

daDPO: Distribution-Aware DPO for Distilling Conversational Abilities

Zhengze Zhang^{12*} Shiqi Wang^{12*} Yiqun Shen¹² Simin Guo³
Dahua Lin⁴ Xiaoliang Wang¹² Cam Tu Nguyen^{12†} Fei Tan^{5†}

¹State Key Laboratory for Novel Software Technology, Nanjing University

²School of Artificial Intelligence, Nanjing University

³University of Chicago ⁴The Chinese University of Hong Kong

⁵East China Normal University

{zzzhang, wangsky}@smail.nju.edu.cn

ncamtunju.edu.cn

tanfei2007@gmail.com

Abstract

Large language models (LLMs) have demonstrated exceptional performance across various applications, but their conversational abilities decline sharply as model size decreases, presenting a barrier to their deployment in resource-constrained environments. Knowledge distillation with Direct Preference Optimization (dDPO) has emerged as a promising approach to enhancing the conversational abilities of smaller models using a larger teacher model. However, current methods primarily focus on “black-box” KD, which only uses the teacher’s responses, overlooking the output distribution offered by the teacher. This paper addresses this gap by introducing *daDPO* (Distribution-Aware DPO), a unified method for preference optimization and distribution-based distillation. We provide rigorous theoretical analysis and empirical validation, showing that *daDPO* outperforms existing methods in restoring performance for pruned models and enhancing smaller LLM models. Notably, in in-domain evaluation, our method enables a 20% pruned Vicuna1.5-7B to achieve near-teacher performance (-7.3% preference rate compared to that of dDPO’s -31%), and allows Qwen2.5-1.5B to occasionally outperform its 7B teacher model (14.0% win rate).

1 Introduction

Large language models (LLMs) have demonstrated remarkable success across various applications (Achiam et al., 2023; Claude, 2024; Touvron et al., 2023; Yang et al., 2024; Liu et al., 2024a; Gu et al.; Sun et al., 2025a,b). However, a key challenge in their deployment is the sharp decline in conversational ability as model size decreases. As

*Zhang and Wang contributed equally to this work: Zhang leads experiments and refined the methodology, while Wang originates the idea and leads the theoretical proofs. All authors contributed to the writing, discussion, and overall development of the paper.

†Corresponding authors

shown in Figure 1, smaller models experience a sharp performance drop. For example, Qwen2.5-0.5B achieves an AlpacaEval score of only 37.8%, compared to 93.8% for Qwen2.5-7B. Since conversational ability is essential for effective human interaction across tasks and domains—requiring both intent understanding and response generation aligned with human values—this limitation poses a major barrier to the adoption of LLMs in resource-constrained environments.

Knowledge Distillation (KD) with DPO (Direct Preference Optimization) (Tunstall et al., 2023) has recently emerged as a promising approach to enhance the conversational abilities of a student model (small model) using a teacher model (large model). Specifically, responses generated by the teacher can be treated as the preferred (winning) responses, while those from the student serve as the less preferred (losing) responses, forming a preference dataset. The student model is then aligned using DPO (Rafailov et al., 2023) based on the dataset. Prior work (Tunstall et al., 2023; Lee et al., 2024) has demonstrated that DPO-based distillation can significantly improve the conversational performance of weaker models. A key advantage of this approach is the automatic generation of a preference dataset, reducing the reliance on large-scale human-annotated preference data. However, existing studies primarily focus on *black-box* KD, leveraging only the teacher’s responses while overlooking the rich information embedded in the teacher’s probability distribution or internal representations.

The increasing availability of powerful open-source LLMs presents a unique opportunity to explore white-box distillation with broad practical implications. For instance, small models of various sizes can be developed through pruning or quantization of LLM, followed by knowledge distillation (Lee et al., 2024; Muralidharan et al., 2024). By doing so, one can avoid the need to build small models by from scratch, significantly reducing devel-

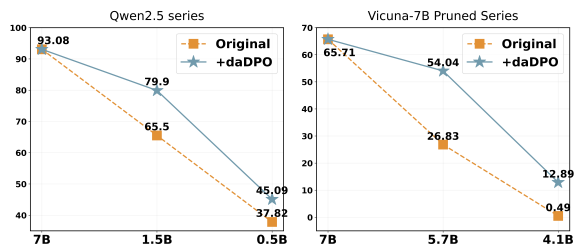


Figure 1: Conversation ability (AlpacaEval scores) drops significantly when the model gets smaller.

opment costs. White-box distillation is potentially useful as shown in knowledge distillation literature for text classification (Sanh, 2019), image classification (Hinton et al., 2015), LLM (Gu et al., 2024), however, there is still a research gap in leveraging white-box distillation to enhance the conversational ability of small language models. Specifically, existing studies primarily focus on either supervised knowledge distillation (Muralidharan et al., 2024) or online reinforcement learning (Agarwal et al., 2024; Gu et al., 2024). The former is suboptimal for aligning conversational ability (Ouyang et al., 2022), whereas the latter—though adaptable for alignment (e.g., by training the reward model)—is more computationally expensive and challenging to optimize compared to DPO (Rafailov et al., 2023).

To address this gap, we focus on the critical research question: “How can we adapt DPO-based distillation to incorporate the teacher’s distribution for enhancing the student’s conversational ability?” Our preliminary results suggest that directly combining the Kullback–Leibler (KL) divergence between the student’s and teacher’s distributions—a common metric in knowledge distillation—into the final DPO loss does not yield optimal results. Instead, we begin by incorporating the KL metric into the RL objective. Our goal is to maximize the reward that favors the teacher’s response over the student’s response, while also regulating the student’s distribution (over its own response) more closely with that of the teacher. We then derive a new DPO-style loss for distribution-aware DPO (termed *daDPO*) from this RL objective. Additionally, we explore further considerations, such as the role of the reference model in DPO, the size of the teacher’s model, and the possibility to distill across LLM families.

Our contributions: 1) **Novel Integration Framework**: *daDPO* offers a unified objective that simultaneously facilitates preference optimization and “white-box” KD; 2) **Theoretical Analysis**: We pro-

vide theoretical analysis and show how we should balance between training stability and the “white-box” KD objective. Finally, 3) **Empirical Validation**: we evaluate the effectiveness of *daDPO* with various scenarios and empirical analysis.

2 Related Work

Language Model Alignment Aligning language models to human preferences is crucial for their conversational ability. Reinforcement Learning with Human Feedback (RLHF) has been introduced as an effective tool for LLM alignment (Ouyang et al., 2022; Stiennon et al., 2020). To mitigate the cost and complexity of RLHF, numerous alternative methods have been proposed, such as SLic-HF (Zhao et al., 2023), DPO (Rafailov et al., 2023), and GRPO (Liu et al., 2024a). Among these, DPO and its variants (Azar et al., 2024; Ethayarajh et al., 2024; Meng et al., 2024; Wang et al., 2024b) have gained prominence due to their strong theoretical foundations and robust performance. Despite these advancements, the conversational ability of smaller models remains limited (see Figure 1). In this work, we focus on distilling the conversational capabilities of a large LM into smaller models and propose an effective distillation method based on DPO.

Language Model Distillation Knowledge Distillation (KD) is a technique for training a smaller model (the student) by leveraging knowledge from a larger model (the teacher). In the era of LLMs, KD is widely used to reduce the computational demands of large models by developing smaller LLMs under the guidance of a larger teacher model.

Recent KD methods for LLMs can be broadly categorized into two types: *black-box KD*, where only the teacher-generated texts are accessible, and *white-box KD*, where the teacher model’s output distribution or intermediate hidden states are also available. Most existing KD approaches for LLMs fall under black-box KD, including (Xu et al., 2023; Taori et al., 2023; Zheng et al., 2023; Tunstall et al., 2023; Lee et al., 2024; Zhang et al., 2024; Ravi et al., 2024). White-box KD methods have been comparably less explored (Gu et al., 2024; Agarwal et al., 2024).

