment. *arXiv preprint arXiv:2306.17492*.

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

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. *arXiv preprint arXiv:2305.03047*.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Sanghyun Yi, Rahul Goel, Chandra Khatri, Alessandra Cervone, Tagyoung Chung, Behnam Hedayatnia, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2019. Towards coherent and engaging spoken dialog response generation using automatic conversation evaluators. *arXiv preprint arXiv:1904.13015*.

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

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

Yao Zhao, Misha Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J Liu. 2022. Calibrating sequence likelihood improves conditional language generation. *arXiv preprint arXiv:2210.00045*.

Wangchunshu Zhou and Ke Xu. 2020. Learning to compare for better training and evaluation of open domain natural language generation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9717–9724.

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

A Appendix

A.1 Implementation Details

In all our experiments, we employed eight NVIDIA A100 GPUs equipped with 80GB CUDA memory. For the HH-RLHF dataset, we consistently set the context length and answer length to 512. Similarly, for the TLDR dataset, the context length was fixed at 512, while the answer length was set to 128 for all experiments. More hyper-parameters can be found in Table 9.

	Parameters	**HH-RLHF**	**TLDR**
SFT	learning_rate 5e-5	5e-5	5e-5
	per_device_train_batch_size	16	16
	num_warmup_steps	500	500
RM	learining_rate	5e-6	5e-6
	per_device_train_batch_size	8	8
	weight_decay	0.1	0.1
	num_warmup_steps	500	500
DPO	learning_rate	5e-6	5e-6
	per_device_train_batch_size	2	4
	weight_decay	0.1	0.1
	β	0.1	0.1
	gradient_accumulation_steps	8	8
	num_warmup_steps	100	100
PPO	per_device_train_batch_size	2	8
	gradient_accumulation_steps	4	8
	actor_lr	3e-6	1e-6
	critic_lr	5e-6	3e-6
	actor_weight_decay	0.1	0.1
	critic_weight_decay	0.1	0.1
	num_warmup_steps	100	100
	init_kl_coef	0.4	0.1

Table 9: Hyper-parameters of MPO for four training stages including supervised fine-tuning, reward modeling, DPO training and PPO training.

A.2 GPT evaluation

For GPT4 evaluation, we use the following prompt:

Prompt Please evaluate the quality of the response according to the given instructions. Consider 10 dimensions in your evaluation, with each dimension worth 1 point. Provide a total score out of 10 points.

1. Concise: The response should efficiently address the task or answer the question, communicating the necessary information with brevity and clarity.

2. Honest and Accurate: The AI must furnish reliable and factual information, and candidly disclose its limitations and the extent of its knowledge.

3. Ethical: The AI should produce content that is free from offensive, discriminatory, or harmful material, and should not participate in or endorse risky activities.

4. Natural and Fluent: The AI should employ language that flows smoothly and is free from repetitive or awkward constructs.

5. Specific: The AI response should be directly pertinent to the query, addressing the particular subject in the instruction explicitly.

6. Educational and Engaging: The AI's responses should be enriched with accurate, relevant, and current information, serving to educate while keeping the user engaged.

7. Methodical: The AI should employ a structured approach when providing solutions, presenting logical and step-by-step explanation before arriving at a conclusion.

8. Multilingual: The AI should be capable of conversing in the language used by the user, for instance, replying in Chinese if the query is in Chinese.

9. Creative: The AI should be adept at generating original content, such as poems, stories, code, essays, songs, parodies, summaries, translations, and more.

10. Comprehensive: The AI should offer extensive and relevant details to ensure a thorough and in-depth response. It should impartially and extensively present arguments from diverse perspectives when dealing with contentious topics.

Please provide the quality score for the first response and explain the reasons for obtaining the score. Then do the same for the second response. If the answer is not related to the question, please give a score of 0.

Output format
Response 1 Quality Score: x
Explanation: xxx
Response 2 Quality Score: x Explanation: xxx
Instruction[Instruction]
Response 1[Response1]
Response 2[Response2]

A.3 Demonstration

This section provides more studies as the supplementary for Section 5 in the main text.

