RAISING THE FLOOR, LOWERING THE CEILING: ADAPTIVE REWARD BOUNDS FOR DIRECT PREFERENCE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

An increase in the preference margin optimized by Direct Preference Optimization (DPO) does not imply an increase in the probability assigned to preferred behavior. The margin can also increase if the chosen and rejected rewards both decrease, provided that the rejected decreases more rapidly. This behavior reflects an under-specification in the DPO objective. It constrains only the difference between the policy's and reference model's response scores, rather than their individual evolution. Moreover, the shared prompt, token representations, and parameters induce coupled gradients for the two responses, allowing a simultaneous decrease to constitute a favorable descent direction for the pairwise objective. We propose Adaptive Reward Bounds (ARB), a trajectory-aware regularization framework that augments DPO with explicit directional control. At each step, ARB computes length-normalized changes in the reward, aggregates chosen and rejected values over the batch, and tracks them with sliding-windows. It imposes a lower bound on the chosen statistic and an upper bound on the rejected statistic using detached soft penalties only upon violations. Different from existing methods, ARB adapts its constraints to the training trajectory, tolerates local reversals, and preserves directional pressure through optional biases. Experiments confirm that ARB effectively mitigates the undesired joint decline of chosen and rejected rewards while preserving the preference learning objective.

1 INTRODUCTION

Aligning large language models (LLMs) with human preferences has become a standard stage of post-training, turning capable but unaligned base models into helpful and reliable assistants (Christiano et al., 2017; Ouyang et al., 2022). While the classical pipeline of reinforcement learning from human feedback (RLHF) relies on an explicit reward model and on-policy optimization, Direct Preference Optimization (DPO) (Rafailov et al., 2023) recasts alignment as a single supervised objective on pairs of preferred (chosen) and dispreferred (rejected) responses, implicitly parameterizing the reward through the log-ratio between the policy and a frozen reference model. Owing to its simplicity and stability, DPO and its direct-alignment variants have been adopted widely and now underpin the preference-tuning stage of many open LLMs (Azar et al., 2024; Meng et al., 2024).

Despite this success, the objective optimizes only the margin between the chosen and rejected implicit rewards, and places no constraint on their absolute levels. A growing body of empirical evidence shows that this freedom is routinely exploited in an undesirable way. Over the course of training the reward the model assigns to the chosen response often decreases (Pal et al., 2024; Rafailov et al., 2023). The reward margin still widens, because the rejected reward is pushed down even faster, but the intended "chosen up, rejected down" dynamics collapse into "both down".

Recent analyses trace the effect to the structure of the margin-based gradient itself. Yuan et al. (2025) show that a margin objective under-specifies the desired behavior on the two responses individually, and formalize a gradient entanglement effect. The updates to the chosen and rejected log-probabilities are coupled through the inner product of their gradients, so that whenever the two responses share enough of their representation, suppressing the rejected probability drags the chosen probability down in lockstep, producing a synchronized decrease. Crucially, Razin et al. (2025) fur-

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.

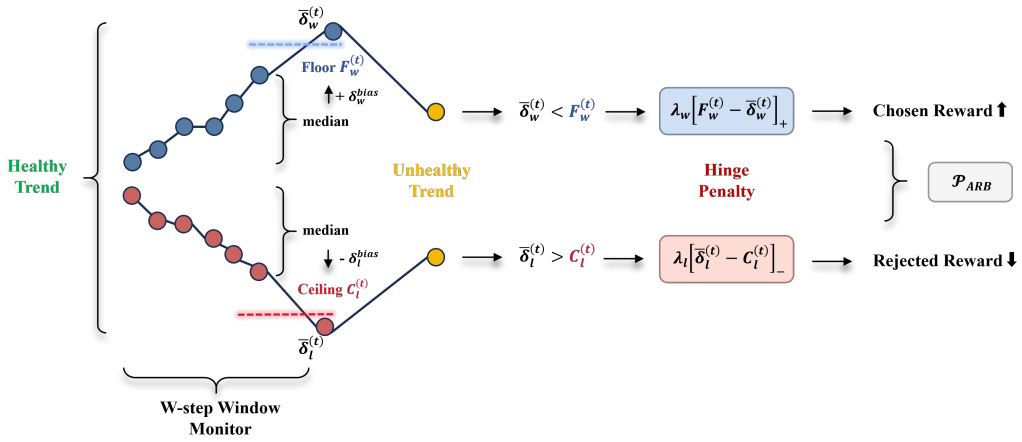


Figure 1: Adaptive Reward Bounds (ARB). Each branch is tracked by the reward offset at optimizer step t. A healthy trend separates the two, the chosen offset moving upward while the rejected moves downward, whereas an unhealthy step (yellow) drives either branch in the wrong direction. From the recent trajectory ARB derives a floor $F_w^{(t)}$ under the chosen offset and a ceiling $C_l^{(t)}$ over the rejected one, each formed as a sliding-window median displaced by a directional bias, so that both bounds track the current training level instead of a fixed anchor. A violated bound activates the corresponding one-sided hinge.

