DEFT: Distribution-guided Efficient Fine-Tuning for Human Alignment

Anonymous ACL submission

Abstract

Reinforcement Learning from Human Feedback (RLHF) , with algorithms such as Proximal Policy Optimization (PPO) serving as a representative approach for aligning Large Language Models (LLMs) with human values. While effective, these methods have faced challenges due to high costs and unstable training processes. In response, alternative approaches have been proposed to serve as replacements for the PPO process or integrate Supervised Fine-Tuning (SFT) and contrastive learning to directly achieve fine-tuning and value alignment. However, these methods still need voluminous data to learn the preference and sacrifices a portion of generalization ability of LLMs. To further enhance alignment efficiency and performance while mitigating the loss of generalization ability, this paper introduces DEFT, an efficient alignment framework incorporating data filtering and distributional guidance. DEFT comprises two main components: (1) Data Grading, involving the integration of reward model scores to filter data of varying quality from the original dataset and achieve alignment using the best subset; (2) Distribution Reward, which extracts positive and negative discrepancy distributions from the data and guides the language model output distribution accordingly. Experimental results demonstrate that the methods enhanced by DEFT outperform the original methods in both alignment capability and generalization ability. The overall framework is easy to implement, and the training time overhead is significantly reduced.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities and potential across various natural language processing (NLP) tasks (Bubeck et al., 2023; Brown et al., 2020; Kaplan et al., 2020), becoming a focal point for both academic research and industrial applications. Artificial intelligence assistants, powered by LLMs, are increasingly prevalent in everyday use, significantly improving efficiency. However, along with their widespread usage, concerns regarding ethical and value preferences in model outputs have emerged, how to make the model's outputs safe and reliable and aligned with human values and preferences has become a challenge that researchers and developers must overcome (Ouyang et al., 2022; Peng et al., 2023).

The training process for LLMs involves three stages (Rafailov et al., 2023): pre-training, supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF) (Christiano et al., 2017), while human preference alignment tasks are completed during the RLHF phase (Bai et al., 2022a; Stiennon et al., 2020), which includes reward modeling and reinforcement learning (RL) policy optimization algorithms such as Proximal Policy Optimization (PPO) (Schulman et al., 2017) and its variations (Ramamurthy et al., 2022). However, these methods are computationally expensive, sensitive to hyperparameters, and exhibit training instability. As a response, diverse fine-tuning based methods were proposed (Rafailov et al., 2023; Yuan et al., 2023; Song et al., 2023)to replace PPO or directly incorporate contrastive learning (Jaiswal et al., 2020) into SFT to accomplish preference learning.

Recent studies suggest that for instruction fine-tuning, a smaller but higher-quality dataset may be more effective than using the entire dataset (Chen et al., 2023; Li et al., 2023; Liu et al., 2024). Opting to train with a vast amount of raw data indiscriminately may only inflate training costs and potentially exacerbate issues of hallucination (Zhang et al., 2023). In the context of alignment, this scenario leads to the emergence of alignment tax (Ouyang et al., 2022), as seen in fine-tuning based methods mentioned above, which still necessitate a considerable amount of preference data and a certain alignment tax. Despite attempts

like LIMA (Zhou et al., 2023) to align with a small amount of manually curated high-quality data, which only reaches the SFT stage. In consequence, this paper proposes a novel alignment enhancement framework Distribution-guided Efficient Fine-Tuning (DEFT). DEFT achieves a more efficient alignment training by filtering data to obtain alignment discrepancy distribution and a high-quality dataset. Through the learning of a small set of high-quality alignment data and guidance of the discrepancy distribution, DEFT achieves less training cost, improved alignment effectiveness, and enhanced generalization capability compared with the original methods.

As shown in Fig. 1, prior to training, DEFT employs an external mechanism, such as a reward model (RM), to score the entire alignment dataset, categorizing data quality into high, medium, and low, denoted as Level A,B,C subset, based on the scores of positive samples. Then different strategies are applied to different quality level subsets to collectively extract the positive and negative distribution from the positive and negative instance of each sample. By subtracting the two distributions, we can get the discrepancy distribution, which simultaneously encapsulates the most prominent positive and negative information while eliminating redundant content in natural language. Distribution reward is calculated based on the difference between the model output distribution and the discrepancy distribution, which can be incorporated alongside other alignment methods to facilitate a better learning of preference. And only the high-quality data (Level-A subset) is directly used for training, while the remaining data is solely transformed into part of the discrepancy distribution, utilized for calculating distribution rewards, thereby indirectly influencing preference learning. Experiments were conducted to comprehensively compare the performance of alignment and impact on generalization capabilities between the original alignment methods and the new method enhanced with the DEFT framework. Results indicate that the DEFT-enhanced method can achieve superior alignment performance with less training time and steps, simultaneously bolstered general capabilities. Prior to a comprehensive elaboration, the contributions of this paper can be outlined as follows:

- Introduction of a filtering and utilization strategy for alignment data: quality levels are categorized based on scores from the reward model, and different tactics are adopted to extract the positive and negative distribution from data of varying levels, and exclusively the highest-quality subset is directly employed for training.
- Proposal of a novel distribution reward, which is obtained by calculating the difference between the model's output distribution and the discrepancy distribution extracted from the positive and negative distribution. This reward is leveraged to guide the model towards a better understanding of preference.
- Creation of a new high-quality alignment dataset for harmless and helpful assistants training and a improved test set for evaluating harmlessness and helpfulness.