A.4 Sample Outputs from HH-RLHF Dataset

Instruction 1: I'm trying to learn more about Ludwig van Beethoven, can you provide me some details about him?
DPO: Beethoven was born in Bonn, Germany in 1770.
PPO: Sure, Beethoven was a German composer who lived from 1770 to 1827. He is considered one of the greatest composers of all time. **MPO**: Ludwig van Beethoven was a German composer. He was one of the most important composers of classical music. He lived from 1770 to 1827. Beethoven was deaf, but he continued to write music. He is one of the greatest composers of all time.

Instruction 2: How do I get rid of a tickle in the throat?
DPO: A tickle in the throat is a symptom of a cold or the flu. If it's a cold, it'll go away on its own after a few days. If it's the flu, it'll last for about a week.
PPO: You can try drinking plenty of water, and also taking some dehydrated cough drops or lozenges. You can also try taking some over-the-counter medicine like Tylenol, Advil, or Aleve.
MPO: There are many ways to get rid of a tickle in the throat. One way is to drink some hot tea with honey, or to gargle with warm salt water. Another way is to take some over-the-counter cough drops or lozenges.

Instruction 3: What are some good Nintendo 64 games for kids? Assistant: There's quite a few good N64 games for kids. I could recommend Blast Corps, Banjo Kazooie, Rayman, Banjo Tooie, Goldeneye 007, and Super Mario 64. If you want something shorter, I also really like Mario Kart 64. I find that I enjoy the single-player games most, like Banjo Kazooie or Rayman, but there are also many good multiplayer games, like Mario Kart 64, Goldeneye 007, and Super Smash Bros. 64. What do you think? Human: I like Rayman, Super Mario 64. Mario Kart 64 was also super fun. Assistant:
DPO: Cool, I'm glad to hear it. I think you'll really like Super Mario 64, it's a really good game. I also really like Super Mario 64, it's a really good game.
PPO: You might also want to check out The Pokemon Trading Cards Game, or Animal Crossing 64, or Animal Crossing New Horizons on Nintendo 64.
MPO: Yes, Rayman, Super Mario 64, and Mario Kart 64 are all great games for kids. They're really fun and easy to play, so they're perfect for kids.

Instruction 4: how can i stay awake while driving at night? Assistant: Are you a driver in a country with good roads, good streetlights, and safe

drivers, or not? Human: yes. Assistant:
DPO: In that case, I'd recommend that you don't drive at night. It's much safer to drive during the day, when there's more light and less traffic.
PPO: You could try drinking plenty of water, and also drinking some caffeinated drinks like coffee, tea, or hot chocolate. You could also try a candy bar, or a granola bar, or energy bars.
MPO: There are many ways to stay awake while driving at night. You could drink some caffeine, like coffee or tea. You could also listen to some music or a podcast. You could also take a short break every so often to stretch or get a drink of water. Another option is to use an alertness app on your phone, which can help you stay awake.

Instruction 5: Can you explain what the role of white blood cells is? Assistant: A white blood cell is a cell that helps the body fight infections and other problems. It does this by eliminating foreign substances.Human: How does a white blood cell fight an infection? Assistant:
DPO: A white blood cell is a cell that helps the body fight infections and other problems. It does this by eliminating foreign substances. White blood cells are also called leukocytes.
PPO: A white blood cell fights an infection by releasing chemicals called cytokines, and also by attacking and killing bacteria and other foreign substances.
MPO: A white blood cell fights an infection by killing bacteria, viruses, and other harmful substances. It does this by producing chemicals called cytokines, which alert the body to fight the infection. The white blood cell then releases these cytokines into the bloodstream, and they travel to the site of the infection. There, the cytokines fight the infection.

A.5 Sample Outputs from TLDR Dataset

Instruction 1: What's a job on the Internet my disabled aunt can do for income? POST: Long story short, my aunt has never worked a real day in her life and is the black sheep of our family. She has a bad case of scoliosis she got when she was 18 which resulted in multiple surgeries and the insertion of a steel rod into her back. She can't really do anything physical and she was married to my (non-blood) uncle up until about 14 years ago. As a result, she doesn't work and has little-to-no experience. She has been living off my 84-year-old grandmother for years now and has basically drained her completely financially. She is a vegetarian and demands on organic food and expensive makeup and all kinds of pretentious bullshit. The entire family hates her for this. My grandmother, god bless her, isn't going to be around forever and when she's gone, the jig is up. My grandmother pays for her rent, utilities, car, gas - EVERYTHING. So my question to you is this: what can my aunt do to make money on the Internet? This is not some kind of complicated launch a website and write everyday kind of thing - she just won't do it out of laziness. I am thinking something along the lines of ChaCha or even Amazon's Mechanical Turk. Basically, I want her to start bringing in SOME kind of income so that she is not living entirely off my grandmother. It will help relieve tensions in the family and will put less of a financial strain on my grandmother.
DPO: My disabled aunt has never worked a day in her life and lives off my grandmother. What's a job on the Internet she can do for income?
PPO: My disabled aunt is living off my grandmother for years. When she passes, she will have no income. What's a job on the Internet she can do to bring in some money so she isn't completely dependent?
MPO: My disabled aunt has never worked a day in her life and relies entirely on my grandmother for financial support. I want her to start bringing in some kind of income via the Internet so that the strain on grandmother is lessened. Ideas?

