Pre-DPO: Improving Data Utilization in Direct Preference Optimization Using a Guiding Reference Model

Junshu Pan[1,2,3], **Wei Shen**[4], **Shulin Huang**[1,2], **Qiji Zhou**[2], **Yue Zhang**[2*]
[1]Zhejiang University
[2]School of Engineering, Westlake University
[3]Shanghai Innovation Institute
[4]Independent Researcher
{panjunshu,zhangyue}@westlake.edu.cn

Abstract

Direct Preference Optimization (DPO) simplifies reinforcement learning from human feedback (RLHF) for large language models (LLMs) by directly training on offline preference data to align with human preferences. During DPO training, the reference model serves as a data weight adjuster. However, the common practice of initializing the policy and reference models identically in DPO can lead to inefficient data utilization and impose a performance ceiling. Meanwhile, the absence of a reference model in Simple Preference Optimization (SimPO) reduces training robustness and requires stricter conditions to prevent catastrophic forgetting. In this work, we propose **Pre-DPO**, a simple yet effective DPO-based training paradigm that improves preference optimization by introducing a *guiding reference model*. This reference model provides *foresight* into the desired policy state achievable through the training preference data, serving as a guiding mechanism that adaptively assigns higher weights to samples more suitable for the model and lower weights to those less suitable. Extensive experiments on the AlpacaEval 2 and Arena-Hard v0.1 benchmarks demonstrate that Pre-DPO consistently improves the performance of both DPO and SimPO, without relying on external models or additional data.

Code — https://github.com/DtYXs/Pre-DPO
Extended version — https://arxiv.org/abs/2504.15843

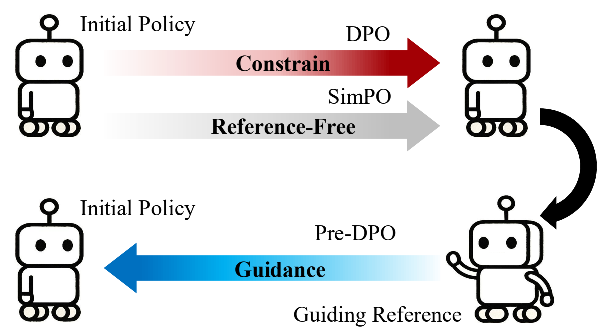

Figure 1: Pre-DPO introduces a guiding reference model derived from the optimized policy to guide re-optimization, transforming the reference from a constraint into an informed guide with foresight.

1 Introduction

Preference-based training has become a widely adopted and effective paradigm for aligning large language models (LLMs) with human values and preferences. Direct Preference Optimization (DPO) (Rafailov et al. 2023), as a representative of reference-based preference optimization methods (Rafailov et al. 2023; Ethayarajh et al. 2024; Azar et al. 2024; Gorbatovski et al. 2025), directly trains LLMs on preference data under the constraint of a reference model, without relying on an explicit reward model or complex online reinforcement learning.

Recently, it has been shown that a reference model is not necessary for achieving effective preference optimization. Simple Preference Optimization (SimPO) (Meng, Xia, and Chen 2024), as a representative of reference-free preference optimization methods (Xu et al. 2023; Xu et al. 2024; Hong, Lee, and Thorne 2024; Meng, Xia, and Chen 2024; Nath et al. 2025), eliminates the need for a reference model and yields better performance and efficiency, though at the cost of an increased risk of catastrophic forgetting (Meng, Xia, and Chen 2024). Meanwhile, other studies have empirically demonstrated that DPO can benefit either from relaxing the constraints imposed by the reference model (Gorbatovski et al. 2025) or from stronger external reference models (Liu, Liu, and Cohan 2024). However, decreasing reliance on a reference model imposes stricter practical requirements to ensure effective learning, and clear methodologies for obtaining an appropriate reference model are still lacking.

Despite the empirical efforts, the role of the reference model in DPO and its impact on training dynamics remain insufficiently explored. During DPO training, the reference model serves as a data weight adjuster (see Section 4.1). It adaptively tends to assign higher weights to data aligned with itself while reducing weights for conflicting data. However, due to the common practice in DPO of initializing the policy and reference models identically (Rafailov et al. 2023), as training progresses, the reference model increasingly constrains the policy model by penalizing deviations, potentially introducing a performance ceiling. Moreover, identical initialization of policy and reference models results

*Corresponding author.

in the nearly uniform weighting of training examples during the early training stages, which is in contrast to prior studies showing that assigning non-uniform weights to training data can lead to improved learning and performance (Lin et al. 2017; Ren et al. 2018; Shu et al. 2019).

In light of the limitations of conventional reference models, we hypothesize that an ideal reference model for DPO should originate from the initial policy model and provide insights into potential directions for policy improvement based on the preference data. We define this type of reference model as a *guiding reference model*, which can better support learning by transforming the role of the reference model from a constraint into a guide with *foresight* (see Section 4.2). Building on this insight, we propose Pre-DPO, a simple yet effective training paradigm that enhances data utilization and improves the performance of existing preference optimization methods without relying on external models or additional data, as shown in Figure 1. Pre-DPO first optimizes the initial policy using a standard preference optimization method (e.g., DPO or SimPO). The resulting optimized policy is then employed as the guiding reference model. Finally, the initial policy is re-optimized using DPO with this guiding reference model, yielding a better-optimized policy through more effective data reweighting. The guiding reference model in Pre-DPO essentially serves as an adaptive guiding mechanism that naturally assigns higher weights to samples more suitable for the model and lower weights to those less suitable. In practice, these suitable cases typically correspond to examples that are easier to learn, allowing the model to efficiently leverage data that aligns well with its learning trajectory.

