UNLOCKING INTRINSIC SELF-REFLECTION FOR LLM PREFERENCE POLICY OPTIMIZATION

Yu Li, Tian Lan, Zhengling Qi
George Washington University
{yul, tlan, qizhengling}@gwu.edu

ABSTRACT

Direct Preference Optimization (DPO) and its variants have become the standard for aligning Large Language Models (LLMs). However, we identify two fundamental limitations. First, the optimized policy lacks invariance since it varies with modeling choices such as scalarization function or reference policy, whereas an optimal policy should remain invariant. Second, most existing methods yield theoretically suboptimal policies by not fully exploiting the comparative information in pairwise preference data, thus missing an opportunity for self-reflection through comparing and contrasting responses. To address both limitations, we propose Intrinsic Self-reflective Preference Optimization (InSPO), which derives a globally optimal policy conditioned on both context and alternative response, explicitly formalizing self-reflection. We prove this formulation surpasses standard DPO and RLHF targets while guaranteeing invariance. InSPO serves as a plug-and-play enhancement for DPO-family algorithms, decoupling alignment from modeling constraints without architectural changes. Using privileged information learning, InSPO requires no alternative response at inference since the self-reflective mechanism is distilled during training, incurring zero overhead. Experiments show InSPO consistently improves win rates and length-controlled metrics across DPO variants, yielding more robust and human-aligned LLMs.

1 INTRODUCTION

Large language models (LLMs) are fine-tuned after pretraining through post-training alignment, combining supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF) (Kumar et al., 2025). SFT fine-tunes the model to generate the correct response given an instruction, typically by minimizing the negative log-likelihood of the target tokens, which effectively imparts formatting conventions, stylistic preferences, and basic task-following behaviors (Dong et al., 2023). RLHF aligns models using pairwise preference data through a two-stage pipeline (Ouyang et al., 2022). First, a reward model is fitted to approximate human preferences via the Bradley-Terry (BT) model, which links preference probability to reward contrast through a logistic function. Second, the LLM is fine-tuned by maximizing the learned reward with KL divergence regularization toward a reference policy, motivated by Proximal Policy Optimization (PPO) (Schulman et al., 2017). While effective, RLHF requires iterative on-policy sampling, coordinated reward training, and careful hyperparameter tuning, complicating practical deployment (Zhong et al., 2024).

Direct Preference Optimization (DPO) simplifies this by eliminating explicit reward modeling and online rollouts. Given a preference pair, DPO associates preference scores with log-likelihood ratios between the trainable and reference models and maximizes margins favoring preferred responses under logistic loss, thus preserving the KL regularization intrinsic while operating offline on the static preference pairs. The success of DPO has inspired variants refining loss curvature (Azar et al., 2024), length normalization (Meng et al., 2024), or reference formulations (Ethayarajh et al., 2024).

Despite impressive performance, existing methods suffer two fundamental limitations. First, the derived optimal policy lacks invariance to arbitrary modeling choices—the scalarization function and reference policy. As we formally demonstrate, the "optimal" target shifts with these design choices, producing behavior reflecting parameterization artifacts rather than genuine human preferences. Second, these methods fail to exploit comparative information in pairwise data, leaving the

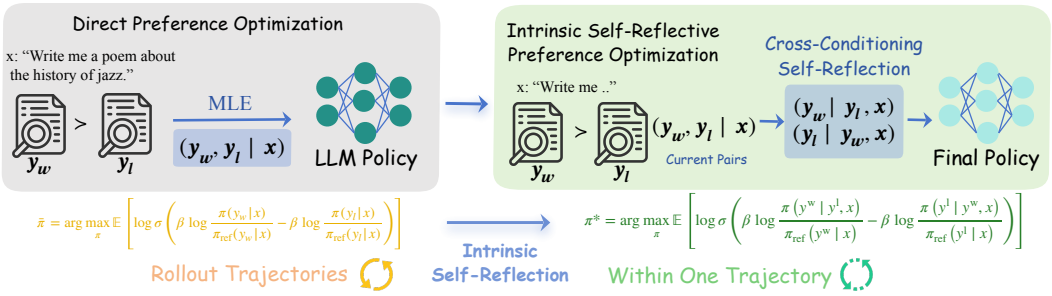


Figure 1: From pairwise preference to our proposed INSPO. Standard DPO (left) learns from response comparisons where both the preferred and dispreferred responses are evaluated based solely on the prompt. INSPO (right) unleashes intrinsic self-reflection through symmetric cross-conditioning: the policy generates the preferred response while seeing the dispreferred one as context, and vice versa, allowing the model to leverage alternative responses as in-context guidance for improvement. Green terms highlight the conditioning mechanism in the objectives.

