LSC-DPO: LEARNING-SIGNAL-CONTROLLED DIRECT PREFERENCE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Direct Preference Optimization (DPO) has become a standard reward-model-free approach for aligning language models with preference data. However, as the scaled preference margin grows during training, the logistic DPO loss becomes progressively less sensitive to further changes. We study DPO from a loss-level geometric perspective and identify the sigmoid factor as a learning signal that characterizes the local sensitivity of the objective. Based on this view, we propose Learning-Signal-Controlled Direct Preference Optimization (LSC-DPO), which dynamically regulates the learning signal near a target regime. A log-space analysis establishes conditions for stable tracking of the target learning-signal regime. Experiments on AlpacaEval 2, MT-Bench, and Anthropic-HH show that LSC-DPO consistently improves over DPO and strong preference-optimization baselines. We further observe that learning-signal dynamics are systematically related to the initial coefficient. We then investigate this relationship and derive a signal-budget compensation rule that substantially reduces performance variation across coefficient initializations.

1 INTRODUCTION

Large language models (LLMs) have achieved remarkable performance across a broad range of tasks. However, their outputs are not always aligned with human preferences. Reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022; Stiennon et al., 2020) has become a widely used paradigm for improving alignment. In the standard RLHF pipeline, approaches such as Proximal Policy Optimization (PPO) (Schulman et al., 2017) optimize a policy based on a learned reward model that captures human preferences, while a KL penalty constrains the policy from deviating excessively from a reference model.

Recent work has explored reward-model-free approaches to reduce the computational cost of RLHF. Direct Preference Optimization (DPO) (Rafailov et al., 2023) directly optimizes a policy from offline preference data without explicitly training a reward model. In DPO, preference learning is formulated through the policy-reference log-ratio difference between chosen and rejected responses. Under the Bradley–Terry model (Bradley & Terry, 1952), the scaled log-ratio difference can be interpreted as an implicit reward difference. The coefficient β scales this quantity and thereby affects the resulting optimization dynamics.

A central question in DPO is how to characterize and regulate its evolving optimization state during training. The DPO objective contains a logistic loss term whose sensitivity decreases as optimization progresses (Rafailov et al., 2023; 2024; Gorbatovski et al., 2024). As this sensitivity decreases, the same preference pairs receive progressively smaller gradient weight, weakening their contribution to subsequent policy updates. Importantly, the initial choice of the coefficient β also affects the sensitivity of the logistic loss term and can therefore lead to different optimization trajectories and downstream performance (Rafailov et al., 2023). These observations motivate a principled mechanism for characterizing and regulating variations in the training state.

In this paper, we analyze the DPO objective from a loss-level geometric perspective and identify a sigmoid factor that characterizes the local sensitivity of the objective during training. This factor reflects both the evolving optimization state and the effect of the coefficient β, making it a useful quantity for diagnosing and regulating DPO training. Based on this observation, we propose **Learning-Signal-Controlled Direct Preference Optimization (LSC-DPO)**, which dynamically

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

regulates the learning signal to keep it close to a target level during preference optimization. We further derive a local stability condition for tracking the desired learning-signal regime. Empirically, LSC-DPO achieves strong performance compared with existing preference-optimization methods (Rafailov et al., 2023; Wu et al., 2024b; Lee et al., 2025; Park et al., 2024; Azar et al., 2024; Ethayarajh et al., 2024; Xu et al., 2024; Hong et al., 2024; Zhao et al., 2023; Meng et al., 2024) across standard alignment benchmarks and preference-optimization experiments. We further use the learning-signal view to analyze sensitivity to the initial choice of β and derive a signal-budget compensation rule that reduces initialization-induced performance degradation and improves robustness across different initial coefficients.

Our contributions are summarized in three aspects. (1) We provide a loss-level geometric characterization of DPO, showing that the sigmoid factor serves as a learning signal for tracking the local sensitivity of the objective and the evolving training state. (2) We propose **LSC-DPO**, a learning-signal-controlled preference optimization method that dynamically adjusts the DPO coefficient according to the evolving training state. We further derive a local stability condition that characterizes stable tracking around the target learning-signal regime. Empirically, LSC-DPO achieves strong performance compared with existing preference-optimization baselines across standard alignment benchmarks. (3) We investigate the relationship between the initial coefficient and learning-signal dynamics, and derive a signal-budget compensation rule that reduces initialization-induced performance degradation and improves robustness across different initial coefficients.

2 PRELIMINARIES

In RLHF, a policy is optimized against a reward function while being regularized toward a reference policy (Christiano et al., 2017; Jaques et al., 2017; 2020):

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

where π_θ denotes the policy and π_{ref} denotes the reference model. DPO (Rafailov et al., 2023) uses the closed-form solution (Ziegler et al., 2019; Peng et al., 2019; Peters & Schaal, 2007; Go et al., 2023; Rafailov et al., 2023) to reparameterize the reward function as:

$$r(x,y) = \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)} + \beta \log Z(x), \tag{2}$$

where $Z(x)$ is the partition function. Substituting the reparameterized reward into the Bradley–Terry model (Bradley & Terry, 1952) yields the DPO objective:

$$\mathcal{L}_{\mathrm{DPO}}(\pi_\theta;\pi_{\mathrm{ref}}) = -\mathbb{E}_{(x,y_+,y_-)}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_+ \mid x)}{\pi_{\mathrm{ref}}(y_+ \mid x)} - \beta\log\frac{\pi_\theta(y_- \mid x)}{\pi_{\mathrm{ref}}(y_- \mid x)}\right)\right], \tag{3}$$

where y_+ and y_- denote the chosen and rejected responses, respectively. The DPO objective depends on the difference between the policy log-ratio of the chosen and rejected responses and the corresponding reference model log-ratio. We define this difference as the preference margin:

$$m = \log\frac{\pi_\theta(y_+ \mid x)}{\pi_\theta(y_- \mid x)} - \log\frac{\pi_{\mathrm{ref}}(y_+ \mid x)}{\pi_{\mathrm{ref}}(y_- \mid x)}. \tag{4}$$

This margin measures how much more the policy favors the chosen response over the rejected response, relative to the reference model. We further define the scaled preference margin as

$$z = \beta m. \tag{5}$$

Using the margin formulation above, the DPO objective for a single preference pair can be written compactly as

$$\ell_{\mathrm{DPO}} = -\log\sigma(\beta m) = -\log\sigma(z). \tag{6}$$

Thus, the optimization behavior of DPO depends on the scaled margin $z = \beta m$, which couples the evolving preference margin with the coefficient β.

3 LEARNING-SIGNAL-CONTROLLED DIRECT PREFERENCE OPTIMIZATION

In this section, we introduce Learning-Signal-Controlled Direct Preference Optimization (LSC-DPO). We first identify the sigmoid factor $\sigma(-z)$ in the DPO objective as a learning signal that reflects the local sensitivity of the DPO loss during training and illustrate how different initial choices of β affect the evolution of this signal. We then use this signal to characterize the training state of DPO and derive an update rule that regulates the signal during preference optimization. We analyze the stability of the resulting dynamics in log-space.

3.1 LEARNING SIGNAL IN DPO

To understand the training behavior of DPO, we analyze the local structure of the DPO objective in Eq. 6. The gradient can be written as

$$\nabla_\theta \ell = -\sigma(-z_\theta)\nabla_\theta z_\theta. \tag{7}$$

Taking another derivative gives the Hessian of the DPO objective:

$$\nabla_\theta^2 \ell = \sigma(z_\theta)\sigma(-z_\theta)\nabla_\theta z_\theta \nabla_\theta z_\theta^\top - \sigma(-z_\theta)\nabla_\theta^2 z_\theta. \tag{8}$$

Equations 7 and 8 show that $\sigma(-z_\theta)$ plays a central role in the training dynamics of DPO. From the gradient perspective, $\sigma(-z_\theta)$ directly measures the first-order sensitivity of the DPO loss. From the curvature perspective, the Hessian contains two scaled terms: the outer-product term is weighted by $\sigma(z_\theta)\sigma(-z_\theta)$, while the second-order term is weighted by $\sigma(-z_\theta)$. During training, z_θ typically increases for preference pairs that become better aligned with the preference ordering. As z_θ grows, the sigmoid factor $\sigma(-z_\theta)$ decreases. Since these factors scale the Hessian terms that determine local curvature, their decrease indicates reduced local sensitivity of the objective and a weaker preference-learning tendency. Figure 1 illustrates this loss geometry.

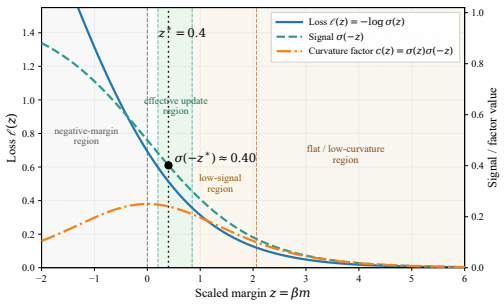


Figure 1: Illustration of the DPO loss geometry as a function of the scaled margin $z = \beta m$. The learning signal $\sigma(-z)$ and curvature factor $c(z) = \sigma(z)\sigma(-z)$ decrease as z grows.

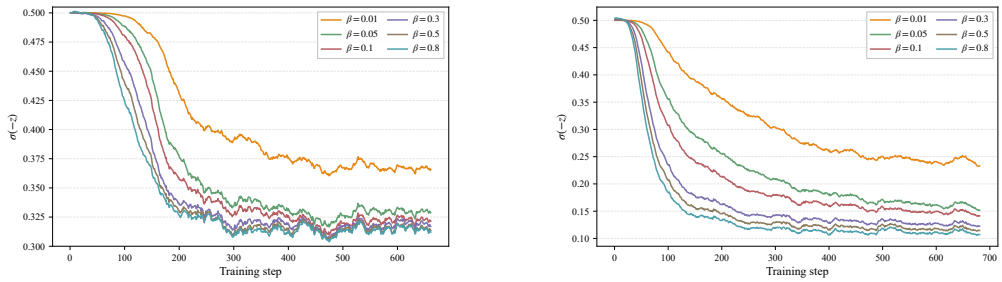
We therefore use $\sigma(-z_\theta)$ as a learning signal for DPO training. When this signal is relatively large, the DPO loss remains more sensitive to changes in the scaled margin, corresponding to a high-signal regime. As z_θ increases, $\sigma(-z_\theta)$ decreases and the loss becomes progressively less sensitive to further changes in the scaled margin. We refer to this as a low-signal regime, which progressively approaches a lower-curvature region as z_θ grows further. Figures 2a and 2b further show that $\sigma(-z_\theta)$ evolves systematically over training. It remains relatively high during the early stage and gradually decreases before stabilizing at a lower level. This evolution is consistent with our interpretation of $\sigma(-z_\theta)$ as reflecting the local responsiveness of the DPO loss during training. The figures further show that different choices of the DPO coefficient lead to markedly different learning-signal trajectories $\sigma(-z)$. This observation suggests that the learning-signal dynamics are related to the coefficient setting and motivates our later analysis of coefficient initialization in Section 4.4.