Our work also focuses on the white-box KD setting. However, instead of using online RL like (Gu et al., 2024; Agarwal et al., 2024), we aim to enhance conversational capabilities through a simple yet effective method based on the DPO framework.

3 Preliminaries

This paper investigates the problem of enhancing a **student model** π_{st} under the guidance of a **teacher model** π_{te} . Unlike (Tunstall et al., 2023; Lee et al., 2024), we consider the white-box setting, where the teacher model’s distribution, π_{te} , is fully accessible. Specifically, for any given context x and token t in the vocabulary space, we can directly obtain the probability $\pi_{te}(t|x)$ from the teacher model.

3.1 Supervised Knowledge Distillation

Supervised Finetuning (SFT) Given a set $\mathcal{C}_{sft} = \{(x_1, y_1), \dots, (x_N, y_N)\}$, where (x, y) are respectively the prompt and the ground-truth output, the SFT loss is calculated as follows:

$$\pi_{SFT} = \arg \max_{\pi} \mathbb{E}_{(x, y) \sim \mathcal{C}_{sft}} \log \pi(y|x) \quad (1)$$

Alternatively, one can treat the teacher-generated outputs as ground truth and perform distillation by applying the SFT loss to these outputs. We refer to this as *dSFT* (Distillation by Supervised Fine-Tuning) and denote the resulting model as π_{dSFT} .

Supervised KD Given the white-boxed assumption, supervised KD is widely used to distill information from teacher models (Gu et al., 2024; Muralidharan et al., 2024). Specifically, given a dataset $\mathcal{C}_{skd} = \{(x, y^t)\}$, where y^t is the teacher model’s output for the input x , the following KL-divergence loss is minimized:

$$\begin{aligned} \ell_{KL} &= \mathbb{E}_{(x, y^t) \sim \mathcal{C}_{skd}} \mathcal{D}_{KL}[\pi_{\theta} || \pi_{te}] \quad (2) \\ &= -\mathbb{E}_{x, y^t} \left[\frac{1}{L_y} \sum_{b=1}^{L_y} \pi_{\theta}(\cdot | y_{\leq n}^t, x) \log \frac{\pi_{\theta}(\cdot | y_{\leq n}^t, x)}{\pi_{te}(\cdot | y_{\leq n}^t, x)} \right] \end{aligned}$$

Typically, this token-level KL is used along with the SFT loss or pretraining loss (Gu et al., 2024).

3.2 DPO-based Distillation

DPO-based distillation (dDPO) has recently been introduced to enhance the performance of small LLMs, particularly in terms of conversational ability. It typically involves two stages: dSFT (Distillation by Supervised Fine-Tuning) followed by DPO (Direct Preference Optimization). Specifically, given a set of prompts $\mathcal{X} = [x_1, x_2, \dots, x_N]$, for each $x_i \in \mathcal{X}$, we sample a response $y_i^t \sim \pi_{te}(x_i)$ from the teacher model and another response $y_i^s \sim \pi_{st}(x_i)$ from the student model.

In the dSFT stage, the student model is finetuned on these teacher-generated outputs to obtain π_{dSFT} , as described in Section 3.1. In the DPO stage, we treat y^t as *the winning response* and y^s as *the losing response*, forming a preference dataset $\mathcal{C}_{dpo} = \{(x_1, y_1^t, y_1^s), \dots, (x_N, y_N^t, y_N^s)\}$. The preference for y^t over y^s is justified under the assumption that the teacher model π_{te} represents an optimal policy, as outlined in prior work (Lee et al., 2024). Finally, distillation is conducted by optimizing the DPO loss (Rafailov et al., 2023):

$$\ell_{dDPO} = -\mathbb{E} \left[\log \sigma \left(\beta \frac{\log \pi_{\theta}(y^t|x)}{\log \pi_{ref}(y^t|x)} - \beta \frac{\log \pi_{\theta}(y^s|x)}{\log \pi_{ref}(y^s|x)} \right) \right] \quad (3)$$

The expectation is computed over $(x, y^t, y^s) \sim \mathcal{C}_{dpo}$. Here, the reference policy π_{ref} is set to π_{dSFT} , while π_{θ} represents the student model being optimized. The student model is initialized with π_{dSFT} .

4 Distribution-aware DPO for Distillation

4.1 A Direct Approach: DPO with Reference Replacement

Similar in some previous studies (Li et al., 2024; Liu et al., 2024b), our initial attempt was to directly incorporate the KL-loss from Eq. 2 into the DPO loss from Eq. 3. We refer to this approach as *dDPO+KL*. This method, however, does not preserve the theoretical guarantees offered by *DPO*. As an alternative, we start with the Reinforcement Learning (RL) objective, which, as shown by Rafailov et al. (2023), yields the same optimal solution π^* as optimizing the DPO loss (Eq. 3):

$$\max \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_{\theta}} [r(x, y)] - \beta \mathcal{D}_{KL}[\pi_{\theta} || \pi_{ref}] \quad (4)$$

where $r(x, y)$ represents the reward model that favors the teacher’s response. In the context of DPO, $r(x, y)$ corresponds to the implicit reward model (Rafailov et al., 2023). To integrate the teacher’s distribution, we replace the reference model in Eq. 4 with the teacher model, resulting in the new RL loss as follows:

$$\max \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_{\theta}} [r(x, y)] - \beta \mathcal{D}_{KL}[\pi_{\theta} || \pi_{te}] \quad (5)$$

This loss can be broken down into two parts: the first part optimizes the reward that prioritizes the teacher response over the student response (the

alignment objective), while the second part distills knowledge based on the teacher model’s distribution. Simple algebra demonstrates that this approach effectively replaces the reference model in Eq. 3 with the teacher model’s distribution.

Lemma 1 *The loss in Eq. 5 has the same optimal policy π^* as the optimization of the following loss:*

$$\ell_{rDPO} = -\mathbb{E} \left[\log \sigma \left(\beta \frac{\log \pi_\theta(y^t|x)}{\log \pi_{te}(y^t|x)} - \beta \frac{\log \pi_\theta(y^s|x)}{\log \pi_{te}(y^s|x)} \right) \right] \quad (6)$$

where the expectation is computed over $(x, y^t, y^s) \sim \mathcal{C}_{dpo}$, similar to DPO. This loss (*rDPO*), however, discards the reference policy constraint from the original RL loss. This reference model helps prevent large policy updates and ensure stable learning (Schulman et al., 2017; Rafailov et al., 2023).

4.2 The Joint Force of Two Distributions

The limitation of *rDPO* motivates us to propose Distillation-Aware DPO (*daDPO*), which reintroduces the reference distribution into the *rDPO* equation. Specifically, we seek to optimize the following RL objective.

$$\max_{x, y \sim \pi_\theta} \mathbb{E} [r(x, y)] - \beta_1 D_{KL}[\pi_\theta(y|x) || \pi_{ref}(y|x)] - \beta_2 D_{KL}[\pi_\theta(y|x) || \pi_{te}(y|x)] \quad (7)$$

where π_{te} represents the teacher LLM and π_{ref} for the reference LLM (usually the student LLM).

Theorem 1 *The optimal solution of the objective defined in Eq.7 is:*

$$\pi_\theta^*(y|x) = \frac{1}{Z(x)} \pi_{ref}(y|x)^{\frac{\beta_1}{\beta_1+\beta_2}} \pi_{te}(y|x)^{\frac{\beta_2}{\beta_1+\beta_2}} \times \exp\left(\frac{1}{\beta_1+\beta_2} r(x, y)\right) \quad (8)$$

where $Z(x) = \sum_y \pi_{ref}(y|x)^{\frac{\beta_1}{\beta_1+\beta_2}} \times \pi_{te}(y|x)^{\frac{\beta_2}{\beta_1+\beta_2}} \exp\left(\frac{1}{\beta_1+\beta_2} r(x, y)\right)$ is a constant that depends solely on the input x .

