Confidence-Weighted Preference Optimization for Instruction-Following RLHF

Aadit Sood and Aditya Hariharan

Research manuscript prepared for portfolio publication

Abstract

Reinforcement learning from human feedback (RLHF) has become a central approach for adapting language models to human preferences, yet different preference-optimization and online policy-optimization methods can behave differently under the same instruction-following evaluation. This study compares offline preference optimization, reward modeling, and online RLHF methods under a controlled experimental setup using a fixed base policy, preference dataset, and external judge evaluation. We evaluate Direct Preference Optimization (DPO), Identity Preference Optimization (IPO), distribution-level preference optimization through AOT-style quantile matching, a Bradley-Terry reward model, and online methods including GRPO, DrGRPO, and GSPO. We further investigate a confidence-weighted variant of DPO that moderately upweights preference pairs with clearer reference-corrected margins. DPO was the strongest offline baseline, achieving a 0.8298 external judge win rate against the base model, while GSPO was the strongest online method at 0.7857. Confidence-weighted DPO achieved the highest overall score, 0.8372, suggesting that bounded confidence weighting can modestly improve a strong DPO baseline. In contrast, label-smoothed variants were substantially weaker, indicating that softening the preference target can reduce useful learning signal in this benchmark. These results highlight the importance of external generation-quality evaluation, since reward-model scores and internal preference diagnostics were useful but incomplete predictors of final judged performance.

Keywords: RLHF; preference optimization; DPO; reward modeling; instruction following; language model alignment

1. Introduction

Reinforcement learning from human feedback (RLHF) provides a practical framework for aligning language models with human preferences. Instead of optimizing only next-token likelihood, RLHF-style methods use preference comparisons, reward models, or online policy updates to improve the quality of model responses under human- or judge-defined criteria.

This paper studies RLHF for instruction following in a controlled setting. The experiments use a fixed base language model, a fixed preference dataset, and a common external judge evaluation, allowing several offline and online alignment methods to be compared under the same conditions. The primary goal is to understand which methods produce the strongest judged generations and how well internal diagnostics, such as reward-model score or preference margin, predict external performance.

The empirical results show that DPO is a strong offline baseline. A confidence-weighted DPO variant provides a small additional improvement by assigning moderately larger weights to clearer preference pairs, while retaining all examples. By contrast, label smoothing and weighted label smoothing underperform substantially. Among online RLHF methods, GSPO performs best, suggesting that sequence-level clipping can be more effective than token-level clipping in this instruction-following benchmark.

Contributions. This manuscript makes three contributions: (i) a compact comparison of offline preference optimization, reward modeling, and online RLHF methods under a shared instruction-following evaluation; (ii) an ablation study of confidence weighting and label smoothing for DPO; and (iii) an analysis of when internal training diagnostics do and do not align with external judge win rate.

2. Background and Methods

2.1 Offline preference optimization

Offline preference optimization uses pairwise comparisons of the form (x, y+, y-), where x is an instruction prompt, y+ is the preferred response, and y- is the rejected response. Let pi_theta denote the trainable policy and pi_ref denote the frozen reference policy. DPO-style methods optimize a reference-corrected log-probability margin:

```
m_theta = beta[(log pi_theta(y+|x) - log pi_theta(y-|x)) - (log pi_ref(y+|x) - log pi_ref(y-|x))].
```

DPO applies a logistic preference loss to this margin, encouraging the policy to increase the relative likelihood of preferred responses while limiting deviation from the reference model. IPO uses a related reference-corrected margin but regresses toward a finite target gap, making the update more conservative. AOT-style optimization changes the comparison from pairwise examples to distribution-level separation by sorting chosen and rejected reference-corrected rewards within a batch and applying a DPO-style loss to quantile gaps.

2.2 Reward modeling and online RLHF

A Bradley-Terry reward model was trained to assign scalar rewards to chosen and rejected responses. The model optimizes the probability that a chosen response receives a higher reward than its paired rejected response. The learned reward model is then used to score sampled generations during online RLHF.