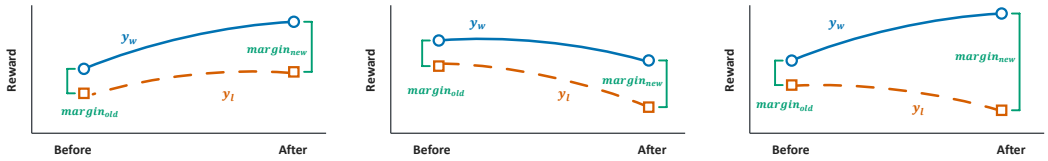
ther demonstrate that this displacement can be catastrophic: probability mass is transferred not to the intended response but to unrelated or even contradictory outputs, so that optimizing the preference objective can unintentionally unalign the model and, in safety-critical settings, raise the likelihood of undesirable responses. A widening margin is therefore not a sufficient indicator of a healthy update; when the chosen likelihood erodes, the very behavior preference learning is meant to reinforce is being weakened. This motivates intervening directly on the absolute reward levels of each response, rather than on their margin alone.

Therefore, a line of related works augment the preference objective to guard the chosen likelihood. DPO-Positive (Pal et al., 2024) adds a gated penalty that discourages the chosen log-probability from dropping below its value under the reference model, and a negative log-likelihood term on the chosen response is used to similar effect in large-scale recipes (Grattafiori et al., 2024). Others regularize the reject penalty for robustness (Xie et al., 2024), or rethink the role of the rejected response to balance the two sides of the update (Cho et al., 2025). These methods confirm that constraining absolute levels is beneficial, but the ones that place an explicit bound share two limitations. First, the anchor is static that the chosen reward is held against a fixed reference, which cannot express the goal of a reward that should keep rising as training proceeds, and which ignores how far training has already moved. Second, the bound is typically one-sided, protecting the chosen response while leaving the trajectory of the rejected reward unmanaged, even though it is the joint dynamics of both quantities that characterize the failure.

We propose **Adaptive Reward Bounds (ARB)** in Fig. 1, which addresses both limitations by bounding the two implicit rewards directly and dynamically. For each side we maintain a bound derived from the policy model's own recent training trajectory, a lower floor on the chosen reward and an upper ceiling on the rejected reward, each computed as the median of the batch-level reward over a sliding window of optimizer steps and offset by a directional bias. The window follows the overall trend while tolerating the local fluctuation inherent to stochastic optimization, and the bias supplies a persistent corrective pressure so that the constraint continues to push in the intended direction rather than dissolving as the rewards drift.

We summarize our contributions as follows:

- We reformulate the objective to bound each branch individually. A floor on the chosen reward and a ceiling on that of the rejected response are operated in the per-token level and we also render the two bounds comparable across batches.



(a) Chosen & Rejected Up (b) Chosen & Rejected Down (c) Chosen Up & Rejected Down

Figure 2: Three ways to widen the reward margin. In all three the margin grows identically. (a) Both rise, the chosen faster, where the rejected reward is inadvertently reinforced even as the margin widens. (b) Both fall, the rejected faster, where the chosen reward collapses even as the margin widens. (c) The intended dynamics, the chosen rising while the rejected falls.

- Instead of fixing the bounds apriori, we estimate them online from the recent optimization history through a sliding-window median, offset by a directional bias that preserves corrective pressure. The resulting band follows the overall trend while tolerating local oscillation, and is imposed by a one-sided hinge penalty.
- The gradient-level analysis explains that a DPO update realizes the intended dynamics only when the chosen and rejected gradients are weakly aligned, whereas the two bounds of ARB supply an extra degree of freedom that provably enlarges the set of configurations on which this dynamics is attainable, exactly covering the regime in which DPO fails.
- Results show that ARB restores the intended "chosen up, rejected down" trajectory and widens the margin in a healthy regime. This recovered dynamics translates into improved downstream performance, leaving ARB a crucial addition to existing preference optimization pipelines.

2 Preliminary

2.1 Failure Mode in DPO

We consider preference optimization over a dataset $\mathcal{D} = \{(x, y_w, y_l)\}$, where each prompt x is paired with a preferred (chosen) response y_w and a dispreferred (rejected) response y_l. DPO trains a policy π_θ that is initialized from a frozen reference model π_{ref} (typically the supervised fine-tuned checkpoint). Instead of learning an explicit reward model, DPO defines an implicit reward through the log-ratio between the policy and the reference,

$$r_\theta(x, y) \;=\; \beta \,\log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)}, \tag{1}$$

where $\beta > 0$ scales the permitted deviation from the reference. Fitting this reward to the pairwise preferences under a Bradley–Terry model yields the objective

$$\mathcal{L}_{\text{DPO}}(\theta) \;=\; -\,\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log \sigma\big(r_\theta(x, y_w) - r_\theta(x, y_l)\big)\Big], \tag{2}$$

where σ is the logistic function. The two responses enter the loss only through the margin $m_\theta = r_\theta(x, y_w) - r_\theta(x, y_l)$, which the objective drives to be large and positive.

