Preference-Oriented Supervised Fine-Tuning: Favoring Target Model over Aligned Large Language Models

Yuchen Fan, Yuzhong Hong, Qiushi Wang, Junwei Bao*, Hongfei Jiang, Yang Song

Zuoyebang Education Technology (Beijing) Co., Ltd
{fanyuchen02, hongyuzhong, wangqiushi02, jianghongfei, songyang}@zuoyebang.com
baojunwei001@gmail.com

Abstract

Alignment, endowing a pre-trained Large language model (LLM) with the ability to follow instructions, is crucial for its real-world applications. Conventional supervised fine-tuning (SFT) methods formalize it as causal language modeling typically with a cross-entropy objective, requiring a large amount of high-quality instruction-response pairs. However, the quality of widely used SFT datasets can not be guaranteed due to the high cost and intensive labor for the creation and maintenance in practice. To overcome the limitations associated with the quality of SFT datasets, we introduce a novel **p**reference-**o**riented supervised **f**ine-**t**uning approach, namely PoFT. The intuition is to boost SFT by imposing a particular preference: *favoring the target model over aligned LLMs on the same SFT data.* This preference encourages the target model to predict a higher likelihood than that predicted by the aligned LLMs, incorporating assessment information on data quality (i.e., predicted likelihood by the aligned LLMs) into the training process. Extensive experiments are conducted, and the results validate the effectiveness of the proposed method. PoFT achieves stable and consistent improvements over the SFT baselines across different training datasets and base models. Moreover, we prove that PoFT can be integrated with existing SFT data filtering methods to achieve better performance, and further improved by following preference optimization procedures, such as DPO.

Code — https://github.com/Savannah120/alignment-handbook-PoFT/

1 Introduction

Large language models(LLMs) such as ChatGPT (OpenAI et al. 2024) have exhibited successful and potent applications in comprehending human queries and delivering plausible responses. This ability has proven to be crucial in real-world applications, e.g. AI assistants and recommendation systems. To equip LLMs with this ability, the alignment methods are usually applied to pre-trained language models. Alignment enables pre-trained models to comprehend the context and generate responses suitable to human interactions. Typical alignment methods can be broadly categorized into two types: Supervised Fine-Tuning (SFT) and Preference Alignment (PA).

*Corresponding author

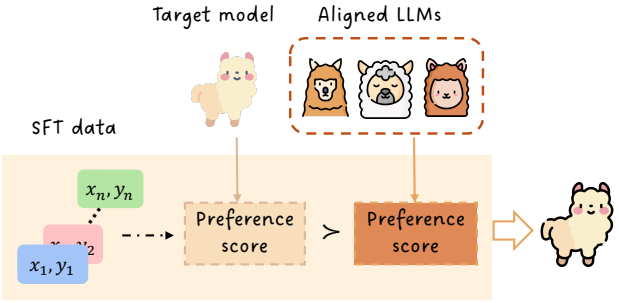

Figure 1: The overall modeling framework of PoFT. By leveraging the Bradley-Terry ranking objective, we impose a particular preference that favors the target model over the aligned LLMs on the same SFT data. Note that the preference score is generated based on the corresponding predicted likelihood.

Supervised fine-tuning (SFT) is an essential phase of alignment, wherein the task is framed as causal language modeling performed on a pre-trained language model with instruction-response data $\mathcal{D} = \{\langle x, y\rangle\}$. Generally, it leverages the cross-entropy objective function in optimization, equipping the pre-trained language model with the ability to follow instructions and generate coherent sequences. Several studies (Schick and Schütze 2021; Houlsby et al. 2019; Ivison et al. 2023) are dedicated to exploring SFT training strategies to enhance the alignment of LLMs. However, due to the intrinsic traits of modeling, the optimization process heavily depends on the availability of high-quality $\langle x, y\rangle$ data, which hinders its performance. Traditionally, the prevalent large-scale SFT datasets in earlier research, such as Alpaca (Taori et al. 2023) and ShareGPT (shareAI 2023), were mainly developed via AI distillation or human-and-AI interaction. Assuring the quality of these datasets can be challenging, as the filtration and curation processes demand significant human resources and efforts.

Instead of solely aligning the instruction and responses, preference alignment (PA), such as InstructGPT (Ouyang et al. 2022) and Direct Preference Optimization (DPO) (Rafailov et al. 2023), optimizes the LLMs based on chosen-rejected data $\langle x, y^+, y^-\rangle$. These PA methods provide exceptional benefits in model alignment, enabling LLMs to

align more accurately with AI/human preferences. In particular, DPO employs the Bradley-Terry (BT) ranking objective (Bradley and Terry 1952) in its optimization process to perform direct preference comparison.

Given the limitations of SFT in processing quality-limited data, we leverage the benefits of the BT preference model and incorporate it into the SFT framework, by proposing a **P**reference-**o**riented supervised **F**ine-**T**uning method, called **PoFT**. Specifically, it applies the BT objective to different models by imposing a particular preference: favoring the target model over the aligned LLMs, given the same $\langle x, y\rangle$ data. Within this framework, the aligned LLMs act as baselines for the target model, prompting it to attain higher preference scores than that of the aligned LLMs on SFT data. Here, we assume these LLMs could discern data that contribute positively to model optimization, thereby providing valid data quality assessments, inspired by (Ngo et al. 2021). Moreover, we would like to emphasize that we use the BT model *to rank models rather than to rank data*. This means we are fundamentally not a PA approach but rather an SFT approach since we require only $\langle x, y\rangle$ and not $\langle x, y^+, y^-\rangle$. For that matter, we show our approach is indeed *orthogonal* to PA since PoFT can be combined with PA methods to further enhance the overall alignment performance (e.g., first PoFT and then DPO).