The online methods sample groups of completions for each prompt, score them with the reward model, compute group-relative advantages, and update the policy using clipped policy-gradient objectives. GRPO normalizes rewards within each group by subtracting the group mean and dividing by the group standard deviation. DrGRPO removes the standard-deviation normalization and sequence-length normalization, making the update more sensitive to raw reward differences and completion length. GSPO instead clips a sequence-level likelihood ratio, which directly constrains the probability of the full generated response rather than each token independently.

2.3 Confidence-weighted DPO

The main proposed modification is a bounded confidence-weighted version of DPO. The motivation is that preference pairs with clearer reference-corrected margins may provide stronger learning signals, while highly ambiguous pairs may be noisier. Each DPO loss is multiplied by a detached confidence weight:

```
w = 0.5 + sigmoid(|stopgrad(m_theta)|),     L = w[-log sigmoid(m_theta)].
```

The weight is detached from the computation graph so that it changes only the relative contribution of examples. Its range is approximately [1.0, 1.5], so all examples remain active and the maximum upweighting is deliberately modest. Label-smoothed DPO and weighted label-smoothed DPO were also evaluated as ablations intended to test robustness to noisy preference labels.

3. Experimental Setup

All experiments used the same base policy, dataset splits, and evaluation protocol. The base policy was a Qwen2.5 instruction-following model trained with LoRA adapters. The dataset was a WildChat-derived preference and generation benchmark, with separate preference-pair splits for offline optimization and reward modeling, prompt-only splits for online rollouts, and held-out splits for diagnostics and external judge evaluation.

The primary evaluation metric is external judge win rate against the base model. Offline diagnostics include reference-corrected preference margin and held-out preference accuracy. Reward-model diagnostics include pair accuracy and chosen-minus-rejected reward margin. Online diagnostics include reward-model score on sampled completions, likelihood-ratio behavior, KL divergence, and completion length.

4. Results

4.1 External judge evaluation

Table 1 summarizes the external evaluation results. DPO achieved the best performance among the offline baselines, while GSPO was the strongest online method. The best overall method was confidence-weighted DPO with beta = 0.1, which improved external judge win rate from 0.8298 for DPO to 0.8372.

Method	Category	Result
Reward model	Pair accuracy	0.8281
DPO	Offline preference optimization	0.8298 win rate
IPO	Offline preference optimization	0.6875 win rate
AOT	Offline preference optimization	0.7455 win rate
GRPO	Online RLHF	0.6889 win rate
DrGRPO	Online RLHF	0.6154 win rate
GSPO	Online RLHF	0.7857 win rate
Weighted DPO (beta = 0.1)	DPO modification	0.8372 win rate
Weighted DPO (beta = 0.2)	DPO modification	0.8121 win rate
Weighted DPO (beta = 0.01)	DPO modification	0.7901 win rate
Smooth DPO	DPO modification	0.6875 win rate
Weighted Smooth DPO	DPO modification	0.7018 win rate

Table 1. Summary of reward-model accuracy and external judge win rates across offline, online, and DPO-modified methods.