From Theorem 1, we can derive the following relationship between the reward model and the optimal model π^* :

$$r(x, y) = \log \frac{\pi_\theta^*(y|x)^{\beta_1+\beta_2}}{\pi_{ref}(y|x)^{\beta_1} \pi_{te}(y|x)^{\beta_2}} + (\beta_1 + \beta_2) \log Z(x) \quad (9)$$

Since $\log Z(x)$ is independent of y , under the Bradley-Terry model, the optimal policy π^* satisfies the preference model $p(y^t \succ y^s|x)$, which not only depends on the optimal policy π^* , the reference policy π_{ref} like DPO but also the teacher distribution π_{te} . Please refer to the Appendix Eq. 15 for more details. Accordingly, the loss function becomes:

$$-\mathbb{E} \left[\log \sigma \left(\log \frac{\pi_\theta(y^t|x)^{\beta_1+\beta_2}}{\pi_{ref}(y^t|x)^{\beta_1} \pi_{te}(y^t|x)^{\beta_2}} - \log \frac{\pi_\theta(y^s|x)^{\beta_1+\beta_2}}{\pi_{ref}(y^s|x)^{\beta_1} \pi_{te}(y^s|x)^{\beta_2}} \right) \right] \quad (10)$$

The key difference between *daDPO* and the original DPO loss (Eq. 3) is highlighted in blue. Notably, when $\beta_2 = 0$, the loss in Eq. 10 simplifies to the original DPO loss in Eq. 3. Likewise, when $\beta_1 = 0$, it becomes equivalent to the *rDPO* loss in Eq. 6. Our derivation maintains the simplicity and the theoretical analysis of DPO.

Gradient Analysis The gradient of the *daDPO* loss in Eq. 10 can be derived as follows:

$$\begin{aligned} \nabla \ell &= -\nabla \mathbb{E} \left[\log \sigma \left(\log \frac{\pi_\theta(y^t)^{\beta_1+\beta_2}}{\pi_{ref}(y^t)^{\beta_1} \pi_{te}(y^t)^{\beta_2}} - \log \frac{\pi_\theta(y^s)^{\beta_1+\beta_2}}{\pi_{ref}(y^s)^{\beta_1} \pi_{te}(y^s)^{\beta_2}} \right) \right] \\ &= -\nabla \mathbb{E} \left[(\beta_1 + \beta_2)(-\delta_\theta) + \beta_2 \delta_{te} \right] \times \\ &\quad \left[(\beta_1 + \beta_2)(\nabla_\theta \log \pi(y^t) - \nabla_\theta \log \pi(y^s)) \right] \end{aligned}$$

where $\delta_\theta = \log \frac{\pi_\theta(y^t)}{\pi_{ref}(y^t)} - \log \frac{\pi_\theta(y^s)}{\pi_{ref}(y^s)}$ and $\delta_{te} = \log \frac{\pi_{te}(y^t)}{\pi_{ref}(y^t)} - \log \frac{\pi_{te}(y^s)}{\pi_{ref}(y^s)}$.

Similar to the original DPO, the gradient of the loss function increases the likelihood of the chosen response y^t while decreasing that of the rejected response y^s . The gradient is scaled by a coefficient consisting of two terms: $-(\beta_1 + \beta_2)\delta_\theta$ and $\beta_2\delta_{te}$.

The first term, $-(\beta_1 + \beta_2)\delta_\theta$, represents how much the implicit reward model of the currently optimized policy π_θ favors the dispreferred completions, scaled by $\beta_1 + \beta_2$. The second term, $\beta_2\delta_{te}$, captures how much the implicit reward model of the teacher model π_{te} favors the preferred completions. The more strongly the teacher model prefers the chosen response over the rejected one, the larger this coefficient becomes.

Sentence-Level KL vs. Token-Level KL In both *dDPO* and *daDPO*, the KL divergence used in the RL objective (Eq. 4) is computed at the sentence level, as the problem is framed as a sentence-level Markov Decision Process (MDP). In contrast, KL divergence for KD is typically calculated at the token level. Recent studies (Rafailov et al., 2024; Zeng et al., 2024), however, show that token-level DPO produces a loss function equivalent to the original sentence-level formulation. This establishes that the DPO framework remains mathematically consistent whether applied in a token-level or sentence-level MDP setting. Given this theoretical equivalence, we adopt the sentence-level MDP formulation for its mathematical simplicity.

Advantages of daDPO *daDPO* offers several advantages. First, it preserves the stabilizing effect of the reference policy, preventing large policy updates. Second, it leverages the rich distributional information from the teacher model, enhancing distillation. Additionally, since *daDPO* exploits sentence-level KL, it also has the advantage of being applicable across LLM families with different tokenizers, unlike most existing KD methods that rely on token-level KL. Lastly, *daDPO* preserves the simplicity and theoretical guarantees of DPO. Compared to vanilla DPO, it requires only an additional teacher LLM for sampling and computing logits of sentences in the preference dataset. In other words, implementing *daDPO* merely involves adding a few lines of code to the original DPO. Moreover, since the teacher model can be frozen in a gradient-free mode, the additional memory overhead and training time remain manageable.

5 Experiments

5.1 Datasets and Evaluation Metrics

Our training dataset comprises triplets $\{(x, y^t, y^s)\}$, where x serves as the prompt. These prompts are from ShareGPT*, after being filtered and deduplicated. We use only the prompt from the first round of each dialogue, yielding a total of 49,839 prompts. The response y^t is generated by the teacher model π_{te} via greedy decoding, while y^s corresponds to the response from the student model π_{st} . For evaluation, we conduct both in-domain assessments on ShareGPT and out-of-domain evaluations using MT-Bench and AlpacaEval.

*https://huggingface.co/datasets/Aeala/ShareGPT_Vicuna_unfiltered

In-domain Evaluation To evaluate in-domain performance, we randomly sampled 300 unique prompts from the ShareGPT dataset (excluding those used in training) and conducted a comparative evaluation between the teacher and student models using GPT-4 as a judge (i.e., LLM-as-Judge). The model win rate was then calculated (Ji et al., 2024) as:

$$\omega = \frac{N_w - N_l}{N_w + N_l + N_e} \times 100\%$$

where N_w, N_l, N_e are the win/lose/tie count of the student vs the teacher.

MT-Bench (Zheng et al., 2023) is a multi-turn benchmark consisting of 160 questions spanning eight diverse knowledge domains. A model under evaluation needs to respond to an initial query, followed by a second reply to a predefined follow-up question. GPT-4[†] evaluates each response on a scale of 1 to 10, with the final score being the average of the two turns.

AlpacaEval (Li et al., 2023b) is a single-turn benchmark in which the model generates responses to 805 questions covering various topics, with a focus on helpfulness. Following (Tunstall et al., 2023; Li et al., 2023b), the evaluation metrics for AlpacaEval are the mean and standard deviation of the tested model’s win rate against a baseline model, text-davinci-003, as judged by GPT-4.

5.2 Experimental Settings

Teacher/Student Models We conduct experiments in two settings: (1) capability recovery for pruned models with the help of the full-size LLM; and (2) enhancing the conversational ability of smaller models by distillation from larger LLM within an LLM Series.

For pruned models, we use Vicuna1.5-7B (Zheng et al., 2023), an open-source conversational model fine-tuned from LLaMA2 (Touvron et al., 2023) on user-contributed conversations from ShareGPT, as our teacher model. The student model is obtained by applying LLM-Pruner (Ma et al., 2023), a structured pruning method that identifies groups of hidden states that activate together during the forward pass. We prune 50% and 20% of the parameters, yielding the models Vicuna-7B_{0.5} and Vicuna-7B_{0.8}, respectively.

[†]All mentions of GPT-4 in this paper refer to the OpenAI GPT-4o-08-06 model.