Because σ is monotonic, minimizing Eq. 2 amounts to nothing more than enlarging the margin m_θ. As shown in Fig. 2, the margin grows, however, whenever the chosen reward moves upward relative to the rejected one, irrespective of the absolute direction of either: both may rise, both may fall, or the chosen may rise as the rejected falls. All three are equally optimal in the eyes of the loss, and only the last realizes the intended dynamics. This latitude is amplified by the fact that a single policy π_θ scores both responses through one shared set of model parameters. An update that suppresses y_l is applied to the same network that produces y_w, and especially when the two responses share the prompt and much of their surface form, the same update tends to pull the chosen probability down alongside the rejected one. The margin then widens exactly as the DPO loss demands, yet the likelihood of the preferred response declines.

Table 1: Preference objectives that constrain the absolute reward levels, expressed through the reference-relative log-ratio $s_\theta(x, y) = \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$. For AlphaPO, $\bar{p}_\theta(y) = \pi_\theta(y \mid x)^{1/|y|}$ is the length-normalized sequence probability and the reward is reference-free.

Method	Objective
DPO-Positive (Pal et al., 2024)	$-\log \sigma\big(\beta s_w - \beta s_l - \lambda \operatorname{ReLU}(-s_w)\big)$
Cal-DPO (Xiao et al., 2024)	$-\log \sigma\big(s_w - s_l\big) + \big(s_w - r_w/\beta\big)^2 + \big(s_l - r_l/\beta\big)^2$
DPO-Shift (Yang et al., 2025)	$-\log \sigma\big(\beta s_w - f(\lambda)\,\beta s_l\big), \quad 0 < f(\lambda) < 1$
AlphaPO (Gupta et al., 2025)	$-\log \sigma\big(r_\alpha(y_w) - r_\alpha(y_l) - \gamma\big), \quad r_\alpha(y) = \frac{\beta}{\alpha}\big(1 - \bar{p}_\theta(y)^{-\alpha}\big)$
ARB (ours)	$-\log \sigma\big(\beta s_w - \beta s_l\big) + \lambda_w \operatorname{ReLU}\big(F_w^{(t)} - \bar{\delta}_w^{(t)}\big) + \lambda_l \operatorname{ReLU}\big(\bar{\delta}_l^{(t)} - C_l^{(t)}\big)$

2.2 Related Work

DPO-Positive (Pal et al., 2024) gates a floor on the chosen log-ratio at the fixed reference level $s_w = 0$. A chosen-side floor helps, but the pressure vanishes once $s_w \geq 0$ even if the chosen reward is still being dragged down, and the rejected side is left unmanaged. Cal-DPO (Xiao et al., 2024) regresses both log-ratios onto fixed calibration targets r_w/β and r_l/β with squared penalties. It acts on both sides and pushes the rejected reward down explicitly, but the targets are constants. DPO-Shift (Yang et al., 2025) scales the rejected reward by $f(\lambda) \in (0, 1)$ inside the logistic, easing the downward pull on the chosen side through the shared-parameter coupling. The effect is indirect, gives no explicit control over the chosen level and trades directly against the margin. AlphaPO (Gupta et al., 2025) reshapes the reference-free SimPO reward through an exponent α, recovering the log reward as $\alpha \to 0$ and bending the curve otherwise. It removes the anchor against and reshapes the reward geometry globally. ARB relaxes the fixed choices above along three axes in Tab. 1.

- **Adaptive, not static.** ARB derives $F_w^{(t)}$ and $C_l^{(t)}$ from the policy's own recent trajectory, so the constraint tracks where training currently is, tolerates local reversals, and can keep demanding improvement.
- **Two-sided, not one-sided.** ARB floors the chosen and caps the rejected at once, matching the fact that displacement is a joint property of both trajectories.
- **Violation-triggered, not always-on.** The two penalties are one-sided hinges detached from the gradient of the bound, so they exert force only upon violation and vanish when the update is already healthy.

3 Adaptive Reward Bounds (ARB)

3.1 Floor and Ceiling

A healthy update should raise the chosen reward and lower the rejected one. The "both-down" failure is precisely a violation of this on the chosen side. ARB encodes the requirement directly, as a two-sided constraint on the per-step implicit reward. Let $\bar{\delta}_w^{(t)}$ and $\bar{\delta}_l^{(t)}$ denote a per-step summary of the chosen and rejected reward at optimizer step t. ARB asks that at every step

$$\bar{\delta}_w^{(t)} \;\geq\; F_w^{(t)} \qquad \text{and} \qquad \bar{\delta}_l^{(t)} \;\leq\; C_l^{(t)}, \tag{3}$$

where $F_w^{(t)}$ is a floor under the chosen reward and $C_l^{(t)}$ a ceiling over the rejected reward. The two constraints act in opposite directions and together target the *joint* dynamics of the pair.