model's self-reflection capacity untapped. Current approaches treat response generation as isolated maximization, disregarding that human preferences are contextual, shaped by response interactions rather than standalone quality (Tversky & Simonson, 1993). By preventing conditioning on alternative responses, existing methods limit the ability to "compare and contrast", a mechanism we term *intrinsic self-reflection*, thereby capping policy quality. These two limitations our key question:

A Key Question

*How can we construct a framework that is both **invariant to modeling choices** and capable of **fully exploiting the comparative nature** of human preferences?*

In this work, we propose Intrinsic Self-reflective Preference Optimization (INSPO), a new perspective that fully leverages comparative information in pairwise preference data—a capacity left untapped in current frameworks. We first derive a globally optimal policy conditioning on both context and alternative responses, then prove this formulation yields a target superior to standard DPO and RLHF while guaranteeing invariance to scalarization and reference policy choices.

To realize this, we introduce a generic framework unlocking self-reflective capability for existing methods as a plug-and-play enhancement. Our approach is theoretically rigorous and computationally efficient, decoupling alignment goals from arbitrary modeling constraints. Crucially, deployment remains standard with no additional inference overhead since the self-reflective mechanism is incorporated into the policy during training without generating alternative responses during inference. Comprehensive experiments demonstrate consistent improvements in win rates and length-controlled metrics across benchmarks, while our analysis reveals that INSPO induces dense reward shaping and scales effectively with model capacity. See Figure 1 for an overview.

2 Background and Notations

To align an LLM with human preference, most existing pipelines first collect pairwise preference data generated by a pre-trained or SFT policy π_{ref}. Each sample in the preference data consists of a context $x \in \mathcal{X}$ generated by some distribution ρ, and two responses $y_w \in \mathcal{Y}$ and $y_\ell \in \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are the context and response spaces respectively. After collecting such preference data, one can apply RLHF or DPO, two primary approaches, to fine-tune an LLM.

RLHF consists of two steps. In the first step, it uses the Bradley-Terry (BT) probabilistic model to understand the human preference on (x, y_w, y_ℓ) and learn a reward function. Specifically, BT model assumes that

$$\mathbb{P}(y_w \succ y_\ell \mid x) = \sigma\big(r(x, y_w) - r(x, y_\ell)\big), \tag{1}$$

where $y_w \succ y_\ell$ indicates y_w is preferred to y_ℓ, $\sigma(z) = 1/(1 + e^{-z})$, and r is the unknown reward function to evaluate the quality of each response to the context x. Then given preference data, one

can implement maximum likelihood estimation (MLE) to estimate the reward function. In the second step, a policy optimization is executed to find a better LLM policy that maximizes the learned reward. For example, PPO (Ouyang et al., 2022) solves

$$\max_{\pi} \mathbb{E}_{x\sim\rho, y\sim\pi}\left[r(x,y)\right] - \beta D_{\mathrm{KL}}(\pi \,\|\, \pi_{\mathrm{ref}}), \tag{2}$$

where $\beta > 0$ is a regularization parameter that controls the strength of the KL divergence toward the reference policy π_{ref}, and the divergence $D_{\mathrm{KL}}(\pi \,\|\, \pi_{\mathrm{ref}})$ is defined as

$$D_{\mathrm{KL}}(\pi \,\|\, \pi_{\mathrm{ref}}) = \mathbb{E}_{x\sim\rho}\Big[\mathrm{KL}\big(\pi(\cdot \mid x) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid x)\big)\Big].$$

As an alternative approach to RLHF, DPO shows that solving (2) is equivalent to modeling

$$r(x, y_w) - r(x, y_\ell) = \beta\left(\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\mathrm{ref}}(y_w \mid x)} - \log\frac{\pi_\theta(y_\ell \mid x)}{\pi_{\mathrm{ref}}(y_\ell \mid x)}\right).$$

Then based on the MLE loss derived under BT model (1), DPO fine-tunes an LLM via solving

$$\max_{\pi} \mathbb{E}_{(x,y_w,y_\ell)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi(y_w \mid x)}{\pi_{\mathrm{ref}}(y_w \mid x)} - \beta\log\frac{\pi(y_\ell \mid x)}{\pi_{\mathrm{ref}}(y_\ell \mid x)}\right)\right], \tag{3}$$

where $\mathcal{D}$ is the joint distribution of (x, y_w, y_ℓ). Here without loss of generality, we assume y_w is always preferred to y_ℓ after rearrangement of the preference data.