3.2 LEARNING-SIGNAL-CONTROLLED UPDATE

Since $\sigma(-z)$ reflects the learning signal in DPO, we aim to keep it near a target level rather than allowing it to decay rapidly during training. Because $\sigma(-z)$ decreases monotonically with the scaled margin z, regulating z provides a direct way to control the signal. Since $z = \beta m$, we use β as the control variable.

At training step k, we compute the per-example preference margins on the current batch and use their mean as the batch-level margin m_k. We then obtain the corresponding scaled margin:

$$z_k = \beta_k m_k. \tag{9}$$



(a) HH-harmless-base (b) HH-helpful-base

Figure 2: $\sigma(-z)$ under fixed β values during DPO training with Llama3-8B on two Anthropic-HH subsets: (a) harmless-base and (b) helpful-base. EMA-smoothed trajectories are shown for readability; raw trajectories are provided in Appendix A.

Thus, the controller computes the signal from the scaled batch-mean margin z_k. We then measure its deviation from a target scaled margin z^*:

$$u_k = z_k - z^*. \tag{10}$$

The target z^* determines the learning-signal regime that LSC-DPO tracks, equivalently corresponding to the target signal level $\sigma(-z^*)$. LSC-DPO updates β using the following multiplicative rule:

$$\beta_{k+1} = \beta_k \exp(-\eta u_k) = \beta_k \exp\left[-\eta(\beta_k m_k - z^*)\right], \tag{11}$$

where $\eta > 0$ controls the adjustment rate. The multiplicative form preserves the positivity of β and implements feedback control over the scaled margin. In the positive-margin regime, when $z_k > z^*$, $\sigma(-z_k)$ is below the target level and the update decreases β_k, moving the scaled margin downward. Conversely, when $z_k < z^*$, the update increases β_k, moving the scaled margin upward toward the target. The complete procedure is provided in Algorithm 1 in Appendix B.

3.3 Stability Analysis of the Update Process

We analyze the LSC-DPO update in log-space to characterize controller-level tracking of the target z^*, rather than convergence of the full policy optimization problem. Starting from Eq. 11, let

$$x_k := \log \beta_k. \tag{12}$$

Substituting $\beta_k = e^{x_k}$ into Eq. 11 gives

$$x_{k+1} = x_k - \eta\left(e^{x_k} m_k - z^*\right). \tag{13}$$

Since the margin statistic m_k evolves during training, the induced system is non-autonomous. Here, m_k denotes the realized batch-level margin at step k and may depend on the current policy, minibatch sampling, and previous coefficient updates. Accordingly, our analysis is conditional on the realized margin sequence and characterizes controller-level tracking rather than joint convergence of the full policy-training dynamics. Conditioned on m_k, the step-dependent update map is

$$F_k(x) = x - \eta\left(e^x m_k - z^*\right). \tag{14}$$

Instantaneous equilibrium. For a realized m_k, the map F_k admits an instantaneous fixed point

$$x_k^f = \log \frac{z^*}{m_k}, \qquad \beta_k^f = \frac{z^*}{m_k}, \tag{15}$$

provided that $m_k > 0$ and $z^* > 0$. Because m_k evolves during training, this equilibrium moves over time. We next characterize its local stability.

Local stability. Linearizing the step-dependent map F_k around the instantaneous equilibrium x_k^f gives

$$F_k'(x_k^f) = 1 - \eta z^*. \tag{16}$$

Hence, the instantaneous equilibrium is locally asymptotically stable whenever

$$0 < \eta z^* < 2. \tag{17}$$

Notably, this condition is independent of the realized margin m_k. The formal proposition and proof are provided in Appendix C.1.

Bounded tracking. Because m_k evolves during training, the corresponding equilibrium x_k^f also moves over time. Under local contraction and bounded equilibrium drift, the resulting tracking error satisfies

$$|x_k - x_k^f| \le \rho^{k-T}|x_T - x_T^f| + \frac{\bar{\delta}}{1-\rho}, \qquad k \ge T, \tag{18}$$

where $\rho \in (0,1)$ denotes a local contraction factor and $\bar{\delta}$ bounds the step-to-step motion of the instantaneous equilibrium. The first term decays geometrically, while the second captures the persistent error induced by the moving target. Thus, when the realized margin evolves sufficiently gradually, the controller remains close to the target scaled-margin regime and keeps the learning signal near $\sigma(-z^*)$. The complete theorem, assumptions, and proof are provided in Appendix C.2.

4 Experiments

In this section, we evaluate LSC-DPO on standard alignment benchmarks and analyze its training dynamics. We first compare LSC-DPO with representative preference-optimization baselines on AlpacaEval 2 (Dubois et al., 2024) and MT-Bench (Zheng et al., 2023b), and then study its performance and learning-signal dynamics on Anthropic-HH (Bai et al., 2022). We further validate the proposed learning-signal control through optimization-dynamics analyses and comparisons with predetermined β schedules. Finally, we investigate robustness to the initial DPO coefficient and the proposed signal-budget compensation rule. Additional analyses of target-signal selection and hyperparameters are provided in Appendix J, control stability in Appendix G, and downstream capability evaluations in Appendix K.

4.1 Experimental Setup

Models and training datasets. We evaluate on two open-source models, meta-llama/Meta-Llama-3-8B (Llama3-8B) (Grattafiori et al., 2024) and Qwen/Qwen3-8B-Base (Qwen3-8B) (Yang et al., 2025). We first perform supervised fine-tuning on UltraChat-200k (Ding et al., 2023), and then perform preference optimization on UltraFeedback (Cui et al., 2023). We additionally use the harmless-base and helpful-base subsets of Anthropic-HH (Bai et al., 2022) for controlled analyses. Full training settings are provided in Appendix E.1.

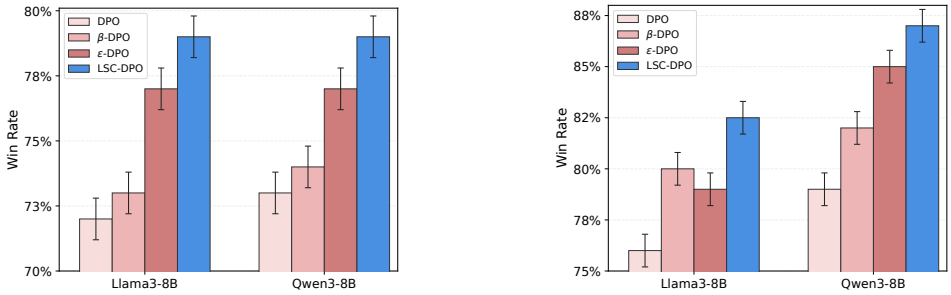
Evaluation benchmarks and baselines. For the UltraFeedback experiments, we evaluate on AlpacaEval 2 (Dubois et al., 2024) using win rate and length-controlled (LC) win rate, and on MT-Bench (Zheng et al., 2023b). On Anthropic-HH, we report pairwise win rates on both harmless-base and helpful-base. We compare our method with representative preference-optimization baselines, including DPO (Rafailov et al., 2023), β-DPO (Wu et al., 2024b), ε-DPO (Lee et al., 2025), R-DPO (Park et al., 2024), IPO (Azar et al., 2024), KTO (Ethayarajh et al., 2024), CPO (Xu et al., 2024), ORPO (Hong et al., 2024), SLiC-HF (Zhao et al., 2023), and SimPO (Meng et al., 2024). For the controlled Anthropic-HH analyses, we focus on DPO, β-DPO, and ε-DPO as the most relevant coefficient-control baselines. Detailed evaluation protocols and baseline configurations are provided in Appendix E.2.

4.2 Main Results on Standard Alignment Benchmarks

LSC-DPO achieves the strongest overall performance. Table 1 reports that LSC-DPO achieves the best AlpacaEval 2 performance on both Llama3-8B and Qwen3-8B, while attaining top or tied-top MT-Bench scores. Compared with DPO, LSC-DPO improves the length-controlled win rate

Table 1: Evaluation results on AlpacaEval 2 and MT-Bench. For AlpacaEval 2, we report both win rate (WR) and length-controlled win rate (LC). For MT-Bench, we report the average score on a 1–10 scale.

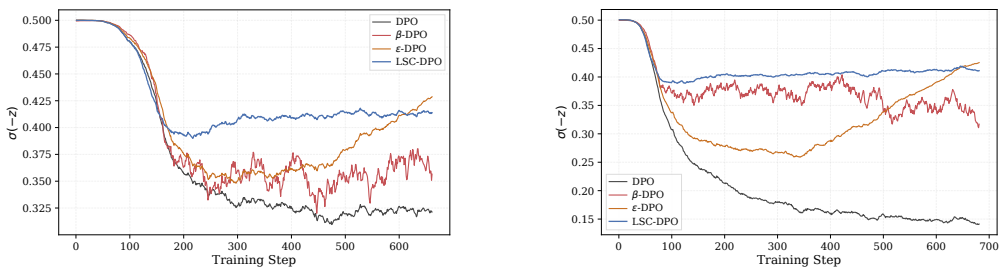
Method	Llama3-8B			Qwen3-8B		
	AlpacaEval 2		MT-Bench	AlpacaEval 2		MT-Bench
	LC (%)	WR (%)	Score	LC (%)	WR (%)	Score
SFT	6.8	3.7	6.3	8.5	4.8	6.9
DPO	19.2	17.7	6.6	17.2	13.1	7.6
ε-DPO	19.6	17.6	**6.9**	19.4	14.0	7.6
β-DPO	16.6	16.5	6.8	16.4	10.3	7.2
CPO	11.2	10.3	6.4	16.5	14.5	7.6
IPO	8.6	8.5	6.6	14.9	9.2	7.4
KTO	16.9	15.1	6.8	17.2	11.3	7.4
ORPO	6.4	5.3	5.5	9.4	6.8	6.9
R-DPO	19.5	18.2	6.8	19.3	13.2	7.6
SLiC-HF	10.1	8.7	5.9	12.5	10.5	7.1
SimPO	19.7	19.2	**6.9**	16.8	10.9	7.4
LSC-DPO	**22.7**	**21.1**	**6.9**	**22.1**	**16.1**	**7.8**



(a) HH-harmless-base (b) HH-helpful-base

Figure 3: Performance comparison on Anthropic-HH. We report win rates of DPO, β-DPO, ε-DPO, and LSC-DPO using Llama3-8B and Qwen3-8B. Error bars denote approximate 95% confidence intervals computed from prompt-level GPT-4 judgment results.

from 19.2 to 22.7 on Llama3-8B and from 17.2 to 22.1 on Qwen3-8B, with similar gains in standard win rate. It also consistently outperforms β-DPO and ε-DPO and remains competitive with broader preference-optimization baselines, including R-DPO, KTO, CPO, ORPO, SLiC-HF, IPO, and SimPO. We next use Anthropic-HH for controlled comparisons of coefficient-regulation methods and their learning-signal behavior.



(a) harmless-base (b) helpful-base

Figure 4: Learning signal $\sigma(-z)$ on Anthropic-HH using Llama3-8B. The corresponding raw curves and Qwen results are provided in Appendix F.