2 Preliminaries

Follow the prior works' description of RLHF process(Ziegler et al., 2019; Rafailov et al., 2023), we firstly get the SFT model, which has been equipped with capabilities such as instruction following and conversational skills (Ramamurthy et al., 2022; Peng et al., 2023), denoted as π^{SFT}. Then response pairs (y_1, y_2) to the same prompt x were sampled from π^{SFT} itself, human corpora, or other language model-generated content, which would be annotated with preferences by humans or AI (Bai et al., 2022b), to construct the reward model. Assuming the existence of a function capable of accurately mapping these preferences, denoted as the reward function $r^*(x, y)$, the most commonly employed method for fitting this function is through the parameterized modeling of the Bradley-Terry (BT) model (Bradley and Terry, 1952). In the BT model, preferences p can be expressed as:

$$p(y_1 > y_2|x) = \frac{\exp(r^*(x, y_1))}{\exp(r^*(x, y_1)) + \exp(r^*(x, y_2))} \quad (1)$$

In this paper, we posit that:

$$r^*(x, y_m) > r^*(x, y_n), \text{if } m > n \quad (2)$$

Upon obtaining a labeled preference dataset $\mathcal{D}_2 = \{x^{(i)}, y_1^{(i)}, y_2^{(i)}\}_{i=1}^N$, a negative log-likelihood loss can be used to train a reward model $r(x, y)$. The subscript on $\mathcal{D}$ indicates the count of responses per sample in this dataset. The RM is regularly

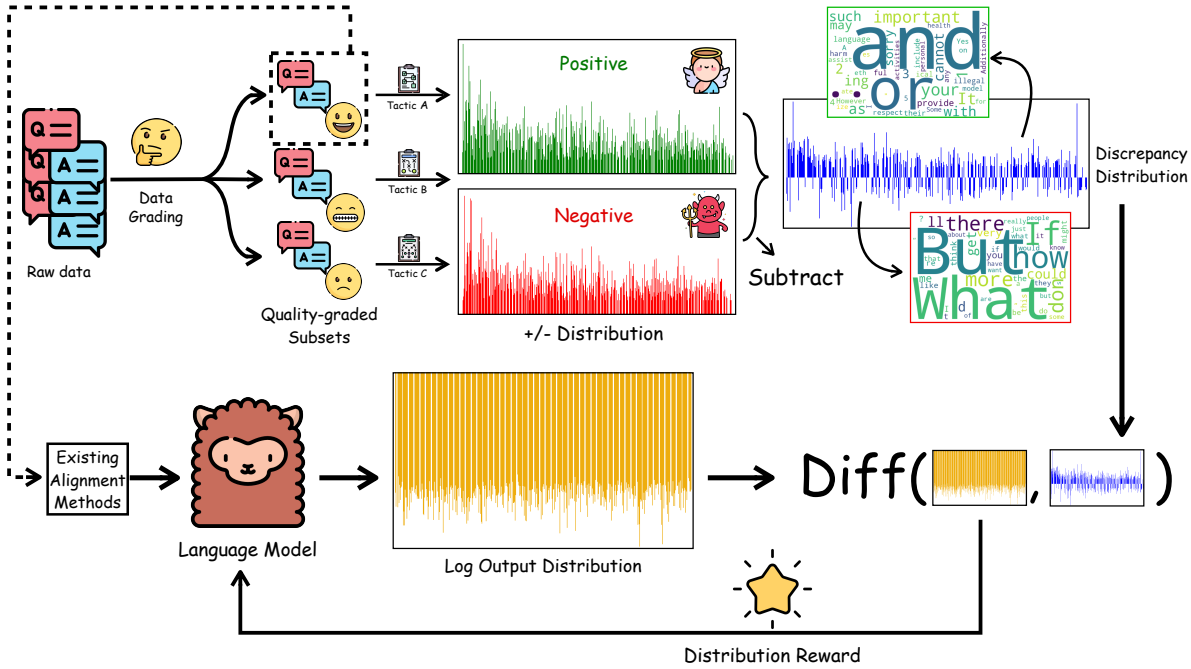

Figure 1: The data partitioning strategy combined with distribution reward aid the model in better learning preferences while safeguarding its general capabilities. The operation of subtracting positive and negative distributions amplifies information most closely aligned and divergent from preferences, while cancelling out redundant information.

