# Boosting Direct Preference Optimization with Penalization

**Pengwei Sun** [1]

## Abstract

Offline preference optimization has become a practical substitute for reinforcement learning from human feedback, but pairwise objectives such as Direct Preference Optimization (DPO) and its variants use only the chosen and rejected responses stored in a static dataset. This leaves a useful signal unused: the response that the reference model itself would generate for the same prompt. We propose Direct Preference Optimization with Penalization (DPOP), a simple extension of DPO that augments the base preference loss with a gated penalty on reference-greedy responses. DPOP activates this penalty only when the current policy still assigns a lower likelihood to the preferred response than to the rejected response. On AlpacaEval 2.0, DPOP improves length-controlled win rate over DPO, SimPO, and AlphaDPO on both Llama-3-8b-it and Gemma-2-9b-it, achieving relative gains of 5.3% and 4.4% over baselines on the two models, respectively. Ablations further show that a SimNPO-style length-normalized penalty is stronger than NPO and token-level unlikelihood in this setting.

## 1. Introduction

Aligning large language models with human preferences has traditionally relied on Reinforcement Learning from Human Feedback (RLHF), a complex, multi-stage pipeline requiring a separately trained reward model and constrained policy optimization (Ouyang et al., 2022). Direct Preference Optimization (DPO) simplified this pipeline by bypassing the explicit reward model entirely, elegantly re-framing alignment as a purely pairwise classification task over reference-corrected policy likelihoods (Rafailov et al., 2023). This formulation has made offline alignment substantially easier to implement, and it remains a strong baseline because it preserves the regularizing role of a reference model without requiring online rollouts during training.

The same simplification also imposes a restriction: the training signal is confined to the two responses already stored in the preference dataset. This is limiting for modern instruction tuning, where preference data are often off-policy and are frequently produced or filtered by teacher models rather than collected directly from the policy being trained. In this regime, the reference model is not merely a source of KL regularization. It is also a concrete generator whose own greedy response to each prompt can reveal local modes that the policy may continue to imitate unless explicitly corrected. But DPO only updates the policy on $(x, y_w, y_l)$ and leave the reference-greedy output $y_g$ outside the objective.

We introduce Direct Preference Optimization with Penalization (DPOP), a minimal offline preference objective that incorporates this missing response. DPOP keeps the ordinary pairwise preference loss as the base objective, but additionally penalizes the policy likelihood of the reference-greedy response under a detached policy-margin gate. The gate is active only when the policy still ranks the rejected response above the chosen response in likelihood, which targets examples where the current policy has not yet internalized the pairwise preference.

Empirically, this design is especially effective with a SimNPO-style penalty, which is length-normalized and reference-free on the penalized response. On AlpacaEval 2.0, DPOP improves the length-controlled win rate (LC-WR) from 44.01 for SimPO and 41.48 for AlphaDPO to 46.35 on Llama-3-8b-it. On Gemma-2-9b-it, it improves from 73.08 for SimPO and 74.90 for AlphaDPO to 78.22. These results suggest that reference-greedy responses are not only useful as cached artifacts for analysis, but can be turned into an effective offline penalization signal.

## 2. Related Work

**Direct preference optimization.** DPO uses a binary preference loss derived from the Bradley–Terry model (Bradley & Terry, 1952; Rafailov et al., 2023), optimizing $\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y_w,y_l)}\left[\log\sigma\left(\beta\left[\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right]\right)\right]$. Given a prompt, chosen response, and rejected response, DPO compares the policy's log-likelihood ratio with the

[1]Stanford University. Correspondence to: Pengwei Sun <pengwei@stanford.edu>.

Accepted at *ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning.* Copyright 2026 by the author(s).

corresponding reference-model log-likelihood ratio. Its stability and simplicity make it the most natural base loss for DPOP. Our method preserves this pairwise DPO component, but adds a separate penalty on the response generated greedily by the reference model.