Table 2: Optimization diagnostics before and after LSC-DPO control activation. We report the mean curvature factor and mean pre-clipping gradient norm.

	Before control (steps 1–83)		After control (steps 84–477)	
Method	Curvature	Grad. norm	Curvature	Grad. norm
DPO	0.24993	29.861	0.23724	90.330
LSC-DPO	0.24993	29.896	**0.24371**	73.373

Table 3: Comparison with predetermined β schedules.

Method	LC (%)	WR (%)
DPO	19.2	17.7
Linear β schedule	19.7	18.4
Exponential β schedule	20.3	18.5
Cosine β schedule	19.1	18.1
LSC-DPO	**22.7**	**21.1**

Overall performance on Anthropic-HH. Figures 3a and 3b show that LSC-DPO achieves the best pairwise win rates across both model families and Anthropic-HH subsets, reaching approximately 79% on harmless-base and 82–87% on helpful-base. It consistently outperforms DPO, β-DPO, and ε-DPO, further supporting the effectiveness of learning-signal control under controlled preference-optimization settings.

Learning-signal dynamics. Figures 4a and 4b show that DPO exhibits a sustained signal decay and eventually moves toward a low-signal regime, β-DPO shows larger fluctuations, and ε-DPO partially recovers after an initial decline. In contrast, LSC-DPO maintains a higher and more stable signal near its target regime across most of training. These relative dynamics are consistent with the performance trends in Figure 3; on helpful-base, for instance, ε-DPO maintains a lower signal than β-DPO for much of training before recovering later, aligning with its lower win rate. Overall, these results support the intended learning-signal regulation behavior of LSC-DPO.

4.3 Mechanistic Validation of Learning-Signal Control

Learning signal and realized optimization dynamics. Table 2 shows that DPO and LSC-DPO are nearly indistinguishable before control activation: both have the same mean curvature factor and almost identical pre-clipping gradient norms. After control activation, LSC-DPO maintains a larger curvature factor than DPO. The slower reduction of this quantity indicates that LSC-DPO keeps the objective farther from the flatter, low-curvature regime. At the same time, its pre-clipping gradient norm is lower, showing that this increased loss-level responsiveness is not obtained by simply amplifying the overall parameter-gradient magnitude. Together, these results support our interpretation of $\sigma(-z)$ as a meaningful loss-level training signal. Full training trajectories are provided in Appendix H.1.

Feedback versus predetermined β schedules. Table 3 compares LSC-DPO with several predetermined schedules. Linear, exponential, and cosine schedules yield only modest changes relative to DPO, whereas LSC-DPO achieves substantially higher LC and WR. This suggests that the improvement cannot be explained solely by using a time-varying β schedule and supports the role of state-dependent learning-signal feedback in LSC-DPO.

4.4 Initialization Robustness and Signal Compensation

Initialization sensitivity. Table 4 shows substantial sensitivity to β_0 across all methods. LSC-DPO achieves the highest performance at every tested initialization and degrades less severely, but still exhibits a 7.5-point LC spread. Figure 5 further shows how different β_0 values affect the

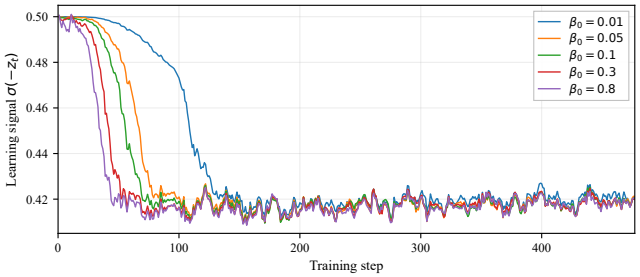


Figure 5: Learning-signal trajectories under different initial β_0. Curves are EMA-smoothed.

Table 4: Initialization sensitivity on AlpacaEval 2. We report length-controlled (LC) win rate (%). † denotes the reference setting. Full LC/WR results are provided in Appendix I.1.

	Initial coefficient β_0				
Method	$0.01^{\dagger}$	0.05	0.10	0.30	0.80
DPO	19.2	10.4	8.7	8.3	6.6
β-DPO	16.6	10.5	10.8	7.3	6.9
ε-DPO	19.6	15.6	12.9	8.8	6.9
LSC-DPO	**22.7**	**19.6**	**19.1**	**15.3**	**15.2**

learning-signal trajectory $\sigma(-z)$. Although LSC-DPO eventually regulates the trajectories toward a similar controlled level, larger β_0 causes a faster decay of $\sigma(-z)$ during the early stage of training.

Signal-budget compensation. We hypothesize that the residual performance variation across different initial coefficients is related to differences in the cumulative learning signal during the transient stage. For an arbitrary anchor run a, we define the cumulative signal deficit as

$$D_H(\beta_0) = \sum_{t=1}^{H} \left[\sigma(-z_t^a) - \sigma(-z_t^{(\beta_0)})\right], \tag{19}$$

where H denotes the end of the transient phase. A positive $D_H(\beta_0)$ indicates that the run accumulates a smaller learning signal than the anchor during this stage. We compensate for this deficit over the remaining $T - H$ steps by setting

$$\sigma(-z^*_{\text{comp}}) = \sigma(-z^*_a) + \frac{D_H(\beta_0)}{T-H}, \qquad z^*_{\text{comp}} = \log \frac{1-\sigma(-z^*_{\text{comp}})}{\sigma(-z^*_{\text{comp}})}, \tag{20}$$

where z^*_{comp} denotes the compensated target scaled margin corresponding to the compensated learning-signal target. Thus, a larger early-stage deficit leads to a higher compensated learning-signal target and, equivalently, a smaller z^*. Under ideal post-transient tracking, this rule matches the cumulative learning signal of the anchor run. We evaluate below whether this compensation also reduces the corresponding performance gap.

Compensation results. Table 5 shows that signal-budget compensation improves the LC win rate for all tested non-anchor initializations, with gains of up to 5.1 points, while reducing the LC spread from 4.4 to 1.8 points. These results indicate that adjusting the learning signal can partially compensate for the performance differences induced by coefficient initialization. Overall, the compensation substantially reduces, but does not eliminate, initialization sensitivity. Full results and additional analyses are provided in Appendix I.2.

5 Related Work

Reinforcement learning from human feedback. Reinforcement learning from human feedback (RLHF) has become a central approach for aligning large language models with human preferences and values (Bai et al., 2022; Christiano et al., 2017; Ouyang et al., 2022; Stiennon et al., 2020;

Table 5: Effect of full signal-budget compensation on AlpacaEval 2. We report length-controlled (LC) win rate (%).

	Baseline		Full compensation		
β_0	z^*	LC	z^*_{comp}	LC	(Δ)
0.05	0.400	19.6	0.359	**21.5**	(+1.9)
0.10	0.400	19.1	0.348	**21.1**	(+2.0)
0.30	0.400	15.3	0.333	**20.4**	(+5.1)
0.80	0.400	15.2	0.324	**19.7**	(+4.5)

Ziegler et al., 2019). A typical RLHF pipeline first collects preference comparisons, trains a reward model from human feedback (Chen et al., 2024; Gao et al., 2023; Havrilla et al., 2024; Luo et al., 2023), and then optimizes the policy with a reinforcement learning algorithm (Anthony et al., 2017; Schulman et al., 2017). Proximal policy optimization (PPO) is commonly used for this policy optimization stage (Schulman et al., 2017). Preference-based alignment has also been applied to improve helpfulness and safety, reduce toxic outputs, improve factuality, and control multiple objectives (Dai et al., 2023; Korbak et al., 2023; Zheng et al., 2023a; Tian et al., 2024; Wang et al., 2024).

Direct preference optimization and variants. DPO (Rafailov et al., 2023) has motivated a broad family of direct alignment methods that optimize language models from offline preference data without training an explicit reward model. Representative extensions include IPO (Azar et al., 2024), KTO (Ethayarajh et al., 2024), RRHF (Yuan et al., 2023), R-DPO (Park et al., 2024), ODPO (Amini et al., 2024), CPO (Xu et al., 2024), ORPO (Hong et al., 2024), SLiC-HF (Zhao et al., 2023), and SimPO (Meng et al., 2024). These methods extend preference optimization through different objective designs, reference-model formulations, and robustness mechanisms.

DPO training dynamics and coefficient control. Recent work has studied DPO optimization from training-dynamics and coefficient-control perspectives. Direct alignment methods can also exhibit overoptimization behavior as the policy moves away from the reference model (Rafailov et al., 2024). A related line of work studies this training behavior and mitigates it by updating the reference policy during training (Gorbatovski et al., 2024). In the coefficient-control direction, β-DPO (Wu et al., 2024b) updates β using empirical reward-margin statistics, while ε-DPO (Lee et al., 2025) perturbs the coefficient to reduce sensitivity to a fixed β. Related margin-adaptive approaches include AlphaDPO (Wu et al., 2024a) and γ-PO (Sun et al., 2025).

6 Conclusion

In this paper, we introduced Learning-Signal-Controlled Direct Preference Optimization (LSC-DPO), a feedback-control approach that regulates the learning signal of DPO during preference optimization. By analyzing the DPO objective from a loss-level geometric perspective, we identified the sigmoid factor as a learning signal that characterizes the local sensitivity of the objective and used it to design a control rule that regulates DPO training dynamics around a target learning-signal level. Our log-space analysis shows that, under a simple local stability condition, the feedback update can stably track the target learning-signal regime. Empirical results further show that LSC-DPO maintains more sustained learning-signal dynamics and achieves stronger alignment performance. We further showed that different coefficient initializations are related to learning-signal dynamics and derived a signal-budget compensation rule that substantially reduces the resulting performance variation.

Limitations and Future Work. Although the best-performing learning-signal levels are similar across the model families studied in this work, we do not yet provide a complete theory explaining why these target ranges emerge or when they transfer across models and datasets. In addition, although signal-budget compensation substantially reduces initialization-induced performance variation, it does not completely eliminate the remaining performance gap. Future work may develop a more complete theory of target learning-signal selection and signal-budget dynamics, and extend learning-signal control beyond DPO to other preference-optimization objectives.

AI USE STATEMENT

Generative AI tools were used in this work to edit the manuscript for grammar and readability, to check the formatting and the bibliography against the original sources, and to assist with writing and debugging the training and evaluation scripts. Generative AI tools were not used to generate synthetic data sets, to develop the theoretical framework, or to formulate or prove the mathematical claims; the remaining tasks with required disclosure are not applicable to this work. All AI-assisted text was reviewed and revised by the authors, AI-assisted code was verified and tested by the authors, and every citation was checked against the original publication. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies preference optimization for language-model alignment. All experiments use publicly available datasets (UltraChat-200k, UltraFeedback, and Anthropic-HH) and open-weight base models under their respective licenses, and no human-subject data were collected. The Anthropic-HH harmless-base subset contains prompts designed to elicit harmful responses; we use it only to evaluate harmlessness. As with any preference-optimization method, LSC-DPO can in principle be used to align a model toward undesirable preferences; our method does not change this risk relative to DPO. We do not release any new dataset.