As established in Proposition 1 of Azar et al. (2024), both methods unify under a general preference optimization framework with the scalarization function $\Psi(q) = \log(q/(1-q))$. The general objective is formulated as:

$$\max_{\pi} \mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\pi(\cdot|x), y'\sim\pi_{\mathrm{ref}}(\cdot|x)}\left[\Psi\big(\mathbb{P}(y \succ y' \mid x)\big)\right]. \tag{4}$$

Here, $\Psi : [0,1] \to \mathbb{R}$ can be any non-decreasing function. Within this framework, existing methods seek an optimal policy $\bar{\pi}$ restricted to the class of context-conditioned policies $\bar{\Pi} = \{\pi : \mathcal{X} \to \mathcal{Y}\}$ defined as

$$\bar{\pi} \in \mathrm{argmax}_{\pi\in\bar{\Pi}}\mathcal{V}(\pi),$$

where the value of a policy $\mathcal{V}(\pi)$ is defined as

$$\mathcal{V}(\pi) \triangleq \mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\pi(\cdot|x), y'\sim\pi_{\mathrm{ref}}(\cdot|x)}\left[\Psi\big(\mathbb{P}(y \succ y' \mid x)\big)\right].$$

To conclude this section, we assume that we have a dataset of n pairwise preferences $\mathcal{D}_n = \{(x^{(i)}, y_w^{(i)}, y_\ell^{(i)})\}_{i=1}^n$, where in i-th sample, a prompt/context $x^{(i)}$ is drawn from distribution ρ, and two responses $(y_w^{(i)}, y_\ell^{(i)})$ are generated by the reference policy π_{ref}, labeled such that $y_w^{(i)} \succ y_\ell^{(i)}$.

3 Limitations of Existing Methods

In this section, we identify two critical limitations of existing frameworks by investigating the properties of the restricted optimal policy $\bar{\pi}$: (i) Is $\bar{\pi}$ invariant to the scalarization function Ψ and the reference distribution π_{ref}? (ii) Is $\bar{\pi}$ theoretically optimal? We demonstrate below that the answer to both questions is negative.

3.1 Is $\bar{\pi}$ invariant to Ψ and π_{ref}?

Ideally, a robust alignment framework should yield an optimal policy invariant to the choice of the scalarization function Ψ and the reference policy π_{ref}. This invariance property is critical for modeling robustness and disentanglement from the reference. Specifically, human preferences are fundamentally ordinal. The optimal policy should reflect the underlying ranking of responses, rather than being an artifact of the specific choices of mathematical transformation, *i.e.,* Ψ, used to process the preference probabilities or the reference policy, *i.e.,* π_{ref}. In addition, the choice of Ψ, which is often selected for numerical stability or concavity rather than semantic relevance, should not dictate the final behavior of the model. If the optimal fine-tuned policy changes based on modeling choices, the alignment process may become brittle and inconsistent. Lastly, a principled optimization

objective should decouple the learned preferences from the reference policy π_{ref}. Dependence on the reference policy implies that the "optimal" behavior is relative and transient, rather than converging toward a global optimal policy that maximizes human preference. However, the following proposition establishes that the existing target $\bar{\pi}$ fails to satisfy this condition.

Proposition 3.1. *The form of $\bar{\pi}$ is not invariant to Ψ and π_{ref}.*

In the proof of Proposition 3.1, we provide counter-examples demonstrating that $\bar{\pi}$ shifts when either Ψ or π_{ref} is varied respectively. This lack of invariance raises a fundamental question about the quality of the resulting policy $\bar{\pi}$. If the "optimal" solution shifts based on different choices Ψ and π_{ref}, it suggests that $\bar{\pi}$ is an artifact of the objective function rather than a faithful reflection of human preferences. Consequently, in the following, we demonstrate that the current target $\bar{\pi}$ is, in fact, technically suboptimal.

3.2 Is $\bar{\pi}$ Optimal?

In this subsection, we investigate the theoretical optimality of $\bar{\pi}$ in terms of maximizing $\mathcal{V}(\pi)$. While $\bar{\pi}$ maximizes $\mathcal{V}(\pi)$ over $\bar{\Pi}$, we show that it is generally suboptimal compared to the globally optimal policy π^*, defined as:

$$\pi^* \in \text{argmax}_{\pi \in \Pi} \mathcal{V}(\pi), \tag{5}$$

where $\Pi = \{\pi : \mathcal{X} \times \mathcal{Y} \to \Delta(\mathcal{Y})\}$ represents a broader policy class that conditions not only on the context x but also on an auxiliary response y'. The following theorem establishes that π^* is not only superior to $\bar{\pi}$ but also possesses the desirable invariance property.

Theorem 3.2. *(i) π^* is invariant to any increasing function Ψ and π_{ref}; (ii) The following inequality holds, which implies π^* is superior to $\bar{\pi}$.*

$$\begin{aligned}\mathcal{V}(\pi^*) &= \mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\pi^*(\cdot|x,y'),y'\sim\pi_{\text{ref}}(\cdot|x)}\left[\Psi\big(\mathbb{P}(y \succ y' \mid x)\big)\right] \\ &\geq \mathcal{V}(\bar{\pi}) = \mathbb{E}_{x\sim\rho}\mathbb{E}_{y\sim\bar{\pi}(\cdot|x),y'\sim\pi_{\text{ref}}(\cdot|x)}\left[\Psi\big(\mathbb{P}(y \succ y' \mid x)\big)\right].\end{aligned}$$

(iii) Furthermore, given a fixed Ψ, π^* coincides with $\bar{\pi}$ if the transformed preference probability satisfies the condition that $\Psi\big(\mathbb{P}(y \succ y' \mid x)\big) \propto c(x, y) - c(x, y')$ for some function c.