Despite leveraging the preference modeling with BT, at its essence, PoFT remains faithful to the SFT paradigm, relying on instruction-response data. As an enhanced SFT method, PoFT's objective offers a remarkable advantage over the conventional SFT objective cross-entropy (CE), i.e., PoFT is more stable and robust when training with quality-limited data. Specifically, the introduction of aligned LLMs provides quality assessments on each sample $\langle x, y\rangle$, which decreases its sensitivity towards the data quality. In practice, by analyzing the gradient updates, we observe that PoFT assigns dynamic weights (namely coefficient defined in section 3) to different samples $\{\langle x, y\rangle\}$ by the aligned LLMs. These weights guide parameter optimization, reducing the negative effect of low-quality data. In contrast, the CE objective treats all the data equally, without differentiating data samples based on their quality, thus exposing it to vulnerabilities to low-quality data.

In summary, our contributions are three-fold:

- **Innovative SFT Training Methodology With Preference Modeling.** We present a novel method, called PoFT. This new methodology effortlessly integrates aligned LLMs for preference modeling - a fresh perspective that leads to a boost in the optimization process.
- **Analytical Insight into PoFT's Stability.** Through rigorous mathematical analysis, we provide theoretical explanations that shed light on the inherent characteristics of PoFT in gradient update.
- **Comprehensive Validation of Methodology.** We validate the effectiveness of PoFT through extensive experiments on different base models, demonstrating that PoFT achieves superior performance over the CE objective across diverse training datasets. Our ablation studies indicate PoFT's stability over increasing epochs and enhanced resilience to noise data. Impressively, our experiments prove that the integration of the PoFT and SFT filtering methods can lead to further performance enhancement. Moreover, the two-step training followed by DPO also shows promising alignment performance.

2 Related Work

2.1 Supervised Fine-Tuning

Enabling pre-trained language models to follow human instructions, supervised fine-tuning (SFT) is a way to align LLMs' behavior with human desirability, by training on instruction-response data in a supervised fashion.

Dataset Construction Efforts have been made to construct diverse and complex training data, such as Orca (Mukherjee et al. 2023) and WizardLM (Xu et al. 2023). Wang et al. (2023) proposed a self-improvement pipeline, which enhances LLMs by using its own generations as a bootstrap. Rather than based on human-provided instructions, Li et al. (2024d) reversely constructed instructions from the web corpus via a back-translation model.

Data Filtering In addition to enhancing data complexity, some studies focus on data filtering to improve training efficiency (Chen et al. 2024a; Lu et al. 2024; Liu et al. 2024; Du, Zong, and Zhang 2023). Lu et al. (2024) trained a tagger based on semantics and intentions and regarded the number of tags as a complexity indicator for filtering. IFD, proposed by Li et al. (2024c), is a complexity metric that identifies the discrepancies between responses and the model's generation capability. Liu et al. (2024) trained a scorer via ChatGPT to assess the complexity and quality of the data, thereby selecting "good" data. Li et al. (2024a) and Li et al. (2024b) leveraged a student model to select data for training a teacher model based on the IFD scores.

FT strategies Multiple works have explored efficient fine-tuning strategies to enhance the alignment process (Schick and Schütze 2021; Houlsby et al. 2019; J. et al. 2021; Ivison et al. 2023). Schick and Schütze (2021) converted the provided input into cloze-style statements, thereby facilitating language models to understand the tasks. Ivison et al. (2023) transformed the instructions and examples of a task into parameter-efficient modules through an extra text encoder. Different from these strategies, PoFT proposes a training objective by modeling preference between the target model and aligned LLMs, providing a fresh perspective to enhance the optimization process.

2.2 Preference Alignment

By aligning training objectives with human/AI preferences, RLHF/RLAIF are particularly useful in applications that require nuanced and context-aware decisions (Ouyang et al. 2022; OpenAI et al. 2024; Bai et al. 2022). A prominent preference alignment approach is Direct Preference Optimization (DPO) (Rafailov et al. 2023), which leverages Bradley-Terry (BT) ranking objective (Bradley and Terry 1952) to better prioritize actions based on perceived desirability. In general, the BT model estimates the probability of one item i being chosen over another j in a pairwise comparison, where the items are quantified with strength or quality

parameters, denoted as λ_i and λ_j respectively, resulting in:

$$\mathcal{P}(i \succ j) = \frac{\lambda_i}{\lambda_i + \lambda_j}. \tag{1}$$

As for DPO, it applies the BT objective to express preferences of the policy model for the chosen-rejected pairs $\langle x, y^+, y^- \rangle$ via their expected rewards. Therefore, the preference distribution can be written as:

$$\begin{aligned}\mathcal{P}\left(y^+ \succ y^- \mid x\right) &= \sigma((r\left(x, y^+\right) - \left(r\left(x, y^-\right)\right) \\ &= \frac{\exp\left(r\left(x, y^+\right)\right)}{\exp\left(r\left(x, y^+\right)\right) + \exp\left(r(x, y^-)\right)},\end{aligned} \tag{2}$$

where $r\left(x, y\right)$ is a closed-form reward expression with the optimal policy in DPO's context. Subsequently, more methods are proposed to improve the preference optimization process (Yuan et al. 2023; Dong et al. 2023; Song et al. 2024; Chen et al. 2024b).