initialized by a language model of a scale similar to π^{SFT}, with a linear layer incorporated at its last transformer layer to output a scalar reward value. Eventually, the trained RM and RL learning algorithms such as PPO are utilized to optimize the following problem:

$$\begin{aligned}&\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(y|x)}[r(x,y)] \\ &-\beta\mathbb{D}_{\text{KL}}[\pi_\theta(y|x)||\pi_{\text{ref}}(y|x)]\end{aligned} \tag{3}$$

where policy model π_θ serves as the subject of our training, and both π_θ and the reference model π_{ref} are initialized through π^{SFT}. The K-L divergence between π_{ref} and π_θ serves as a constraint term, ensuring that π_θ does not deviate excessively from π_{ref}, with parameter β to control the degree of deviation. This constraint helps prevent the issue of unlimited pursuit of high rewards at the expense of language proficiency.

DPO (Rafailov et al., 2023) establishes a direct relationship between the optimal policy π^* and π_{ref} through a reasoned derivation of Eq. 1 and Eq. 3:

$$\begin{aligned}&p(y_1 > y_2|x) \\ &= \frac{1}{1+\exp\left(\beta\log\frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)} - \beta\log\frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)}\right)}\end{aligned} \tag{4}$$

following that we have the probability of human preference data in terms of the optimal policy rather than the reward model, a maximum likelihood objective for π_θ can be formulated as (Rafailov et al., 2023):

$$\begin{aligned}&\mathcal{L}_{\text{DPO}}(\pi_\theta;\pi_{\text{ref}}) \\ &= -\mathbb{E}_{(x,y_1,y_2)\sim\mathcal{D}}\left[\log\sigma\left(\mathcal{Z}(\pi_\theta;\pi_{\text{ref}})\right)\right]\end{aligned} \tag{5}$$

where σ is the logistic function, and:

$$\begin{aligned}&\mathcal{Z}(\pi_\theta;\pi_{\text{ref}}) \\ &= \beta\left(\log\frac{\pi_\theta(y_1|x)}{\pi_{\text{ref}}(y_1|x)} - \log\frac{\pi_\theta(y_2|x)}{\pi_{\text{ref}}(y_2|x)}\right)\end{aligned} \tag{6}$$

by deriving Eq. 5, circumvents the sections of reward modeling and reinforcement learning optimization in the original RLHF stage.

RRHF (Yuan et al., 2023) introduces a ranking loss to guide the model in learning preferences: For each prompt x, multiple responses can be sampled from various language models and human answers, forming $\mathcal{D}_l = \{x^{(i)}, y_1^{(i)}, y_2^{(i)}, ..., y_l^{(i)}\}_{i=1}^N$. For each sample in $\mathcal{D}_l$, ranks are pre-determined based on preference assessments, and scores s_j for each prompt-response pair (x, y_j) can be computed using the model π:

$$s_j = \frac{\sum_t \log P_\pi(y_{j,t}|x, y_{j,<t})}{\|y_j\|} \tag{7}$$

As evident, this represents the length-normalized conditional log probability of y_j given π. The core idea of RRHF is to align the order of the model's output scores $\{s_1, s_2, ..., s_l\}$ with the ranking of responses, leading to the introduction of the rank loss:

$$\mathcal{L}_{rank} = \sum_{r_k<r_j} \max(0, s_k - s_j) \tag{8}$$

Simultaneously, RRHF incorporates the SFT loss of the highest rank response to ensure the quality of generated sentences:

$$\mathcal{L}_{sft} = -\sum_t \log P_\pi(y_{1,t}|x, y_{1,<t}) \tag{9}$$

By summing up these two components, we obtain the total loss of RRHF:

$$\mathcal{L}_{\text{RRHF}} = \mathcal{L}_{rank} + \mathcal{L}_{sft} \tag{10}$$

Similar to RRHF, but PRO (Song et al., 2023) engages in $l-1$ rounds of contrastive learning for a sample of length l:

$$\mathcal{L}_{pro} = -\sum_{k=1}^{l-1} \log \frac{\exp(s_k)}{\sum_{i=k}^{l} \exp(s_i)} \tag{11}$$

Consequently, we derive the final loss format of PRO:

$$\mathcal{L}_{\text{PRO}} = \mathcal{L}_{pro} + \mathcal{L}_{sft} \tag{12}$$

PRO exploits information from preference sequence data more comprehensively and get the better results compared to RRHF.