3.2 Adaptive Window

A fixed threshold cannot express the expected goal. The chosen reward should keep rising rather than merely stay above a constant floor, and how far training has already moved should inform where the bound sits. ARB therefore derives each bound from the recent trajectory itself, combining an adaptive window that follows the training level with a directional bias that supplies persistent corrective pressure.

Concretely, ARB maintains a sliding window of the most recent W optimizer steps for each branch and sets

$$F_w^{(t)} \;=\; \mathrm{median}\big(\bar{\delta}_w^{(t-W+1)}, \ldots, \bar{\delta}_w^{(t)}\big) \;+\; \delta_w^{\mathrm{bias}}, \tag{4}$$

$$C_l^{(t)} \;=\; \mathrm{median}\big(\bar{\delta}_l^{(t-W+1)}, \ldots, \bar{\delta}_l^{(t)}\big) \;-\; \delta_l^{\mathrm{bias}}. \tag{5}$$

The window statistic uses the median, so that an occasional hard batch does not drag the bound around. First, because each bound is recomputed from the current window with no running maximum or minimum, it is free to move downward as well as upward. This tolerates the local fluctuation inherent to stochastic optimization during training and tracks only the overall trend. Second, the directional bias $\delta_w^{\mathrm{bias}}, \delta_l^{\mathrm{bias}} \geq 0$ offsets the band away from the current level. They lift the floor under the chosen reward and press the ceiling over the rejected reward. Hence the constraint keeps pushing in the intended direction as the rewards drift. Because the band is anchored to the model's own recent level, the constraint remains meaningful throughout training and adapts to how far optimization has already moved.

3.3 Normalized Offset

Reward control. Restricting the raw policy log-probability directly is undesirable. It may be confounded by how intrinsically a response is. We follow the log-probability relative to the reference model. Because the reference and the policy share the same intrinsic presentation for a given response, their difference isolates the movement induced by preference optimization, which exactly the displacement ARB seeks to control.

Length control. The reference-relative offset scales with response length. We divide by the number of tokens, yielding a length-normalized offset for a response y.

$$\delta_\theta(x, y) \;=\; \frac{1}{|y|}\Big(\log \pi_\theta(y \mid x) - \log \pi_{\mathrm{ref}}(y \mid x)\Big), \tag{6}$$

is comparable across responses with different lengths.

Batch control. ARB aggregates it to the batch mean at each optimizer step t,

$$\bar{\delta}_w^{(t)} \;=\; \frac{1}{|\mathcal{B}_t|} \sum_{(x,y_w)\in\mathcal{B}_t} \delta_\theta(x, y_w), \qquad \bar{\delta}_l^{(t)} \;=\; \frac{1}{|\mathcal{B}_t|} \sum_{(x,y_l)\in\mathcal{B}_t} \delta_\theta(x, y_l), \tag{7}$$

for the chosen and rejected branches, where $\mathcal{B}_t$ is the batch consumed at step t.

These are the per-step signals $\bar{\delta}_w^{(t)}, \bar{\delta}_l^{(t)}$ used throughout. In these cases, the intended healthy dynamics is simply that $\bar{\delta}_w$ trends upward and $\bar{\delta}_l$ downward as training proceeds.

3.4 Hinge Penalty

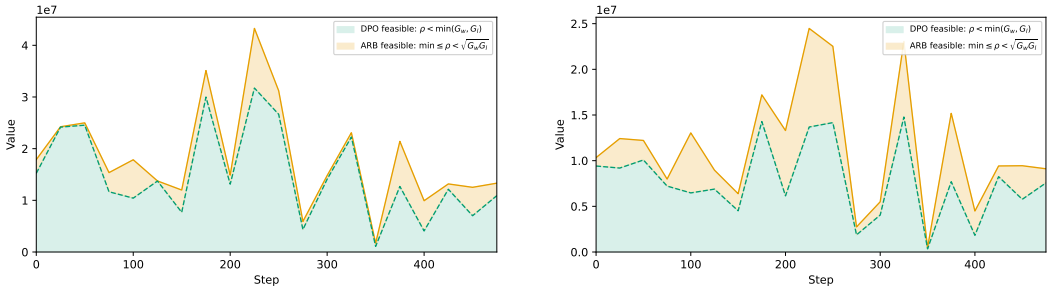
The constraint in Eq. 3 is enforced softly, by a hinge penalty on the offset that is active only when a bound is violated:

$$\mathcal{P}^{(t)} \;=\; \lambda_w \,\mathrm{ReLU}\big(F_w^{(t)} - \bar{\delta}_w^{(t)}\big) \;+\; \lambda_l \,\mathrm{ReLU}\big(\bar{\delta}_l^{(t)} - C_l^{(t)}\big), \tag{8}$$

here $\lambda_w, \lambda_l \geq 0$ weight the two penalties. The hinge applies force only upon violation and otherwise leaves the DPO gradient untouched, so the regularizer corrects the pathological direction without stalling normal preference learning. To avoid acting on unreliable statistics, the penalty is disabled during an initial warmup and until the window is full.