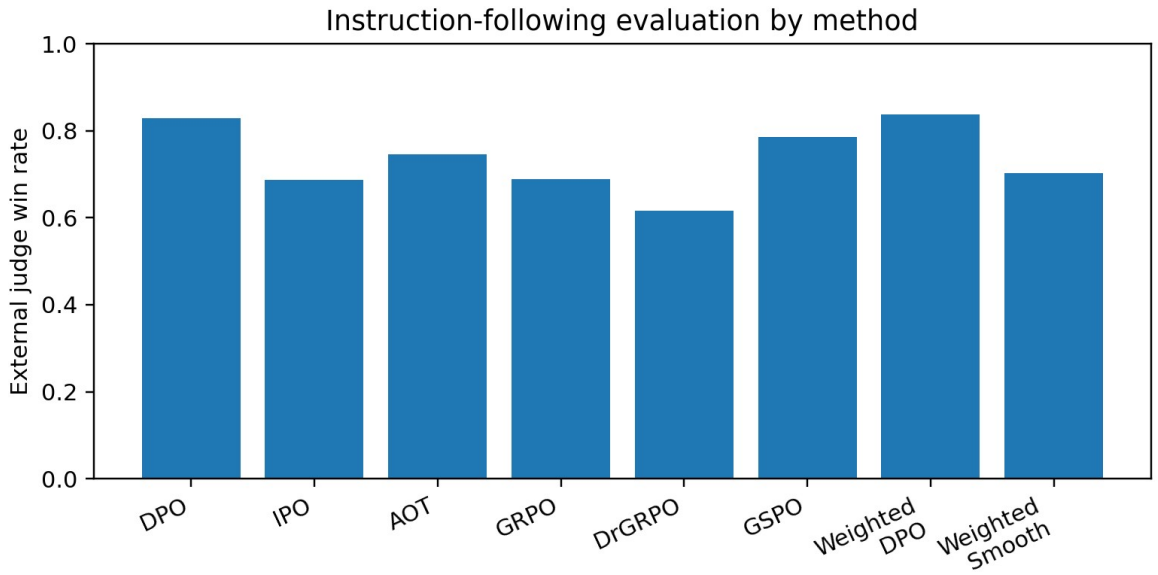

Figure 1. External judge win rates for the main offline, online, and modified DPO methods. Confidence-weighted DPO produced the highest win rate in this comparison.

4.2 Offline preference optimization

Among the offline methods, DPO achieved the strongest performance. Its reference-corrected margin increased steadily during training, and its held-out preference accuracy reached approximately 0.8. AOT achieved intermediate performance, improving distribution-level separation but not matching DPO. IPO was more conservative and exhibited weaker held-out margins, consistent with its bounded target-gap objective.

These results indicate that direct pairwise margin improvement was the most effective offline strategy in this benchmark. Although IPO and AOT operate on similar preference-pair data, their objective structures induce different optimization behavior and different external win rates.

4.3 Online RLHF

The online methods made controlled policy updates, with likelihood ratios remaining close to one and KL divergence increasing gradually. Reward-model scores improved for GRPO, DrGRPO, and GSPO, but these internal improvements did not translate equally into external judge performance. GSPO achieved the strongest online win rate at 0.7857, outperforming GRPO and DrGRPO.

This gap between reward-model diagnostics and external judge win rate suggests that reward-model score is a useful but incomplete proxy for generation quality. In this setting, sequence-level clipping in GSPO appeared to preserve response-level behavior more effectively than token-level clipping.

4.4 DPO modification ablations

The DPO ablations tested whether preference-pair confidence or label-noise robustness could improve a strong DPO baseline. Weighted DPO was the most successful modification. The beta = 0.1 configuration achieved a 0.8372 external judge win rate, exceeding both the vanilla DPO baseline and the other weighted-DPO beta settings.

Label-smoothed DPO was considerably weaker. Smooth DPO and weighted smooth DPO produced lower external win rates even when internal preference accuracy was reasonable. This suggests that the chosen/rejected labels in the benchmark were sufficiently informative that softening the target reduced useful gradient signal. Combining confidence weighting with label smoothing did not preserve the benefit of confidence weighting alone.