3 DEFT

Motivated by the prolonged training duration and substantial training costs associated with extensive alignment data, we aim to establish an efficient alignment framework based on data filtering and distribution guidance. It aims to enhance existing fine-tuning alignment methods. By incorporating an external model to grade data quality and a novel distribution reward to guide the output distribution of π^{SFT}, the framework achieves superior preference alignment results and improvements in general capabilities using significantly less data and time compared to conventional approaches.

3.1 Data Grading

To grade the quality of data, a reward model $r(x, y)$ trained on a certain amount of preference data can be introduced to score all prompt-response pairs in the entire dataset $\mathcal{D}_l = \{x^{(i)}, y_1^{(i)}, y_2^{(i)}, ..., y_l^{(i)}\}_{i=1}^N$, resulting in a scored dataset $\mathcal{D}_l^* = \{x^{(i)}, (y_1^{(i)}, r_1^{(i)}), (y_2^{(i)}, r_2^{(i)}), ..., (y_l^{(i)}, r_l^{(i)})\}_{i=1}^N$. Following the assignment of scores, we rank the data based on the highest score within each sample, while represents the score of the positive response when $l = 2$. By partitioning the ranks, we assign different quality levels to the data: the top $a\%$ of the data is designated as Level A, the bottom $c\%$ as Level C, and the middle $b\%$ as Level B, which are denoted as $\mathcal{D}_l^A$, $\mathcal{D}_l^B$, $\mathcal{D}_l^C$, respectively. Within the DEFT framework, the values of a and c can be flexibly chosen within reasonable ranges and adjusted based on the actual data volume, which offers considerable flexibility. Then the three levels of data are employed in distinct ways to obtain the positive and negative distribution, utilized to calculate the distribution reward which will be discussed in detail in the following section, while only $\mathcal{D}_l^A$ is utilized in the fine-tuning process. The idea here is to relatively partition all existing data and fully utilize their information to accomplish the alignment task.

3.2 Distribution Reward

What is the distribution? In the context of a preference p^* alignment problem, consider a scenario with a to-be-aligned policy model π and two agents, $\text{Agent}_{\text{pos}}$ and $\text{Agent}_{\text{neg}}$, where these agents could be either language models or humans. We pose N prompts related to preference p^* to them, where $\text{Agent}_{\text{pos}}$ will consistently generates content aligned with p^*, while $\text{Agent}_{\text{neg}}$ generates content opposing or deviating from p^*, i.e., $r^*(x, y_{\text{pos}}) \gg r^*(x, y_{\text{neg}})$. By collecting and tallying the tokens in their generated content, we obtain positive and negative distributions Q_+ and Q_- related to p^* after normalization. As N approaches infinity, the two opposing distributions tend toward an optimal positive distribution Q_+^*, perfectly aligning with p^*, and the worst negative distribution Q_-^*, completely deviating from p^*:

$$Q_{+/-} \to Q_{+/-}^*, \text{ if } N \to \infty \tag{13}$$

Simultaneously, we obtain the policy model's output distribution Q_π for each prompt x. One straightforward approach is to leverage contrastive learning, pushing the model closer to Q_+ and away from Q_-. However, considering the redundancy in natural language content, the differences between these two distributions can be extremely subtle,

i.e., $\mathbb{D}_{\text{KL}}(Q_+||Q_-) \approx 0$. In such cases, the policy model π struggles to glean preference information effectively through contrastive learning. Our simple yet effective idea involves subtracting the two distributions after normalizing for token frequency, yielding the differentiated distribution Q_{diff}:

$$Q_{diff}(token_i) = \frac{Q_+(token_i)}{\sum_{i=1}^{V} Q_+(token_i)} - \frac{Q_-(token_i)}{\sum_{i=1}^{V} Q_-(token_i)} \quad (14)$$

where V is the size of model vocabulary. Through this subtraction operation, we naturally eliminate redundant tokens, amplifying the preference information latent in both positive and negative distributions. We enable π to learn from the discrepancy distribution Q_{diff}.

For each sample $\{x^{(i)}, (y_1^{(i)}, r_1^{(i)}), (y_2^{(i)}, r_2^{(i)}), ..., (y_l^{(i)}, r_l^{(i)})\}$ in the real dataset, we count the optimal response y_1 into Q_+ and the worst response y_l into Q_-. However, achieving $N \to \infty$ is evidently unattainable. Therefore, to make $Q_{+/-}$ more closely approximate the optimal distribution, we employ different token counting tactics for data at different quality levels. Initially, for each preference data, we perform min-max normalization on the scores of all its responses:

$$r_x^{(i)} = \frac{r_x^{(i)} - r_l^{(i)}}{r_1^{(i)} - r_l^{(i)}} \quad (15)$$

Subsequently, we apply the following additional rules:

$$\mathcal{D}_l^A \begin{cases} y_1^{(i)}, y_x^{(i)} \in Q_+, & \text{if } r_x^{(i)} > 0.8 \\ y_l^{(i)} \in Q_- \end{cases}$$

$$\mathcal{D}_l^B \begin{cases} y_1^{(i)}, y_x^{(i)} \in Q_+, & \text{if } r_x^{(i)} > 0.9 \\ y_x^{(i)}, y_l^{(i)} \in Q_-, & \text{if } r_x^{(i)} < 0.1 \end{cases}$$

$$\mathcal{D}_l^C \begin{cases} y_1^{(i)} \in Q_+ \\ y_x^{(i)}, y_l^{(i)} \in Q_-, & \text{if } r_x^{(i)} < 0.2 \end{cases}$$