We evaluate Pre-DPO on the Llama3.2-3B (Grattafiori et al. 2024) and Qwen2.5-7B (Yang et al. 2024) model series across AlpacaEval 2 (Li et al. 2023; Dubois et al. 2024) and Arena-Hard v0.1 (Li et al. 2024) benchmarks. The experimental results show that Pre-DPO consistently improves the performance of both DPO and SimPO, achieving average gains of 2.5 points in length-controlled win rate (LC) on AlpacaEval 2 and 2.6 points in win rate (WR) on Arena-Hard v0.1. By introducing the guiding reference model, Pre-DPO can further improve the performance of existing well-tuned preference optimization methods, effectively overcoming the performance ceiling caused by inefficient data utilization under traditional reference model settings. Notably, Pre-DPO does not rely on external models or additional data, making it highly flexible and easy to deploy.

2 Related Work

Reinforcement learning from human feedback (RLHF) (Christiano et al. 2017; Stiennon et al. 2020) has become an effective approach for aligning LLMs with human values and preferences (Ouyang et al. 2022; Achiam et al. 2023; Grattafiori et al. 2024; Yang et al. 2024; Shen et al. 2025). Typically, pretrained LLMs first undergo supervised fine-tuning (SFT) to learn instruction-following behavior. Subsequently, reinforcement learning is conducted using external reward models and policy optimization algorithms (Schulman et al. 2017; Shao et al. 2024; Hu, Liu, and Shen 2025), which are typically performed online. While RLHF has demonstrated strong performance, its full pipeline remains complex and resource-intensive. DPO (Rafailov et al. 2023) simplifies the RLHF process by directly optimizing on preference data in an offline setting. In this work, we focus on improving data utilization in such offline preference optimization.

Offline preference optimization eliminates the need for an explicit reward model and avoids the complex online learning optimization process. A well-optimized model obtained via offline preference optimization can also serve as a strong initial policy for subsequent online reinforcement learning optimization (Yang et al. 2024). Depending on whether or not a reference model is required, offline preference optimization can be classified into two categories: reference-based preference optimization methods (Rafailov et al. 2023; Ethayarajh et al. 2024; Azar et al. 2024; Kim et al. 2025; Gorbatovski et al. 2025) and reference-free preference optimization methods (Xu et al. 2023; Xu et al. 2024; Hong, Lee, and Thorne 2024; Meng, Xia, and Chen 2024; Nath et al. 2025). DPO (Rafailov et al. 2023), as a representative of reference-based preference optimization methods, directly trains on preference data and implicitly optimizes the same objective as existing reinforcement learning algorithms. SimPO (Meng, Xia, and Chen 2024), as a representative of reference-free preference optimization methods, removes the need for a reference model and can achieve better results than DPO at the cost of lower training robustness (Meng, Xia, and Chen 2024). In this paper, we experimentally demonstrate that DPO can also benefit from reference-free preference optimization methods by leveraging a guiding reference model.

For the reference models in DPO, prior work (Liu, Liu, and Cohan 2024) empirically suggested that DPO can benefit from a stronger and more suitable reference model in certain cases. However, it mainly focuses on stronger external reference models and does not provide a theoretical explanation for why they can be beneficial. Gorbatovski et al. (2025) proposed a dynamic update strategy to reset the reference model based on the current policy, which could weaken the regularization effect of the reference model and tends to assign more identical weights to the data samples with more frequent updates. In this work, we introduce the concept of a guiding reference model, analyze its role in enhancing data utilization in DPO through better data reweighting, and propose an effective methodology for leveraging it in the DPO framework.

3 Preliminaries

Given a text prompt x, the RLHF training stage aims to increase the probability that an LLM generates a response y that is better aligned with human values. Specifically, the objective is to maximize the expected reward $r(x, y)$ while controlling the deviation between the policy probability distribution $\pi_\theta(y \mid x)$ and the reference probability distribution $\pi_{\text{ref}}(y \mid x)$. The optimization objective is formulated as follows:

$$\max_{\pi_\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}\times\pi_\theta} \left[r(x, y)\right] - \beta\mathbb{D}_{\text{KL}} \left[\pi_\theta(y|x) \parallel \pi_{\text{ref}}(y|x)\right], \tag{1}$$

where $\mathcal{D} \times \pi_\theta$ denotes the joint distribution of the prompt x and the response y from $\pi_\theta(y \mid x)$. The KL divergence term constrains the deviation of the policy π_θ from the reference model, and β controls the strength of this constraint.

DPO (Rafailov et al. 2023) is a widely used reference-based preference optimization method that eliminates the need for explicit reward signals and has become a component in the post-training pipeline of many popular open-source LLMs (Bi et al. 2024; Jiang et al. 2024; Yang et al. 2024; Xu et al. 2025). It reformulates Eq. 1 into a direct optimization process on a preference dataset $\mathcal{D} = \{(x_i, y_i^+, y_i^-)\}_{i=1}^{|\mathcal{D}|}$, where x is the prompt, y_i^+ is the preferred response, and y_i^- is the less-preferred response. The objective of DPO is as follows:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\,\mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}}\Big[\log\sigma\Big(\beta\log\frac{\pi_\theta(y^+ \mid x)}{\pi_{\text{ref}}(y^+ \mid x)} - \beta\log\frac{\pi_\theta(y^- \mid x)}{\pi_{\text{ref}}(y^- \mid x)}\Big)\Big] \tag{2}$$

where $\sigma(\cdot)$ is the sigmoid function, and β is a hyperparameter that controls the strength of the reference model's constraint.