Theorem 3.2 implies that the superiority of π^* over $\bar{\pi}$ stems from its dependence on the comparator response y', which effectively triggers a novel notion of intrinsic *self-reflection* within the LLM. This capacity for self-reflection, which existing fine-tuning methods lack, is a critical property for enhancing alignment. Furthermore, this mechanism renders π^* invariant to both Ψ and π_{ref}, ensuring that the policy targets the ground-truth human preference probability rather than optimization artifacts as $\bar{\pi}$. While Theorem 3.2 (iii) suggests that self-reflection yields no improvement if the preference function is separable, the separable condition is restrictive as it requires a correctly specified (link) function Ψ. In other words, in the current framework of RLHF and DPO, $\bar{\pi}$ is optimal only if the BT model (1) is correctly specified. More importantly, existing literature (Tversky & Simonson, 1993; Bordalo et al., 2013) demonstrates that human preferences are inherently non-separable and determined by the interaction between options. It is the context x, and the self-reflection by comparing y with y' that fundamentally shape the choice. Consequently, to leverage this comparative property, we aim to learn π^* from the preference data $\mathcal{D}_n$, and since our preference data are paired, we do not consider a larger categories of Π beyond pairwise comparison.

4 Intrinsic Self-reflective Preference Optimization

In this section, we introduce our InSPO method for learning π^*. Thanks to the invariant property of π^* as shown in Theorem 3.2 (i), we consider $\Psi(q) = \log(q/(1-q))$. First of all, we impose the following choice model.

$$\mathbb{P}(y_w \succ y_\ell \mid x) = \sigma(2(r(x, y_w, y_\ell) - \beta \log Z(x, y_\ell))), \tag{6}$$

for some generic reward function $r : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and $Z(x, y') := \sum_y \pi_{\text{ref}}(y \mid x) \exp\Big(\frac{1}{\beta} r(x, y, y')\Big)$. Then by (5),

$$\begin{aligned}\pi^* \in &\text{argmax}_\pi \, 2\mathbb{E}_{x\sim\rho,y\sim\pi(\cdot|x,y'),y'\sim\pi_{\text{ref}}(\cdot|x)}\left[r(x, y, y') - \beta \log Z(x, y')\right] \\ = &\text{argmax}_\pi \, \mathbb{E}_{x\sim\rho,y\sim\pi(\cdot|x,y'),y'\sim\pi_{\text{ref}}(\cdot|x)}\left[r(x, y, y')\right],\end{aligned}$$

Table 1: InSPO variants of six representative preference optimization methods. The Original formulations of DPO-based approaches are in black, while INSPO-enhanced terms are in green.

Method	Objective Function for Minimization
DPO (Rafailov et al., 2023)	$-\log\sigma\Big(\beta\log\frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)}-\beta\log\frac{\pi_\theta(y_\ell\|x)}{\pi_{\text{ref}}(y_\ell\|x)}\Big)$
	$-\log\sigma\Big(\beta\log\frac{\pi_\theta(y_w\|y_\ell,x)}{\pi_{\text{ref}}(y_w\|x)}-\beta\log\frac{\pi_\theta(y_\ell\|y_w,x)}{\pi_{\text{ref}}(y_\ell\|x)}\Big)$
IPO (Azar et al., 2024)	$\Big(\log\frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)}-\log\frac{\pi_\theta(y_\ell\|x)}{\pi_{\text{ref}}(y_\ell\|x)}-\frac{1}{2\tau}\Big)^2$
	$\Big(\log\frac{\pi_\theta(y_w\|y_\ell,x)}{\pi_{\text{ref}}(y_w\|x)}-\log\frac{\pi_\theta(y_\ell\|y_w,x)}{\pi_{\text{ref}}(y_\ell\|x)}-\frac{1}{2\tau}\Big)^2$
RDPO (Park et al., 2024)	$-\log\sigma\Big(\beta[\log\frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)}-\log\frac{\pi_\theta(y_\ell\|x)}{\pi_{\text{ref}}(y_\ell\|x)}]+\alpha(\|y_w\|-\|y_\ell\|)\Big)$
	$-\log\sigma\Big(\beta[\log\frac{\pi_\theta(y_w\|y_\ell,x)}{\pi_{\text{ref}}(y_w\|x)}-\log\frac{\pi_\theta(y_\ell\|y_w,x)}{\pi_{\text{ref}}(y_\ell\|x)}]+\alpha(\|y_w\|-\|y_\ell\|)\Big)$
ORPO (Hong et al., 2024)	$-\log p_\theta(y_w\mid x)\;-\;\lambda\log\sigma\Big(\text{logit}\big(p_\theta(y_w\mid x)\big)\;-\;\text{logit}\big(p_\theta(y_\ell\mid x)\big)\Big)$
	$-\log p_\theta(y_w\mid y_\ell,x)\;-\;\lambda\log\sigma\Big(\text{logit}\big(p_\theta(y_w\mid y_\ell,x)\big)\;-\;\text{logit}\big(p_\theta(y_\ell\mid y_w,x)\big)\Big)$
SimPO (Meng et al., 2024)	$-\log\sigma\Big(\beta\frac{1}{\|y_w\|}\log\pi_\theta(y_w\|x)-\beta\frac{1}{\|y_\ell\|}\log\pi_\theta(y_\ell\|x)-\gamma\Big)$
	$-\log\sigma\Big(\beta\frac{1}{\|y_w\|}\log\pi_\theta(y_w\|y_\ell,x)-\beta\frac{1}{\|y_\ell\|}\log\pi_\theta(y_\ell\|y_w,x)-\gamma\Big)$