REPRODUCIBILITY STATEMENT

The complete LSC-DPO procedure is given in Algorithm 1 (Appendix B). Models, datasets, hyperparameters, and compute are listed in Appendix E.1, and the evaluation protocols, judge models, and decoding settings in Appendix E.2. The assumptions and full proofs of the stability results are in Appendices C and D. The predetermined β schedules and the signal-budget compensation targets are specified in Appendices H.2 and I.2.

REFERENCES

Afra Amini, Tim Vieira, and Ryan Cotterell. Direct preference optimization with an offset. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9954–9972, 2024.

Thomas Anthony, Zheng Tian, and David Barber. Thinking fast and slow with deep learning and tree search. *Advances in Neural Information Processing Systems*, 30, 2017.

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.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

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

Lichang Chen, Chen Zhu, Jiuhai Chen, Davit Soselia, Tianyi Zhou, Tom Goldstein, Heng Huang, Mohammad Shoeybi, and Bryan Catanzaro. ODIN: Disentangled reward mitigates hacking in RLHF. In *International Conference on Machine Learning*, volume 235, pp. 7935–7952. PMLR, 2024.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. UltraFeedback: Boosting language models with scaled AI feedback. *arXiv preprint arXiv:2310.01377*, 2023.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe RLHF: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3029–3051, 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. Model alignment as prospect theoretic optimization. In *International Conference on Machine Learning*, volume 235, pp. 12634–12651. PMLR, 2024.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.

Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, Nahyeon Ryu, and Marc Dymetman. Aligning language models with preferences through f-divergence minimization. In *International Conference on Machine Learning*, volume 202, pp. 11546–11583. PMLR, 2023.

Alexey Gorbatovski, Boris Shaposhnikov, Alexey Malakhov, Nikita Surnachev, Yaroslav Aksenov, Ian Maksimov, Nikita Balagansky, and Daniil Gavrilov. Learn your reference model for real good alignment. *arXiv preprint arXiv:2404.09656*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Alex Havrilla, Sharath Raparthy, Christoforus Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, and Roberta Raileanu. GLoRe: When, where, and how to improve LLM reasoning via global and local refinements. *arXiv preprint arXiv:2402.10963*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Jiwoo Hong, Noah Lee, and James Thorne. ORPO: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11170–11189, 2024.

Natasha Jaques, Shixiang Gu, Dzmitry Bahdanau, José Miguel Hernández-Lobato, Richard E Turner, and Douglas Eck. Sequence tutor: Conservative fine-tuning of sequence generation models with KL-control. In *International Conference on Machine Learning*, pp. 1645–1654. PMLR, 2017.

Natasha Jaques, Judy Hanwen Shen, Asma Ghandeharioun, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Human-centric dialog training via offline reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3985–4003, 2020.

Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Vinayak Bhalerao, Christopher Buckley, Jason Phang, Samuel R Bowman, and Ethan Perez. Pretraining language models with human preferences. In *International Conference on Machine Learning*, pp. 17506–17533. PMLR, 2023.

Sangkyu Lee, Janghoon Han, Hosung Song, Stanley Jungkyu Choi, Honglak Lee, and Youngjae Yu. KL penalty control via perturbation for direct preference optimization. *Advances in Neural Information Processing Systems*, 38:168095–168121, 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.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, 2022.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. WizardMath: Empowering mathematical reasoning for large language models via reinforced Evol-Instruct. *arXiv preprint arXiv:2308.09583*, 2023.

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. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4998–5017, 2024.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

Jan Peters and Stefan Schaal. Reinforcement learning by reward-weighted regression for operational space control. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 745–750, 2007.

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.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, W Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. *Advances in Neural Information Processing Systems*, 37:126207–126242, 2024.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial Winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021. doi: 10.1145/3474381.

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

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Jie Sun, Junkang Wu, Jiancan Wu, Zhibo Zhu, Xingyu Lu, Jun Zhou, Lintao Ma, and Xiang Wang. Robust preference optimization via dynamic target margins. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5399–5416, 2025.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. Fine-tuning language models for factuality. In *The Twelfth International Conference on Learning Representations*, 2024.

Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. Arithmetic control of LLMs for diverse user preferences: Directional preference alignment with multi-objective rewards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8642–8655, 2024.

Junkang Wu, Xue Wang, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. AlphaDPO: Adaptive reward margin for direct preference optimization. *arXiv preprint arXiv:2410.10148*, 2024a.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. β-DPO: Direct preference optimization with dynamic β. *Advances in Neural Information Processing Systems*, 37:129944–129966, 2024b.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In *International Conference on Machine Learning*, pp. 55204–55224. PMLR, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. RRHF: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*, 2023.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, 2019.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. SLiC-HF: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023.

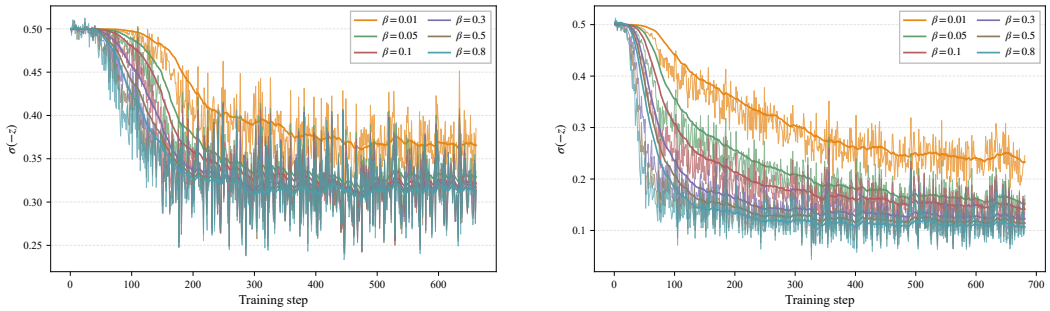
Chujie Zheng, Pei Ke, Zheng Zhang, and Minlie Huang. Click: Controllable text generation with sequence likelihood contrastive learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1022–1040, 2023a.

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, 2023b.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A Learning Signal $\sigma(-z)$ under Fixed β

We report the raw per-step trajectories of the learning signal under fixed-β DPO with different β values. Figures 6a and 6b show that $\sigma(-z)$ generally decreases during training on both Anthropic-HH subsets.



(a) harmless-base (b) helpful-base

Figure 6: Learning signal under DPO. Raw per-step trajectories of $\sigma(-z)$ are shown for different fixed β values on Anthropic-HH: (a) harmless-base and (b) helpful-base.

B LSC-DPO Algorithm

Algorithm 1 summarizes the complete training procedure of LSC-DPO, including computation of the learning signal, controller activation, and the subsequent coefficient update described in Section 3.2.

Algorithm 1 Learning-Signal-Controlled Direct Preference Optimization

Require: policy π_θ, reference π_{ref}, initial β_0, z^*, η, τ_{start}
1: $\beta \leftarrow \beta_0$; control is inactive
2: **while** not converged **do**
3: Sample a batch of preference triples $(x, y_+, y_-) \sim \mathcal{D}$
4: Compute margin m, scaled margin $z = \beta m$, and signal $\sigma(-z)$
5: Activate control if $\sigma(-z) \leq \tau_{\text{start}}$
6: If control is active, update $\beta \leftarrow \beta \exp[-\eta(z - z^*)]$
7: Update π_θ using the DPO loss with the current β
8: **end while**
9: **return** aligned policy π_θ

C Proofs for the Stability Analysis

In this appendix, we provide the formal statements and detailed proofs for the local stability and bounded tracking results summarized in Section 3.3. Recall that the log-space update is

$$x_{k+1} = F_k(x_k), \qquad F_k(x) := x - \eta\,(e^x m_k - z^*)\,, \tag{21}$$

and that the instantaneous equilibrium is

$$x_k^f = \log \frac{z^*}{m_k}. \tag{22}$$

C.1 Proof of Local Stability

The formal local-stability result is stated as follows.

Proposition 1 (Local stability)**.** *Assume $m_k > 0$ and $z^* > 0$. The instantaneous equilibrium $x_k^f = \log(z^*/m_k)$ is locally asymptotically stable whenever*

$$0 < \eta z^* < 2. \tag{23}$$

The corresponding equilibrium linearization factor is

$$\rho_0 := |1 - \eta z^*| < 1. \tag{24}$$

This result follows by linearizing F_k around x_k^f, where

$$F_k'(x_k^f) = 1 - \eta z^*. \tag{25}$$

Proof. Since $m_k > 0$ and $z^* > 0$, the instantaneous equilibrium is well defined and satisfies

$$e^{x_k^f} m_k = z^*. \tag{26}$$

The derivative of the step-dependent map is

$$F_k'(x) = 1 - \eta e^x m_k. \tag{27}$$

Evaluating this derivative at the equilibrium gives

$$F_k'(x_k^f) = 1 - \eta e^{x_k^f} m_k = 1 - \eta z^*. \tag{28}$$

For a one-dimensional discrete-time dynamical system, a fixed point is locally asymptotically stable if the derivative at the fixed point has magnitude smaller than one. Hence,

$$|F_k'(x_k^f)| < 1 \quad \Longleftrightarrow \quad |1 - \eta z^*| < 1. \tag{29}$$

Solving this inequality yields

$$0 < \eta z^* < 2. \tag{30}$$

Under this condition, the equilibrium linearization factor is

$$\rho_0 := |F_k'(x_k^f)| = |1 - \eta z^*| < 1, \tag{31}$$

which proves the proposition. □

C.2 Bounded Tracking Theorem and Proof

Theorem 1 (Bounded tracking)**.** *Assume the local stability condition in Eq. 17 holds, and define*

$$\rho_0 := |1 - \eta z^*| < 1.$$

For any ρ satisfying

$$\rho_0 < \rho < 1,$$

there exists a radius $r > 0$ such that, for every k,

$$|F_k'(x)| \le \rho \qquad \textit{whenever } |x - x_k^f| \le r. \tag{32}$$

Suppose that there exists a step T such that

$$|x_T - x_T^f| \le r, \tag{33}$$

and that, for some constant $\bar{\delta} \ge 0$ and all $k \ge T$,

$$|x_{k+1}^f - x_k^f| \le \bar{\delta}, \qquad \bar{\delta} \le (1 - \rho) r. \tag{34}$$

Then the iterates remain within the local contraction neighborhood,

$$|x_k - x_k^f| \le r, \qquad \forall k \ge T,$$

and satisfy

$$|x_k - x_k^f| \le \rho^{k-T} |x_T - x_T^f| + \frac{\bar{\delta}}{1 - \rho}, \qquad \forall k \ge T. \tag{35}$$

The complete proof is given below. An interpretation of the resulting tracking bound is provided in Appendix C.3.