Method	MT-Bench	AlpacaEval	Indomain
Vicuna-7B	4.74	65.71% _{1.67}	0.0%
Vicuna-7B _{0.8}	3.06	26.83% _{1.56}	-50.6%
+dSFT	3.68	43.97% _{1.75}	-40.0%
+dDPO	3.78	50.94% _{1.77}	-31.0%
+rDPO	3.82	53.73% _{1.77}	-9.0%
+daDPO	3.84	54.04% _{1.76}	-7.3%
Vicuna-7B _{0.5}	1.31	0.49% _{0.23}	-86.6%
+dSFT	2.00	8.71% _{0.99}	-81.6%
+dDPO	2.24	12.32% _{1.16}	-73.6%
+rDPO	2.24	11.28% _{1.12}	-77.0%
+daDPO	2.33	12.89% _{1.16}	-68.0%

Table 1: Performance of daDPO on pruned Vicuna-7B compared to different baselines without distribution. Here, Vicuna-7B_{0.8} and Vicuna-7B_{0.5} refer to the model with 20% and 50% of the parameters pruned respectively.

For the LLM Series setting, we use the Qwen2.5 Series (Yang et al., 2024) for experiments. Specifically, Qwen2.5-7B-Instruct serves as the teacher LLM, while the 1.5B and 0.5B versions are used as student LLMs for distillation.

Baselines To assess the impact of the teacher’s distribution on the distillation process, we compare our proposed methods, *rDPO* and *daDPO*, against dSFT and dDPO, which do not leverage the teacher’s distribution. Additionally, to evaluate the effectiveness of our approach, we compare *daDPO* with alternative methods for incorporating the teacher’s distribution, including dSFT+KL and dDPO+KL. More details on the baseline methods are provided in Section 3 and Appendix D.

Implementation Details Like dDPO, *daDPO* is applied after dSFT. We implement the *daDPO* algorithm using the OpenRLHF repository[‡], requiring only a few additional lines of code. We tune β_1 in *daDPO* and β in all DPO baselines within the range [0.01, 0.1, 1.0], while β_2 in *daDPO* is tuned from [0.001, 0.01, 0.1, 1.0]. Based on preliminary experiments, we set the batch size to 32 and the learning rate to 5×10^{-7} for all experiments.

5.3 Experimental Results

5.3.1 How does the teacher’s distribution help the student model?

Performance of *daDPO* Tables 1 and 2 demonstrate that *daDPO* consistently outperforms other

Method	MT-Bench	AlpacaEval	Indomain
Qwen2.5-7B	7.42	93.08% _{1.67}	0.0%
Qwen2.5-1.5B	5.28	65.50% _{1.56}	-61.7%
+dSFT	5.36	74.35% _{1.75}	-7.3%
+dDPO	5.97	78.85% _{1.44}	0.7%
+rDPO	5.74	78.94% _{1.52}	-5.0%
+daDPO	6.10	81.49% _{1.36}	14.0%
Qwen2.5-0.5B	3.65	37.82% _{1.71}	-75.0%
+dSFT	3.71	41.15% _{1.73}	-66.3%
+dDPO	3.72	42.27% _{1.74}	-63.3%
+rDPO	3.70	43.41% _{1.74}	-48.7%
+daDPO	3.80	45.09% _{1.75}	-55.3%

Table 2: Performance of daDPO in enhancing smaller LLMs within an LLM Series. In this case, the teacher model is Qwen2.5-7B, while the student models are Qwen2.5-1.5B and Qwen2.5-0.5B, respectively.

baselines in both experimental settings. For Vicuna-7B_{0.8} (20% pruned) in Table 1, *daDPO* shows significant improvements over dDPO on MT-Bench (3.84 vs. baseline dDPO’s 3.78), AlpacaEval (54.04% vs. 50.94%), and in-domain evaluation (-7.3% vs. -31.0%). For the more challenging Vicuna-7B_{0.5} (50% pruned), *daDPO* maintains its advantage, achieving the best results across all three metrics. In Table 2, for Qwen2.5-1.5B, *daDPO* achieves the best performance across datasets, with a notable in-domain win rate of 14.0% (versus dDPO’s 0.7%). This positive win rate is especially significant, as it shows that, without additional labeling cost, the daDPO-enhanced 1.5B model can occasionally outperform its 7B teacher model. This could be due to the rank information itself providing valuable signals that the teacher model had not encountered before. Similar improvements are observed for the smaller Qwen2.5-0.5B model, where *daDPO* achieves an AlpacaEval score of 45.09% (vs. dDPO’s 42.27%) and an in-domain win rate of -53.3% (compared to dDPO’s -63.3%).

Performance of *rDPO* While integrating the teacher’s distribution proves beneficial, the role of the student’s distribution as the reference model is also important. As indicated in Table 1, *rDPO* outperforms dDPO at a 20% pruning rate but suffers from degraded performance at 50% pruning. To further investigate this, we analyzed the in-domain performance on 100 test samples for both dDPO and *rDPO* across varying values of β . The results in Figure 2 show that while replacing the refer-

[‡]<https://github.com/OpenRLHF/OpenRLHF>

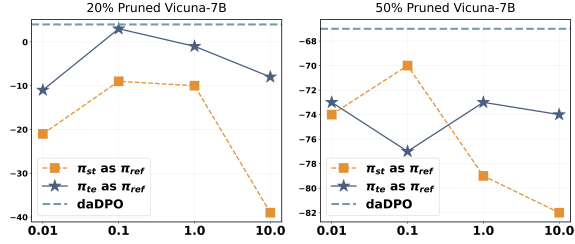


Figure 2: Comparison of in-domain performance of 100 test samples between DPO with different models (teacher v.s. student) as reference policy. Here x is different β values and y is in-domain score

ence policy (i.e., using π_{te} as π_{ref}) can improve performance in some cases, discarding the reference policy constraint leads to instability, especially as the model difference increases (as in the heavily pruned case). This suggests that as the difference between the two models increases, the importance of the student distribution as the reference policy for ensuring stable training becomes more pronounced.

5.3.2 What is the effective way to integrate the teacher’s distribution?

As discussed in Section 4, a straightforward approach to integrating the teacher’s distribution is by introducing KL loss (Eq. 2) into dSFT and dDPO. We investigate whether our approach, *daDPO*, offers a more effective way to incorporate the teacher’s distribution for recovering the capability of pruned models. Figure 3 illustrates the performance gap between the pruned Vicuna-7B models using dDPO, dDPO+KL, dSFT, and dSFT+KL. The results show that adding KL loss to SFT and DPO can improve performance in some cases, but not consistently. For example, with the 20% pruned Vicuna-7B, dSFT+KL results in a lower in-domain score than dSFT alone, and dDPO+KL yields a worse MT-Bench score than dDPO. In contrast, our proposed *daDPO* method, with its integrated approach, achieves the best scores across all metrics.

5.3.3 Does enhancing chat capability make the student forget other abilities?

Alignment tax refers to the phenomenon where enhancing conversational ability can lead to degraded performance on other abilities. This issue has been widely observed with RLHF (Lin et al., 2024) and with distillation for quantized models (Lee et al., 2024). In this study, we examine whether our proposed method exhibits a higher alignment tax compared to dSFT and dDPO when applied to the

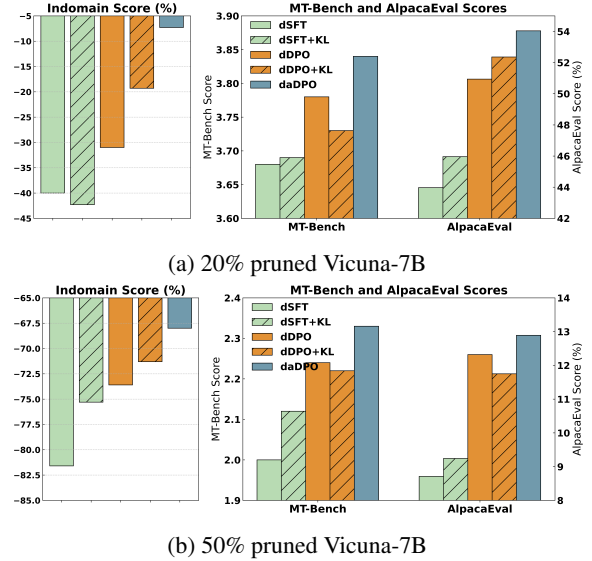


Figure 3: Baselines with distribution information on Vicuna7B-20%pruned and Vicuna7B-50%pruned. It shows that *daDPO* consistently outperforms directly adding distillation KL loss into SFT or DPO loss.