$p_\theta(y|\cdot)=\exp(\frac{1}{|y|}\log\pi_\theta(y|\cdot))$; $\beta,\tau,\alpha,\lambda,\gamma$: hyperparameters.

which is independent of Z. Therefore model assumption in (6) is mild as r is unspecified and can be generic. Then following the paradigm of RLHF, we can estimate π^* via

$$\max_\pi \mathbb{E}_{x\sim\rho,y\sim\pi(\cdot|x,y'),y'\sim\pi_{\text{ref}}(\cdot|x)}\left[r(x,y,y')\right]-\beta D_{\text{KL}}(\pi\,\|\,\pi_{\text{ref}}),\tag{7}$$

where

$$D_{\text{KL}}(\pi\,\|\,\pi_{\text{ref}})=\mathbb{E}_{x\sim\rho,y'\sim\pi_{\text{ref}}(\cdot|x)}\Big[\text{KL}\big(\pi(\cdot\mid y',x)\,\|\,\pi_{\text{ref}}(\cdot\mid x)\big)\Big].$$

While this is a promising approach, the reward function r may be hard to estimate and PPO is known for instability. In the following, we propose a family of DPO-based approaches for learning π^*.

To begin with, we have the following proposition that establishes the connection between the general reward function r and π_r, which is denoted as an optimal solution to (7).

Theorem 4.1. *Solving the optimization problem* (7) *gives*

$$r(x,y,y')=\beta\Big[\log\frac{\pi_r(y|x,y')}{\pi_{\text{ref}}(y|x)}+\log Z(x,y')\Big].\tag{8}$$

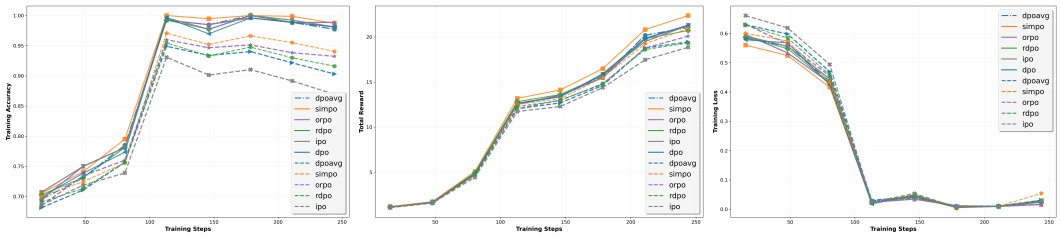
Furthermore, π_r can be obtained by solving

$$\max_\pi \mathbb{E}_{(x,y_w,y_\ell)\sim\mathcal{D}}\Big[\log\sigma\Big(\beta\Big(\log\frac{\pi(y_w|x,y_\ell)}{\pi_{\text{ref}}(y_w|x)}-\log\frac{\pi(y_\ell|x,y_w)}{\pi_{\text{ref}}(y_\ell|x)}\Big)\Big)\Big].\tag{9}$$

Then based on Theorem 3.2 and the preference dataset $\mathcal{D}_n$, InSPO estimates π^* via

$$\min_\theta -\frac{1}{n}\sum_{i=1}^n\Big[\log\sigma\Big(\beta\Big(\log\frac{\pi_\theta(y_w^{(i)}|x,y_\ell^{(i)})}{\pi_{\text{ref}}(y_w^{(i)}|x)}-\log\frac{\pi_\theta(y_\ell^{(i)}|x,y_w^{(i)})}{\pi_{\text{ref}}(y_\ell^{(i)}|x)}\Big)\Big)\Big],\tag{10}$$

where we parametrize the trainable policy π by θ. Similar to DPO, the proposed InSPO based on (10) aligns an LLM by directly shifting probability mass toward human preferred responses and away



