Adaptive KL Control for Direct Preference Optimization in Instruction-Following LLMs

Yi Khuen Chai

Singapore Management University
yk.chai.2024@computing.smu.edu.sg

Abstract

The scaling parameter β in Direct Preference Optimization governs a fundamental trade-off: low β produces weak gradients that fail to learn from ambiguous preferences, while high β amplifies updates and causes excessive drift from the reference policy. Prior work treats β as fixed or scheduled throughout training. We introduce DualLoop-DPO, which modulates β via dual feedback: a fast loop raises β temporarily on high-uncertainty batches to enforce stronger preference margins, while a slow loop uses EMA-smoothed KL tracking to regulate policy drift. Experiments on preference alignment benchmarks show consistent improvements over existing static-β, β-scheduling, and dynamic-β baselines. These findings suggest that dual-loop β control—responding to uncertainty for learning and divergence for stability—offers a promising direction for preference-based fine-tuning.

Introduction

Direct Preference Optimization (DPO) (Rafailov et al. 2023) aligns language models via a Bradley-Terry loss, where the KL penalty β balances alignment strength against reference-policy drift. A static β creates a dilemma: low values cause *gradient laziness* on uncertain prompts, while high values induce *global drift* (Xu et al. 2024). Most DPO variants fix β or anneal it across training steps.

We introduce DualLoop-DPO, which adjusts β online via dual-loop feedback: a fast loop reacts to batch entropy, and a slow loop tracks KL divergence via exponential moving average (EMA). We compare this controller against tuned Static/Annealed baselines and recent dynamic-β methods (Beta-DPO, Epsilon-DPO) on Qwen2.5-7B-Instruct, demonstrating consistent wins across UltraFeedback, HelpSteer2, and Sycophancy-Eval under a 120-step budget.

Method

The DPO loss gradient for model parameters θ is proportional to the implicit reward margin M:

$$\|\nabla_\theta \mathcal{L}\| \propto \beta \cdot |\sigma(-\beta M)| \cdot \|\nabla M\|, \tag{1}$$

where $M = \log \frac{\pi_\theta(y_w)}{\pi_{ref}(y_w)} - \log \frac{\pi_\theta(y_l)}{\pi_{ref}(y_l)}$ and σ is the sigmoid. When the model is uncertain ($M \approx 0$), $\sigma(-\beta M) \approx 0.5$ and the effective gradient magnitude scales as $0.5\beta\|\nabla M\|$. A static low β therefore produces gradient laziness, while a high β amplifies all gradients and induces global drift away from the reference policy.

DualLoop-DPO treats β as a dynamic gain controlled by two lightweight feedback loops. The **fast loop** reacts to high batch entropy $\tilde{H}$ via

$$\beta_{total} = \beta_{base}(1 + \lambda \tilde{H}), \tag{2}$$

upweighting high-uncertainty examples to enforce larger preference margins, thereby improving the model's discriminative capacity on ambiguous prompts. The **slow loop** provides a stabilizing counterpart by tracking the exponential moving average of KL divergence, KL_{ema}, and adjusting the baseline:

$$\beta_{base}^{t+1} \leftarrow \beta_{base}^{t} \exp\left(\eta \frac{KL_{ema} - KL_{target}}{KL_{target}}\right), \tag{3}$$

constraining long-term policy drift to ensure the model retains its reference capabilities even as the fast loop applies increased penalties to challenging examples. We implement the controller with EMA smoothing ($\alpha = 0.1$), a 10% deadband around the target, β_{base} clipping to $[0.05, 2.0]$, and default hyperparameters $KL_{target} = 0.04$, $\beta_{init} = 0.1$, $\lambda = 4.0$, $\eta = 0.01$.

We train on a 0.5% UltraFeedback (Cui et al. 2023) subsample ($\sim$300 pairs) using 4-bit QLoRA on Qwen2.5-7B-Instruct for 120 steps (effective batch size 4, learning rate 5×10^{-6}, seed 42). All baselines share this configuration, differing only in β schedules.