4 Feasibility Condition

App. A shows that ARB offers a successful paradigm by examining the gradient step. Because the chosen and rejected responses share a prompt, many overlapping tokens, and one set of parameters, their log-probability gradients g_w and g_l are typically aligned. We denote their norms and inner product by $G_w = \|g_w\|^2$, $G_l = \|g_l\|^2$, and $\rho = g_w^\top g_l$, so that ρ measures this alignment. A



(a) Llama3.1-8B Instruct (b) Qwen3-8B Instruct

Figure 3: The two feasibility bounds measured along a DPO run. At steps we probe the gradients and record DPO threshold $\min(G_w, G_l)$ and ARB threshold $\sqrt{G_w G_l}$. The shaded region below is where a DPO or ARB step attains the intended dynamics.

plain DPO step raises the chosen log-probability while lowering the rejected one, if and only if $\rho < \min(G_w, G_l)$. This condition is broken when the alignment is strong and the suppressed rejected response carries the larger gradient. The chosen likelihood is then driven down together with the rejected one, which is precisely the displacement failure.

ARB relaxes this condition. Its floor and ceiling contribute separate, non-negative weights to the chosen and rejected gradients, so the update need no longer treat the two sides symmetrically; this single extra degree of freedom widens the range over which the intended dynamics can be realized to

$$\underbrace{\rho < \min(G_w, G_l)}_{\text{DPO feasible}} \quad \longrightarrow \quad \underbrace{\rho < \sqrt{G_w G_l}}_{\text{ARB feasible}}. \tag{9}$$

The two thresholds differ whenever $G_w \neq G_l$, and the intervening region $\min(G_w, G_l) \le \rho < \sqrt{G_w G_l}$ is precisely the displacement regime where DPO fails. The gap closes only when the gradients are exactly collinear, which the strict Cauchy–Schwarz inequality rules out for genuinely distinct responses. Thus ARB does not merely penalize the symptom, but also enlarges the feasible set of the underlying optimization.

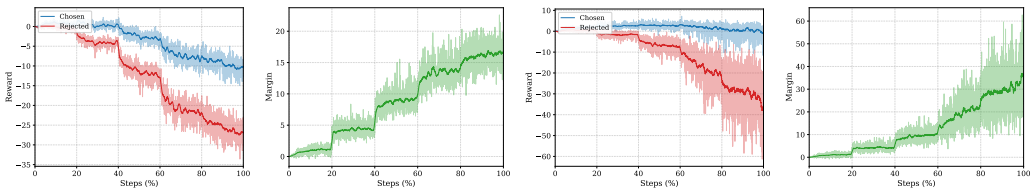
Fig. 3 verifies Eq. 9 empirically by probing G_w, G_l on two instruction-tuned models. The ARB threshold $\sqrt{G_w G_l}$ lies strictly above the DPO threshold $\min(G_w, G_l)$ at essentially every step, and the band between them is the feasibility condition gap that ARB admits but DPO excludes. This relaxation is therefore substantial throughout training.

5 Experiments

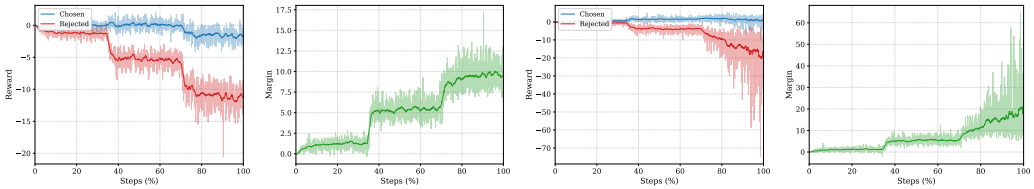
5.1 Main Results

We first explore whether ARB removes the failure mode during training. Fig. 4 reports under DPO the margin grows steadily. The chosen reward does not rise but declines, and it declines alongside the rejected reward. ARB produces a qualitatively different trajectory. The floor prevents the chosen reward from falling below its adaptive lower bound and the ceiling presses the rejected reward beneath its upper bound. The chosen reward stays at or above its original level throughout training, which confirms that the recovered margin reflects a genuine gain in the preferred likelihood rather than a slower joint collapse. The two reward curves separate in the direction the objective was always meant to encourage.