A large β imposes a strong constraint, which can limit the model's performance improvement, while a small β may result in insufficient constraints, leading to model degradation (Liu, Liu, and Cohan 2024). In practical LLM training, π_θ and π_{ref} are typically initialized as the supervised fine-tuning (SFT) model. π_{ref} remains fixed during training.

SimPO (Meng, Xia, and Chen 2024) is a reference-free preference optimization method that removes the reference model in DPO while introducing length normalization and target reward margin. Its loss function is defined as:

$$\mathcal{L}_{\text{SimPO}}(\pi_\theta) = -\,\mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}}\Big[\log\sigma\Big(\frac{\beta}{|y^+|}\log\pi_\theta(y^+ \mid x) - \frac{\beta}{|y^-|}\log\pi_\theta(y^- \mid x) - \gamma\Big)\Big] \tag{3}$$

where $|y^+|$ and $|y^-|$ denote the lengths, β is a hyperparameter constant, and γ is the target reward margin. SimPO has the potential to surpass DPO in performance but suffers from reduced robustness due to the lack of reference constraints (Meng, Xia, and Chen 2024).

4 Method

In this section, we first present the motivation behind Pre-DPO, focusing on the limitations of the reference setting in DPO and the specific challenges Pre-DPO addresses. Then, we provide a detailed explanation of Pre-DPO, outlining how it adaptively reweights the training examples and describing its overall process.

4.1 Challenges in Data Reweighting with Vanilla DPO's Reference Model

From the loss function of DPO (Eq. 2), we can derive the gradient with respect to the parameters θ (the detailed derivation is provided in the Appendix of the extended version):

$$\nabla_\theta\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\,\beta\mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}}\left[\lambda\cdot\nabla_\theta\log\frac{\pi_\theta(y^+ \mid x)}{\pi_\theta(y^- \mid x)}\right], \tag{4}$$

where λ is defined as:

$$\lambda = \sigma\left(\beta\log\frac{\pi_{\text{ref}}(y^+|x)}{\pi_{\text{ref}}(y^-|x)} - \beta\log\frac{\pi_\theta(y^+|x)}{\pi_\theta(y^-|x)}\right). \tag{5}$$

From the perspective of example reweighting, DPO learns from preference pairs with weights λ, where the reference model π_{ref} controls the training process by adjusting λ.

When the policy π_θ and the reference π_{ref} are initialized from an identical SFT model, λ starts around the constant 0.5 in the early stage of training due to $\sigma(0) = 0.5$. However, a constant λ can lead to degeneracy (Rafailov et al. 2023), and more importantly, previous research in the field of example reweighting has shown that assigning appropriate and varying weights to training samples can improve model performances and data efficiency (Lin et al. 2017; Ren et al. 2018; Shu et al. 2019).

On the other hand, as training progresses, the reference continuously constrains the policy deviation by adjusting the value of λ. Specifically, when $\frac{\pi_{\text{ref}}(y^+|x)}{\pi_{\text{ref}}(y^-|x)}$ is large, it encourages a larger value of λ, promoting learning from the corresponding preference pair. Conversely, a smaller ratio typically leads to a lower λ, which in turn reduces the influence of that sample on the learning process. The difference in λ between DPO and Pre-DPO is largely driven by the reference model, particularly in the early stage of training. Therefore, a suboptimally configured reference model can result in suboptimal weighting of training samples.

4.2 Pre-DPO: Improving Data Utilization in DPO Using a Guiding Reference Model

One straightforward solution is to employ a reference model that differs from the initial policy and provides foresight into promising directions for policy improvement based on the preference data $\mathcal{D}$, enabling more effective data reweighting and guidance during training.

Notably, a model that has already undergone preference optimization contains information about the entire training process. More importantly, it reflects the outcomes that the initial policy can achieve through the available preference data. Specifically, when the reference model in DPO is set to a guiding reference model π_{guide}, the weight λ becomes:

$$\lambda = \sigma\left(\beta\log\frac{\pi_{\text{guide}}(y^+|x)}{\pi_{\text{guide}}(y^-|x)} - \beta\log\frac{\pi_\theta(y^+|x)}{\pi_\theta(y^-|x)}\right). \tag{6}$$

The foresight of the guiding reference model is reflected in its way to modulate λ: assigning higher weights to samples that the policy model can learn effectively, while down-weighting those that are difficult to learn or potentially conflicting. This behavior naturally aligns with findings suggesting that avoiding ambiguous, mislabeled, or overly difficult preference data can benefit alignment (Houliston et al.