Evaluation spans five phases: (i) fixed-β sweep; (ii) main UltraFeedback comparison; (iii) ablations; (iv) generalization (HelpSteer2 (Wang et al. 2024), Anthropic HH (Bai et al. 2022), Sycophancy-Eval (Sharma et al. 2023)); (v) dynamic-β baselines (Beta-DPO, Epsilon-DPO). Win rates are computed via blind pairwise judging by GPT-4o-mini (primary) and Gemini 2.0 Flash (secondary), evaluating response pairs generated for the same $N = 50$–200 prompts per comparison. Cross-judge agreement is 83.3% (Phase 2) and 76.8% overall (Cohen's $\kappa \approx 0.38$), with disagreement concentrated on near-tie cases.

Experiments and Results

Phase 1: Baseline selection. A hyperparameter sweep over $\beta \in \{0.05, 0.1, 0.2\}$ identifies $\beta = 0.2$ as the strongest

Comparison	Win Rate	95% CI
DualLoop-DPO vs. Base	87.0%	[83.7, 90.3]
DualLoop-DPO vs. Static	86.3%	[82.9, 89.6]
DualLoop-DPO vs. Annealed	86.0%	[82.6, 89.4]

Table 1: UltraFeedback win rates aggregated over GPT-4o-mini and Gemini 2.0 Flash judges. Models generate responses to 200 held-out prompts; both judges perform blind pairwise comparisons on all pairs, yielding $N = 400$ decisions per model comparison.

Dataset	vs. Static	vs. Annealed	vs. Base
HelpSteer2	86.0%	85.8%	84.8%
Sycophancy-Eval	75.0%	76.0%	77.0%
Anthropic HH	83.3%	79.8%	30.5%

Table 2: Generalization win rates for DualLoop-DPO on three out-of-distribution datasets. Each dataset uses 200 evaluation prompts with dataset-specific judging criteria. Win rates are aggregated over GPT-4o-mini and Gemini 2.0 Flash judges.

static baseline (92% vs. base under GPT-4o-mini, 84% under Gemini) via judge evaluation of model responses to 50 UltraFeedback prompts.

Phase 2: Main comparison. On 200 held-out UltraFeedback prompts, Table 1 shows that DualLoop-DPO achieves 87% win rate vs. base and ≈86% vs. Static/Annealed baselines.

Phase 3: Ablations. Removing EMA drops win rate from 86% to 78%, confirming EMA as the critical stabilizer. Fast-loop or clipping removal yields moderate degradation (80%); deadband removal has negligible effect.

Phase 4: Generalization. We test whether the UltraFeedback-trained adaptive model generalizes to three out-of-distribution evaluation sets targeting distinct behaviors: HelpSteer2 (Wang et al. 2024) for helpfulness and instruction-following, Sycophancy-Eval (Sharma et al. 2023) for resisting flattery and correcting false premises, and Anthropic HH (Bai et al. 2022) for safety and red-teaming.

For each dataset, we construct a 200-prompt evaluation set. All four models (DualLoop-DPO, Static, Annealed, Base) generate one response per prompt, and two LLM judges perform binary A/B comparisons with dataset-specific criteria: "Which response is more helpful and well-structured?" on HelpSteer2, "Which response is more truthful and less sycophantic?" on Sycophancy-Eval, and "Which response is safer?" on Anthropic HH.

Table 2 shows strong transfer to HelpSteer2 (86% vs. Static) and Sycophancy-Eval (77% vs. Base). On Anthropic HH, all DPO variants underperform the base model (DualLoop-DPO 30.5% vs. Base), yet DualLoop-DPO degrades least among DPO variants (83.3% vs. Static, 79.8% vs. Annealed).

Phase 5: Dynamic-β baselines. We compare against Beta-DPO (Wu et al. 2024) (batch-margin adaptation) and Epsilon-DPO (Lee et al. 2025) (perturbation-based KL control) under the same UltraFeedback training and evaluation setup as Phase 2. Table 3 shows results split by judge. Primary judge (GPT-4o-mini) consistently favors DualLoop-DPO (>91%); secondary (Gemini) shows moderate agreement (∼64%). Aggregated across both judges, DualLoop-DPO achieves 77.6% vs. Beta-DPO and 77.5% vs. Epsilon-DPO. Both judges prefer DualLoop-DPO; disagreement reflects stylistic divergence on near-ties.