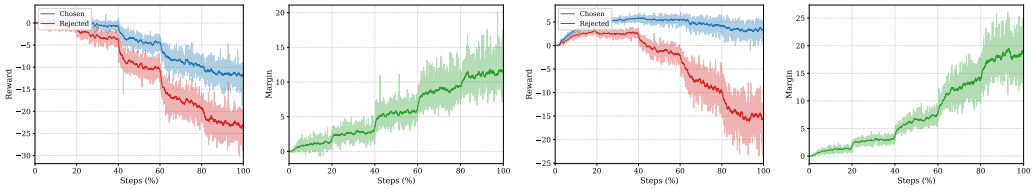
We evaluate the DPO and ARB objective with the two policy probabilities $\pi_\theta(y_w \mid x)$ and $\pi_\theta(y_l \mid x)$. In Fig. 5, brighter regions are lower loss. Compared with DPO, ARB enlarges this low-loss region. Actually ARB shifts where the optimization settles. Since descent follows the loss downward in training, widening the low-loss region toward large $\pi_\theta(y_w \mid x)$ makes the configurations with a rising chosen probability the ones the model is more inclined to converge on.



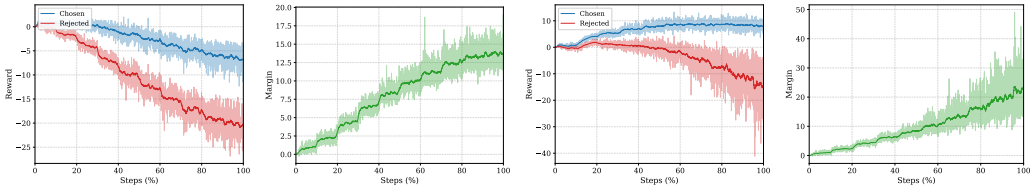
(a) Llama3.1-8B Base



(b) Llama3.1-8B Instruct



(c) Qwen3-8B Base



(d) Qwen3-8B Instruct

Figure 4: Reward dynamics under DPO (left) and ARB (right) across models. Under DPO the margin grows while both rewards fall together. Under ARB the floor holds the chosen reward above its original level and the ceiling presses the rejected reward down, so the same margin growth is realized through the intended "chosen up, rejected down" dynamics.

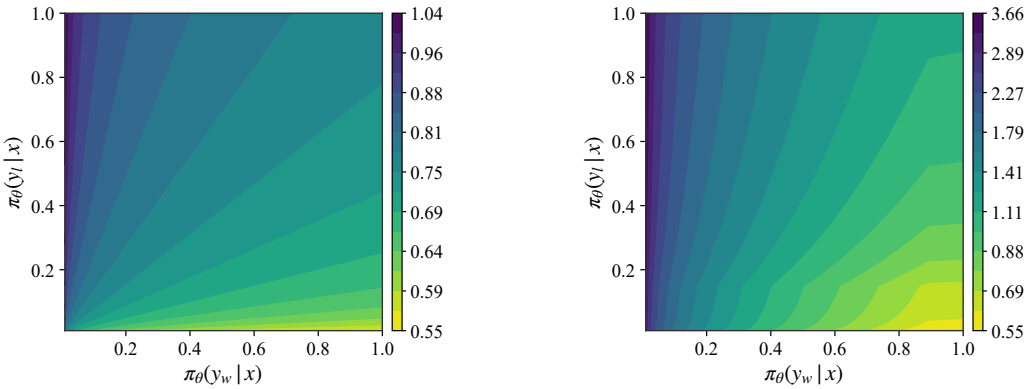
App. C repeats the comparison on HH-RLHF, where DPO exhibits the opposite pathology in some steps: both rewards rise, the chosen faster, so the margin again grows while the policy simultaneously increases the reward it assigns to the rejected response. Both unhealthy patterns follow from Fig. 2, and the gradient geometry determines which one appears. From Eq. 15 a DPO step moves the chosen offset with the sign of $G_w - \rho$ and the rejected offset with the sign of $\rho - G_l$, so the two branches move together exactly when ρ lies between the two gradient norms,

$$\underbrace{G_w \le \rho < G_l}_{\text{both down}} \quad \text{or} \quad \underbrace{G_l < \rho < G_w}_{\text{both up}}. \tag{10}$$

5.2 Downstream Performance

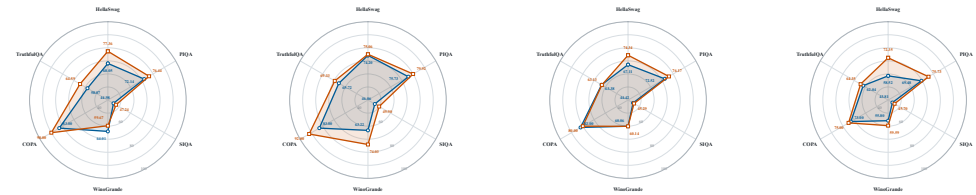
The reward trajectories establish that ARB changes how the margin is obtained. We therefore evaluate these models on six benchmarks, namely HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019), WinoGrande (Sakaguchi et al., 2021), COPA (Roemmele et al., 2011) and TruthfulQA (Lin et al., 2022), to examine whether the recovered reward of the chosen response corresponds to a genuinely more capable model.