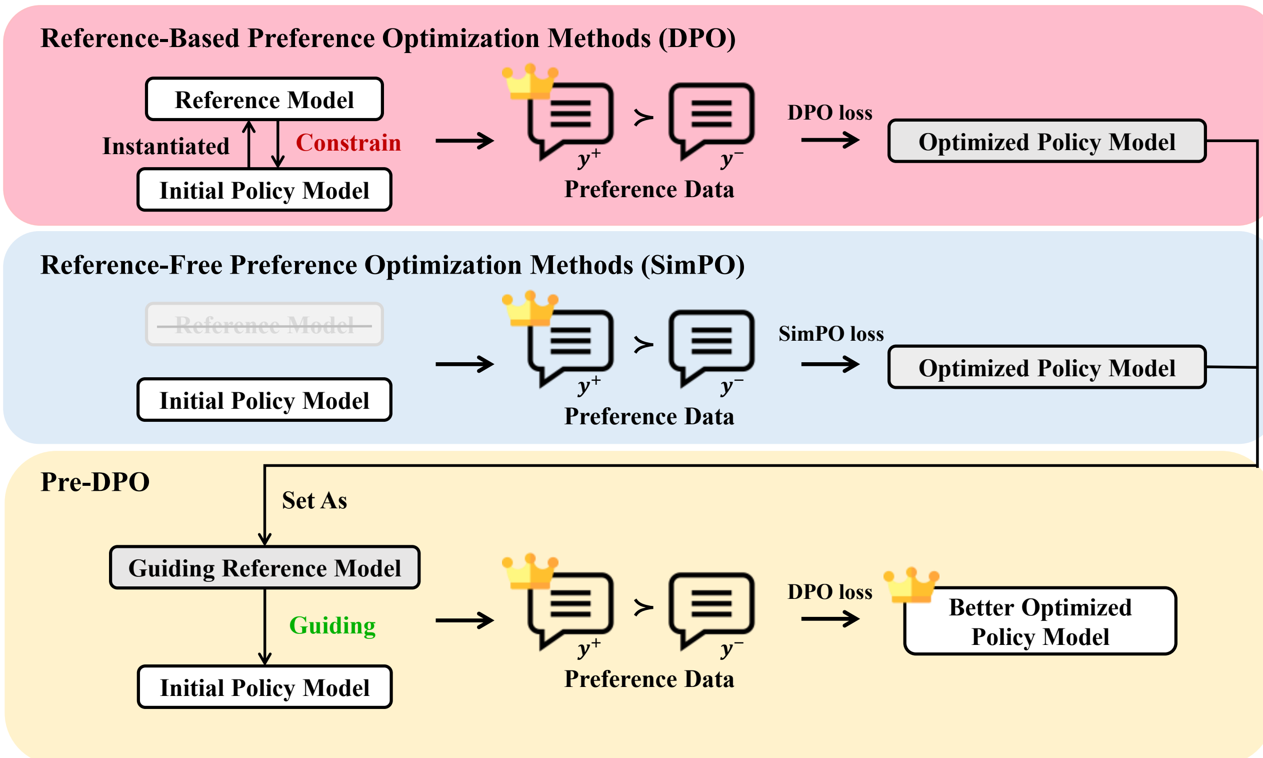

Figure 2: An overview of Pre-DPO. DPO constrains training using the initial policy model as the reference, while SimPO is reference-free. Pre-DPO first optimizes a policy model using DPO or SimPO, then resets it as a guiding reference model, and re-optimizes the initial policy using DPO. This process enhances data utilization and results in a better-optimized policy model.

2024; Gao et al. 2025). Compared to the reference model used in standard DPO, which merely serves as a constraint without foresight, the guiding reference model enables more informed and data-dependent reweighting, leading to more efficient and targeted policy improvement.

Hence, we propose a simple yet effective paradigm for obtaining a suitable reference without needing external models. Specifically, we reset the optimized policy model as the reference for the next training iteration. Since the reference model retains all the information from prior policy training, its role shifts from a constraint to that of a guide, which we refer to as the *guiding reference model*. Employing this guiding reference model in DPO adaptively assigns higher weights to training data that aligns with it while reducing the weights of conflicting samples.

Let π_θ denote the policy model to be optimized, π_{ref} represent the reference model (which can also be set as None), and $\mathcal{M}(\pi_\theta; \pi_{\text{ref}})$ indicate the preference optimization method. The procedure of Pre-DPO (illustrated in Figure 2) is described in detail below:

Step 1: Instantiate the initial reference model. If $\mathcal{M}$ is reference-based, set the reference π_{ref} to π_{SFT}. Otherwise, for reference-free methods, π_{ref} is set to None:

$$\pi_{\text{ref}} = \begin{cases} \pi_{\text{SFT}} & \text{for reference-based } \mathcal{M}, \\ \texttt{None} & \text{for reference-free } \mathcal{M}. \end{cases} \tag{7}$$

Step 2: The first round of preference optimization. Perform preference optimization $\mathcal{M}$ on π_{SFT} with the preference dataset $\mathcal{D}$:

$$\pi_{\mathcal{M}} = \mathcal{M}(\pi_{\text{SFT}}; \pi_{\text{ref}}). \tag{8}$$

Step 3: Set the guiding reference model. After the first round of optimization, reset π_{ref} to the optimized model $\pi_{\mathcal{M}}$ obtained from the previous round. This optimized model now serves as the guiding reference model π_{guide}:

$$\pi_{\text{guide}} = \pi_{\mathcal{M}}. \tag{9}$$

Step 4: Preference optimization with the guiding reference model. Apply DPO to π_{SFT} using the guiding reference π_{guide} on the same dataset $\mathcal{D}$ to obtain the better optimized model $\pi_{\text{Pre-DPO}}$:

$$\pi_{\text{Pre-DPO}} = \mathcal{M}_{\text{DPO}}(\pi_{\text{SFT}}; \pi_{\text{guide}}). \tag{10}$$

5 Experiments

We empirically evaluate the effectiveness of Pre-DPO in enhancing existing preference optimization methods through a guiding reference model. To ensure a comprehensive and fair assessment, we conduct experiments on the Llama3.2-3B (Grattafiori et al. 2024) and Qwen2.5-7B (Yang et al. 2024) model series, including both Base and Instruct versions. We evaluate the models' performance on two widely-used preference optimization benchmarks: AlpacaEval 2 (Li et al. 2023; Dubois et al. 2024) and Arena-Hard v1.0 (Li et al. 2024). Given the sensitivity of preference optimization to hyperparameters (Meng, Xia, and Chen 2024), we conduct an extensive hyperparameter search to ensure reliable results. All experiments are conducted based on the LlamaFactory (Zheng et al. 2024) repository, and all models and datasets used are publicly available.

5.1 Experimental Setup

Models and datasets. In our experiments, we primarily consider two widely recognized series of open-source models, Llama3.2-3B and Qwen2.5-7B, including both Base and Instruct versions. The diversity in model types and scales enables a more comprehensive evaluation of our method's effectiveness.