Method	MMLU	ARC_C	GSM8K	Hellaswag
Qwen2.5-7B	74.2	66.9	82.4	81.4
Qwen2.5-1.5B	60.2	54.8	53.1	67.0
+dSFT	60.3	54.1	53.1	67.0
+dDPO	60.4	55.6	52.4	68.1
+daDPO	60.4	56.1	52.4	68.3
Qwen2.5-0.5B	46.7	36.1	31.8	51.0
+dSFT	46.6	35.8	34.2	50.9
+dDPO	46.4	35.4	30.6	50.7
+daDPO	46.8	35.9	30.5	51.1

Table 3: The alignment tax is maintained at a low rate.

Qwen2.5 Series of LLMs.

The results in Table 3 show that improving conversational abilities through *daDPO* does not come at much expense of academic task performance. Specifically, compared to dDPO, *daDPO* preserves the performance of Qwen2.5-1.5B on MMLU and GSM8K while slightly improving the scores on ARC_C (56.1% vs. 55.5% for dDPO) and Hellaswag (68.3% vs. 68.1% for dDPO). Similarly, *daDPO* enhances Qwen2.5-0.5B on three out of four tasks, including MMLU (46.8% vs. 46.4% for dDPO), ARC_C (35.9% vs. 35.4% for dDPO), and Hellaswag (51.1% vs. 50.7% for dDPO). On GSM8K, alignment tax is observed with both dDPO and *daDPO*. However, *daDPO* does not introduce any additional penalty, maintaining performance close to that of dDPO.

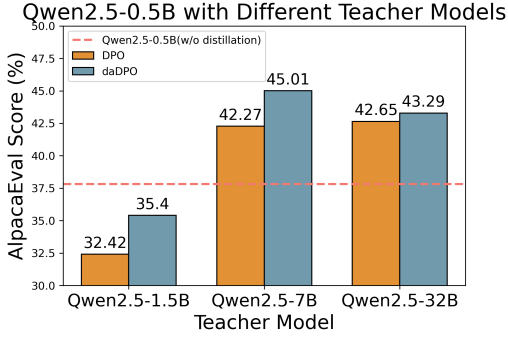


Figure 4: Comparison of AlpacaEval scores for Qwen2.5-0.5B student model using different teacher models (Qwen2.5-1.5B, Qwen2.5-7B, Qwen2.5-32B) with dDPO and *daDPO* methods. Results show that *daDPO* outperforms dDPO, with the highest win rate achieved using Qwen2.5-7B as the teacher model.

Method	MT-Bench	AlpacaEval	Indomain
Qwen2.5-7B	7.42	93.08% _{1.67}	0.0%
LLama3.2-1B	4.37	17.64% _{1.34}	-72.3%
+dSFT	4.68	63.10% _{1.70}	-38.0%
+dDPO	4.71	65.34% _{1.67}	-36.3%
+rDPO	4.69	64.10% _{1.69}	-39.3%
+daDPO	4.99	65.71% _{1.67}	-36.0%

Table 4: Performance of daDPO when the teacher is Qwen2.5-7B, and the student model is LLama3.2-1B.

5.3.4 How does the teacher’s size matter?

We conducted experiments using Qwen2.5-0.5B as the student model and the 1.5B, 7B, and 32B versions as teacher models. The results in Figure 4 show that the 7B version is a more suitable teacher for the 0.5B one in comparison with the 1.5B and the 32B. This highlights the importance of balancing the teacher’s capability with the teacher-student knowledge gap. More specifically, since the 32B and 7B teachers achieve similar AlpacaEval scores (95.46 vs. 93.08), the smaller 7B teacher—exhibiting less distribution shift relative to the 1.5B student—facilitates more effective distillation. In contrast, the 1.5B teacher’s advantage over the 0.5B student (65.50 vs. 37.82) is less significant, indicating a higher noise rate in the preference dataset that may hinder distillation performance. Nevertheless, *daDPO* consistently maintains its advantage over dDPO.

5.3.5 Ablation study on β_1 and β_2

We adjust the values of β_1 and β_2 on the Qwen2.5 Series to observe how they affect the model’s performance. Figure 5 provides a visual comparison

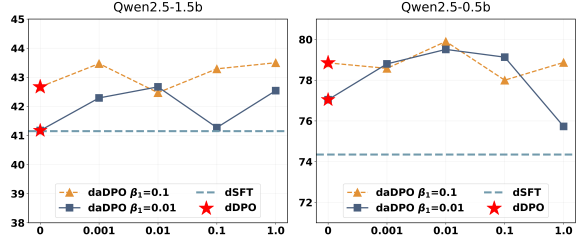


Figure 5: Ablation study on the impact of β_1 (for reference model KL) and β_2 (for teacher model KL) in *daDPO*. Here y-axis shows the AlpacaEval scores and x-axis shows the value of β .

of the AlpacaEval scores for the 0.5B and 1.5B versions of Qwen2.5.

The results from Figure 5 suggests several insights: (1) Both dDPO and *daDPO* outperform dSFT across different values of β_1 and β_2 , demonstrating the effectiveness of preference learning in distillation tasks; (2) In most cases, gradually increasing β_2 up to 1 consistently leads to better performance compared to dDPO (i.e., $\beta_2=0$). However, the ratio between two KL terms should be carefully controlled and not be too large. When β_2 is excessively high, the reference model imposes weaker constraints, leading to less stable training and ultimately affecting performance.

5.3.6 Dose daDPO work with models from different families?

Our previous experiments focused on student-teacher pairs from the same model family. To evaluate the broader applicability of our method, we conducted cross-family distillation experiments between Qwen2.5-7B and Llama3.2-1B. Results in Table 4 demonstrate that *daDPO* outperforms dDPO and dSFT on all three benchmarks. The suboptimal performance of rDPO aligns with previous findings (Liu et al., 2024b), which can be attributable to the increased model distance between student and teacher architectures across families. The results demonstrate that while rDPO becomes ineffective, *daDPO maintains its functionality with models from different model families.*

6 Conclusion

In this work, we developed Distribution Aware DPO (*daDPO*), a novel framework that integrates the rich distributional information of white-box teacher LLMs into the Direct Preference Optimization (DPO) distillation process. Our approach offers several advantages, including more stable

“white-box” distillation, the potential for cross-LLM family distillation, and the ability to maintain DPO’s simplicity and theoretical guarantees. Our experiments on pruned models (Vicuna-7B) and smaller models in the Qwen2.5 Series demonstrated that *daDPO* consistently outperforms existing baselines, achieving significant improvements in conversational performance without compromising downstream task capabilities (compared to dDPO). This work provides a flexible and theoretically grounded approach to knowledge distillation, enabling more efficient deployment of smaller, yet highly capable, language models in resource-constrained environments.

7 Limitation

Despite the effectiveness and novelty of the methodology in this paper, it has several limitations as follows:

1. Even though we have made a step forward compared to the traditional distribution-based knowledge distillation method which requires the teacher model and student model to share the same tokenization vocabulary space, the proposed method still requires the teacher model to be white-box as we need the sentence possibility of it to calculate the loss.
2. The Scope of this paper is mostly limited to general chat ability distillation. In the experiment part, both training datasets and evaluations are limited to general chat ability, whether our method can be generalized in other specific areas like reasoning ability distillation is still unknown and needs further study.