Method	Primary	Secondary
vs. Beta-DPO	91.3%	63.8%
vs. Epsilon-DPO	91.0%	63.9%

Table 3: Win rates of DualLoop-DPO vs. dynamic-β baselines on UltraFeedback under the Phase 2 evaluation setup.

Discussion and Conclusion

DualLoop-DPO consistently outperforms Static/Annealed baselines (86% win rate) and dynamic-β methods (77.5% vs. Beta-DPO/Epsilon-DPO) across UltraFeedback, HelpSteer2, and Sycophancy-Eval. Ablations confirm EMA smoothing as the critical stabilizer. These results suggest that dual-loop KL control is a promising alternative for DPO-style alignment.

Anthropic HH reveals a broader limitation: all DPO variants underperform the base model on this out-of-distribution safety set, though DualLoop-DPO degrades least among DPO variants (83% vs. Static). Closing this gap likely requires multi-objective control or richer safety signals. Notably, these gains were achieved with just ∼300 pairs over 120 steps, highlighting DualLoop-DPO's potential for resource-limited practitioners.

Limitations

Results rely on automated judges (GPT-4o-mini, Gemini) over $N = 50$–200 prompts; no formal human evaluation study is performed. Cross-judge agreement (Cohen's $\kappa \approx 0.38$) reflects stylistic divergence on near-ties. Controller hyperparameters were tuned on UltraFeedback; transfer to larger scales is untested. The entropy-based β adaptation assumes high-entropy examples reflect epistemic uncertainty rather than aleatoric label noise; preliminary flip-rate diagnostics (re-judging 90 prompts $\times 3$) showed stable ∼7% flip rates across entropy buckets, supporting this assumption, though formal verification on other datasets is needed. Additional baselines (IPO, SimPO, KTO) remain future work.

Ethical Statement

Datasets (UltraFeedback, HelpSteer2, Anthropic HH, Sycophancy-Eval) reflect specific annotator demographics and may encode biases. Improved scores do not imply broad safety. Automated judges inherit their own biases; human evaluation is encouraged for future work.

References

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862.

Cui, G.; Yuan, L.; Ding, N.; Yao, G.; He, B.; Zhu, W.; Ni, Y.; Xie, G.; Xie, R.; Lin, Y.; et al. 2023. UltraFeedback: Boosting Language Models with High-quality Feedback. arXiv:2310.01377.

Lee, S.; Han, J.; Song, H.; Choi, S. J.; Lee, H.; and Yu, Y. 2025. KL Penalty Control via Perturbation for Direct Preference Optimization. arXiv:2502.13177.

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

Sharma, M.; Tong, M.; Korbak, T.; Duvenaud, D.; Askell, A.; Bowman, S. R.; Cheng, N.; Durmus, E.; Hatfield-Dodds, Z.; Johnston, S. R.; Kravec, S.; Maxwell, T.; McCandlish, S.; Ndousse, K.; Rausch, O.; Schiefer, N.; Yan, D.; Zhang, M.; and Perez, E. 2023. Towards Understanding Sycophancy in Language Models. arXiv:2310.13548.

Wang, Z.; Dong, Y.; Delalleau, O.; Zeng, J.; Shen, G.; Egert, D.; Zhang, J. J.; Sreedhar, M. N.; and Kuchaiev, O. 2024. HelpSteer2: Open-source dataset for training top-performing reward models. arXiv:2406.08673.

Wu, J.; Xie, Y.; Yang, Z.; Wu, J.; Gao, J.; Ding, B.; Wang, X.; and He, X. 2024. β-DPO: Direct Preference Optimization with Dynamic β. arXiv:2407.08639.

Xu, S.; Fu, W.; Niu, J.; Yang, Y.; Wang, S.; Zhang, H.; Liu, P.; Lin, Z.; Chen, Q.; and Wu, W. 2024. Is DPO Superior to PPO for LLM Alignment? A Comprehensive Study. *arXiv preprint arXiv:2404.10719*.