For Base models, we first train them on the UltraChat-200k (Ding et al. 2023) dataset to obtain their corresponding SFT models. We then directly perform preference optimization on the existing binarized UltraFeedback (Cui et al. 2024) dataset, which is widely used in prior work on offline preference optimization training (Ethayarajh et al. 2024; Hong, Lee, and Thorne 2024; Meng, Xia, and Chen 2024; Liu, Liu, and Cohan 2024; Kim et al. 2025; Gorbatovski et al. 2025), using the SFT model as the initialization.

For Instruct models, we directly use them as SFT models. During the preference optimization stage, we construct two additional on-policy preference datasets for Llama3.2-3B-Instruct and Qwen2.5-7B-Instruct, respectively. Specifically, for each prompt in the binarized UltraFeedback dataset, we sample six responses from each Instruct model using a temperature of 0.8 and a top-p of 0.95 as sampling parameters. Subsequently, following prior work(Meng, Xia, and Chen 2024), we use the ArmoRM-Llama3-8B-v0.1 (Wang et al. 2024) reward model to score each response and select the highest-scoring and lowest-scoring responses to form preference pairs (x, y^+, y^-). We discard prompts for which all sampled responses receive identical scores. This process results in two new preference datasets, one for each of the Llama3.2-3B-Instruct and Qwen2.5-7B-Instruct models.

Baselines. Prior work (Meng, Xia, and Chen 2024) shows that, with sufficient hyperparameter tuning, both DPO (Rafailov et al. 2023) and SimPO (Meng, Xia, and Chen 2024) are highly competitive. Therefore, we adopt them as representative baselines for the reference-based and reference-free methods in our large-scale main experiments. Additionally, we conduct experiments with ORPO (Hong, Lee, and Thorne 2024), sDPO (Kim et al. 2025) and TR-DPO (Gorbatovski et al. 2025) under the Llama3.2-3B-Base setting.

Implementation details. For the SFT stage in Base models, we train for 3 epochs using a batch size of 32, a maximum sequence length of 4096, a learning rate of 2×10^{-6}, and a cosine learning rate schedule with a 6% warmup ratio.

All preference optimization experiments use a batch size of 128, a maximum sequence length of 4096, and a cosine learning rate schedule with a 6% warmup ratio, training for 1 epoch. All models are fine-tuned using full parameter tuning.

Given the importance of hyperparameter tuning in offline preference optimization (Meng, Xia, and Chen 2024), we perform extensive hyperparameter searches for all preference optimization experiments to ensure fairness. Specifically, for the key hyperparameters, including the learning rate, β (for DPO-based methods and SimPO), and γ (for SimPO), a two-stage tuning strategy is employed. We first fix the learning rate and search for the optimal β or γ. Then, with the best β or γ fixed, we search for the optimal learning rate. More details of hyperparameter tuning and the best hyperparameter setting can be found in the Appendix of the extended version.

Evaluation benchmarks. We evaluate methods primarily on two open-source instruction-following benchmarks: AlpacaEval 2 (Li et al. 2023; Dubois et al. 2024) and Arena-Hard v0.1 (Li et al. 2024), which are widely adopted in the community for evaluating the instruction-following capabilities of LLMs (Meng, Xia, and Chen 2024). We report the raw win rate (WR) and length-controlled win rate (LC) on AlpacaEval 2, and the WR on Arena-Hard v0.1, using their respective official repositories. More evaluation details can be found in the Appendix of the extended version .

5.2 Main Results

Pre-DPO further improves DPO and SimPO by leveraging guiding reference models. In Table 1, we report the performance on the AlpacaEval 2 and Arena-Hard v0.1 across the Llama3.2-3B-Base, Llama3.2-3B-Instruct, Qwen2.5-7B-Base, Qwen2.5-7B-Instruct settings. Compared with baselines, Pre-DPO achieves better performance on the AlpacaEval 2 LC and WR benchmarks, yielding average improvements of 2.5 and 2.8 points, respectively. On the Arena-Hard v0.1 benchmark, Pre-DPO also consistently demonstrates improvements across most settings. For instance, on Qwen2.5-7B-Instruct, Pre-DPO achieves an improved WR of 68.8 compared to 62.9 of the DPO baseline. These results indicate that Pre-DPO is effective in further improving both reference-based and reference-free methods by leveraging guiding reference models.

Pre-DPO improves performance without significantly increasing the response length. Although Pre-DPO continuously improves performance, we observe that it does not significantly increase the response length compared to the baselines. Notably, with SimPO as the guiding reference model, Pre-DPO achieves the best performance and the shortest average response length in the Qwen2.5-7B-Base setting.

Pre-DPO is compatible with the iterative preference optimization framework. Note that the DPO and SimPO experiments for Llama3.2-3B-Instruct and Qwen2.5-7B-Instruct use on-policy preference datasets constructed by sampling from the current policy, corresponding to the first round of iterative preference optimization (Xiong et al. 2024; Yuan et al. 2024; Rosset et al. 2024; Zhang et al. 2025). Under this setting, Pre-DPO achieves better optimization performance, indicating that Pre-DPO is comple-