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*.
- Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*.
- Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.
- Claude. 2024. Conversation with claude 3.5 sonnet. Anthropic. Anthropic AI assistant, version 3.5 Sonnet, knowledge cutoff April 2024.
- Daya Guo DeepSeek-AI, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.
- Jiawei Gu, Zacc Yang, Chuanghao Ding, Rui Zhao, and Fei Tan. CMR scaling law: Predicting critical mixture ratios for continual pre-training of language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.
- Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Tianyi Qiu, Juntao Dai, and Yaodong Yang. 2024. Aligner: Efficient alignment by learning to correct. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Janghwan Lee, Seongmin Park, Sukjin Hong, Minsoo Kim, Du-Seong Chang, and Jungwook Choi. 2024. Improving conversational abilities of quantized large language models via direct preference alignment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11346–11364.
- Dedong Li, Rui Zhao, and Fei Tan. 2023a. Cwseg: An efficient and general approach to chinese word segmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 1–10.
- Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023b. AlpacaEval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

- Yixing Li, Yuxian Gu, Li Dong, Dequan Wang, Yu Cheng, and Furu Wei. 2024. Direct preference knowledge distillation for large language models. *arXiv preprint arXiv:2406.19774*.
- Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, et al. 2024. Mitigating the alignment tax of rlhf. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 580–606.
- Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Daya Guo, et al. 2024a. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*.
- Yixin Liu, Pengfei Liu, and Arman Cohan. 2024b. Understanding reference policies in direct preference optimization. *arXiv preprint arXiv:2407.13709*.
- Jinghui Lu, Dongsheng Zhu, Weidong Han, Rui Zhao, Brian Mac Namee, and Fei Tan. 2022. What makes pre-trained language models better zero-shot learners? *Association for Computational Linguistics ACL 2023*.
- Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2023. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36:21702–21720.
- Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*.
- Saurav Muralidharan, Sharath Turuvekere Sreenivas, Raviraj Bhuminand Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. 2024. Compact language models via pruning and knowledge distillation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- 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*.
- Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. 2024. From r to q: Your language model is secretly a q-function. *First Conference on Language Modeling*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.
- Sahithya Ravi, Patrick Huber, Akshat Shrivastava, Aditya Sagar, Ahmed Aly, Vered Shwartz, and Arash Einolghozati. 2024. Small but funny: A feedback-driven approach to humor distillation. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.
- V Sanh. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings of Thirty-third Conference on Neural Information Processing Systems (NIPS2019)*.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- 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*.
- Hai-Long Sun, Zhun Sun, Houwen Peng, and Han-Jia Ye. 2025a. Mitigating visual forgetting via take-along visual conditioning for multi-modal long cot reasoning. In *ACL*.
- Hai-Long Sun, Da-Wei Zhou, Yang Li, Shiyin Lu, Chao Yi, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, et al. 2025b. Parrot: Multilingual visual instruction tuning. In *ICML*.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- 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. *Conference on Language Modeling*.
- Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. 2024a. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints. In *The Twelfth International Conference on Learning Representations*.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. In *the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

- Shiqi Wang, Zhengze Zhang, Rui Zhao, Fei Tan, and Nguyen Cam-Tu. 2024b. Reward difference optimization for sample reweighting in offline RLHF. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.
- Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6268–6278.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.
- Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. 2024. Token-level direct preference optimization. *Proceedings of the 41st International Conference on Machine Learning*.
- Hengyuan Zhang, Yanru Wu, Dawei Li, Sak Yang, Rui Zhao, Yong Jiang, and Fei Tan. 2024. Balancing speciality and versatility: a coarse to fine framework for supervised fine-tuning large language model. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7467–7509.
- Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. 2023. Slic-hf: Sequence likelihood calibration with human feedback. *Advances in Neural Information Processing Systems*.
- 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. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Proof of Theorem 1

Assume that $\beta_1 + \beta_2 > 0, \beta_1 \geq 0, \beta_2 \geq 0$, we have:

$$\begin{aligned}
& \arg_{\theta} \max \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_{\theta}(x)} [r(x, y)] - \beta_1 D_{KL}[\pi_{\theta}(y|x) || \pi_{ref}(y|x)] - \beta_2 D_{KL}[\pi_{\theta}(y|x) || \pi_{te}(y|x)] \\
&= \arg_{\theta} \max \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi_{\theta}(x)} \left[r(x, y) - \beta_1 \log \frac{\pi_{\theta}(y|x)}{\pi_{ref}(y|x)} - \log \beta_2 \frac{\pi_{\theta}(y|x)}{\pi_{te}(y|x)} \right] \\
&= \arg_{\theta} \min \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi_{\theta}(x)} \left[(\beta_1 + \beta_2) \log \frac{\pi_{\theta}(y|x)}{\pi_{ref}(y|x)^{\frac{\beta_1}{\beta_1+\beta_2}} * \pi_{te}(y|x)^{\frac{\beta_2}{\beta_1+\beta_2}}} - r(x, y) \right] \\
&= \arg_{\theta} \min \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi_{\theta}(x)} \left[\log \frac{\pi_{\theta}(y|x)}{\pi_{ref}(y|x)^{\frac{\beta_1}{\beta_1+\beta_2}} * \pi_{te}(y|x)^{\frac{\beta_2}{\beta_1+\beta_2}} \exp(\frac{1}{\beta_1+\beta_2} r(x, y))} \right] \\
&= \arg_{\theta} \min \mathbb{E}_{x \sim \mathcal{D}} \mathbb{E}_{y \sim \pi_{\theta}(x)} \left[\log \frac{\pi_{\theta}(y|x)}{\frac{1}{Z(x)} \pi_{ref}(y|x)^{\frac{\beta_1}{\beta_1+\beta_2}} * \pi_{te}(y|x)^{\frac{\beta_2}{\beta_1+\beta_2}} \exp(\frac{1}{\beta_1+\beta_2} r(x, y))} - \log Z(x) \right]
\end{aligned} \tag{11}$$

where $Z(x) = \sum_y \pi_{ref}(y|x)^{\frac{\beta_1}{\beta_1+\beta_2}} * \pi_{te}(y|x)^{\frac{\beta_2}{\beta_1+\beta_2}} \exp(\frac{1}{\beta_1+\beta_2} r(x, y))$

Note that the partition function is a function of only x and the reference policy π_{ref} and the teacher policy π_{te} , but does not depend on the policy π_{θ} . We can now define

$$\pi^*(y|x) = \frac{1}{Z(x)} \pi_{ref}(y|x)^{\frac{\beta_1}{\beta_1+\beta_2}} * \pi_{te}(y|x)^{\frac{\beta_2}{\beta_1+\beta_2}} \exp(\frac{1}{\beta_1 + \beta_2} r(x, y)) \tag{12}$$

which is a valid distribution as $\pi^*(y|x) > 0$ for all y and $\sum_y \pi^*(y|x) = 1$. For $x \sim \mathcal{D}, y \sim \pi_{\theta}(x)$ We can then re-organize the objective in eq.11 as:

$$\begin{aligned}
& \max_{x,y} \mathbb{E} [r(x, y)] - \beta_1 D_{KL}[\pi_{\theta}(y|x) || \pi_{ref}(y|x)] - \beta_2 D_{KL}[\pi_{\theta}(y|x) || \pi_{te}(y|x)] \\
&= \min \mathbb{E}_{x \sim \mathcal{D}} \left[\mathbb{E}_{y \sim \pi_{\theta}(x)} \left[\log \frac{\pi_{\theta}(y|x)}{\pi^*(y|x)} \right] - \log Z(x) \right] \\
&= \min \mathbb{E}_{x \sim \mathcal{D}} [D_{KL}[\pi_{\theta}(y|x) || \pi^*(y|x)] - \log Z(x)]
\end{aligned} \tag{13}$$

Because $\log Z(x)$ is independent of π_{θ} thus the minimum is only decided by the KL term. According to Gibbs' inequality, the KL-divergence is minimized at 0 if and only if the two distributions are identical. Hence we have proven that the optimal policy of the original objective in eq.7 is

$$\pi^*(y|x) = \frac{1}{Z(x)} \pi_{ref}(y|x)^{\frac{\beta_1}{\beta_1+\beta_2}} \pi_{te}(y|x)^{\frac{\beta_2}{\beta_1+\beta_2}} * \exp(\frac{1}{\beta_1 + \beta_2} r(x, y))$$