We define the tracking error by

$$e_k := x_k - x_k^f. \tag{36}$$

Lemma 1 (One-step tracking recursion)**.** *For each step k, there exists some ξ_k between x_k and x_k^f such that*

$$e_{k+1} = F_k'(\xi_k)e_k + \left(x_k^f - x_{k+1}^f\right). \tag{37}$$

Consequently,

$$|e_{k+1}| \le |F_k'(\xi_k)|\,|e_k| + |x_{k+1}^f - x_k^f|. \tag{38}$$

Proof. By definition,

$$e_{k+1} = x_{k+1} - x_{k+1}^f. \tag{39}$$

Using the update $x_{k+1} = F_k(x_k)$ and adding and subtracting x_k^f, we obtain

$$\begin{aligned} e_{k+1} &= F_k(x_k) - x_{k+1}^f \\ &= \left(F_k(x_k) - F_k(x_k^f)\right) + \left(x_k^f - x_{k+1}^f\right), \end{aligned} \tag{40}$$

where we used $F_k(x_k^f) = x_k^f$. By the mean value theorem, there exists some ξ_k between x_k and x_k^f such that

$$F_k(x_k) - F_k(x_k^f) = F_k'(\xi_k)(x_k - x_k^f) = F_k'(\xi_k)e_k. \tag{41}$$

Substituting this identity gives Eq. 37. Taking absolute values gives Eq. 38. □

Proof of Theorem 1. From Proposition 1, the derivative at the instantaneous equilibrium satisfies

$$\rho_0 := |F_k'(x_k^f)| = |1 - \eta z^*| < 1. \tag{42}$$

Choose any ρ such that

$$\rho_0 < \rho < 1. \tag{43}$$

For a displacement e from the instantaneous equilibrium,

$$F_k'(x_k^f + e) = 1 - \eta e^{x_k^f + e} m_k = 1 - \eta z^* e^e. \tag{44}$$

The right-hand side is independent of m_k. Since it is continuous in e and equals $1 - \eta z^*$ at $e = 0$, there exists a radius $r > 0$ such that

$$|F_k'(x_k^f + e)| \le \rho, \qquad \text{for all } |e| \le r \tag{45}$$

and for all $k \ge T$.

We now show that this neighborhood is invariant. By assumption,

$$|e_T| \le r. \tag{46}$$

Suppose inductively that $|e_k| \le r$. Since ξ_k lies between x_k and x_k^f, it also satisfies

$$|\xi_k - x_k^f| \le |x_k - x_k^f| = |e_k| \le r. \tag{47}$$

Therefore,

$$|F_k'(\xi_k)| \le \rho. \tag{48}$$

Using Lemma 1 and the bounded-drift assumption gives

$$|e_{k+1}| \le \rho|e_k| + \bar{\delta}. \tag{49}$$

Since $\bar{\delta} \le (1 - \rho)r$,

$$|e_{k+1}| \le \rho r + (1 - \rho)r = r. \tag{50}$$

Hence, by induction,

$$|e_k| \le r, \qquad \forall k \ge T. \tag{51}$$

Thus, the local contraction bound used in Eq. 49 remains valid along the entire tracked trajectory.

Unrolling Eq. 49 gives

$$|e_k| \le \rho^{k-T}|e_T| + \bar{\delta} \sum_{j=0}^{k-T-1} \rho^j. \tag{52}$$

Since $0 \le \rho < 1$,

$$\sum_{j=0}^{k-T-1} \rho^j = \frac{1-\rho^{k-T}}{1-\rho} \le \frac{1}{1-\rho}. \tag{53}$$

Substituting this bound into Eq. 52 yields

$$|e_k| \le \rho^{k-T}|e_T| + \frac{\bar{\delta}}{1-\rho}, \qquad \forall k \ge T. \tag{54}$$

Since $e_k = x_k - x_k^f$, this proves the theorem. □

C.3 Interpretation of the Bound

The bound

$$|e_k| \le \rho^{k-T}|e_T| + \frac{\bar{\delta}}{1-\rho} \tag{55}$$

contains two terms. The first term, $\rho^{k-T}|e_T|$, is a transient error that decays geometrically. The second term, $\bar{\delta}/(1-\rho)$, is a persistent tracking-error bound controlled by the maximum equilibrium drift. Therefore, the log-space β dynamics remains close to the instantaneous equilibrium trajectory as long as the equilibrium drift is bounded and the trajectory stays within the local contraction neighborhood.

D A Sufficient Condition for Bounded Equilibrium Drift

The bounded tracking theorem assumes that the moving equilibrium satisfies

$$|x_{k+1}^f - x_k^f| \le \bar{\delta}.$$

We show that this condition follows from a simple regularity condition on the margin statistic.

Lemma 2 (Sufficient condition for bounded equilibrium drift)**.** *Assume that there exist constants $T \in \mathbb{N}$, $m_{\min} > 0$, and $\varepsilon \ge 0$ such that for all $k \ge T$,*

$$m_k \ge m_{\min}, \qquad |m_{k+1} - m_k| \le \varepsilon.$$

Then

$$|x_{k+1}^f - x_k^f| \le \frac{\varepsilon}{m_{\min}}, \qquad \forall k \ge T.$$

Proof. Using $x_k^f = \log(z^*/m_k) = \log z^* - \log m_k$, we have

$$|x_{k+1}^f - x_k^f| = |\log m_{k+1} - \log m_k|.$$

By the mean value theorem, for some ζ_k between m_k and m_{k+1},

$$|\log m_{k+1} - \log m_k| = \frac{1}{\zeta_k}|m_{k+1} - m_k|.$$

Since $m_k, m_{k+1} \ge m_{\min} > 0$, we have $\zeta_k \ge m_{\min}$. Therefore,

$$|x_{k+1}^f - x_k^f| \le \frac{1}{m_{\min}}|m_{k+1} - m_k| \le \frac{\varepsilon}{m_{\min}}.$$

□

This lemma provides an explicit bound on the equilibrium drift whenever the margin statistic stays bounded away from zero and changes gradually across training steps. In particular, the drift condition in Theorem 1 is satisfied whenever $\varepsilon/m_{\min} \le (1-\rho)r$.

E Implementation Details

E.1 Models and Training Datasets

UltraFeedback experiments. For the UltraFeedback experiments, we follow the SimPO (Meng et al., 2024) training setup. We first perform supervised fine-tuning of the Meta-Llama-3-8B[1] and Qwen3-8B-Base[2] models on UltraChat-200k[3] (Ding et al., 2023); the resulting SFT checkpoints are used as the policy initializations and reference models for preference optimization on the UltraFeedback binarized[4] (Cui et al., 2023) dataset. We also follow the SimPO protocol for common training settings, including the learning rate, scheduler, and optimizer, and keep them fixed across experiments. Table 6 reports these fixed training settings and the LSC-DPO control parameters, with bold entries denoting the main settings used in Table 1.

Table 6: Training configuration of LSC-DPO on UltraFeedback.

Configuration	Llama3-8B	Qwen3-8B
Model	meta-llama/Meta-Llama-3-8B	Qwen/Qwen3-8B-Base
Optimizer	AdamW	AdamW
Epoch	1	1
Batch size	128	128
Learning rate	5e-7	5e-7
Scheduler	cosine	cosine
Warmup ratio	0.1	0.1
Weight decay	0	0
Precision	BF16	BF16
β_0	0.01	0.01
η	0.1	0.1
z^*	$\{0.20, 0.30, 0.35, \mathbf{0.40}, 0.45, 0.50, 0.60, 0.70\}$	$\{0.20, 0.30, \mathbf{0.35}, 0.40, 0.45, 0.50, 0.60, 0.70\}$
τ_{start}	$\{0.40, 0.43, \mathbf{0.45}\}$	$\{0.40, 0.43, \mathbf{0.45}\}$
Device	4×H200 GPUs	4×H200 GPUs

Anthropic-HH experiments. For the Anthropic-HH (Bai et al., 2022)[5] experiments, we use both the harmless-base and helpful-base subsets. We follow DPO (Rafailov et al., 2023) and β-DPO (Wu et al., 2024b) for the batch size (64) and single epoch, ε-DPO (Lee et al., 2025) for the optimizer, scheduler and warmup, and keep the learning rate of 5×10^{-7} shared with our UltraFeedback runs. For method-specific hyperparameters of β-DPO and ε-DPO, we use the recommended values reported in the respective papers. Tables 7 and 8 summarize the SFT-stage and preference-optimization-stage configurations, respectively.

Table 7: SFT training configuration for Anthropic-HH.

Configuration	Llama3-8B	Qwen3-8B
Model	meta-llama/Meta-Llama-3-8B	Qwen/Qwen3-8B-Base
Epochs	1	1
Batch size	64	64
Scheduler	cosine	cosine
Warmup ratio	0.1	0.1
Weight decay	0	0
Learning rate	2e-5	2e-5
Precision	BF16	BF16
Device	4×H200 GPUs	4×H200 GPUs

[1] `https://huggingface.co/meta-llama/Meta-Llama-3-8B`; Llama 3 Community License.
[2] `https://huggingface.co/Qwen/Qwen3-8B-Base`; Apache 2.0 License.
[3] `https://huggingface.co/datasets/HuggingFaceH4/ultrachat_200k`; MIT License.
[4] `https://huggingface.co/datasets/HuggingFaceH4/ultrafeedback_binarized`; MIT License.
[5] `https://huggingface.co/datasets/Anthropic/hh-rlhf`; MIT License.

Table 8: Training configurations of LSC-DPO on Anthropic-HH.

Configuration	Llama3-8B	Qwen3-8B
Model	meta-llama/Meta-Llama-3-8B	Qwen/Qwen3-8B-Base
Epochs	1	1
Batch size	64	64
Scheduler	cosine	cosine
Warmup ratio	0.1	0.1
Weight decay	0	0
Learning rate	5e-7	5e-7
Precision	BF16	BF16
β_0	0.1	0.1
τ_{start}	$\{0.40, 0.43, \mathbf{0.45}, 0.48, 0.50\}$	$\{0.40, 0.43, \mathbf{0.45}, 0.48, 0.50\}$
z^*	$\{0.35, \mathbf{0.40}, 0.45, 0.60, 0.80, 0.85\}$	$\{0.35, \mathbf{0.40}, 0.45, 0.60, 0.80, 0.85\}$
η	{0.01, 0.05, 0.1, 0.3, 0.5, 1, 5, 8}	{0.01, 0.05, 0.1, 0.3, 0.5, 1, 5, 8}
Device	4×H200 GPUs	4×H200 GPUs

Computation environment. All training experiments in this paper are based on the Hugging Face alignment-handbook[6] and TRL[7]. In our setup with 4 H200 GPUs, each LSC-DPO training run on UltraFeedback completes within two hours, while each run on the Anthropic-HH helpful-base or harmless-base subset completes within one hour.