**Reference-free and adaptive-margin objectives.** Several variants of DPO have been developed to improve its margin and reference mechanisms. SimPO argues that sequence-average log probability is better aligned with inference than a reference-corrected implicit reward, and removes the reference model from the preference loss while adding a fixed target margin $\gamma$ (Meng et al., 2024), optimizing

$$\mathcal{L} = -\mathbb{E}_{(x,y_w,y_l)}\left[\log\sigma\left(\beta\left[\frac{1}{|y_w|}\log\pi_\theta(y_w \mid x) - \frac{1}{|y_l|}\log\pi_\theta(y_l \mid x) - \gamma\right]\right)\right].$$

AlphaDPO instead keeps a reference-based perspective but adapts the reward margin according to an interpolation between policy and reference behavior (Wu et al., 2024), optimizing

$$\mathcal{L} = -\mathbb{E}_{(x,y_w,y_l)}\left[\log\sigma\left(\frac{\beta}{|y_w|}\log\pi_\theta(y_w \mid x) - \frac{\beta}{|y_l|}\log\pi_\theta(y_l \mid x) - sg(\gamma + \alpha M)\right)\right],$$

where

$$M(x, y_w, y_l) = \beta\left(\log\frac{\pi_\theta(y_w \mid x)\pi_{\text{ref}}(y_l \mid x)}{\pi_{\text{ref}}(y_w \mid x)\pi_\theta(y_l \mid x)}\right).$$

**Negative preference optimization.** NPO was introduced for machine unlearning as a preference-style alternative to direct gradient ascent on undesirable data $y$ (Zhang et al., 2024), optimizing $\mathcal{L}_{\text{NPO}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\frac{2}{\beta}\log\sigma\left(-\beta\log\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}\right)\right]$. SimNPO later revisited this setting and showed that removing the reference model from the negative objective can improve stability and effectiveness (Fan et al., 2024), optimizing $\mathcal{L}_{\text{SimNPO}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\frac{2}{\beta}\log\sigma\left(-\beta\frac{1}{|y|}\log\pi_\theta(y|x) - \gamma\right)\right]$. DPOP adapts these penalty families to preference alignment rather than unlearning. The reference-greedy response is not private or harmful data; it is a model-generated response that may represent an undesirable local mode for the current prompt. This makes NPO and SimNPO useful penalty templates inside an otherwise standard preference-optimization pipeline.

# 3. Methodology

## 3.1. Base Preference Objective

Let $x$ denote a prompt, $y_w$ a preferred response, and $y_l$ a rejected response. Let $\pi_\theta$ be the trainable policy and $\pi_{\text{ref}}$ the fixed reference model. The DPO base loss uses the reference-corrected logit

$$z_{\text{dpo}} = \log\frac{\pi_\theta(y_w \mid x)\pi_{\text{ref}}(y_l \mid x)}{\pi_{\text{ref}}(y_w \mid x)\pi_\theta(y_l \mid x)}. \tag{1}$$

and optimizes

$$\mathcal{L}_{\text{base}}(x, y_w, y_l) = -\log\sigma(\beta z_{\text{dpo}}), \tag{2}$$

where $\beta$ is the preference temperature.

## 3.2. Reference-Greedy Penalization

DPOP augments each preference triple with a reference-greedy response

$$y_g = \text{GreedyDecode}\big(\pi_{\text{ref}}(\cdot|x)\big). \tag{3}$$

This response is generated offline and cached in the training data. The core design choice is to penalize $y_g$ selectively, not universally. We first compute the policy likelihood margin

$$r = \text{sg}\left(\log\pi_\theta(y_w|x) - \log\pi_\theta(y_l|x)\right), \tag{4}$$

where $\text{sg}(\cdot)$ denotes stop-gradient. The penalty gate is

$$g_\pi = \text{sg}\left(\mathbf{1}[r < 0]\right), \tag{5}$$

so the reference-greedy penalty is active only when the policy still assigns higher likelihood to the rejected response than to the chosen response. DPOP then defines a detached penalty weight

$$w_{\text{pen}} = g_\pi f(r), \tag{6}$$

where $f$ is chosen from simple nonnegative families: linear, $f(r) = \max(-r, 0)$; constant; square root, $f(r) = \sqrt{\max(-r, 0) + \epsilon}$; and quadratic, $f(r) = \max(-r, 0)^2$.