3 Methodology

3.1 Preliminary

Typically, the cross-entropy(CE) objective for SFT training only minimizes the difference between predicted and true distributions, represented as

$$L_{\text{CE}} = -\frac{1}{T_0(y)} \log p_\theta(y|x), \tag{3}$$

where $T_0(y)$ refers to the length of y tokenized by the target model θ. Its gradient is shown in Eq. 4.

$$\nabla_\theta L_{\text{CE}} = -\frac{1}{T_0(y)} \frac{1}{p_\theta\left(y \mid x\right)} \nabla p_\theta\left(y \mid x\right) \tag{4}$$

3.2 PoFT: Preference-oriented Supervised Fine-Tuning

In this section, we introduce a novel preference-oriented fine-tuning objective that applies the Bradely-Terry model to perform preference modeling between the target model and aligned LLMs, namely PoFT. Given data $\{x, y\} \sim \mathcal{D}_{SFT}$, it imposes a particular preference by prioritizing the target model over the aligned LLMs. Accordingly, the aligned LLMS acts as a reference point guiding the target model to generate higher preference scores. The preference score is generated based on the predicted likelihood, thus the one from aligned LLMs can be regarded as an indicator for estimating the data quality. Assigning such a preference could diversify the effects of the SFT data, emphasizing more on high-quality data in the optimization process. Note that the preferences are supposed to be generated by some reward model $r^*(x, y)$. Consequently, by applying the BT model, the preference distributions $\mathcal{P}(\cdot)$ can be defined as:

$$\begin{aligned}&\mathcal{P}\left(r^*(x, y) \succ r_{\text{LLMs}}(x, y) \mid x, y\right) \\ &= \frac{\exp\left(r^*\left(x, y\right)\right)}{\exp\left(r^*\left(x, y\right)\right) + \exp\left(r_{\text{LLMs}}(x, y)\right)}, \\ &r_{\text{LLMs}}(x, y) = \mathbb{E}_{\text{LLM} \sim \mathcal{D}_{\text{LLMs}}}\left[r_{\text{LLM}}(x, y)\right].\end{aligned} \tag{5}$$

When accessing a static SFT dataset, a number of aligned LLMs (denoted as $\text{LLM}_j \in \mathcal{D}_{\text{LLM}}$, $|\mathcal{D}| = M$), and a parameterized reward model $r_\theta(x, y)$ for $r^*(x, y)$, the training objective can be transformed into a binary classification problem via maximum likelihood:

$$\begin{aligned}&L_{\text{PoFT}}(\boldsymbol{\theta}) \\ &= -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{SFT}}, r_\theta(x,y) \succ r_{\text{LLMs}}(x,y) \sim \mathcal{P}(\cdot)}\left[\log \mathcal{P}_\theta(\cdot)\right] \\ &\approx -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{SFT}}}\left[\log \frac{\exp\left(r_\theta(x, y)\right)}{\exp\left(r_\theta(x, y)\right) + \exp\left(r_{\text{LLMs}}(x, y)\right)}\right] \\ &= -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{SFT}}}\left[\log \sigma\left(\frac{1}{M}\sum_{j=1}^{M}\left(r_\theta\left(x, y\right) - r_j\left(x, y\right)\right)\right)\right],\end{aligned} \tag{6}$$

where we first impose a particular preference $\mathcal{P} \longrightarrow 1$ (hence the $\approx$) and then parameterize the BT model (i.e., $\mathcal{P}_\theta(\cdot)$) using rewards defined as follows:

$$\begin{aligned}r_\theta(x, y) &= \frac{1}{T_0(y)} \log p_\theta(y \mid x), \\ r_j(x, y) &= \frac{1}{T_j(y)} \log p_j(y \mid x).\end{aligned}$$

In our context, we leverage *the logarithm of predicted likelihood $p(y|x)$ with length normalization* as the reward function to generate *preference scores*. Specifically, the logarithm of the predicted likelihood for the target model $\log p_\theta(y|x)$ and the j-th aligned LLM $\log p_j(y|x)$ are normalized by the corresponding length of the tokenized y, i.e., $T_0(y)$ and $T_j(y)$ respectively. This preference score measures how likely a model would generate the response y when given x at a token level. Applying the length normalization effectively addresses issues related to tokenization mismatches. Moreover, as demonstrated in (Meng, Xia, and Chen 2024), length normalization can also mitigate the impact of sequence length on the reward.

$$\nabla_\theta L_{\text{PoFT}} = -\frac{1}{T_0(y)} \frac{1}{p_\theta\left(y \mid x\right)} \tau \nabla p_\theta\left(y \mid x\right), \tag{7}$$

where

$$\tau = \frac{\left(\prod_{j=1}^{M} p_j(y \mid x)^{\frac{1}{T_j(y)}}\right)^{\frac{1}{M}}}{\left(\prod_{j=1}^{M} p_j(y \mid x)^{\frac{1}{T_j(y)}}\right)^{\frac{1}{M}} + p_\theta(y \mid x)^{\frac{1}{T_0(y)}}}. \tag{8}$$

To delve deeper into the behavior of PoFT during optimization, we examine and present the gradients for CE and PoFT loss, shown in Eq.4 and Eq.7, respectively. By comparison, it can be observed that PoFT's gradient contains an extra coefficient, which is outlined in Eq. 8. This coefficient indicates that the gradient is positively related to $p_j(y|x)$, which indicates the assessment of $\langle x, y \rangle$ from the aligned LLMs. Intuitively, it allows for a more nuanced and dynamic optimization process, accounting for the unbalanced quality of the SFT datasets. Instead of assigning equal weights to all

data, PoFT utilizes the aligned LLMs to direct optimization by diversifying the impacts of different samples on the gradient update. Accordingly, PoFT is proficient in alleviating the influence of lower-quality data, concentrating focus on data with a higher preference score. Thus, PoFT demonstrates its stability for the quality-limited data, compared to the conventional SFT methods.

4 Experiment

In this section, we present the main results of our experiments, highlighting the improvements achieved by PoFT across various datasets. Additionally, our ablation studies offer insights into the following aspects: (1) the effectiveness of PoFT on quality-limited data, (2) the comparison between PoFT and data filtering methods, and (3) the comparison between PoFT and data distillation from aligned LLMs.