According to theorem 1, with simple algebra, we can derive the following relationship between the reward model and π^*

$$r(x, y) = (\beta_1 + \beta_2) \frac{\log(\pi_{\theta}^*(y|x))}{\log \pi_{ref}(y|x)^{\frac{\beta_1}{\beta_1+\beta_2}} \pi_{te}(y|x)^{\frac{\beta_2}{\beta_1+\beta_2}}} + (\beta_1 + \beta_2) \log Z(x) \tag{14}$$

Similar to DPO (Rafailov et al., 2023), because $\log Z(x)$ is not related to y and thus under the Bradley-Terry model, we will get preference probability in terms of only the optimal policy π^* , reference policy π_{ref} and teacher LLM π_{te} .

$$p^*(y_1 \succ y_2) = \frac{1}{1 + \exp\left(\frac{(\beta_1+\beta_2) \log(\pi_{\theta}^*(y_2|x))}{\log \pi_{ref}(y_2|x)^{\beta_1/(\beta_1+\beta_2)} \pi_{te}(y_2|x)^{\beta_2/(\beta_1+\beta_2)}} - \frac{(\beta_1+\beta_2) \log(\pi_{\theta}^*(y_1|x))}{\log \pi_{ref}(y_1|x)^{\beta_1/(\beta_1+\beta_2)} \pi_{te}(y_1|x)^{\beta_2/(\beta_1+\beta_2)}}\right)} \tag{15}$$

B Future Works

Several promising directions for future research emerge from this work:

- 1. Extended Experimentation on Reasoning Tasks.** Recent developments with DeepSeek-R1 (DeepSeek-AI et al., 2025) have demonstrated remarkable capabilities in reasoning and coding tasks, with the model’s open-source nature enabling distillation experiments. Evidence suggests that single dSFT distillation significantly enhances the Qwen model’s reasoning capabilities. Further investigation is warranted to examine whether dDPO and *daDPO* can provide additional improvements in this context.
- 2. Integration with Alternative Preference Learning Frameworks.** While our method has successfully incorporated teacher distribution information into the original DPO framework, additional theoretical analysis could illuminate potential integrations with DPO variants, such as KTO (Ethayarajh et al., 2024) and f-DPO (Wang et al., 2024a).
- 3. Expansion to Diverse Teacher LLMs.** Our current experiments primarily utilize a single-teacher LLM. Future work should explore scenarios involving multiple teacher responses from diverse sources, similar to Zephyr (Tunstall et al., 2023), to better understand the impact of teacher diversity and ensemble effects.

C Algorithm

The algorithm of our proposed *daDPO* is shown in Algorithm 1. This demonstrates that our algorithm is easy to implement and requires only a few additional lines of code compared to DPO. The algorithm contains three steps: Firstly, we leverage datasets with prompts only and then sample teachers’ and students’ responses to them. After that, we first run dSFT on datasets with the teacher’s responses as golden answers. Finally, we run *daDPO* training with our proposed loss in eq.10.

D Details for baselines

We leverage the same decoding strategy (i.e., generation config) and avoid much tuning of it for all methods with greedy decoding (i.e., sampling=False). Training details are as follows:

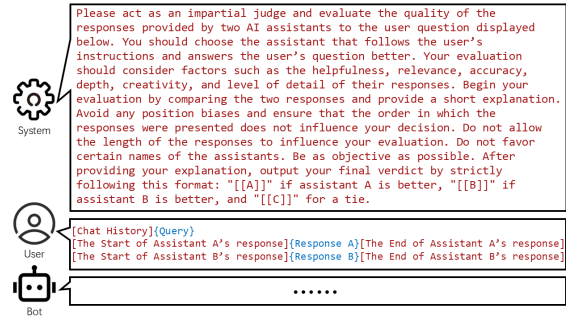


Figure 6: Evaluation prompt for GPT-4o-2024-08-06.

- 1. dSFT (Tunstall et al., 2023):** We use $\{(x, y^t)\}$ for Supervised Fine-Tuning of π_{st} .
- 2. dSFT+KL (Muralidharan et al., 2024; Gu et al., 2024):** We use $\{(x, y^t)\}$ to supervise and fine-tune the student model, and incorporate a token-level KL divergence loss defined in Eq. 2. The weight of KL divergence loss is tuned within the range $[0.1, 0.2, 0.4]$.
- 3. dDPO (Tunstall et al., 2023; Lee et al., 2024; Ravi et al., 2024):** We use $\{(x, y^t, y^s)\}$ to optimize the student model after dSFT by DPO algorithm. We tune β of DPO within the range $[0.01, 0.1, 1.0]$.
- 4. dDPO+KL:** We use $\{(x, y^t, y^s)\}$ to optimize the student model after dSFT, and then add a token-level KL divergence loss defined in Eq. 2 into DPO loss. For hyperparameters, we tune the weight of KL divergence loss within the range $[0.1, 0.2, 0.4]$.

E Supplementary experiments

E.1 Evaluation on More Challenging Benchmarks

Recently, several more challenging benchmarks have been introduced to more accurately assess the capabilities of modern large language models (LLMs), such as Arena-Hard and AlpacaEval 2.0. However, these benchmarks are excessively difficult for weaker models like those in the Vicuna Series. For instance, the Vicuna-7B model achieves only 7.2% on AlpacaEval 2.0[§], and even the larger Vicuna-33B model attains only 8.6% on Arena-Hard[¶].

[§]https://tatsu-lab.github.io/alpaca_eval/

[¶]<https://github.com/lmarena/arena-hard-auto?tab=readme-ov-file#leaderboard>

Algorithm 1 distribution-aware Direct Preference Optimization (daDPO) for distillation

- 1: **Input:** $\mathcal{C}_p = \{x_i\}^N, \pi_{st}, \pi_{te}, \beta_1, \beta_2$ \triangleright Stand for Prompt dataset, student model, teacher model, hyper-parameters respectively
 - 2: **Output:** Optimized policy model π_θ
 - 3: Sample teachers responses $\{y^t\}^N$ from π_{te} on \mathcal{C}_p
 - 4: Sample student responses $\{y^s\}^N$ from π_{st} on \mathcal{C}_p
 - 5: Get the preference dataset $\mathcal{C}_{dpo} = \{(x, y^t, y^s)\}^N$ and sft dataset $\mathcal{C}_{sft} = \{(x, y^t)\}^N$
 - 6: $\pi_{sft} \leftarrow$ Optimize π_{st} on \mathcal{C}_{sft} with sft loss in eq.1
 - 7: $\pi_{ref} \leftarrow \pi_{sft}, \pi_\theta \leftarrow \pi_{sft}$
 - 8: **for** each training batch **do**
 - 9: Sample batch $(x, y^t, y^s) \sim \mathcal{C}_{dpo}$
 - 10: $r^w = \log \pi_\theta(y^t|x) - \log \pi_{ref}(y^t|x)$ \triangleright Log ratio for winning response
 - 11: $r^l = \log \pi_\theta(y^s|x) - \log \pi_{ref}(y^s|x)$ \triangleright Log ratio for losing response
 - 12: $r_t^w = \log \pi_\theta(y^t|x) - \log \pi_{te}(y^t|x)$ \triangleright Log ratio for winning response with teacher model as reference
 - 13: $r_t^l = \log \pi_\theta(y^s|x) - \log \pi_{te}(y^s|x)$ \triangleright Log ratio for losing response with teacher model as reference
 - 14: Compute loss:
 - 15: $\mathcal{L}_{daDPO} = -\mathbb{E}_{(x, y^w, y^l) \sim \mathcal{D}}[\log \sigma(\beta_1(r^w - r^l) + \beta_2(r_t^w - r_t^l))]$
 - 16: Update parameters θ to minimize \mathcal{L}_{daDPO}
 - 17: **end for**
 - 18: **Return** Distilled model π_θ^*
-

Method	MT-Bench	Indomain
Vicuna-7B	4.74	0.0%
Vicuna-7B _{0.8}	3.06	-50.6%
+dSFT	3.68	-40.0%
+dDPO	3.78	-31.0%
+dKTO	3.75	-19.0%
+daDPO	3.84	-7.3%

Table 5: Performance of daDPO on pruned Vicuna-7B compared to different baselines without distribution including RLHF alternatives. Here, Vicuna-7B_{0.8} refer to the model with 20% of the parameters pruned.