(a) Preference Accuracy (b) Reward (c) Training Loss

Figure 2: Training dynamics of INSPO methods (solid) versus baselines (dashed). Our INSPO exhibits stable optimization with smooth loss convergence, consistent accuracy improvement, and enhanced reward margins without requiring additional on-policy rollouts.

from dispreferred responses without performing RL, without rollouts, and without training a reward model. On top of it, our method leverages the self-reflection property embedded in the preference data to further calibrate the policy towards the preferred response by contrasting with the dispreferred one. Therefore, this approach enables the learning of π^*.

During the inference, instead of first generating y' from π_{ref} for self-reflection, which incurs an overhead cost, we directly deploy $\widehat{\pi}$ given a testing query x. While this introduces a distinction between the training context (x, y') and the inference context x, we understand this under the paradigm of Learning Using Privileged Information (LUPI) (Pechyony & Vapnik, 2010). In the LUPI framework, the learning algorithm is provided with additional "privileged" information (here, the alternative response) during the training phase to stabilize the optimization landscape and accelerate the convergence. This privileged context acts as a contrastive scaffold, guiding the gradient updates for the shared weights associated with x. Consequently, the self-reflective capability is distilled into the policy weights (Hinton et al., 2015), allowing the deployed model to retain the optimized decision boundaries even when the privileged scaffolding is removed at test time. This ensures that InSPO incurs zero inference overhead, as the self-reflective mechanism is implicitly encoded in the parameter space rather than requiring explicit rollout generation.

It is worth noting that our proposed formulation for π^* is method-agnostic and can seamlessly integrate with standard preference optimization techniques. Table 1 illustrates the InSPO variants of six representative preference optimization methods, including DPO, demonstrating how each is adapted to unlock the self-reflective capability. In the next section, we study the empirical performance of each method.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Models and data. We employ Mistral-7B-Instruct-v0.2 and Llama-3-8B-Instruct as base models, initializing all methods from identical checkpoints within each family. Training data is sourced from UltraFeedback (Cui et al., 2023), containing approximately 60K preference pairs $(x; y_w, y_\ell)$ with $y_w \succ y_\ell$ after deduplication and safety filtering. The data generating process follows exactly from SimPO (Meng et al., 2024).

Benchmarks. We evaluate on three widely-adopted benchmarks. **AlpacaEval 2** (Dubois et al., 2024) contains 805 diverse instructions; we report both standard win rates (WR) and length-controlled win rates (LC) to mitigate verbosity bias. **Arena-Hard** (Li et al., 2024) features 500 challenging queries that test advanced reasoning; we report WR alongside its style-controlled variant (SC) to account for stylistic preferences. **MT-Bench** (Zheng et al., 2023) comprises 80 multi-turn questions spanning eight capability categories, scored on a 10-point scale.

Baselines. We compare INS-enhanced variants: INS-DPO, INS-SimPO, INS-IPO, INS-RDPO, INS-ORPO against their standard counterparts and additional baselines: RRHF, SLiC-HF, CPO, and KTO. For fair comparison, baseline results are obtained by evaluating publicly released checkpoints from SimPO (Meng et al., 2024) on identical benchmark versions. All INS-enhanced methods are

Table 2: Benchmark performance on AlpacaEval2, Arena-Hard, and MT-Bench for Mistral-Instruct (7B) and Llama-3-Instruct (8B). Shaded rows denote continuation-conditioned variants with deltas computed relative to their corresponding baselines. Bold values indicate the highest performance in each metric. All methods employ DPO-family objectives with identical training configurations.