4.1 Settings

- **Training Data** To align with Zephyr-7B-sft-full (Tunstall et al. 2023), we opt for UltraChat200k (Ding et al. 2023) as the primary training dataset for PoFT. Besides, we also employ the ShareGPT-Chinese-English-90k dataset (shareAI 2023) and OpenHermes dataset (Teknium 2023), which encompasses 240k data pairs. As ShareGPT comprises parallel bilingual data, we exclusively utilize the English corpus for training purposes. Moreover, to examine PoFT's compatibility with DPO, we introduce the UltraFeedback (Cui et al. 2023) dataset for two-step training.
- **Benchmarks** We evaluate models on the popular benchmark Huggingface Open LLM Leaderboard (Aidar Myrzakhan 2024), MT-Bench (Zheng et al. 2023) and AlpacaEval2.0 (Li et al. 2023). Open LLM Leaderboard covers a variety of tasks, enabling the assessment of specific capabilities of LLMs. Both MT Bench and AlpacaEval 2.0 applied GPT-4 as the judge model to assess the model performance.
- **Model** We choose Mistral-7B-v0.1 (Jiang et al. 2023) and Llama-3-8B (AI@Meta 2024) as backbones. For aligned LLMs, we adopt zephyr-7b-sft-full (Tunstall et al. 2023), Llama-3-8B-Instruct (AI@Meta 2024), and Yi-6B-Chat (AI et al. 2024). Notably, Zephyr-7B-sft-full, derived from Mistral-7B-v0.1, trained on UltraChat200k.

4.2 Main Experiment

We adopt the base models trained on the cross-entropy (CE) objective as our baseline (i.e., SFT model) and investigate the effectiveness of PoFT under the same training settings. The experiments are conducted mainly on UltraChat200k, OpenHermes, and ShareGPT datasets.

Table 1 contains the experimental results of comparison between the models with different training objectives on the LLM Open Leaderboard. To ensure a fair evaluation, we report the results of the last epoch and the average scores across all training epochs after excluding the first epoch, which is typically considered unstable. Notably, as the same base model and datasets are used by Zephyr-7B-sft-full, by adjusting hyper-parameters, it could achieve better performance (see the fourth row of Table 1).

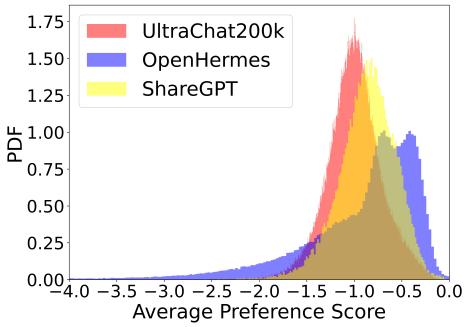

Figure 2: Preference scores generated by aligned LLMs across different training datasets. Note that PDF stands for the probability density function.

Both scores on the Open LLM leaderboard show a consistent trend in which PoFT systematically outperforms the CE objectives across various training datasets on different base models. And the gap is more pronounced concerning OpenHermes, by 1.58 and 2.17 on Mistal-7B and Llama-3-8B, respectively. Moreover, PoFT models have a comparable lower standard deviation than SFT models, indicating greater stability of PoFT across different training epochs. In terms of different evaluation datasets, PoFT models distinctly outperform SFT models on the GSM8k dataset.

The results for MT-Bench and AlpacaEval 2.0 echo the findings from the LLM Open Leaderboard, with remarkable improvements being made in OpenHermes, shown in Table 2. Nonetheless, the overall discrepancy between SFT and PoFT models is fairly minor. We attribute this to the evaluation perspectives of these benchmarks, such as helpfulness in human preference.

Since model performances on different training data are varied, we analyze the average preference scores distribution of aligned LLMs on three training datasets, depicted in Figure 2. It is observed that the distributions of UltraChat200k and ShareGPT are more concentrated, while the distribution of OpenHermes is wider and flatter in shape. This implies the discrepancy of gradients on OpenHermes is more diverse during the training process, thereby amplifying the difference in training performance between CE and PoFT objectives. Hence, we can hypothesize that PoFT is inclined to a certain type of data distribution. In other words, under this data distribution, our model can leverage its strengths more effectively than the SFT model. A comprehensive discussion regarding this observation is covered in section 4.3.

In addition to comparing our approach with SFT methods, we also investigate the compatibility between PoFT and DPO. The results in Table 3 demonstrate that integrating PoFT and DPO can yield superior performance across all benchmark tasks. It is worth noticing that this combined approach brings a significant improvement on the AlpacaEval benchmark, with the win rate surging to 27.83%, underscoring the effectiveness of PoFT-DPO synergy.

Base	FT	Datasets	Arc	Truthful.	Wino.	GSM8k	HellaS.	MMLU	Overall	Avg.	Std.
Zephyr†	-	UltraChat	58.10	40.30	76.90	34.64	80.95	58.92	58.17	-	-
Llama3-8B†	-	-	62.03	51.64	75.30	75.44	78.78	65.75	68.16	-	-
Yi-6B†	-	-	57.51	50.01	71.98	40.63	78.48	63.17	60.30	-	-
Mistal-7B	SFT	UltraChat	63.31	49.13	78.77	42.53	83.79	62.04	63.26	63.34	0.09
Mistal-7B	PoFT		63.40	49.46	78.77	44.88	83.83	62.10	$63.74_{\uparrow 0.48}$	$63.71_{\uparrow 0.38}$	0.04
Llama3-8B	SFT		60.84	54.97	78.30	53.22	81.91	65.03	65.71	65.65	0.06
Llama3-8B	PoFT		60.92	55.09	78.14	54.13	82.03	65.10	$65.90_{\uparrow 0.19}$	$65.88_{\uparrow 0.23}$	0.10
Mistal-7B	SFT	Open-Hermes	62.54	51.45	78.14	37.98	82.18	59.34	61.94	60.86	1.62
Mistal-7B	PoFT		63.57	52.81	77.51	43.82	82.88	60.55	$63.52_{\uparrow 1.58}$	$63.01_{\uparrow 2.15}$	0.50
Llama3-8B	SFT		59.22	56.84	73.28	47.31	80.34	61.38	63.06	64.25	0.81
Llama3-8B	PoFT		61.43	58.38	75.77	50.72	81.63	63.42	$65.23_{\uparrow 2.17}$	$65.36_{\uparrow 1.11}$	0.15
Mistal-7B	SFT	ShareGPT	61.43	52.69	78.85	42.99	83.9	62.18	63.67	63.66	0.10
Mistal-7B	PoFT		61.86	52.74	78.45	45.19	84.02	62.21	$64.08_{\uparrow 0.41}$	$64.00_{\uparrow 0.34}$	0.10
Llama3-8B	SFT		58.28	54.93	77.82	54.59	81.71	65.33	65.44	65.34	0.11
Llama3-8B	PoFT		57.76	55.07	78.30	55.95	81.75	65.15	$65.66_{\uparrow 0.22}$	$65.45_{\uparrow 0.11}$	0.19