## 3.3. Penalty Families

We consider three penalties on $y_g$. Token-level unlikelihood penalizes the next-token probabilities assigned by the policy to the reference-greedy response:

$$\mathcal{L}_{\text{unll}} = -\frac{1}{|y_g|}\sum_t \log\left(1 - \pi_\theta(y_{g,t}|x, y_{g,<t})\right). \tag{7}$$

The NPO penalty uses a reference-relative negative objective,

$$\mathcal{L}_{\text{npo}} = -\frac{2}{\beta_p}\log\sigma\left(-\beta_p\left[\log\pi_\theta(y_g|x) - \log\pi_{\text{ref}}(y_g|x)\right]\right), \tag{8}$$

**Algorithm 1** DPOP loss for one minibatch

Compute $\mathcal{L}_{\text{base}}$ on $(x, y_w, y_l)$.
Compute $r = \text{sg}(\log \pi_\theta(y_w|x) - \log \pi_\theta(y_l|x))$.
Set $w_{\text{pen}} = \text{sg}(\mathbf{1}[r < 0])f(r)$.
Evaluate $\mathcal{L}_{\text{pen}}$ on reference-greedy response $y_g$.
Return $\mathcal{L}_{\text{base}} + \lambda_{\text{pen}} w_{\text{pen}} \mathcal{L}_{\text{pen}}$.

where $\beta_p$ is the penalty temperature. The SimNPO penalty removes the reference from the negative objective and uses length-normalized log probability:

$$\mathcal{L}_{\text{simnpo}} = -\frac{2}{\beta_p} \log \sigma \left( -\beta_p \overline{\log \pi_\theta(y_g|x)} - \gamma_p \right), \quad (9)$$

where $\overline{\log \pi_\theta}$ denotes average sequence log probability and $\gamma_p$ is a penalty margin.

The final DPOP objective is

$$\mathcal{L}_{\text{DPOP}} = \mathcal{L}_{\text{base}} + \lambda_{\text{pen}} w_{\text{pen}} \mathcal{L}_{\text{pen}}, \quad (10)$$

where $\mathcal{L}_{\text{pen}}$ is one of the penalties above and $\lambda_{\text{pen}}$ is a hyperparameter tuning the overall penalty strength. The design is intentionally concise: DPOP preserves the original pairwise loss and adds pressure only on a generated reference response, with a detached gate controlling when that pressure is applied.

# 4. Experiments

## 4.1. Experimental Setup

**Models and data** We optimize preferences using two widely used instruction-tuned models as the SFT initialization: `meta-llama/Meta-Llama-3-8B-Instruct`, and `google/gemma-2-9b-it`.

Training uses UltraFeedback-style preference data (Cui et al., 2024), specifically `https://huggingface.co/datasets/princeton-nlp/llama3-ultrafeedback` for the Llama model and `https://huggingface.co/datasets/princeton-nlp/gemma2-ultrafeedback-armorm` for the Gemma model.

**Evaluation benchmarks.** We evaluate with AlpacaEval 2.0 and report raw win rate (WR), length-controlled win rate (LC-WR), and average response length. We primarily focus on LC-WR because automatic evaluators are known to favor longer responses and length control improves robustness to this confounder (Dubois et al., 2024).

**Baselines.** We select SimPO and AlphaDPO as representative baselines. While a large body of work has proposed variants of DPO, we found SimPO to be one of the strongest practical baselines in our experimental setting. In contrast, several alternative DPO variants did not yield consistent gains over SimPO and were often more sensitive to hyperparameter tuning. Therefore, we use SimPO as the primary reference-free baseline and AlphaDPO as a representative margin-based baseline.