E.2 Evaluation Benchmarks and Baselines

UltraFeedback evaluation. For the UltraFeedback experiments, we follow the SimPO open-alignment evaluation protocol (Meng et al., 2024) and the official benchmark settings of AlpacaEval 2 (Dubois et al., 2024) and MT-Bench (Zheng et al., 2023b). We compare LSC-DPO with SFT, DPO (Rafailov et al., 2023), β-DPO (Wu et al., 2024b), ε-DPO (Lee et al., 2025), R-DPO (Park et al., 2024), IPO (Azar et al., 2024), KTO (Ethayarajh et al., 2024), CPO (Xu et al., 2024), ORPO (Hong et al., 2024), SLiC-HF (Zhao et al., 2023), and SimPO (Meng et al., 2024). All methods are evaluated with the same generation settings and evaluation pipeline to ensure a fair comparison. On AlpacaEval 2, we report both the raw win rate and the length-controlled win rate, with the length-controlled metric used as the primary metric because preference optimization can increase response length. On MT-Bench, we report the average score over both turns.

Anthropic-HH evaluation. For the Anthropic-HH experiments, we follow the evaluation protocol used in DPO (Rafailov et al., 2023) and β-DPO (Wu et al., 2024b). We evaluate harmless-base and helpful-base separately by comparing each model response with the HH chosen response, using `gpt-4-0613` as the judge with randomized response order. We report the win rate of the model response. All methods use the same decoding configuration, with temperature 0.7 and top-p sampling ($p = 0.9$), to ensure a fair comparison.

Control-parameter selection. We selected the LSC-DPO control parameters using a separate Anthropic-HH development subset before the main AlpacaEval 2 evaluation. This preliminary study identified a coarse candidate region around $z^* \in [0.35, 0.45]$ and $\tau_{\text{start}} \in [0.40, 0.45]$. The AlpacaEval 2 results in Appendix J.1 are therefore reported as a target-sensitivity analysis rather than as an unrestricted benchmark-specific hyperparameter search. After the target and training settings were fixed, the same trained checkpoints were evaluated on MT-Bench without additional tuning.

Arena-Hard reproducibility considerations. We also considered Arena-Hard (Li et al., 2024) as an additional open-ended benchmark, but do not include it in the main evaluation because prior Arena-Hard v0.1 results for SimPO (Meng et al., 2024) and ε-DPO (Lee et al., 2025) used `gpt-4-1106-preview` as the judge. OpenAI has since retired this model, so the original protocol cannot be reproduced in our evaluation environment. To examine how the released checkpoints behave under currently available judging setups, we re-evaluated SimPO's released Llama3-8B-base preference-optimization checkpoints under two settings: Arena-Hard v0.1 judged with

[6] `https://github.com/huggingface/alignment-handbook`; Apache 2.0 License.
[7] `https://github.com/huggingface/trl`; Apache 2.0 License.

`gpt-4.1`, and Arena-Hard v2.0 hard-prompt judged with `gpt-4.1`. Table 9 shows that all evaluated preference-optimization methods score below the SFT baseline under these updated settings, in contrast to their behavior on AlpacaEval 2 and MT-Bench. The relative ordering also differs substantially from prior Arena-Hard reports, suggesting that results obtained under the updated judging setups may not be directly comparable to those reported under the original protocol. We therefore use AlpacaEval 2 and MT-Bench as the main evaluations.

Table 9: Re-evaluation of the released Llama3-8B-base preference-optimization checkpoints from the SimPO study on Arena-Hard. Panel (a) reports Arena-Hard v0.1 results using `gpt-4.1`; Panel (b) reports Arena-Hard v2.0 hard-prompt results using `gpt-4.1`. The original v0.1 judge, `gpt-4-1106-preview`, is no longer available in our evaluation environment.

Model	Score (%)	95% CI (%)
`gpt-4-0314` (ref.)	50.0	$(-0.0/+0.0)$
SFT	26.4	$(-1.9/+1.7)$
SFT + CPO	22.4	$(-1.3/+1.6)$
SFT + SimPO	20.1	$(-1.7/+2.1)$
SFT + DPO	16.3	$(-1.4/+1.6)$
SFT + R-DPO	14.5	$(-1.3/+1.3)$
SFT + IPO	13.4	$(-1.3/+1.5)$
SFT + ORPO	12.4	$(-1.2/+1.2)$
SFT + KTO	10.3	$(-1.1/+1.1)$
SFT + SLiC-HF	7.7	$(-0.9/+0.8)$

(a) Arena-Hard v0.1, judge: `gpt-4.1`.

Model	Score (%)	95% CI (%)
`o3-mini` (ref.)	50.0	$(-0.0/+0.0)$
SFT	9.2	$(-0.8/+0.8)$
SFT + CPO	7.4	$(-0.7/+0.9)$
SFT + DPO	2.3	$(-0.4/+0.4)$
SFT + SimPO	1.8	$(-0.4/+0.4)$
SFT + IPO	1.4	$(-0.3/+0.4)$
SFT + R-DPO	1.4	$(-0.3/+0.3)$
SFT + ORPO	1.1	$(-0.4/+0.4)$
SFT + KTO	0.7	$(-0.2/+0.3)$
SFT + SLiC-HF	0.5	$(-0.2/+0.2)$

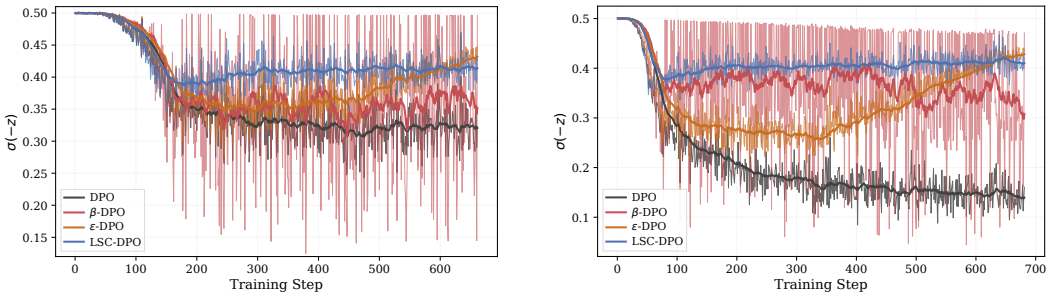
(b) Arena-Hard v2.0 hard-prompt, judge: `gpt-4.1`.

F Additional Learning-Signal Trajectories on Anthropic-HH

This appendix provides additional learning-signal trajectories for the Anthropic-HH analysis in Section 4.2. We first report the raw per-step trajectories corresponding to the EMA-smoothed Llama3-8B results, and then provide the corresponding Qwen3-8B results as a cross-model comparison.

F.1 Raw Learning-Signal Trajectories on Llama3-8B

Figures 7a and 7b show the raw per-step trajectories together with their EMA-smoothed trends for Llama3-8B. Compared with the smoothed curves reported in Section 4.2, the raw trajectories make the short-term fluctuations in the learning signal more visible, especially for β-DPO. This behavior suggests that dynamically adjusting β based on margin statistics can introduce large short-term fluctuations in the learning signal. In contrast, LSC-DPO maintains a more stable smoothed trajectory while avoiding the persistent low-signal regime observed under DPO, particularly on the helpful-base subset.

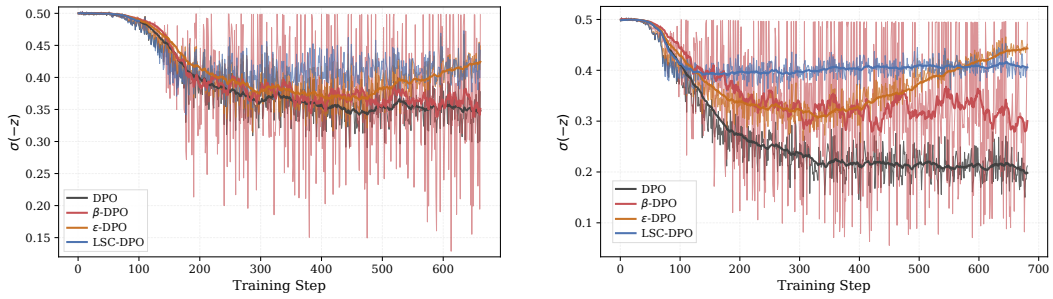


(a) harmless-base (b) helpful-base

Figure 7: Raw and EMA-smoothed trajectories of the learning signal $\sigma(-z)$ on Anthropic-HH with Llama3-8B. Faint curves show raw per-step values, while bold curves show EMA-smoothed trends.

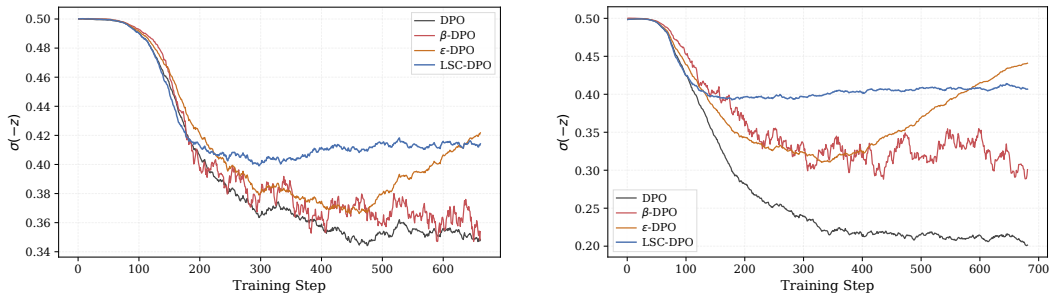
F.2 Learning-Signal Trajectories on Qwen3-8B

Figures 8 and 9 report the corresponding learning-signal trajectories for Qwen3-8B. The same qualitative pattern observed with Llama3-8B also appears for Qwen3-8B. DPO drives the learning signal toward a lower regime as training proceeds, while β-DPO exhibits larger fluctuations. ε-DPO partially recovers the signal in later stages. Across both harmless-base and helpful-base, LSC-DPO maintains a higher and more stable learning signal, providing additional evidence that the proposed control rule is not specific to a single model family.



(a) harmless-base (b) helpful-base

Figure 8: Raw and EMA-smoothed trajectories of the learning signal $\sigma(-z)$ on Anthropic-HH with Qwen3-8B. Faint curves denote raw per-step values, while bold curves denote EMA-smoothed trends.



(a) harmless-base (b) helpful-base

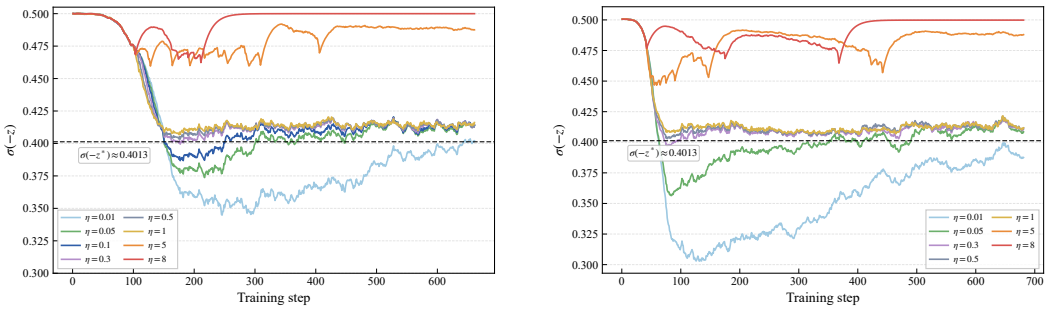
Figure 9: EMA-smoothed trajectories of the learning signal $\sigma(-z)$ on Anthropic-HH with Qwen3-8B.