Method	Ref.	Llama3.2-3B-Base				Llama3.2-3B-Instruct			
		AlpacaEval 2			Arena-Hard	AlpacaEval 2			Arena-Hard
		LC (%)	WR (%)	Len.	WR (%)	LC (%)	WR (%)	Len.	WR (%)
SFT	-	6.1	4.0	1012	2.1	19.0	18.9	1956	18.5
DPO	SFT	10.5	12.0	2042	10.6	36.3	36.9	2026	30.6
Pre-DPO	DPO	12.5 (+19.0%)	13.9 (+15.8%)	2061	11.9 (+12.3%)	**39.3** (+8.3%)	**40.9** (+10.8%)	2095	**34.7** (+13.4%)
SimPO	-	13.1	13.1	1950	11.7	33.8	29.9	1797	28.1
Pre-DPO	SimPO	**18.1** (+38.2%)	**18.4** (+40.5%)	2020	**14.0** (+19.7%)	35.0 (+3.6%)	32.3 (+8.0%)	1846	30.0 (+6.8%)

Method	Ref.	Qwen2.5-7B-Base				Qwen2.5-7B-Instruct			
		AlpacaEval 2			Arena-Hard	AlpacaEval 2			Arena-Hard
		LC (%)	WR (%)	Len.	WR (%)	LC (%)	WR (%)	Len.	WR (%)
SFT	-	18.6	6.9	892	9.4	31.2	31.0	2020	55.9
DPO	SFT	24.8	22.2	1778	33.1	52.2	56.8	2270	62.9
Pre-DPO	DPO	27.4 (+10.5%)	24.5 (+10.4%)	1790	32.6 (-1.5%)	53.3 (+2.1%)	**59.4** (+4.6%)	2322	**68.8** (+9.4%)
SimPO	-	34.7	31.9	1836	38.1	51.7	52.4	2119	62.4
Pre-DPO	SimPO	**37.2** (+7.2%)	**32.6** (+2.2%)	1758	**41.6** (+9.2%)	**54.6** (+5.6%)	55.5 (+5.9%)	2121	64.5 (+3.4%)

Table 1: Performance of Pre-DPO under four different model settings on AlpacaEval 2 and Arena-Hard v0.1. LC and WR denote the length-controlled and raw win rate, respectively. Ref. denotes the reference model and Len. denotes the average response length. The SFT models for the Base settings are trained on the UltraChat-200k dataset, while the Instruct models are used as the SFT models directly for the Instruct settings. The guiding reference models are obtained from DPO and SimPO.

	Method	Ref.	Epoch	AlpacaEval 2		
				LC (%)	WR (%)	Len.
Base	DPO	SFT	1	10.5	12.0	2042
	DPO	SFT	2	11.0 (+4.8%)	12.0 (+0.0%)	1976
	Pre-DPO	DPO_1	1	**12.5** (+19.0%)	**13.9** (+15.8%)	2061
Instruct	DPO	SFT	1	36.3	36.9	2026
	DPO	SFT	2	35.2 (-3.0%)	37.1 (+0.5%)	2113
	Pre-DPO	DPO_1	1	**39.3** (+8.3%)	**40.9** (+10.8%)	2095

Table 2: Ablation studies under the Llama3.2-3B model settings. DPO trained for 2 epochs has the same computational cost as Pre-DPO. DPO_1 denote the guiding reference model trained with DPO for 1 epoch.

mentary to the iterative framework and can be employed as part of the iterative process to enhance the use of newly collected preference data.

5.3 Ablations and More Results

The guiding reference model plays a critical role in the improvement of Pre-DPO. Although Pre-DPO consistently improves the performance of DPO and SimPO, it introduces additional computational cost due to the need to obtain a guiding reference model. To investigate whether the performance gain is simply due to increased training, we compare Pre-DPO with a baseline where DPO is trained for 2 epochs using the original reference configuration. As shown in Table 2, DPO with a larger computational budget

Method	Ref.	AlpacaEval 2		
		LC (%)	WR (%)	Len.
ORPO	-	10.2	7.9	1588
Pre-DPO	ORPO	12.3 (+20.6%)	12.1 (+53.2%)	1907
sDPO	last stage	12.0	11.9	1908
Pre-DPO	sDPO	**12.9** (+7.5%)	13.0 (+9.2%)	1951
TR-DPO	hard update	11.7	12.3	1985
Pre-DPO	TR-DPO	12.8 (+9.4%)	**14.2** (+15.4%)	2087

Table 3: More results of Pre-DPO with diverse guiding reference models under Llama3.2-3B-Base setting.

does not yield a noticeable gain, which aligns with previous findings that a single training epoch generally yields the best results (Meng, Xia, and Chen 2024). However, Pre-DPO with a guiding reference model consistently achieves the best LC and WR on AlpacaEval 2, benefiting from the better data utilization enabled by the guiding reference model and avoiding the excessive constraints imposed by traditional reference model setups.

Pre-DPO can consistently benefit from diverse preference optimization methods. To further validate the generality of Pre-DPO, we obtain guiding reference models from more preference optimization methods under the Llama3.2-3B-Base setting, including ORPO, sDPO, and TR-DPO. For sDPO, we divide the preference dataset into two equal parts and perform a two-stage DPO training procedure. For TR-DPO, we adopt a hard update strategy, where the reference