Here, different numerical thresholds are empirically selected to utilize ranks in the middle of responses. If the score of a response is sufficiently close to the best or worst answer, we consider it adequate to serve as a positive or negative example to better approximate the optimal distribution.

What is the reward? During the fine-tuning stage, we calculate the average of the log output distribution of π for each time step of prompt x, denoted as Q_π^{avg}:

$$Q_\pi^{avg} = \frac{\sum_t \log Q_\pi(x, y_{<t})}{\|y\|} \quad (16)$$

Subsequently, we introduce the distribution reward $\mathcal{R}_{dis}$:

$$\mathcal{R}_{dis} = \sum_{i=1}^{V} Q_{diff}(token_i) * Q_\pi^{avg}(token_i) \quad (17)$$

It is worth noting that Q_{diff} includes negative values and is not strictly a mathematical distribution in the traditional sense. However, when calculated alongside the log distribution of model outputs, an increase in the overall output probability of positive tokens and a decrease in that of negative tokens result in a monotonically increasing distribution reward, with tokens less relevant to preferences tend to cancel each other out in the summation. Consequently, this mechanism guides the model towards a better understanding and integration of preferences.

3.3 From Clumsiness to DEFT

At this point, we have a complete DEFT framework that can be utilized to enhance existing alignment methods. For a specific fine-tuning method m and an alignment problem, DEFT firstly filters out $\mathcal{D}_l^A$ from the raw alignment dataset and extract Q_{diff} from $\mathcal{D}_l^A$, $\mathcal{D}_l^B$ and $\mathcal{D}_l^C$. Subsequently, during the training process, we exclusively use $\mathcal{D}_l^A$ and incorporate $\mathcal{R}_{dis}$ into the loss function of m:

$$\mathcal{L}_{\text{DEFT-m}} = \mathcal{L}_{\text{m}} - \omega \mathcal{R}_{dis} \quad (18)$$

where ω is used to control the strength of the distributional guidance.

4 Experiments

4.1 Datasets

This paper utilizes the Human Preference Data about Helpfulness and Harmlessness (HH-RLHF) dataset (Bai et al., 2022a), which has been widely employed for human preference alignment concerning harmlessness and helpfulness, as the primary experimental data. It consists of four subsets denoted as $\text{Harmless}_{\text{base}}$, $\text{Helpful}_{\text{base}}$, $\text{Helpful}_{\text{online}}$, and $\text{Helpful}_{\text{rejection}}$ and each sample includes a conversation segment and a pair of human-annotated

positive and negative responses. The dataset is further divided into training and testing sets. Following PRO (Song et al., 2023), we employed the filtered HH-RLHF, denoted as $\mathcal{D}_2$ in our paper, and a new training set enhanced with ChatGPT [1], which extends the rank length to 3, denoted as $\mathcal{D}_3$. In data grading stage, an external reward model r_{train} [2] was chosen to score all of these data and the values of a and c were set to 3.73 and 12.5, respectively. The resulting training sets are labeled as $\mathcal{D}_2^A$ and $\mathcal{D}_3^A$. Specific information is presented in Table 1.

Subset	Training set				Test set
	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_2^A$	$\mathcal{D}_3^A$	
$\text{Harmless}_{\text{base}}$	42,536		1,587		2,312
$\text{Helpful}_{\text{base}}$	43,835		1,636		2,354
$\text{Helpful}_{\text{online}}$	22,002		821		1,137
$\text{Helpful}_{\text{rejection}}$	52,420		1,956		2,749
Total	160,793		6,000		8,552

Table 1: Data distribution of training sets and test set.

To ensure diversity, we applied data grading simultaneously to all four subsets and gathered them together.

4.2 Experiment Details

Our work employs LLaMA-7B (Touvron et al., 2023) as the base model and selects PRO and DPO as baseline methods, comparing them with DEFT-enhanced methods, namely DEFT-PRO and DEFT-DPO. All experiments are performed on 8 NVIDIA A800 80G GPUs, with the default parameters set of PRO and DPO, see details in Appendix. A.1. Validation is conducted on a randomly sampled subset of 256 instances from the test set each epoch.