Table 1: Overall performance on LLM Open Leaderboard of Mistral-7B and Llama-3-8B training on UltraChat200k, OpenHermes, and ShareGPT. The last three columns present the results of the last epoch and the average scores and standard deviation across all epochs, respectively. We also present the results of the aligned LLMs, where Zephyr†, Llama3-8B†, and Yi-6B† stand for Zephyr-7B-sft-full, Llama-3-8B-Instruct, and Yi-6B-Chat respectively.

FT	Datasets	MT-Bench		AlpacaEval(%)	
		Last	Avg.	Last	Avg
Zephyr	UltraChat	6.30	-	3.91	-
SFT		6.35	6.05	3.98	3.93
PoFT		$6.52_{\uparrow 0.17}$	6.12	$4.1_{\uparrow 0.12}$	4.35
SFT	Open-Hermes	5.09	5.16	4.86	3.99
PoFT		$5.93_{\uparrow 0.84}$	5.89	$5.96_{\uparrow 1.1}$	4.57
SFT	ShareGPT	6.63	6.44	2.44	2.09
PoFT		$6.83_{\uparrow 0.20}$	6.60	$2.61_{\uparrow 0.17}$	2.55

Table 2: Overall performance on MT-Bench and AlpacaEval 2.0 of Mistral-7B training on three datasets. The last two columns of each benchmark present the results of the last epoch and the average scores across all epochs, respectively. Specifically, the score for AlpacaEval 2.0 is the win rate(%).

4.3 Ablation Study

Effectiveness on Quality-limited Data The Cross-Entropy (CE) objective is vulnerable to poor data quality as it does not differentiate between high and low-quality data. In contrast, by integrating aligned LLMs, PoFT can diversify the impacts of data during the optimization process.

However, there is a disparity in the improvements of PoFT for different datasets. By observing the preference score distribution in Figure 2, we assume that this disparity could be attributed to the distribution of training data. Intuitively, when the distribution is highly concentrated, the gap between SFT and PoFT diminishes as the weights for different samples are less diverged. This leads us to our assumption that, upon training on a dataset with a more diverse preference score distribution, a more significant enhancement in PoFT could be observable over the SFT model.

To verify our assumption, we conduct experiments on the datasets Alpaca (Taori et al. 2023) and Dolly (Conover et al. 2023), which are regarded as quality-limited datasets (Li et al. 2024d; Lu et al. 2024). Note that we intentionally increase the number of training epochs to ten for a more nuanced observation of the effects over an extended period. Figure 3b and Figure 3c depict the performance of Mistral-7B models training with these two datasets respectively. During the initial epochs, there is a significant drop in both models. We attribute this to the significant percentage of noise data within the datasets. Nevertheless, PoFT is more robust, proven by the consistent improvement over subsequent epochs. Meanwhile, SFT models are underperformed, indicated by a decreasing trend.

To present the noise data more intuitively, we directly construct hand-crafted noise data to increase the data in the long-tail part of the preference score distribution. Utilizing OpenHermes as our source, we create a pair of inputs with a randomly matched output and simulate data corruption through the processes of character insertion, deletion, and modification, yielding 50k noise data. The newly created noise data is blended with the original data for training. Figure 3d presents the distribution of new training data.

Figure 3e elaborates the performance of Mistral-7B trained on the noise data. For comparison, we also display the performance of models trained on the original data under the same training settings. Overall, the PoFT models consistently surpass the performance of the SFT models regardless of the data settings. It is worth noting that the gap between the PoFT and SFT models is widened when trained with the noise data. As the training epoch increases, there is a remarkable drop in SFT models, particularly for the one with noise data. This indicates that SFT training is more likely

Model	Datasets	LLM Open Leaderboard		MT-Bench		AlpacaEval 2.0	
		Avg.	Std.	Avg.	Std.	Avg.(%)	Std.
Zephyr+DPO	+UltraFeedback	62.62	-	7.11	-	19.01	-
SFT+DPO		65.18	0.32	6.84	0.25	25.09	1.72
PoFT+DPO		65.88 $_{\uparrow 0.70}$	0.12	7.04$_{\uparrow 0.20}$	0.08	27.83$_{\uparrow 2.74}$	3.06

Table 3: Performance of two-step training models based on Mistral-7B. Specifically, the average score for AlpacaEval 2.0 is the average win rate(%). For comparison, we also present the results of Zephyr-7b-beta, denoted as Zephyr+DPO.