G Stability under Different Update Rates

We further examine the stability behavior of LSC-DPO by varying the control rate η while fixing the target scaled margin to $z^* = 0.4$. Under this setting, the target learning signal is $\sigma(-z^*) \approx 0.4013$, and changing η directly changes the stability-relevant product ηz^* in Proposition 1. Figures 10 and 11 show the EMA-smoothed trajectories of the learning signal $\sigma(-z)$ on Anthropic-HH subsets. Very small values of η react slowly to deviations from the target signal. For example, when $\eta = 0.01$, the signal often drops below the target level and recovers only gradually. In contrast, moderate values such as $\eta = 0.3$, 0.5, and 1.0 provide more responsive correction and keep the signal close to the target after the initial adjustment phase across both model families and subsets.

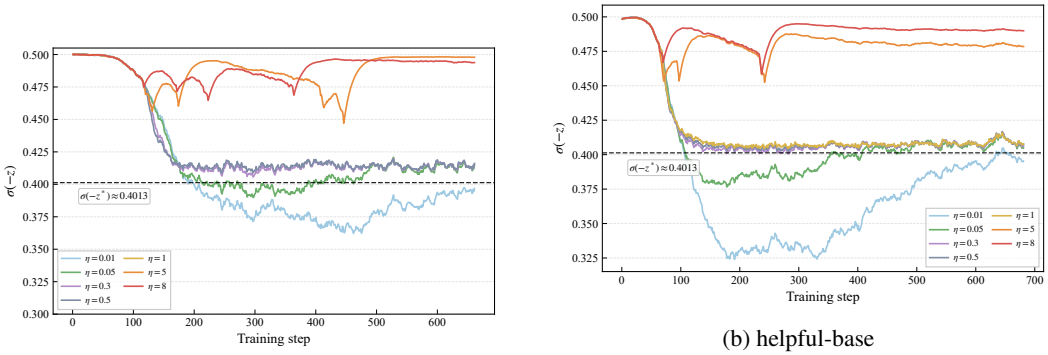
With $z^* = 0.4$, the local stability condition $0 < \eta z^* < 2$ reduces to $\eta < 5$. At $\eta = 5$, which lies exactly on the boundary with $|1 - \eta z^*| = 1$, the local contraction is lost and the learning-signal trajectory exhibits substantially larger fluctuations. In contrast, $\eta = 8$ violates the stability condition and fails to reliably track the target signal. These results support the feedback-control

interpretation of LSC-DPO: appropriately bounded control rates maintain the learning signal near the target regime, whereas overly large rates can destabilize the signal dynamics.



(a) harmless-base (b) helpful-base

Figure 10: Learning signal $\sigma(-z)$ trajectories of LSC-DPO under different control rates η with Llama3-8B on Anthropic-HH subsets. The dashed line marks the target signal $\sigma(-z^*) \approx 0.4013$.



(a) harmless-base (b) helpful-base

Figure 11: Learning signal $\sigma(-z)$ trajectories of LSC-DPO under different control rates η with Qwen3-8B on Anthropic-HH subsets. The dashed line marks the target signal $\sigma(-z^*) \approx 0.4013$.

H Additional Mechanistic Validation of Learning-Signal Control

The mechanistic experiments in Section 4.3 use the same Llama3-8B SFT checkpoint and UltraFeedback training configuration as the main experiments. Unless otherwise specified, all training and evaluation settings are kept fixed across the compared methods.

H.1 Learning Signal and Realized Optimization Dynamics

We compute the curvature factor as $c_t = \sigma(-z_t)(1-\sigma(-z_t))$ and report the gradient norm measured before gradient clipping. LSC-DPO control is activated at step 84. For visualization, solid curves show EMA-smoothed trajectories, whereas all phase-averaged statistics in Table 2 are computed from the original unsmoothed measurements.

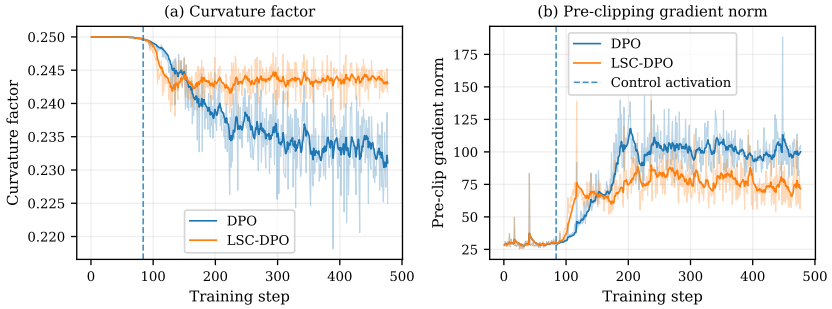


Figure 12: Optimization diagnostics for DPO and LSC-DPO throughout training. Left: curvature factor. Right: pre-clipping gradient norm. The vertical dashed line marks LSC-DPO control activation at step 84.

Before control activation, DPO and LSC-DPO follow nearly identical trajectories. After activation, LSC-DPO consistently maintains a larger curvature factor, whereas DPO continues moving toward a lower-curvature regime. The pre-clipping gradient trajectories also separate after control activation. These full trajectories confirm that the phase-averaged differences reported in Table 2 persist throughout the controlled stage rather than arising from a small number of training steps.

H.2 Feedback versus Predetermined β Schedules

For the predetermined-schedule baselines in Table 3, we vary β according to linear, cosine, and exponential schedules fixed before training. Let $t = 0, \ldots, T-1$ index the β updates and $f_t = t/(T-1)$. All schedules start from β_0 and end at β_{end}, and are defined as

$$\beta_t^{\text{lin}} = \beta_0 + f_t(\beta_{\text{end}} - \beta_0), \tag{56}$$

$$\beta_t^{\cos} = \beta_0 + \frac{1}{2}\left(1 - \cos(\pi f_t)\right)(\beta_{\text{end}} - \beta_0), \tag{57}$$

$$\beta_t^{\exp} = \exp\left(\log \beta_0 + f_t(\log \beta_{\text{end}} - \log \beta_0)\right). \tag{58}$$

All predetermined schedules use the same initial and final coefficient values as LSC-DPO, differ only in how β evolves between these endpoints, and update the coefficient once per gathered microbatch.

I Additional Results on Initialization Robustness and Signal Compensation

The experiments in Section 4.4 use the same Llama3-8B, UltraFeedback, training, and evaluation settings as those in Section 4.3; only β_0 and, where applicable, the compensation target are varied.

I.1 Initialization Sensitivity: Full LC/WR Results

Table 10 reports the complete AlpacaEval 2 results for the initialization-sensitivity study, including both length-controlled (LC) win rate and standard win rate (WR). The $\beta_0 = 0.01$ configuration corresponds to the reference setting used in the main experiments.

Table 10: Full initialization-sensitivity results on AlpacaEval 2. We report length-controlled win rate (LC) and standard win rate (WR), both in percentage points. † denotes the reference setting.

	$\beta_0 = 0.01^\dagger$		$\beta_0 = 0.05$		$\beta_0 = 0.10$		$\beta_0 = 0.30$		$\beta_0 = 0.80$	
Method	LC	WR	LC	WR	LC	WR	LC	WR	LC	WR
DPO	19.2	17.7	10.4	7.7	8.7	5.8	8.3	4.8	6.6	3.9
β-DPO	16.6	16.5	10.5	7.8	10.8	7.8	7.3	4.7	6.9	3.8
ε-DPO	19.6	17.6	15.6	13.2	12.9	9.8	8.8	5.9	6.9	4.5
LSC-DPO	**22.7**	**21.1**	**19.6**	**18.3**	**19.1**	**17.2**	**15.3**	**14.1**	**15.2**	**13.4**

I.2 Additional Signal-Budget Compensation Results

We provide additional results for the signal-budget compensation experiments described in Section 4.4. The $\beta_0 = 0.01$ run is used as the anchor configuration, and the baseline LSC-DPO target is $z^* = 0.400$. For each non-anchor initialization, the full-compensation target is determined by the signal-budget rule in Eq. 20. For $\beta_0 = 0.10$ and $\beta_0 = 0.80$, we additionally evaluate a half-compensation setting, obtained by applying half of the target shift from the baseline $z^* = 0.400$ toward the corresponding full-compensation target. All other training and evaluation settings are kept unchanged.

Table 11: Detailed signal-budget compensation results on AlpacaEval 2. LC and WR denote length-controlled and standard win rates, respectively.

Setting	z^*	LC (%)	WR (%)
$\beta_0 = 0.05$	0.359	21.5	20.1
$\beta_0 = 0.10$ (half)	0.374	20.7	18.9
$\beta_0 = 0.10$ (full)	0.348	21.1	19.5
$\beta_0 = 0.30$	0.333	20.4	19.4
$\beta_0 = 0.80$ (half)	0.362	17.4	15.8
$\beta_0 = 0.80$ (full)	0.324	19.7	18.3

Compared with the corresponding uncompensated runs reported in Table 10, full compensation improves both LC and WR across all tested non-anchor initializations. For the two initializations where both half and full compensation are evaluated, half compensation yields intermediate performance, while full compensation produces larger improvements. This pattern is consistent with the intended effect of progressively correcting the initialization-induced learning-signal deficit.

To directly examine whether the target adjustment produces the intended signal-level correction, we compare the cumulative learning-signal recovery under half and full compensation. The required recovery is defined relative to the anchor run and represents the cumulative signal deficit that must be recovered to match its signal budget.

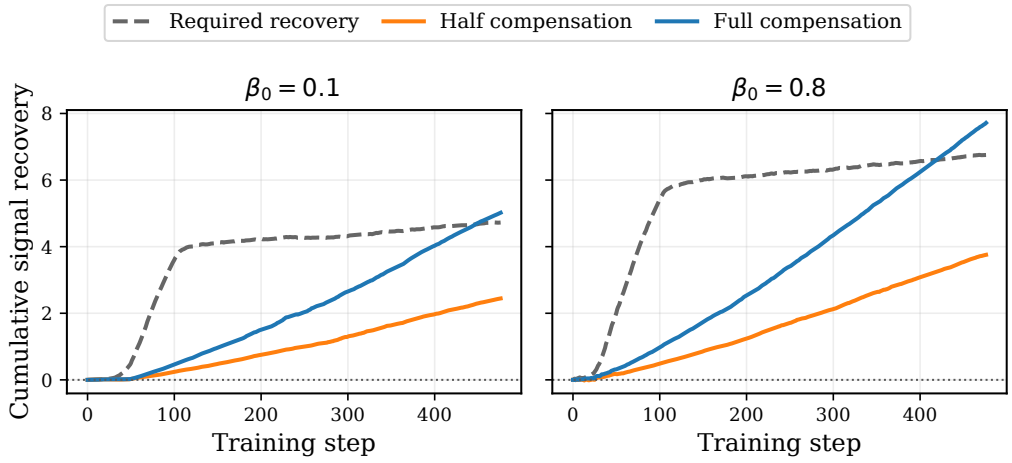


Figure 13: Cumulative learning-signal recovery under half and full signal-budget compensation for $\beta_0 = 0.10$ and $\beta_0 = 0.80$. The dashed curves denote the recovery required to match the anchor run ($\beta_0 = 0.01$), while the solid curves show the recovery achieved by half and full compensation. Full compensation closely matches or slightly exceeds the required recovery, whereas half compensation recovers approximately half of the deficit.