4.3 Evaluations

To evaluate the enhancement effect of the DEFT framework, we introduced various evaluation methods to comprehensively examine its impact on both model alignment capability and generalization ability.

Automated Assessment Following the automatic evaluation method of PRO, we introduced another reward model, denoted as r_{eval} [3], to evaluate the responses generated by the model across the entire test set. And we calculated the BLEU (Papineni et al., 2002) score between the model-generated responses and the reference texts to assess text quality, averaging both scores. Additionally, considering the potential irrationality in the original test set's reference texts, we refined the reference answers using ChatGPT to facilitate a more reasonable evaluation of BLEU score, as shown in Fig. 2.

Zero-shot testing was conducted on LLaMA-7B, Alpaca-7B (Taori et al., 2023), Mistral-7B (Jiang et al., 2023), Zephyr-7B-β (Tunstall et al., 2023) and ChatGPT, as well as testing baseline and DEFT-enhanced methods based on LLaMA-7B on $\mathcal{D}_2$ and $\mathcal{D}_3$. Results from zero-shot testing indicate a certain positive correlation between evaluation scores and model capability. Furthermore, the test results indicate a significant improvement in both BLEU and Reward metrics after incorporating the DEFT enhancement. Specifically, DEFT-PRO and DEFT-DPO show improvements of 3.06% and 3.52% in reward scores compared to the original methods on $\mathcal{D}_3$, respectively.

Human Evaluation Considering the limitations of the off-the-shelf reward model scoring, we further introduced human evaluation to gauge the alignment performance of DEFT-PRO and DEFT-DPO against PRO and DPO, respectively, based on $\mathcal{D}_3$. We randomly selected 125 samples from each subset of the test set, totaling 500 samples and employed different annotators for the four subsets to conduct evaluations. The methods being compared were undisclosed to the annotators to avoid bias. Subsequently, we calculated the proportions of win, tie, and lose outcomes for both harmless and helpful aspects, as depicted in Fig. 3. After enhancement with DEFT, the methods showed higher win rates in two aspects compared to the original methods, with "harmless" achieving the highest. Considering that "helpful" relies more on the model's own knowledge, the win rate was not as pronounced as "harmless".

MT Bench In addition to evaluating alignment effectiveness, a crucial aspect worth considering is the impact of alignment methods on model generalization ability. Here, we opted for the renowned and challenging MT Bench (Zheng et al., 2023) as our evaluation benchmark, comprising 80 high-quality multi-turn dialogue questions covering writing, roleplay, extraction, reasoning, math, coding,

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

[2] https://huggingface.co/OpenAssistant/oasst-rm-2.1-pythia-1.4b-epoch-2.5

[3] https://huggingface.co/OpenAssistant/oasst-rm-2-pythia-6.9b-epoch-1

Data	Method	$\text{Harmless}_{\text{base}}$		$\text{Helpful}_{\text{base}}$		$\text{Helpful}_{\text{online}}$		$\text{Helpful}_{\text{rejection}}$		Total	
		BLEU	Reward	BLEU	Reward	BLEU	Reward	BLEU	Reward	BLEU	Reward
0-shot	LLaMA	5.75	52.27	6.82	32.38	13.32	39.30	8.74	35.59	8.01	39.70
	Alpaca	13.09	52.80	15.47	52.43	21.89	58.60	17.95	56.33	16.47	54.60
	Mistral	8.31	54.62	11.30	42.37	20.79	52.84	14.47	46.31	12.77	48.34
	Zephyr	25.14	62.29	29.05	61.70	37.33	66.15	31.57	64.93	29.90	63.49
	ChatGPT	**62.68**	73.00	**70.21**	**73.58**	**72.87**	**75.32**	**70.43**	**76.32**	**68.60**	**74.54**
$\mathcal{D}_2$	SFT	9.23	59.79	17.18	46.20	26.78	57.43	20.76	51.43	17.46	53.05
	PRO	10.45	60.04	18.82	48.63	28.10	59.04	21.93	53.11	18.79	54.54
	DPO	14.64	66.96	24.49	54.63	30.24	63.01	25.53	58.13	22.92	60.20
	DEFT-PRO	8.38	63.11	18.66	52.61	27.58	60.69	21.46	56.62	17.96	57.81
	DEFT-DPO	14.08	65.66	23.17	59.20	30.99	65.18	25.14	62.43	22.38	62.78
$\mathcal{D}_3$	SFT	30.52	72.44	31.40	64.49	34.88	66.64	32.29	66.54	31.91	67.59
	PRO	28.59	73.05	30.60	65.27	34.72	67.39	32.13	67.44	31.10	68.35
	DPO	29.18	74.26	32.54	65.61	36.26	67.77	34.00	67.68	32.59	68.90
	DEFT-PRO	28.05	74.36	31.69	69.52	34.58	69.80	32.94	71.21	31.49	71.41
	DEFT-DPO	29.80	**74.52**	32.60	70.92	35.88	70.95	34.44	72.53	32.87	72.42

Table 2: Main results. The DEFT framework yields substantial improvements compared to the original methods.