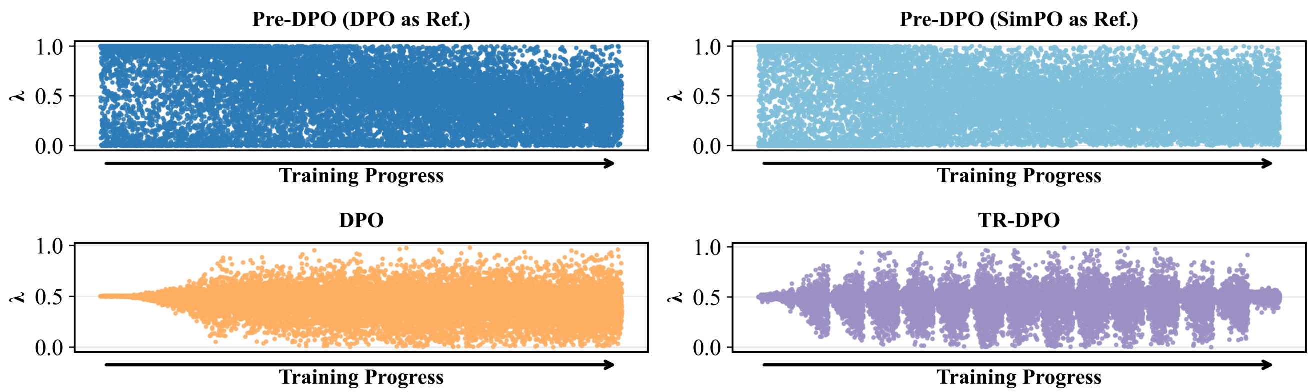

Figure 3: λ distribution dynamics of DPO, TR-DPO, and Pre-DPO under the Llama3.2-3B-Base setting. Pre-DPO maintains a broader distribution during the entire training.

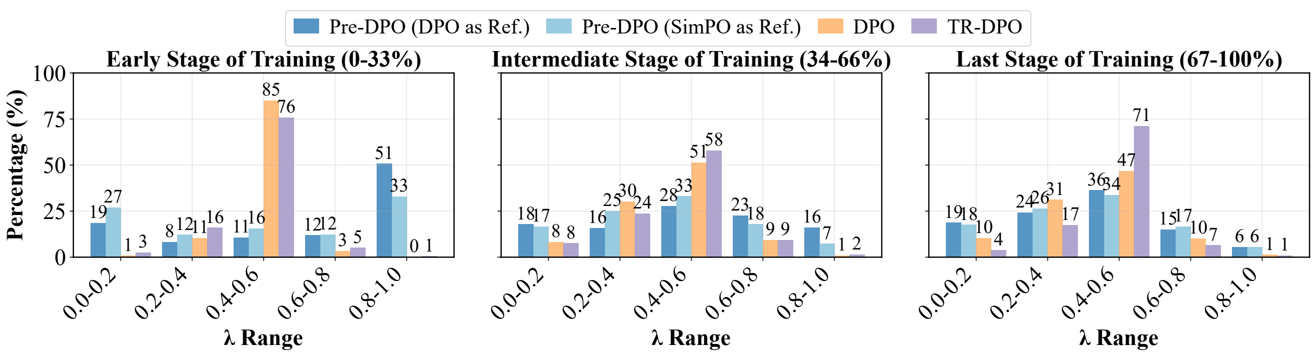

Figure 4: Quantitative analysis of the λ distribution during training for DPO, TR-DPO (hard update), and Pre-DPO under the Llama3.2-3B-Base setting. Numerical values on top of the bars indicate the corresponding percentages.

model is reset to the current policy every 32 training steps. As shown in Table 3, Pre-DPO consistently improves performance on AlpacaEval 2. The results demonstrate that by leveraging guiding reference models, Pre-DPO can consistently benefit from diverse preference optimization methods.

5.4 The λ Distribution during Pre-DPO Training

Pre-DPO demonstrates a clear advantage via adaptive data reweighting, assigning higher weights to samples aligned with the guiding reference model and down-weighting those that may introduce conflicting learning signals.

To better illustrate the training dynamics, we visualize the evolution of the λ distribution throughout training for DPO, TR-DPO, and Pre-DPO under the Llama3.2-3B-Base setting. As intuitively observed in Figure 3, Pre-DPO, equipped with a guiding reference model, maintains a broader λ distribution throughout training. In contrast, DPO initially assigns nearly uniform data weights and gradually adjusts λ to regulate the policy's deviation from the initial policy. TR-DPO with hard reference updates (i.e., resetting the reference model to the current policy every k steps) tends to maintain a more uniform λ distribution throughout training, thereby diminishing the influence of the reference model.

Figure 4 quantitatively demonstrates that in the early stages of training (0–33%), Pre-DPO's λ values are more concentrated at the extremes (i.e., in the ranges 0–0.2 and 0.8–1.0), whereas those of DPO and TR-DPO are more centered around 0.5. As training progresses into the later stages (34–100%), Pre-DPO's λ distribution shifts closer to 0.5. Nevertheless, it consistently maintains a broader and more balanced distribution across the entire training.

6 Conclusion

We proposed Pre-DPO, a simple yet effective DPO-based preference optimization paradigm that enhances data utilization and improves performance by leveraging a guiding reference model. Unlike traditional DPO, which uses a reference identical to the initial policy, Pre-DPO reuses an optimized policy model as the guiding reference model to re-optimize the initial policy model. This shifts the role of the reference model from a constraint to an informed guide, enabling more effective data reweighting. Extensive experiments across multiple models and scales show that Pre-DPO consistently outperforms both DPO and SimPO, without requiring external models or additional data. We hope this work can inspire more exploration and discussion on the role and improvement of reference models in RLHF for LLMs.

Acknowledgments

We would like to sincerely thank all the anonymous reviewers for their valuable feedback. This work is partially funded by the National Natural Science Foundation of China Key Program (Grant No. 62336006).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Azar, M. G.; Guo, Z. D.; Piot, B.; Munos, R.; Rowland, M.; Valko, M.; and Calandriello, D. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, 4447–4455. PMLR.

Bi, X.; Chen, D.; Chen, G.; Chen, S.; Dai, D.; Deng, C.; Ding, H.; Dong, K.; Du, Q.; Fu, Z.; et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.