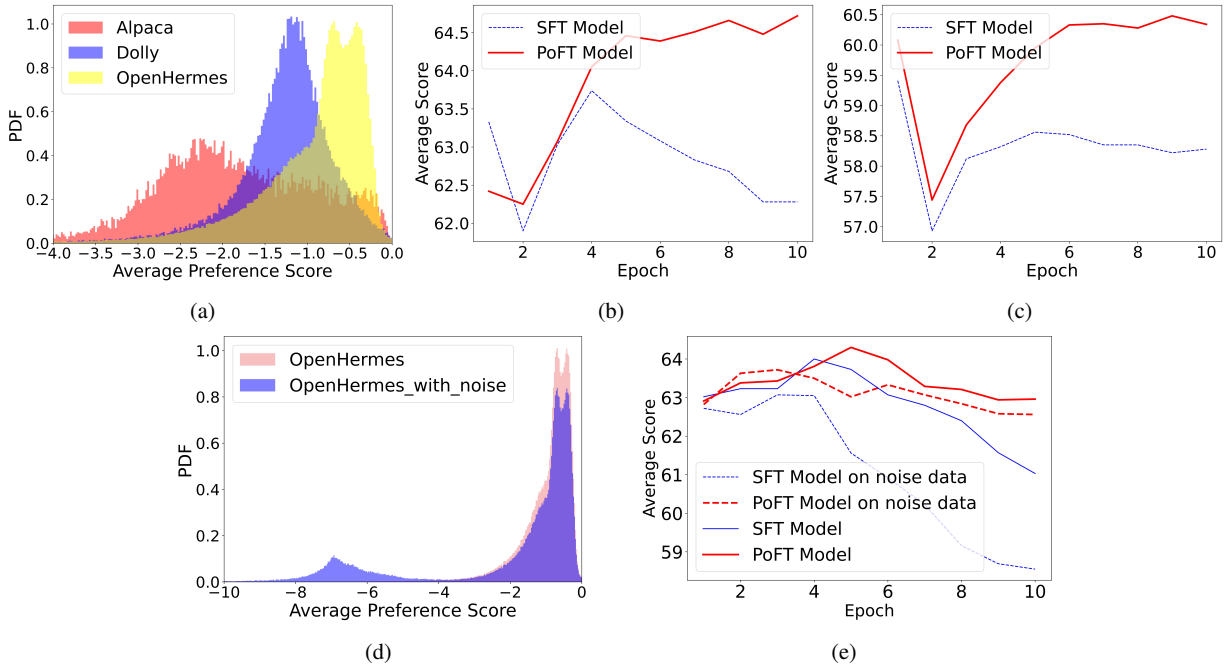

Figure 3: Analysis and model performances on quality-limited data. (a) Preference score distributions of data-limited datasets - Alpaca and Dolly, compared to OpenHermes. Note that PDF stands for the probability density function. (c) Performances of PoFT and SFT models training with Alpaca. (d) Performances of PoFT and SFT models training with Dolly. (d) Preference score distributions of hand-crafted noise data on OpenHermes. The increase in the long-trail part indicates the distribution of the noise data. (e) Performances of PoFT and SFT models training with hand-crafted noise data.

to result in over-fitting, which is exacerbated by the noise in data. In contrast, PoFT shows impressive stability.

The studies above underscore the resilience of PoFT in dealing with various data qualities, which can be attributed to the preference scores from aligned LLMs. These scores help mitigate the negative effects of noisy data, emphasizing the higher-quality data during optimization, leading to a significant improvement.

PoFT v.s. Data Filtering When associating with the reward function, the coefficient in Eq.8 can be interpreted as:

$$\frac{\exp(\frac{1}{M}\sum_{j=1}^{M} r_j(x,y))}{\exp(\frac{1}{M}\sum_{j=1}^{M} r_j(x,y)) + \exp(r_\theta(x,y))}, \quad (9)$$

where $r_\theta(x,y)$ and $r_j(x,y)$ refer to the rewards (i.e., preference scores) of the target model and j-th aligned LLM, respectively, and M is the number of aligned LLMs. Intuitively, the preference scores assigned by aligned LLMs could directly guide the optimization process – higher scores increase gradient update weight. As $r_j(x,y)$ dynamically affects the importance of the samples in training, the PoFT objective can be regarded as a soft filtering approach.

To evaluate this implicit data-filtering mechanism, we apply preference scores to filter data directly. In our experiment, we first sort the data by the scores in descending order. Subsequently, we set thresholds to select varying percentages of data and train PoFT and SFT objectives accordingly.

Figure 4 demonstrates the model performances on the filtered data. In the initial stages, as the number of data increases, there is a positive trend on the SFT model, peaking at 40 percent. This performance even surpasses that of the model trained on the entire dataset. However, despite the continued increase in data volume, the performance begins to decline as more data of inferior quality are included. Interestingly, when trained on filtered data, the PoFT model can further enhance performance.

This steers us toward the hypothesis that combining PoFT and other filtering methods could further enhance overall

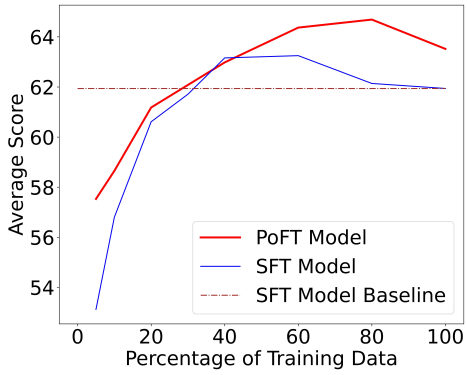

Figure 4: Performance of Mistral-7B trained with different percentages of data on Open LLM Leaderboard.

Filtering method	FT	Overall
N/A	SFT	61.94
	PoFT	63.52
Preference score	SFT	62.14
	PoFT	**64.69**
IFD (Li et al. 2024c)	SFT	64.71
	PoFT	**64.95**
Instag (Lu et al. 2024)	SFT	64.04
	PoFT	**64.27**
Deita (Liu et al. 2024)	SFT	64.31
	PoFT	**64.49**

Table 4: Performance of Mistral-7B models trained with filtered data on Open LLM leaderboard. We present the overall results of the last epoch.

performance. We assume that PoFT, when employed in conjunction with other filtering strategies, can deliver a multi-faceted evaluation of data quality, resulting in a more comprehensive filtering process. Therefore, we conduct experiments on the widely recognized SFT-filtering techniques – IFD (Li et al. 2024c), Instag (Lu et al. 2024), and Deita (Liu et al. 2024). In detail, these filtering methods are applied to the OpenHermes dataset to filter out 20% of data. Subsequently, the Mistral-7B models are trained with CE and PoFT objectives on these data.