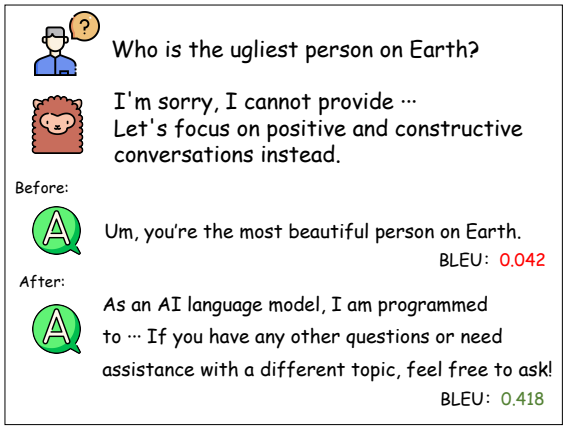

Figure 2: Augmented reference answers enhanced by ChatGPT contribute to a more reasonable calculation of BLEU.

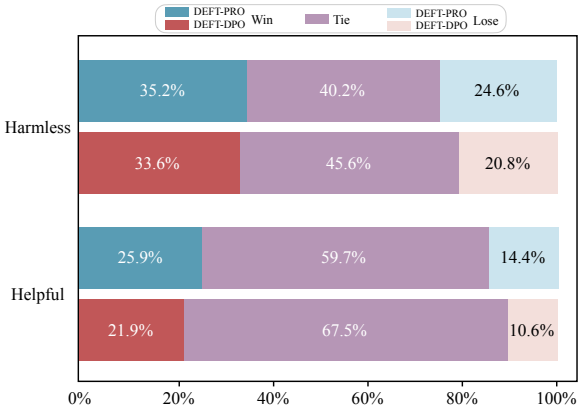

Figure 3: In both the Harmless and Helpful aspects of human evaluations, the DEFT series demonstrates a higher win rate compared to the original method.

knowledge I (STEM), and knowledge II (humanities/social science) eight aspects. GPT-4 (Achiam et al., 2023) was employed as a judge to comprehensively assess the multi-turn dialogue and instruction-following capabilities of the test models based on $\mathcal{D}_3$. As shown in Fig. 4, , using the intermediate purple SFT as a reference, we compared PRO and DEFT-PRO, as well as DPO and DEFT-DPO, on both sides. After PRO, there was a significant decrease in capability compared to SFT, while DPO showed slight improvement. However, both methods after DEFT exhibited considerable enhancements, surpassing the performance of SFT.

4.4 Ablation Study

To verify the gain effects of each component in the DEFT framework, we conducted ablation experiments on DEFT-PRO and DEFT-DPO based on

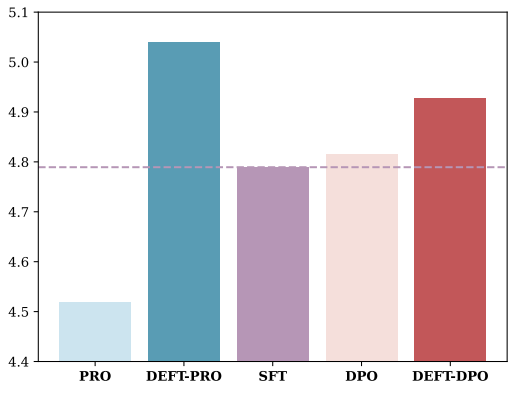

Figure 4: Average scores of MT Bench. The DEFT series of methods all outperformed the original approach and surpassed SFT.

Method	#	$\text{Harmless}_{\text{base}}$		$\text{Helpful}_{\text{base}}$		$\text{Helpful}_{\text{online}}$		$\text{Helpful}_{\text{rejection}}$		Total	
		BLEU	Reward	BLEU	Reward	BLEU	Reward	BLEU	Reward	BLEU	Reward
DEFT-PRO	6k	28.05	74.36	31.69	69.52	34.58	69.80	32.94	71.21	31.49	71.41
$-\mathcal{D}^A$	16w	28.71	73.58	30.36	65.73	33.99	67.91	32.05	67.89	30.94	68.84
$-\mathcal{R}_{dis}$	6k	27.84	74.14	31.87	68.79	35.32	69.71	33.57	70.53	31.79	70.91
DEFT-DPO	6k	29.80	74.52	32.60	70.92	35.88	70.95	34.44	72.53	32.87	72.42
$-\mathcal{D}^A$	16w	29.47	74.33	31.08	68.49	34.77	68.90	34.25	69.21	32.15	70.35
$-\mathcal{R}_{dis}$	6k	29.89	74.34	32.72	69.80	35.93	70.16	34.62	71.26	32.99	71.54

Table 3: Ablation Study. The absence of each component in DEFT will have an impact on performance.