Cui, G.; Yuan, L.; Ding, N.; Yao, G.; He, B.; Zhu, W.; Ni, Y.; Xie, G.; Xie, R.; Lin, Y.; et al. 2024. ULTRAFEEDBACK: Boosting Language Models with Scaled AI Feedback. In *International Conference on Machine Learning*, 9722–9744. PMLR.

Ding, N.; Chen, Y.; Xu, B.; Qin, Y.; Hu, S.; Liu, Z.; Sun, M.; and Zhou, B. 2023. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 3029–3051.

Dubois, Y.; Galambosi, B.; Liang, P.; and Hashimoto, T. B. 2024. Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators. *arXiv preprint arXiv:2404.04475*.

Ethayarajh, K.; Xu, W.; Muennighoff, N.; Jurafsky, D.; and Kiela, D. 2024. Model Alignment as Prospect Theoretic Optimization. In *International Conference on Machine Learning*, 12634–12651. PMLR.

Gao, C.; Li, H.; Liu, L.; Xie, Z.; Zhao, P.; and Xu, Z. 2025. Principled data selection for alignment: The hidden risks of difficult examples. *arXiv preprint arXiv:2502.09650*.

Gorbatovski, A.; Shaposhnikov, B.; Malakhov, A.; Surnachev, N.; Aksenov, Y.; Maksimov, I.; Balagansky, N.; and Gavrilov, D. 2025. Learn Your Reference Model for Real Good Alignment. In *The Thirteenth International Conference on Learning Representations*.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hong, J.; Lee, N.; and Thorne, J. 2024. ORPO: Monolithic Preference Optimization without Reference Model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 11170–11189.

Houliston, S.; Pace, A.; Immer, A.; and Ratsch, G. 2024. Uncertainty-Penalized Direct Preference Optimization. In *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*.

Hu, J.; Liu, J. K.; and Shen, W. 2025. Reinforce++: An efficient rlhf algorithm with robustness to both prompt and reward models. *arXiv preprint arXiv:2501.03262*.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Hanna, E. B.; Bressand, F.; et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Kim, D.; Kim, Y.; Song, W.; Kim, H.; Kim, Y.; Kim, S.; and Park, C. 2025. sDPO: Don't Use Your Data All at Once. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, 366–373.

Li, T.; Chiang, W.-L.; Frick, E.; Dunlap, L.; Wu, T.; Zhu, B.; Gonzalez, J. E.; and Stoica, I. 2024. From Crowdsourced Data to High-Quality Benchmarks: Arena-Hard and Bench-Builder Pipeline. *arXiv preprint arXiv:2406.11939*.

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.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Liu, Y.; Liu, P.; and Cohan, A. 2024. Understanding reference policies in direct preference optimization. *arXiv preprint arXiv:2407.13709*.

Meng, Y.; Xia, M.; and Chen, D. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37: 124198–124235.

Nath, A.; Volozin, A.; Saha, S.; Nanda, A. A.; Grunin, G.; Bhotika, R.; and Krishnaswamy, N. 2025. DPL: Diverse Preference Learning Without A Reference Model. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3727–3747.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36: 53728–53741.

Ren, M.; Zeng, W.; Yang, B.; and Urtasun, R. 2018. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, 4334–4343. PMLR.

Rosset, C.; Cheng, C.-A.; Mitra, A.; Santacroce, M.; Awadallah, A.; and Xie, T. 2024. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shen, W.; Liu, G.; Wu, Z.; Zhu, R.; Yang, Q.; Xin, C.; Yue, Y.; and Yan, L. 2025. Exploring data scaling trends and effects in reinforcement learning from human feedback. *arXiv preprint arXiv:2503.22230*.

Shu, J.; Xie, Q.; Yi, L.; Zhao, Q.; Zhou, S.; Xu, Z.; and Meng, D. 2019. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in Neural Information Processing Systems*, 32.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33: 3008–3021.

Wang, H.; Xiong, W.; Xie, T.; Zhao, H.; and Zhang, T. 2024. Interpretable Preferences via Multi-Objective Reward Modeling and Mixture-of-Experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 10582–10592.

Xiong, W.; Dong, H.; Ye, C.; Wang, Z.; Zhong, H.; Ji, H.; Jiang, N.; and Zhang, T. 2024. Iterative preference learning from human feedback: bridging theory and practice for RLHF under KL-constraint. In *Proceedings of the 41st International Conference on Machine Learning*, 54715–54754.

Xu, H.; Sharaf, A.; Chen, Y.; Tan, W.; Shen, L.; Van Durme, B.; Murray, K.; and Kim, Y. J. 2024. Contrastive Preference Optimization: Pushing the Boundaries of LLM Performance in Machine Translation. In *International Conference on Machine Learning*, 55204–55224. PMLR.

Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; Zhang, B.; Wang, X.; Chu, Y.; and Lin, J. 2025. Qwen2.5-Omni Technical Report. *arXiv preprint arXiv:2503.20215*.

Xu, J.; Lee, A.; Sukhbaatar, S.; and Weston, J. 2023. Some things are more CRINGE than others: Iterative Preference Optimization with the Pairwise Cringe Loss. *arXiv preprint arXiv:2312.16682*.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Yuan, W.; Pang, R. Y.; Cho, K.; Li, X.; Sukhbaatar, S.; Xu, J.; and Weston, J. 2024. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, 57905–57923.

Zhang, Y.; Yu, D.; Peng, B.; Song, L.; Tian, Y.; Huo, M.; Jiang, N.; Mi, H.; and Yu, D. 2025. Iterative Nash Policy Optimization: Aligning LLMs with General Preferences via No-Regret Learning. In *The Thirteenth International Conference on Learning Representations*.

Zheng, Y.; Zhang, R.; Zhang, J.; YeYanhan, Y.; and Luo, Z. 2024. LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 400–410.