In this subsection, we present the performance of our Qwen Series distillation experiments on these challenging benchmarks, where the teacher model, Qwen2.5-7B, demonstrates reasonably strong results. The results are shown in table 6 which makes our work more convincing and sound.

E.2 Comparison with the existing RLHF alternative

We compare with one more offline RLHF method for distillation: dKTO. We use $\{(x, y^t, y^s)\}$ to optimize the student model after dSFT by KTO algorithm. We tune β of KTO (Ethayarajh et al., 2024) within the range of [0.01, 0.1, 0.5, 1.0, 10.0]. Results are shown in table 5.

Method	Arena-Hard	AlpacaEval-2.0
Qwen2.5-7B	54.1%	28.34%
Qwen2.5-1.5B	10.7%	6.44%
+dDPO	21.5%	14.66%
+daDPO	22.4%	16.41%

Table 6: Performance of daDPO on Qwen2.5 Series models on more challenging benchmarks.

F LLM-as-judge evaluation details

We use "GPT-4o-2024-08-06" for In-domain evaluation. Following some previous studies (Zheng et al., 2023; Wang et al., 2023; Li et al., 2023a; Lu et al., 2022), we ask the LLMs to determine which of the two responses under the same query is better. The prompt for evaluation is as Figure 6

F.1 In-domain evaluation results details

This section presents the details of the in-domain evaluation results in section 5, including Vicuna-7B, Qwen2.5 Series, and LLama3.2-1B. The detailed results are shown in Tables 7, 8, and 9.

G Case Study

We show cases in Table 10 and Table 11 in this section to compare responses on the MTbench of

Model	Original			+dSFT			+dDPO			+rDPO			+daDPO		
	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
Vicuna-7B	0	300	0	-	-	-	-	-	-	-	-	-	-	-	-
Vicuna-7B _{0.8}	62	24	214	71	38	191	82	43	175	117	39	144	120	38	142
Vicuna-7B _{0.5}	15	10	275	17	10	262	27	25	248	23	23	254	37	22	241

Table 7: Details of In-domain Evaluation Versus Vicuna-7B in Table 1.

Model	Original			+dSFT			+dDPO			+rDPO			+daDPO		
	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
Qwen2.5-7B	0	300	0	-	-	-	-	-	-	-	-	-	-	-	-
Qwen2.5-1.5B	41	33	226	124	30	146	138	26	136	130	25	145	161	20	119
Qwen2.5-0.5B	29	17	254	43	15	242	47	16	237	68	18	214	75	16	209

Table 8: Details of In-domain Evaluation Versus Qwen2.5-7B in Table 2.

dDPO and *daDPO*. The student model is Vicuna-7B-20% pruned model and the teacher model is Vicuna-7B. These cases show that integrating distribution information into DPO helps the distilled model better understand the instruction and thus output better responses.

Model	Original			+dSFT			+dDPO			+rDPO			+daDPO		
	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
Qwen2.5-7B	0	300	0	-	-	-	-	-	-	-	-	-	-	-	-
LLama3.2-1B	38	7	255	79	28	193	81	29	190	76	30	194	82	28	190

Table 9: Details of In-domain Evaluation Versus Qwen2.5-7B in Table 4.

Round 1 Human: Provide insights into the correlation between economic indicators such as GDP, inflation, and unemployment rates. Explain how fiscal and monetary policies affect those indicators.?

+daDPO Economic indicators such as GDP, inflation, and unemployment rates are closely correlated as they provide insights into the health and performance of an economy. GDP (Gross Domestic Product) is the total value of all goods and services produced within a country over a specific period of time...

+dDPO Economic indicators such as GDP, inflation, and unemployment rates are closely interconnected and can influence each other in a complex manner. Here are some insights into the correlation between economic indicators and their impact on fiscal and monetary policies: 1. GDP: GDP stands for gross domestic product, which is...

Round 2 Human: Now, *explain them again like I'm five*

+daDPO Economic indicators such as GDP, inflation, and unemployment rates tell us how well our economy is doing. **GDP is like a report card** for how much our country is making from producing and selling things. Inflation is like a report card for how much things are getting more expensive or cheaper. Unemployment rate is like a report card for how many people are having trouble finding a job. **Fiscal policies are like the rules our government makes to help or hurt our economy...** Rating: [[8]]

+dDPO Sure! Let's start with the first important factor, GDP (Gross Domestic Product): GDP stands for Gross Domestic Product, which is the total output of a country's economy. It's a measure of the country's economic health and well-being, and it's important because it reflects the production and consumption of goods and services in the economy...(Still too professional for a kid) Rating: [[4]]

Table 10: Case Study on MT-bench: Responses comparison between dDPO and daDPO; Integrating distribution information into DPO helps on instruction following ability.

Round 1 Human: Write a persuasive email to convince your introverted friend, who dislikes public speaking, to volunteer as a guest speaker at a local event. Use compelling arguments and address potential objections. Please be concise.

+daDPO Subject: Convince Your Introverted Friend to Volunteer as a Guest Speaker at a Local Event
\n\nDear [Friend's Name],\n\nI hope this email finds you well. I know how much you dislike public speaking, but I also know how much your presence and insights can make a difference at a local event. That's why I'm reaching out to convince you to volunteer as a guest speaker at this upcoming event. Here are some compelling arguments to consider:...

+dDPO Subject: Encouraging Your Introverted Friend to Be a Speaker at a Local Event
\n\nDear [Friend's Name],\n\nI hope this email finds you well. I wanted to reach out and see if you're open to considering being a speaker at one of our local events. I believe that you have a unique perspective and voice that could add a lot to the conversation, and I think that being a part of this event would be a great way for you to connect with others and share your ideas...

Round 2 Human: Can you rephrase your previous answer and *incorporate a metaphor or simile in each sentence*?

+daDPO Subject: Convince Your Introverted Friend to Volunteer as a Guest Speaker at a Local Event with Metaphor and Simile
\n\nDear [Friend's Name],\n\nI hope this email finds you well. I know how much you dislike public speaking, but **think of it as climbing a mountain** - it can be intimidating, but the view from the top is worth it. By volunteering as a guest speaker at this local event, you have the opportunity to make a positive impact and grow personally, **like polishing a diamond with constant refinement**. Here are some compelling arguments to consider, along with a metaphor and simile:... [Rating: \[\[8\]\]](#)

+dDPO Sure, here's a revised version of the email that incorporates metaphors and similes:\n\nSubject: Encouraging Your Introverted Friend to Be a Speaker at a Local Event\n\nDear [Friend's Name],\n\nI hope this email finds you well. I wanted to reach out and see if you're open to considering being a speaker at one of our local events. I believe that you have a unique perspective and a valuable voice, and I'm excited to see what you can do to contribute to this event.\n\nAs you may know, public speaking can be a daunting task, especially for introverts who may feel more comfortable being a part of a group. However, I want to assure you that there are many ways to overcome this fear and make the most of your public speaking experience.\n\nOne way to prepare for a speech is to practice, practice, and surround yourself with supportive people. It's important to have a plan in place and to be confident in your ability to deliver a great speech... **(No metaphor or simile in this response)** [Rating: \[\[3\]\]](#)

Table 11: Case Study on MT-bench: Responses comparison between dDPO and daDPO; Integrating distribution information into DPO helps on instruction following ability.