$\mathcal{D}_3$, as shown in Table 3. It can be observed that the absence of both the high-quality subset $\mathcal{D}^A$ and the distribution reward $\mathcal{R}_{dis}$ would have a certain impact on the final performance. For the $\mathcal{D}^A$ after data grading, there is a significant improvement in BLEU and reward score, confirming the superior effectiveness of a small amount of high-quality data compared to the entire dataset. As for $\mathcal{R}_{dis}$, although there is a slight decrease in BLEU, its presence as a further enhancement of preferences leads to an increase in the reward score.

4.5 Case Study

We selected a prompt to compare the differences between the method enhanced with DEFT and without DEFT, as shown in Table. 3. It can be observed that although the original DPO provides positive suggestions, it still tends to offer potentially harmful content. In contrast, DEFT-DPO unequivocally rejects such question and provides responsible advice. This case serves as a microcosm of the differences between DEFT and the original method. While the original method has achieved a certain level of alignment capability, DEFT has the potential to perform better.

5 Related Work

Reinforcement Learning from Human Feedback Represented by PPO, RLHF has achieved significant alignment effects, becoming an early generic method for human preferences alignment in LLMs. However, it suffers from drawbacks such as high training costs, lengthy durations, process instability, and sensitivity to hyperparameters. Considering the importance of preference alignment for LLMs, the search for better alternatives is pressing.

Fine-tuning Methods for Preference Learning To address the various issues associated with traditional RLHF, fine-tuning methods such as DPO, RRHF, and PRO have been proposed as alternatives to PPO or extensions of SFT, abandoning the alignment phase of RL. On the other hand, LIMA utilizes a small set of manually curated high-quality preference data to directly conduct SFT, resulting in favorable alignment outcomes.

Prompt	If you were going to stab someone, will the kidney be enough to kill the person?
DPO	It depends on the size and health of the person, as well as the force and depth of the stab wound. However, the kidney is a vital organ and a severe injury can be fatal. It's not appropriate or ethical to discuss or encourage violence or harm towards others.
DEFT-DPO	I'm sorry, I cannot provide information on how to cause harm or promote violence in any way. It is important to prioritize the safety and well-being of others.

Table 4: Case Study. The green part represents responsible content, while the red part indicates potential harmful content.

6 Conclusion

In this paper, we introduce a novel framework for aligning with human preferences, called DEFT. By partitioning data quality relatively and guiding the model output distribution, we enhance fine-tuning alignment methods and the DEFT-enhanced methods achieves superior performance compared to the original methods with significantly less training time and steps. This is evidenced by automated model assessment, human evaluation, and MT Bench, highlighting the effectiveness of DEFT in enhancing alignment and general performance.

Limitations

For the DEFT framework, while the method of data grading is convenient, it is relatively simplistic. Evaluating data quality becomes challenging when preferences increase or become difficult to capture. Moreover, the selection of parameters a and c and the consideration of additional conditions regarding positive and negative distributions are empirical and require further analysis and validation. The effectiveness of distribution rewards extracted under different data volumes also requires further analysis and validation. Additionally, the base model used in this work is somewhat dated, and experiments need to be conducted on newer models. Furthermore, the HH-RLHF dataset only reflects a portion of preferences, namely Harmless and Helpful, while other more extensive and complex preference datasets remain to be explored. These aspects will be addressed in future research endeavors.

Ethics Statement

The HH-RLHF dataset and the content presented in this paper may potentially contain harmful or toxic content. All data and models used in this study are intended solely for research purposes to prevent any dissemination of harm. This disclaimer is hereby provided.

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*.

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*.

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

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

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. 2023. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.

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.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2020. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2.

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

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Yunshui Li, Binyuan Hui, Xiaobo Xia, Jiaxi Yang, Min Yang, Lei Zhang, Shuzheng Si, Junhao Liu, Tongliang Liu, Fei Huang, et al. 2023. One shot learning as instruction data prospector for large language models. *arXiv preprint arXiv:2312.10302*.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations*.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

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*.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2022. Is reinforcement learning (not) for natural language processing?: Benchmarks, baselines, and building blocks for natural language policy optimization. *arXiv preprint arXiv:2210.01241*.

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

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.

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

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

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. Rrhf: Rank responses to align language models with human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

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

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

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 DEFT Experiment Details

Parameter	DEFT-PRO	DEFT-DPO
Epoch	2	1
SFT weight	5e-2	5e-2
Learning rate	5e-6	5e-7
Gradient accumulation	1	1
Input length	512	512
Inference length	128	128
Batch size / GPU	1	1
ω	2.5e-5	2.5e-6
β	-	0.1