Fig. 6 reports the comparison on all four models. ARB improves over DPO on the large majority of the axes. It can be seen that preventing the joint decline does not come at the cost of downstream



(a) DPO (b) ARB

Figure 5: Contour maps of the loss. Under ARB, the lowest loss is now concentrated more where the chosen probability is large and the rejected small.



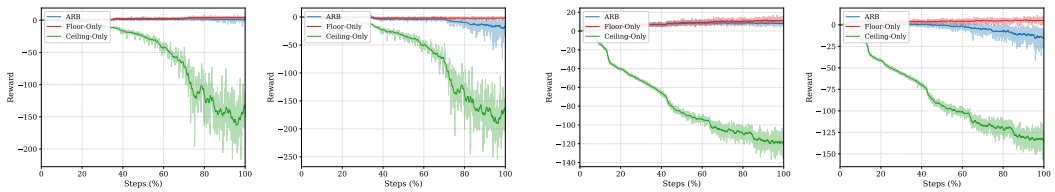
(a) Llama3.1-8B Base (b) Llama3.1-8B Instruct (c) Qwen3-8B Base (d) Qwen3-8B Instruct

Figure 6: Downstream capability across HellaSwag (normalized), TruthfulQA (MC2), Winogrande, PIQA (normalized), SIQA and COPA on DPO (blue) and ARB (orange). Each axis reports accuracy on one benchmark. ARB almost matches or exceeds DPO overall, indicating that its healthier reward dynamics translate into better downstream task performance. More results are in App. D.

ability, which rules out the concern that ARB purchases a healthier reward trajectory by restraining the policy.

5.3 Ablations

We disable one bound at a time to isolate the contribution of each. Fig. 7 shows that each one-sided variant moves both branches in the direction of its own force. With the floor alone the chosen reward rises as intended, and the rejected reward rises with it. With the ceiling alone the rejected reward falls steeply as intended, and the chosen reward falls with it over a comparable range. Neither one-



(a) Llama3.1-8B Instruct (b) Qwen3-8B Instruct

Figure 7: Ablation of the two bounds in ARB. For each model we plot the chosen reward (left) and the rejected reward (right) over training under three settings: the floor alone ($\lambda_l = 0$), the ceiling alone ($\lambda_w = 0$), and full ARB.

sided result means that the active bound has failed on the branch it constrains. However, updating the reward along one branch's gradient necessarily moves the other branch's reward in the same direction.

Removing a penalty restricts which values of t in Eq. 19 remain reachable. Since $t = \frac{a+b_w}{a+b_l}$ with $a > 0$, setting $b_l = 0$ leaves $t \in [1, \infty)$ and setting $b_w = 0$ leaves $t \in (0, 1]$, so each variant retains one half-line rather than the full ray $(0, \infty)$. For $\rho > 0$ the conditions require $t \in (\rho/G_w,\, G_l/\rho)$, and a one-sided variant is feasible exactly when its half-line meets that interval. The interval is nonempty iff $\rho^2 < G_w G_l$. It extends above 1 iff $G_l/\rho > 1$ and below 1 iff $\rho/G_w < 1$. Hence

$$\underbrace{\rho < \min(G_l, \sqrt{G_w G_l})}_{\text{floor-only feasible}}, \qquad \underbrace{\rho < \min(G_w, \sqrt{G_w G_l})}_{\text{ceiling-only feasible}}, \tag{11}$$

Each lies between the DPO set $\rho < \min(G_w, G_l)$ and the ARB set $\rho < \sqrt{G_w G_l}$.

Eq. 11 shows that a one-sided variant is not a uniformly weaker version of ARB. If the rejected response carries the larger gradient, $G_w \le G_l$ gives $\sqrt{G_w G_l} \le G_l$, so the floor alone attains the full ARB set while the ceiling alone collapses to $\rho < G_w = \min(G_w, G_l)$. The chosen is the same. A single bound thus carries the configurations that match its direction. Each supplies the degree of freedom the other lacks. This is the sense in which the two are not separable improvements in ARB.

6 Discussion

The margin is not the truth. Preference learning inherits its objective from a model of comparison. The Bradley–Terry likelihood is a statement about which of two responses is preferred, and a difference of rewards is a faithful expression of that statement. What it cannot express in training is the behavior we actually want from the trained policy, which is that the preferred response become more likely and the rejected less. These two requirements are not equivalent. The margin rises monotonically while the policy either erodes the behavior it was meant to reinforce or strengthens the behavior it was meant to suppress. The consequence we would emphasize is that the margin has been asked to serve as both the objective and the evidence. Preference optimization therefore needs a target that is stated on the two responses individually.

Monitoring and Prevention. Focusing on the target on the two responses individually invites a natural but mistaken reading, that every step should be inspected and every wrong step repaired. Local oscillation is not the failure but the ordinary behavior of optimization. What separates a healthy run from an unhealthy one is the direction the two branches travel over many steps, and a criterion evaluated step by step cannot see that direction. Monitoring and prevention therefore share the same object that the quantity worth watching is the overall trend of each branch.