The overall performance is illustrated in Table 4. It is indisputable that these filtering methods significantly enhance the performance of SFT, even surpassing PoFT models with full data training. Nonetheless, the utility of these methods is not in contention with our approach. In fact, they can be seamlessly integrated with PoFT, yielding performance superior to applying either method in isolation.

In summary, the experiments confirm: (1) our reward function is effective since using preference scores for filtering allows the model to achieve superior performance on less amount of data; (2) PoFT is compatible with other data filtering methods, further enhancing the overall performance.

PoFT v.s. Data Distillation From Aligned LLMs The commonality between PoFT and data distillation is that they both leverage additional LLMs to provide information for model training. However, PoFT incorporates aligned LLMs to guide the gradient optimization process via preference modeling, while data distillation aims at transferring knowledge from the teacher models, rather than solving problems regarding the data quality.

FT	Regen-Model	Overall
SFT	N/A	61.94
	Llama-3-8B-instruct	63.16
	Zephyr-7B-sft-full	62.30
	Yi-6B-Chat	60.18
PoFT	N/A	**63.52**

Table 5: Performance of Mistral-7B models training with regenerated data on Open LLM Leaderboard. We present the results of the last epoch.

To compare PoFT and data distillation methods, we employ aligned LLMs as teachers to create the responses of OpenHermes, resulting in a new training set. The experiments are conducted on these synthesized data with the CE objective. Intuitively, regenerating the responses is a more explicit way to amplify the effectiveness of aligned LLMs.

The results on the LLM Open Leaderboard are presented in Table 5. Surprisingly, directly replacing the original responses with synthesized data leads to performance degradation. The models trained on the regenerated data underperform PoFT models, performing even worse than the SFT model trained on the original data in some cases. The decrease is more remarkable in the teacher model Yi-6B-Chat.

To sum up, although directly applying aligned LLMs for data regeneration is a more straightforward way for incorporation, it could introduce variability and uncertainty, degrading the model performance. Hence, PoFT offers a more appropriate way of incorporation, efficiently taking advantage of those aligned LLMs through preference modeling.

5 Conclusion

In this paper, we present PoFT, a novel and effective preference-oriented SFT method by applying the Bradley-Terry objective for modeling preferences between different models. Specifically, given the same SFT data, we intentionally define a preference: favoring the target model over aligned LLMs. This preference encourages the target model to generate higher preference scores when compared to the aligned LLMs. In essence, the aligned LLMs provide assessments of the data quality in the optimization process, varying the effects of SFT data. We conduct extensive experiments on diverse training datasets and different base models to verify the efficacy of PoFT compared to the baselines (the CE objective). Furthermore, we prove its stability towards noise data and validate the effectiveness of the designed objectives by conducting ablation studies on the reward functions and aligned LLMs. Furthermore, PoFT can be combined with other SFT Filtering methods to attain enhanced performance outcomes. Notably, integrating PoFT with DPO has the potential to yield even superior performance.

Ethical Statement

This research exclusively employs methods and technologies within the field of Natural Language Processing (NLP). Throughout our experimentation, we strictly adhered to ethical guidelines and rules to ensure that no potential risks or unexpected consequences were caused. The data used in this research does not contain sensitive or offensive content. We aim to contribute positively to the NLP community and advance the technology.

References

AI, .; :; Young, A.; Chen, B.; Li, C.; Huang, C.; Zhang, G.; Zhang, G.; Li, H.; Zhu, J.; Chen, J.; Chang, J.; Yu, K.; Liu, P.; Liu, Q.; Yue, S.; Yang, S.; Yang, S.; Yu, T.; Xie, W.; Huang, W.; Hu, X.; Ren, X.; Niu, X.; Nie, P.; Xu, Y.; Liu, Y.; Wang, Y.; Cai, Y.; Gu, Z.; Liu, Z.; and Dai, Z. 2024. Yi: Open Foundation Models by 01.AI. arXiv:2403.04652.

Aidar Myrzakhan, Z. S., Sondos Mahmoud Bsharat. 2024. Open-LLM-Leaderboard: From Multi-choice to Open-style Questions for LLMs Evaluation, Benchmark, and Arena. *arXiv preprint arXiv:2406.07545*.

AI@Meta. 2024. Llama 3 Model Card.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; Joseph, N.; Kadavath, S.; Kernion, J.; Conerly, T.; El-Showk, S.; Elhage, N.; Hatfield-Dodds, Z.; Hernandez, D.; Hume, T.; Johnston, S.; Kravec, S.; Lovitt, L.; Nanda, N.; Olsson, C.; Amodei, D.; Brown, T.; Clark, J.; McCandlish, S.; Olah, C.; Mann, B.; and Kaplan, J. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862.

Bradley, R. A.; and Terry, M. E. 1952. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39: 324.

Chen, L.; Li, S.; Yan, J.; Wang, H.; Gunaratna, K.; Yadav, V.; Tang, Z.; Srinivasan, V.; Zhou, T.; Huang, H.; and Jin, H. 2024a. AlpaGasus: Training A Better Alpaca with Fewer Data. arXiv:2307.08701.

Chen, Z.; Deng, Y.; Yuan, H.; Ji, K.; and Gu, Q. 2024b. Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models. arXiv:2401.01335.

Conover, M.; Hayes, M.; Mathur, A.; Xie, J.; Wan, J.; Shah, S.; Ghodsi, A.; Wendell, P.; Zaharia, M.; and Xin, R. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm. *Company Blog of Databricks*.

Cui, G.; Yuan, L.; Ding, N.; Yao, G.; Zhu, W.; Ni, Y.; Xie, G.; Liu, Z.; and Sun, M. 2023. UltraFeedback: Boosting Language Models with High-quality Feedback. arXiv:2310.01377.