Figure 13 shows that the realized cumulative recovery follows the intended compensation strength. Full compensation approaches the required cumulative recovery by the end of training, whereas half compensation produces a substantially smaller correction. These results verify that the target

adjustment acts on the cumulative learning signal in the intended direction. Matching the cumulative signal budget does not imply identical optimization trajectories or policies, but provides a direct mechanism-level check of the proposed compensation rule.

J Ablation Analysis of Parameters

J.1 Target Learning Signal

Table 12 analyzes the target learning signal in LSC-DPO on AlpacaEval 2. In our update rule, the target scaled margin z^* specifies the target signal level $\sigma(-z^*)$ and therefore determines the learning-signal regime that the method attempts to track. We also include DPO as a reference point; its reported signal is the final stabilized value observed during training rather than a controlled target.

The results indicate that stronger performance is achieved in an intermediate target-signal region rather than at either extreme. When z^* is too small, the target signal remains high and the loss stays in a relatively high-sensitivity, high-curvature region. The empirical results show that maintaining an overly high target signal does not yield the best performance. Conversely, when z^* is too large, the target signal becomes small and the sigmoid loss moves toward a flatter, lower-sensitivity region. For Llama3-8B, increasing z^* from 0.40 to 0.70 reduces the LC win rate from 22.7 to 19.5; for Qwen3-8B, increasing z^* from 0.35 to 0.70 reduces the LC win rate from 22.1 to 13.7.

The best results are obtained at moderate target signal levels. Llama3-8B performs best at $z^* = 0.40$, corresponding to $\sigma(-z^*) = 0.4013$, while Qwen3-8B performs best at $z^* = 0.35$, corresponding to $\sigma(-z^*) = 0.4134$. Compared with DPO, the best LSC-DPO settings improve the LC win rate from 19.2 to 22.7 on Llama3-8B and from 17.2 to 22.1 on Qwen3-8B. Overall, these results suggest that LSC-DPO benefits from an intermediate target learning signal: in our experiments, the strongest settings concentrate around $\sigma(-z^*) \approx 0.40$–$0.41$, where the loss retains meaningful local responsiveness without staying too close to the overly high-signal region or drifting into saturation.

Table 12: Target learning signal on AlpacaEval 2. The DPO row reports the final stabilized signal under fixed-β training. Scores are official AlpacaEval 2 point estimates; standard errors for WR/LC are typically 1.0–1.2 percentage points in our runs.

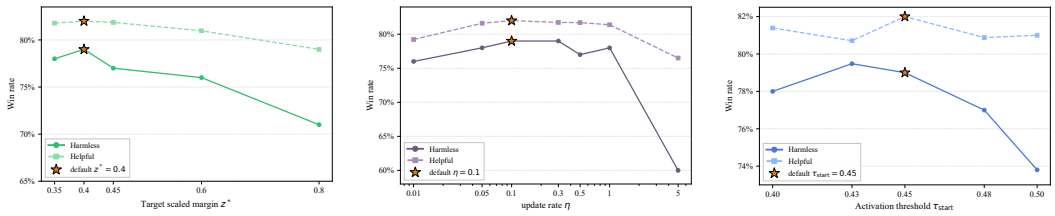
Model	z^* / Method	Target signal $\sigma(-z^*)$	WR (%)	LC (%)
Llama3-8B	0.20	0.4502	3.8	4.3
	0.30	0.4256	12.8	14.7
	0.35	0.4134	20.1	20.9
	0.40	**0.4013**	**21.1**	**22.7**
	0.45	0.3894	19.7	21.1
	0.50	0.3775	18.6	20.7
	0.60	0.3543	18.7	20.2
	0.70	0.3318	17.5	19.5
	DPO	0.3429	17.7	19.2
Qwen3-8B	0.20	0.4502	12.9	18.6
	0.30	0.4256	15.2	21.1
	0.35	**0.4134**	**16.1**	**22.1**
	0.40	0.4013	13.5	18.9
	0.45	0.3894	12.9	18.6
	0.50	0.3775	12.3	17.8
	0.60	0.3543	10.9	18.2
	0.70	0.3318	8.8	13.7
	DPO	0.3787	13.1	17.2

J.2 Ablation of Control Parameters

We further study the parameters of LSC-DPO on Anthropic-HH using Llama3-8B and Qwen3-8B. Figures 14 and 15 vary the target scaled margin z^*, the control rate η, and the start threshold τ_{start}.

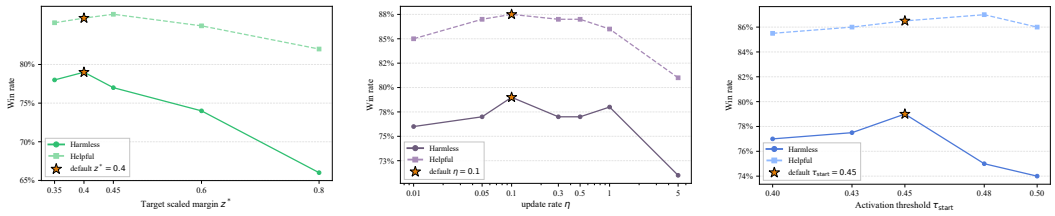
The target scaled margin z^* specifies the desired learning-signal regime through $\sigma(-z^*)$. Moderate values of z^* generally achieve stronger win rates, whereas overly large values reduce performance, consistent with the view that a small target signal moves the loss toward a lower-sensitivity regime. The control rate η determines the responsiveness of the feedback update. Moderate values perform well, while overly large rates can degrade performance, consistent with the stability dependence on ηz^* in Proposition 1. The start threshold τ_{start} determines when the dynamic update is activated. Compared with z^* and η, performance is less sensitive to τ_{start} within the tested range, although moderate activation thresholds tend to perform best.

Overall, these ablations suggest that LSC-DPO tolerates moderate variations in its control parameters, while benefiting from target margins and control rates that maintain a responsive learning-signal regime.



(a) Target scaled margin z^* (b) Control rate η (c) Start threshold τ_{start}

Figure 14: Ablation analysis of key control parameters with Llama3-8B on Anthropic-HH subsets: (a) target scaled margin z^*, (b) control rate η, and (c) start threshold τ_{start}. The star marks the default setting used in the main experiments.



(a) Target scaled margin z^* (b) Control rate η (c) Start threshold τ_{start}

Figure 15: Ablation analysis of key control parameters with Qwen3-8B on Anthropic-HH subsets: (a) target scaled margin z^*, (b) control rate η, and (c) start threshold τ_{start}. The star marks the default setting used in the main experiments.

K Downstream Evaluation

Following the downstream evaluation tasks used in SimPO (Meng et al., 2024) and ε-DPO (Lee et al., 2025), we further evaluate whether preference optimization preserves general task capabilities beyond open-ended chatbot benchmarks. We consider MMLU (Hendrycks et al., 2020), ARC-Challenge (Clark et al., 2018), HellaSwag (Zellers et al., 2019), TruthfulQA (Lin et al., 2022), WinoGrande (Sakaguchi et al., 2021), and GSM8K (Cobbe et al., 2021), covering knowledge, reading comprehension, commonsense reasoning, truthfulness, and mathematical reasoning. Tables 13 and 14 report the results for Llama3-8B and Qwen3-8B, respectively.

Overall, LSC-DPO maintains downstream task performance at a level comparable to strong preference-optimization baselines and does not exhibit a consistent degradation relative to SFT. The effect of preference optimization varies across tasks and model families, consistent with prior observations that downstream task changes can depend on the pretrained model, preference data, and optimization objective. Notably, on GSM8K, LSC-DPO does not show a score drop relative to the SFT initialization on either model family, suggesting that the alignment improvements do not

come from a substantial loss of mathematical reasoning capability. As an additional reproducibility check, we also evaluate the released SimPO checkpoints on the same set of downstream tasks under the same evaluation protocol and obtain highly similar results. This consistency strengthens the reliability of our capability-preservation conclusion. We therefore interpret these downstream results mainly as a capability-preservation check rather than as the primary evidence for alignment improvement.

Table 13: Downstream capability evaluation on Llama3-8B. Scores are reported as percentages. Avg. denotes the mean over the displayed task scores.

Method	**MMLU**	**ARC-C**	**HellaSwag**	**TruthfulQA**	**WinoGrande**	**GSM8K**	**Avg.**
SFT	62.54	60.24	81.30	45.50	74.43	49.20	62.20
DPO	63.15	64.51	83.14	53.12	75.14	54.97	65.67
β-DPO	62.92	63.99	83.04	51.37	75.14	53.45	64.99
ε-DPO	62.98	64.33	83.29	52.94	75.14	53.15	65.30
IPO	62.96	62.63	81.68	49.02	73.88	54.36	64.09
CPO	62.73	60.32	81.44	52.97	75.06	48.22	63.46
KTO	62.88	64.51	83.39	53.93	75.85	53.37	65.65
ORPO	62.80	55.29	80.70	47.72	74.03	44.35	60.82
SLiC-HF	62.69	57.85	80.57	50.82	74.35	46.47	62.13
R-DPO	62.41	61.95	80.91	43.74	74.35	39.58	60.49
SimPO	62.85	65.78	83.12	58.48	74.25	52.54	66.17
LSC-DPO	63.02	65.25	83.34	52.57	74.82	53.98	65.50

Table 14: Downstream capability evaluation on Qwen3-8B. Scores are reported as percentages. Avg. denotes the mean over the displayed task scores.

Method	**MMLU**	**ARC-C**	**HellaSwag**	**TruthfulQA**	**WinoGrande**	**GSM8K**	**Avg.**
SFT	75.99	67.49	80.17	53.87	75.85	82.18	72.59
DPO	76.54	68.69	81.30	57.39	76.01	84.46	74.06
β-DPO	76.48	68.77	81.13	56.40	76.72	83.40	73.82
ε-DPO	76.63	68.34	81.44	57.57	76.01	84.53	74.09
IPO	76.48	67.83	80.92	56.30	76.09	84.15	73.63
CPO	75.63	68.09	80.03	56.72	75.61	83.55	73.27
KTO	76.41	68.77	81.12	57.37	76.09	84.99	74.12
ORPO	75.84	66.38	79.66	53.83	75.22	67.48	69.73
SLiC-HF	75.01	67.06	79.33	55.03	74.51	79.98	71.82
R-DPO	75.92	68.00	80.60	54.07	75.93	84.23	73.13
SimPO	76.37	68.69	81.24	56.73	76.80	84.84	74.11
LSC-DPO	76.50	68.52	81.38	57.50	76.01	84.53	74.07