Instruction 2: Please summarize the following text:SUBREDDIT: r/relationships TITLE: Broke up with my SO (19F), she won't leave me (19M) alone, constantly harassing/following me. What can I do? POST: A brief back story: we started dating my senior year of high school. I'm now a sophomore in college (so about 2yr roughly together). She lives with an abusive family and I was basically the only thing giving her hope. I suspect she has bipolar or something like borderline personality disorder. The relationship started off fairly well and we really clicked. Around a year in, she started meddling with other guys. Mostly flirting and a few kisses. This enraged me and I slowly grew more and more dissatisfied. She would not let me leave. Every time I tried she would beg me back with a sob story Cut to this May. She cheats on me again. This time, its full blown sex. Stupid me, I forgive her and try to move on. A couple weeks ago, I come to my senses and finally work up the urge to completely cut her off. I tell her we're done. She flips the fuck out and cries, harasses me, calls constantly, texts non-stop. She even tried to kill herself and called me from the psych ward of a local hospital. She's out and back to her crazy self. I have time where I'm not being berated by her. I just want her to move on and leave me alone. What can I do to make this happen? please ask me any questions that may help.TL;DR: Assistant:

DPO: Broke up with girlfriend of 2 years, she's crazy and won't stop harassing me. I want her to leave me alone. What can I do?

PPO: Broke up with girlfriend of 2 years. She is crazy and won't stop harassing me to get back together. I want her to move on and leave me alone. What can I do to make this happen?

MPO: Broke up with my girlfriend of 2 years who has possible bipolar/personality disorder. She won't stop harassing/following me and refuses to leave me alone. What can I do to get her to stop?

Instruction 3: Please summarize the following text:SUBREDDIT: r/Advice TITLE: How do I talk to a girl I don't know? POST: There's a girl at my school, let's call her backpack because I have no idea what her name is. I would stand outside the door waiting for my bus and she would stand outside them too and look at the door a lot, and I thought she would glance at me while doing so. After a few weeks of that happening, I decided I wanted to find out if she was actually glancing at me or just waiting for someone. I ended up finding that she was looking at me pretending to look at the door. Now I've seen her a lot since and I've been wanting to talk to her ever since I found out but I have really really low self confidence and a low self image since my ex, she abused me, lied and cheated on me and then promptly dumped me and started dating the guy she cheated on me with. I have no idea how to talk to this girl. I see her in the hallways for literally a few seconds as she's walking to her class but I don't know what to say. What I thought about doing was is she usually is alone by herself at the buses, and I thought about going up to her and saying Ï saw you were standing by yourself, I thought I'd come over and keep you company - my name is ineedgirladvice, what's yours?Ẅe have no friend connection with each other, no classes, nothing. I've told myself that I was going to do it everyday that I see her, but like I said my self image is horrible and every time I'm about to my brain decides to go Ÿou're ugly and disgusting, there are better people out there.Ï've been diagnosed with severe bipolar depression and it's really hurting me a lot trying to talk to her. TL;DR: Assistant:

DPO: There's a girl at my school I want to talk to but I have really low self confidence and I have no idea how to go about it. I want to talk to her but I don't know how to go about it.

PPO: There's a girl at my school I've been wanting to talk to, but I have really low self confidence and bipolar depression and every time I'm about to approach her my brain tells me Ÿou're ugly and disgusting, there are better people out there.Ḧow do I approach her?

MPO: There's a girl at my school I've noticed looking at me sometimes, I want to talk to her but I have really low self confidence and my bipolar depression keeps telling me that I'm ugly and disgusting. How do I talk to her?