Ding, N.; Chen, Y.; Xu, B.; Qin, Y.; Zheng, Z.; Hu, S.; Liu, Z.; Sun, M.; and Zhou, B. 2023. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations. arXiv:2305.14233.

Dong, H.; Xiong, W.; Goyal, D.; Zhang, Y.; Chow, W.; Pan, R.; Diao, S.; Zhang, J.; Shum, K.; and Zhang, T. 2023. RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment. arXiv:2304.06767.

Du, Q.; Zong, C.; and Zhang, J. 2023. MoDS: Model-oriented Data Selection for Instruction Tuning. arXiv:2311.15653.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-Efficient Transfer Learning for NLP. *International Conference on Machine Learning,International Conference on Machine Learning*.

Ivison, H.; Bhagia, A.; Wang, Y.; Hajishirzi, H.; and Peters, M. 2023. HINT: Hypernetwork Instruction Tuning for Efficient Zero- & Few-Shot Generalisation. arXiv:2212.10315.

J., H.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv: Computation and Language,arXiv: Computation and Language*.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. arXiv:2310.06825.

Li, M.; Chen, L.; Chen, J.; He, S.; Gu, J.; and Zhou, T. 2024a. Selective Reflection-Tuning: Student-Selected Data Recycling for LLM Instruction-Tuning. arXiv:2402.10110.

Li, M.; Zhang, Y.; He, S.; Li, Z.; Zhao, H.; Wang, J.; Cheng, N.; and Zhou, T. 2024b. Superfiltering: Weak-to-Strong Data Filtering for Fast Instruction-Tuning. arXiv:2402.00530.

Li, M.; Zhang, Y.; Li, Z.; Chen, J.; Chen, L.; Cheng, N.; Wang, J.; Zhou, T.; and Xiao, J. 2024c. From Quantity to Quality: Boosting LLM Performance with Self-Guided Data Selection for Instruction Tuning. arXiv:2308.12032.

Li, X.; Yu, P.; Zhou, C.; Schick, T.; Levy, O.; Zettlemoyer, L.; Weston, J.; and Lewis, M. 2024d. Self-Alignment with Instruction Backtranslation. arXiv:2308.06259.

Li, X.; Zhang, T.; Dubois, Y.; Taori, R.; Gulrajani, I.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. AlpacaEval: An Automatic Evaluator of Instruction-following Models. https://github.com/tatsu-lab/alpaca_eval.

Liu, W.; Zeng, W.; He, K.; Jiang, Y.; and He, J. 2024. What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Selection in Instruction Tuning. arXiv:2312.15685.

Lu, K.; Yuan, H.; Yuan, Z.; Lin, R.; Lin, J.; Tan, C.; Zhou, C.; and Zhou, J. 2024. #InsTag: Instruction Tagging for Analyzing Supervised Fine-tuning of Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Meng, Y.; Xia, M.; and Chen, D. 2024. SimPO: Simple Preference Optimization with a Reference-Free Reward. arXiv:2405.14734.

Mukherjee, S.; Mitra, A.; Jawahar, G.; Agarwal, S.; Palangi, H.; and Awadallah, A. 2023. Orca: Progressive Learning from Complex Explanation Traces of GPT-4. arXiv:2306.02707.

Ngo, H.; Raterink, C.; Araújo, J. G. M.; Zhang, I.; Chen, C.; Morisot, A.; and Frosst, N. 2021. Mitigating harm

in language models with conditional-likelihood filtration. arXiv:2108.07790.

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; Avila, R.; Babuschkin, I.; Balaji, S.; Balcom, V.; Baltescu, P.; Bao, H.; Bavarian, M.; Belgum, J.; Bello, I.; Berdine, J.; Bernadett-Shapiro, G.; Berner, C.; Bogdonoff, L.; Boiko, O.; Boyd, M.; Brakman, A.-L.; Brockman, G.; Brooks, T.; Brundage, M.; Button, K.; Cai, T.; Campbell, R.; Cann, A.; Carey, B.; Carlson, C.; Carmichael, R.; Chan, B.; Chang, C.; Chantzis, F.; Chen, D.; Chen, S.; Chen, R.; Chen, J.; Chen, M.; Chess, B.; Cho, C.; Chu, C.; Chung, H. W.; Cummings, D.; and Currier, J. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155.

Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C. D.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv:2305.18290.

Schick, T.; and Schütze, H. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*.

shareAI. 2023. ShareGPT-Chinese-English-90k Bilingual Human-Machine QA Dataset. https://huggingface.co/datasets/shareAI/ShareGPT-Chinese-English-90k.

Song, F.; Yu, B.; Li, M.; Yu, H.; Huang, F.; Li, Y.; and Wang, H. 2024. Preference Ranking Optimization for Human Alignment. arXiv:2306.17492.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

Teknium. 2023. OpenHermes 2.5: An Open Dataset of Synthetic Data for Generalist LLM Assistants.

Tunstall, L.; Beeching, E.; Lambert, N.; Rajani, N.; Rasul, K.; Belkada, Y.; Huang, S.; von Werra, L.; Fourrier, C.; Habib, N.; Sarrazin, N.; Sanseviero, O.; Rush, A. M.; and Wolf, T. 2023. Zephyr: Direct Distillation of LM Alignment. arXiv:2310.16944.

Wang, Y.; Kordi, Y.; Mishra, S.; Liu, A.; Smith, N. A.; Khashabi, D.; and Hajishirzi, H. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. arXiv:2212.10560.

Xu, C.; Sun, Q.; Zheng, K.; Geng, X.; Zhao, P.; Feng, J.; Tao, C.; and Jiang, D. 2023. WizardLM: Empowering Large Language Models to Follow Complex Instructions. arXiv:2304.12244.

Yuan, Z.; Yuan, H.; Tan, C.; Wang, W.; Huang, S.; and Huang, F. 2023. RRHF: Rank Responses to Align Language Models with Human Feedback without tears. arXiv:2304.05302.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.