Method	Mistral-Instruct (7B)					Llama-3-Instruct (8B)				
	AlpacaEval2		Arena-Hard	MT-Bench		AlpacaEval2		Arena-Hard	MT-Bench	
	LC (%)	WR (%)	WR (%)	GPT-4o	GPT-4	LC (%)	WR (%)	WR (%)	GPT-4o	GPT-4
Baseline	17.1	14.7	12.6	6.2	7.5	26.0	25.3	22.3	6.9	8.1
Reward-Free Preference Optimization										
RRHF	25.3	24.8	18.1	6.5	7.6	31.3	28.4	26.5	6.7	7.9
SLiC-HF	24.1	24.6	18.9	6.5	7.8	26.9	27.5	26.2	6.8	8.1
CPO	23.8	28.8	22.6	6.3	7.5	28.9	32.2	28.8	7.0	8.0
KTO	24.5	23.6	17.9	6.4	7.7	33.1	31.8	26.4	6.9	8.2
Preference Optimization with Self-Reflection Conditioning										
IPO	20.3	20.3	16.2	6.4	7.8	35.6	35.6	30.5	7.0	8.3
+ INS	28.8 ↑8.5	27.9 ↑7.6	18.5 ↑2.3	6.5 ↑0.1	7.9 ↑0.1	37.7 ↑2.1	41.3 ↑5.7	32.8 ↑2.3	7.1 ↑0.1	8.5 ↑0.2
ORPO	24.5	24.9	20.8	6.4	7.7	28.5	27.4	25.8	6.8	8.0
+ INS	24.6 ↑0.1	24.7 ↓0.2	21.5 ↑0.7	6.3 ↓0.1	7.8 ↑0.1	35.7 ↑7.2	40.3 ↑12.9	25.5 ↓0.3	6.8 – 0.0	8.0 – 0.0
R-DPO	27.3	24.5	16.1	6.2	7.5	41.1	37.8	33.1	7.0	8.0
+ INS	29.1 ↑1.8	27.6 ↑3.1	19.0 ↑2.9	6.4 ↑0.2	7.6 ↑0.1	41.0 ↓0.1	43.8 ↑6.0	35.0 ↑1.9	7.1 ↑0.1	8.1 ↑0.1
DPO	26.8	24.9	16.3	6.3	7.6	40.3	37.9	32.6	7.0	8.0
+ INS	29.9 ↑3.1	29.4 ↑4.5	23.9 ↑7.6	6.5 ↑0.2	7.8 ↑0.2	40.3 – 0.0	41.1 ↑3.2	35.3 ↑2.7	7.1 ↑0.1	8.1 ↑0.1
SimPO	32.1	**34.6**	21.0	6.6	7.6	**44.5**	40.5	33.8	7.0	8.0
+ INS	**32.7** ↑0.6	34.5 ↓0.1	**24.4** ↑3.4	**6.6** – 0.0	**7.7** ↑0.1	44.5 – 0.0	**46.6** ↑6.1	**35.9** ↑2.1	**7.1** ↑0.1	**8.2** ↑0.2

LC: length-controlled win rate; WR: win rate; AE: AlpacaEval2; AH: Arena-Hard. All methods initialize from the same instructed checkpoint within each model family. **INS** applies sequence-level continuation conditioning with symmetric cross-conditioning across preference pairs.

implemented using the OpenRLHF framework (Hu et al., 2024) and trained for three epochs with AdamW optimizer, learning rate 5×10^{-7}, cosine schedule with 10% warmup, and maximum context length 4096. Method-specific hyperparameters follow configurations from prior works.

Inference. All models use standard autoregressive generation $\hat{y} \sim \widehat{\pi}(\cdot \mid x)$ with nucleus sampling at $p = 0.95$ and temperature 0.7, introducing no additional computational overhead versus baseline methods.

5.2 Main Results

Table 2 presents comprehensive benchmark results across both model families. Controlled metrics show substantial improvements. On AlpacaEval(AE) LC, INSPO yields 0.6–8.5 point gains across the DPO family, with the largest improvements observed on INS-IPO: +8.5 points for Mistral-Instruct and +2.1 points for Llama-3-Instruct. Arena-Hard(AH) results follow a similar pattern, with improvements ranging from 0.7 to 7.6 points. Notably, INS-DPO achieves a 7.6-point gain on Mistral-Instruct and 2.7 points on Llama-3-Instruct, demonstrating consistent benefits of enabling intrinsic self-reflection through alternative conditioning. The training dynamic curves in Figure 2 further enhance the persuasiveness of the results.

The gains extend beyond controlled metrics: INS-DPO and INS-SimPO improve raw win rates by 3–6 points on AlpacaEval 2 and 2–3 points on Arena-Hard, indicating genuine quality improvements rather than mere verbosity reduction. MT-Bench results corroborate this finding, with INS-enhanced methods achieving 0.1–0.2 point gains across different judge models, demonstrating improvements where alternative response context provides valuable learning signals.

Table 3: Comparison of conditioning strategies. Deltas are relative to the DPO baseline.

Model	Strategy	AlpacaEval 2		Arena-Hard
		LC (%)	WR (%)	WR (%)
Mistral-7B	DPO (baseline)	26.8	24.9	16.3
	One-sided	28.7 ↑1.9	27.1 ↑2.2	19.2 ↑2.9
	Symmetric	29.9 ↑3.1	29.4 ↑4.5	23.9 ↑7.6
	Averaged	29.6 ↑2.8	29.7 ↑4.8	25.7 ↑9.4
Llama-3-8B	DPO (baseline)	40.3	37.9	32.6
	One-sided	43.4 ↑3.1	40.2 ↑2.3	34.1 ↑1.5
	Symmetric	40.3 ↑0.0	41.1 ↑3.2	35.3 ↑2.7
	Averaged	44.1 ↑3.8	41.5 ↑3.6	35.5 ↑2.9

The relative ranking among DPO family members remains stable under INS enhancement—SimPO variants consistently lead, followed by DPO and R-DPO, with IPO and ORPO showing more variable performance—suggesting that INSPO amplifies rather than disrupts the inherent strengths of each base method. Compared to alternative baselines beyond the DPO family such as RRHF, SLiC-HF, CPO, and KTO, our best INS-enhanced variant INS-SimPO establishes competitive results on controlled metrics while maintaining comparable raw performance.