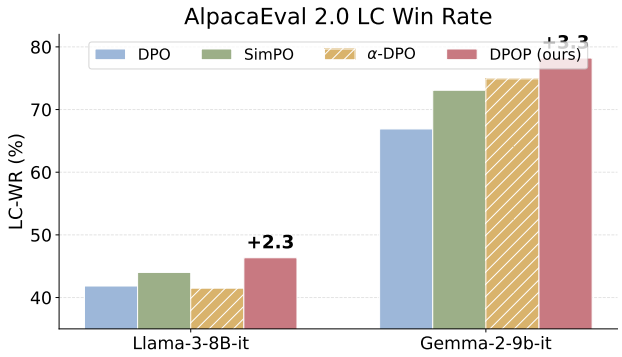


*Figure 1.* AlpacaEval 2.0 length-controlled win rate. DPOP improves over DPO, SimPO, and AlphaDPO on both Llama-3-8b-it and Gemma-2-9b-it. The numbers above DPOP bars are improvements over the best non-DPOP baseline for the same model.

*Table 1.* Main AlpacaEval 2.0 evaluation results. LC-WR is the length-controlled win rate and is the primary metric. The best results are highlighted in bold.

| Model | Method | WR | LC-WR | Length |
|---|---|---|---|---|
| Llama-3-8b-it | DPO | 39.49 | 41.84 | 1909 |
| | SimPO | 37.65 | 44.01 | 1763 |
| | AlphaDPO | 33.05 | 41.48 | 1675 |
| | **DPOP (ours)** | **47.57** | **46.35** | 2057 |
| Gemma-2-9b-it | DPO | 70.40 | 66.90 | 1900 |
| | SimPO | 66.12 | 73.08 | 1754 |
| | AlphaDPO | 65.32 | 74.90 | 1682 |
| | **DPOP (ours)** | **73.76** | **78.22** | 1970 |

**Computational Environment.** All training experiments were conducted using 8×H100 GPUs, as per the procedures detailed in the alignment-handbook repository.

## 4.2. DPOP Consistently Outperforms Preference Optimization Baselines

Figure 1 and Table 1 summarize the main comparison. Across both Llama-3-8b-it and Gemma-2-9b-it, DPOP achieves the strongest LC-WR among all evaluated methods (DPO, SimPO, and AlphaDPO), suggesting that the reference-greedy response provides useful additional signal beyond the original preference pair.

We note that the improvement in raw win rate can partly reflect changes in response length, especially on Llama-3-8b-it. Therefore, we emphasize LC-WR as the primary metric for comparison. Overall, these results support our

*Table 2.* Penalty search on Llama-3-8b-it. For each penalty family, we report the best hyperparameters and evaluation results after a grid search.

| Penalty | $\lambda_{\text{pen}}$ | $\beta_p$ | $\gamma_p$ | WR | LC-WR | Length |
|---|---|---|---|---|---|---|
| SimNPO | 1.0 | 1.0 | 1.0 | 47.57 | **46.35** | 2057 |
| NPO | 0.01 | 0.05 | – | 42.76 | 43.94 | 1960 |
| Unlikelihood | 0.01 | – | – | 40.07 | 41.36 | 1948 |

*Table 3.* Top SimNPO-penalty DPOP hyperparameter settings by AlpacaEval 2.0 LC-WR. The sweep varies the penalty weight $\lambda$, penalty temperature $\beta_p$, and SimNPO margin $\gamma_p$.

| Model | $\lambda$ | $\beta_p$ | $\gamma_p$ | WR | LC-WR | Length |
|---|---|---|---|---|---|---|
| Llama-3-8b-it | 1.0 | 1.0 | 1.0 | 47.57 | **46.35** | 2057 |
| | 1.0 | 1.0 | 0.0 | 44.77 | 45.14 | 1999 |
| | 2.0 | 1.0 | 0.0 | 47.08 | 44.47 | 2121 |
| | 1.0 | 2.0 | 1.0 | 44.78 | 44.28 | 2016 |
| | 1.0 | 2.0 | 2.0 | 45.21 | 43.52 | 2074 |
| Gemma-2-9b-it | 2.0 | 1.0 | 0.0 | 73.76 | **78.22** | 1970 |
| | 4.0 | 1.0 | 0.0 | 74.13 | 77.47 | 1999 |
| | 2.0 | 1.0 | 1.0 | 74.01 | 77.22 | 2014 |
| | 1.0 | 1.0 | 1.0 | 72.88 | 77.22 | 1981 |
| | 4.0 | 1.0 | 1.0 | 74.69 | 76.84 | 2073 |

hypothesis that incorporating the reference-greedy response leads to more effective preference optimization.