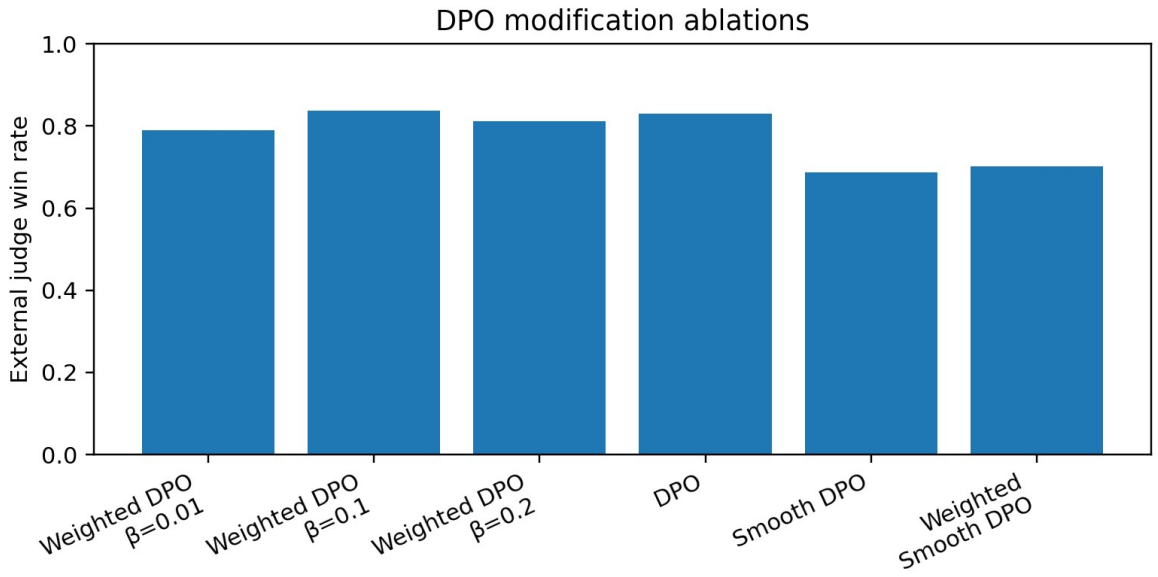

Figure 2. Ablation results for weighted and label-smoothed variants of DPO. Moderate confidence weighting improved over DPO, while label smoothing reduced external judge performance.

5. Discussion

The strongest overall method in these experiments was confidence-weighted DPO. The improvement over DPO was modest, but it is meaningful because DPO was already a strong baseline. The bounded weighting scheme likely helped by emphasizing clearer preference pairs without discarding ambiguous examples or allowing weights to dominate the objective.

The failure of weighted smooth DPO is also informative. Label smoothing is motivated by possible preference-label noise, but in this benchmark it appears to have weakened the learning signal more than it improved robustness. This

result emphasizes that plausible objective modifications must be validated using external generation-quality metrics, not only internal preference accuracy or margin.

The online results show a similar pattern. GRPO, DrGRPO, and GSPO all improved reward-model scores, yet GSPO was much stronger under the external judge. This suggests that sequence-level policy constraints may better preserve coherent response-level quality than token-level clipping in this setting.

6. Limitations and Future Work

- The experiments were conducted under a single controlled benchmark and base model, so the results should not be interpreted as universal rankings of RLHF methods.
- The confidence-weighted DPO study used a small set of weighting schedules and beta values. A broader sweep could test adaptive weights, slower saturation, or reward-model-derived confidence estimates.
- The external judge evaluation is more informative than internal diagnostics, but it is still an automated evaluation. Human evaluation or domain-specific preference analysis would provide a stronger assessment of response quality.
- Future work could jointly study confidence weighting, checkpoint selection, reward-model calibration, and online updates to determine whether weighted offline preference optimization can consistently improve downstream online alignment.

7. Conclusion

This study compared offline and online RLHF methods for instruction following and evaluated objective-level modifications to DPO. DPO was the strongest offline baseline, GSPO was the strongest online method, and a Bradley-Terry reward model achieved strong held-out pair accuracy. The best overall result came from confidence-weighted DPO, which modestly improved external judge win rate over vanilla DPO. In contrast, label smoothing and weighted label smoothing underperformed, showing that robustness-oriented modifications can reduce performance when the underlying preference labels are informative. Overall, the results support confidence weighting as a simple and stable refinement to DPO, while also highlighting the need to evaluate alignment methods using external generation-quality metrics rather than internal diagnostics alone.

References

Christiano, P. F., et al. Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 2017.

Ouyang, L., et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 2022.

Rafailov, R., et al. Direct Preference Optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 2023.

Schulman, J., et al. Proximal Policy Optimization algorithms. arXiv preprint, 2017.