5.3 Analysis and Insights

We conduct ablation studies to understand the mechanisms underlying sequence-level conditioning. Unless otherwise specified, all experiments use INS-DPO with Llama-3-8B-Instruct.

Impact of candidate length and context window. We vary the maximum context window (MaxLen $\in \{1024, 2048, 4096\}$) and the dispreferred candidate length cap $\alpha \in \{30\%, 40\%, 50\%, \text{None}\}$, where $|y_\ell| \leq \alpha \cdot \text{MaxLen}$. Three findings emerge: (i) longer context windows consistently improve performance, with 4096-token contexts yielding the best results across all metrics, confirming that avoiding truncation of (x, y_ℓ, y_w) is crucial; (ii) a moderate draft cap of 40% achieves optimal performance, as uncapped drafts can destabilize training when extremely long dispreferred responses dominate the context; (iii) the gains from longer contexts are most pronounced on challenging benchmarks (e.g., +2.6 Arena-Hard WR from 1024 to 4096).

Conditioning strategies. We compare four variants (Table 3): (i) standard DPO, (ii) one-sided conditioning where only y_w is conditioned on y_ℓ, (iii) symmetric cross-conditioning (INSPO), and (iv) a simple average $\frac{1}{2}\mathcal{L}_{\text{DPO}} + \frac{1}{2}\mathcal{L}_{\text{sym}}$. All use MaxLen=4096 with 40% draft cap.

Symmetric conditioning substantially outperforms the baseline on Arena-Hard (+7.6 WR for Mistral-7B, +2.7 for Llama-3-8B), while even one-sided conditioning provides meaningful gains, confirming that sequence-level context is beneficial regardless of symmetry. The averaged variant achieves favorable trade-offs: best Arena-Hard on Mistral-7B (+9.4 WR) and best AlpacaEval LC on Llama-3-8B (+3.8). In practice, symmetric conditioning is preferred for raw quality, while averaging better balances quality with length control.

Computational overhead. Sequence-level conditioning incurs 18–25% training time overhead due to processing longer concatenated inputs, with max sequence length increasing from 2048 to 4096. Crucially, inference remains identical to standard DPO with zero additional latency, as generation uses standard sampling $\hat{y} \sim \pi_\theta(\cdot \mid x)$ without draft-revise loops.

6 Conclusion

In this work, we addressed two critical limitations in preference optimization: the lack of theoretical invariance to arbitrary modeling choices and the suboptimality of treating response generation in isolation. To overcome these challenges, we introduced INSPO, a framework that conditions the optimal policy on both context and alternative responses, achieving a target guaranteed invariant to scalarization and reference policy choices. Comprehensive experiments confirm consistent improvements across benchmarks and effective scaling with model size, offering a robust, theoretically grounded path toward human-aligned LLMs. Future work may extend this paradigm to online training or multi-turn settings.

REFERENCES

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455. PMLR, 2024.

Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. Salience and consumer choice. *Journal of Political Economy*, 121(5):803–843, 2013.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. 2023.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. How abilities in large language models are affected by supervised fine-tuning data composition. *arXiv preprint arXiv:2310.05492*, 2023.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*, 2024.

Jian Hu, Xibin Wu, Wei Shen, Jason Klein Liu, Zilin Zhu, Weixun Wang, Songlin Jiang, Haoran Wang, Hao Chen, Bin Chen, et al. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. *arXiv preprint arXiv:2405.11143*, 2024.

Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip HS Torr, Fahad Shahbaz Khan, and Salman Khan. Llm post-training: A deep dive into reasoning large language models. *arXiv preprint arXiv:2502.21321*, 2025.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*, 2024.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. *arXiv preprint arXiv:2403.19159*, 2024.

Dmitry Pechyony and Vladimir Vapnik. On the theory of learnining with privileged information. *Advances in neural information processing systems*, 23, 2010.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

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

Amos Tversky and Itamar Simonson. Context-dependent preferences. *Management Science*, 39(10): 1179–1189, 1993.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Han Zhong, Zikang Shan, Guhao Feng, Wei Xiong, Xinle Cheng, Li Zhao, Di He, Jiang Bian, and Liwei Wang. Dpo meets ppo: Reinforced token optimization for rlhf. *arXiv preprint arXiv:2404.18922*, 2024.