### 4.3. Comparing Penalty Designs

Table 2 compares penalty families on Llama-3-8b-it. SimNPO is the strongest variant, reaching 46.35 LC-WR. NPO improves over the DPO baseline but remains below SimNPO at 43.94 LC-WR, while unlikelihood reaches 41.36 LC-WR, performing worse than DPO without penalization.

### 4.4. SimNPO Hyperparameter Search

Table 3 summarizes the top SimNPO-penalty settings for both model families. On Llama-3-8b-it, the best setting uses $\lambda = 1.0$, $\beta_p = 1.0$, and $\gamma_p = 1.0$. On Gemma-2-9b-it, the best setting uses $\lambda = 2.0$, $\beta_p = 1.0$, and $\gamma_p = 0.0$. In both cases, the top several configurations remain above their corresponding SimPO and AlphaDPO baselines, indicating that the observed gains are not isolated to a single brittle setting.

### 4.5. Weight Function Search

Table 4 compares the four choices for the detached weight function $f(r)$. Among these runs, linear weighting gives the strongest LC-WR, followed closely by the best constant-weight setting.

The best setting makes the penalty strength proportional to the severity of the policy-margin error. When the policy is only mildly wrong, the rejected response has a small advantage over the chosen response and the additional penalty on the reference-greedy response remains modest. When the policy is strongly biased toward the rejected response, the margin violation is larger and the penalty weight increases accordingly. Thus, linear weighting concentrates optimization pressure on examples where the current policy most clearly fails to represent the observed preference, while avoiding the uniform pressure of a constant weight. The weaker squareroot and quadratic variants suggest that overly compressing or over-amplifying the margin signal is less effective than preserving this direct proportional relationship.

*Table 4.* Weight-function search for $f(r)$ on Llama-3-8b-it.

| $f(r)$ | WR | LC-WR | Length |
|---|---|---|---|
| linear | 47.57 | **46.35** | 2057 |
| constant | 47.27 | 45.16 | 1958 |
| square root | 45.95 | 44.17 | 1984 |
| quadratic | 45.27 | 45.12 | 1925 |

## 5. Conclusion

DPOP extends offline direct preference optimization by using the reference model's own greedy response as a selectively penalized third response. The method keeps the original pairwise DPO loss, adds a gated penalty only when the policy margin is still wrong, and uses detached weights to avoid turning the gate into a learned reward signal. On AlpacaEval 2.0, DPOP improves LC-WR over DPO, SimPO, and AlphaDPO on both Llama-3-8B-Instruct and Gemma-2-9B-IT, with the strongest results coming from the SimNPO-style penalty.

Future work should evaluate DPOP on a broader set of instruction-following, reasoning, factuality, and safety benchmarks to verify that penalizing reference-greedy responses does not degrade other model capabilities. It would also be important to better characterize when reference-greedy responses correspond to harmful local modes that should be penalized, versus acceptable responses that should be preserved.

## References

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Cui, G., Yuan, L., Ding, N., Yao, G., He, B., Zhu, W., Ni, Y., Xie, G., Xie, R., Lin, Y., Liu, Z., and Sun, M. UltraFeedback: Boosting language models with scaled AI feedback. In *International Conference on Machine Learning*, 2024.

Dubois, Y., Galambosi, B., Liang, P., and Hashimoto, T. B. Length-controlled alpacaeval: A simple way to debias automatic evaluators. In *Conference on Language Modeling*, 2024.

Fan, C., Liu, J., Lin, L., Jia, J., Zhang, R., Mei, S., and Liu, S. Simplicity prevails: Rethinking negative preference optimization for LLM unlearning, 2024.

Meng, Y., Xia, M., and Chen, D. SimPO: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems*, 2024.

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

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

Wu, J., Wang, X., Yang, Z., Wu, J., Gao, J., Ding, B., Wang, X., and He, X. AlphaDPO: Adaptive reward margin for direct preference optimization, 2024.

Zhang, R., Lin, L., Bai, Y., and Mei, S. Negative preference optimization: From catastrophic collapse to effective unlearning, 2024.