Limitation. ARB reads both bounds off a reference-relative quantity. We remove the reference model and apply the same construction to a reference-free objective in the manner of SimPO. Fig. 10 shows the two branches no longer separate as they do under a reference, and the bounds provide markedly less directional control even though the margin continues to behave as the objective requires. We guess once the reference anchor is removed, the trend the bound is meant to follow is estimated from a noisier quantity. We scope our claims to objectives that retain a reference model. Whether ARB can be carried into the reference-free setting is an open question. We leave this to future work.

7 Conclusion

A preference objective built on a comparison certifies only the order of two rewards, whereas alignment asks that the preferred response become more likely and the rejected less. Inside it the margin may rise exactly as the loss demands while the pair moves together in either direction. ARB follows from this diagnosis and bounds each branch against the policy's own recent trajectory. The results show that ARB restores the intended "chosen up, rejected down" trajectory in both directions of the failure and the recovered dynamics carries over to downstream performance.

AI USE STATEMENT

In this work, we used generative AI tools for assistance in the writing of proofs. Additionally, we used generative AI tools to polish the paper to improve readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human-subject experiments or the collection of personally identifiable or sensitive information. The datasets and resources used in our experiments are obtained from publicly available sources and are used in accordance with their respective licenses and intended research purposes. The proposed method is developed for research purposes and contribute to improving the capabilities of AI systems.

REPRODUCIBILITY STATEMENT

The method is specified in full in Sec. 3.1–Sec. 3.4. The theoretical claims are supported in App. A. App. B restates the pseudocode. On the experimental side, App. F reports the training configuration and the hyperparameters, and all runs use the public implementation with publicly available checkpoints and the public preference datasets. The anonymous source code is included in the supplementary material.

REFERENCES

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

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.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

Jae Hyeon Cho, JunHyeok Oh, Myunsoo Kim, and Byung-Jun Lee. Rethinking dpo: The role of rejected responses in preference misalignment. In *EMNLP (Findings)*, pp. 8159–8176, 2025.

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.

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.

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.

Aman Gupta, Shao Tang, Qingquan Song, Sirou Zhu, Jiwoo Hong, Ankan Saha, Viral Gupta, Noah Lee, Eunki Kim, Siyu Zhu, et al. Alphapo: Reward shape matters for llm alignment. *arXiv preprint arXiv:2501.03884*, 2025.

Jian Hu, Xibin Wu, Wei Shen, Jason Klein Liu, Weixun Wang, Songlin Jiang, Haoran Wang, Hao Chen, Bin Chen, Wenkai Fang, et al. Openrlhf: A ray-based easy-to-use, scalable and high-performance rlhf framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 656–666, 2025.

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.

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.

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*, 2024.

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.

Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. Unintentional unalignment: Likelihood displacement in direct preference optimization. In *International Conference on Learning Representations*, volume 2025, pp. 24791–24834, 2025.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pp. 90–95, 2011.

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.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 4463–4473, 2019.

Teng Xiao, Yige Yuan, Huaisheng Zhu, Mingxiao Li, and Vasant G Honavar. Cal-dpo: Calibrated direct preference optimization for language model alignment. *Advances in Neural Information Processing Systems*, 37:114289–114320, 2024.

Shiming Xie, Hong Chen, Fred Yu, Zeye Sun, Xiuyu Wu, and Yingfan Hu. Minor dpo reject penalty to increase training robustness. *arXiv preprint arXiv:2408.09834*, 2024.

Xiliang Yang, Feng Jiang, Qianen Zhang, Lei Zhao, and Xiao Li. Dpo-shift: Shifting the distribution of direct preference optimization. *arXiv preprint arXiv:2502.07599*, 2025.

Hui Yuan, Yifan Zeng, Yue Wu, Huazheng Wang, Mengdi Wang, and Liu Leqi. A common pitfall of margin-based language model alignment: Gradient entanglement. In *International Conference on Learning Representations*, volume 2025, pp. 5278–5297, 2025.

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.

A PROOF

We work to first order in the learning rate η under plain gradient descent, fix a single preference triple (x, y_w, y_l) at parameters θ with π_{ref} held constant, and assume $g_w, g_l \neq 0$. With the notation of Sec. 4,

$$g_w = \nabla_\theta \log \pi_\theta(y_w \mid x), \quad g_l = \nabla_\theta \log \pi_\theta(y_l \mid x), \quad G_w = \|g_w\|^2, \quad G_l = \|g_l\|^2, \quad \rho = g_w^\top g_l, \tag{12}$$

the Cauchy–Schwarz inequality gives $|\rho| \leq \sqrt{G_w G_l}$, with equality iff g_w and g_l are collinear. For an update $\Delta\theta$, a first-order expansion yields

$$\Delta \log \pi_\theta(y_w \mid x) = g_w^\top \Delta\theta + o(\|\Delta\theta\|), \qquad \Delta \log \pi_\theta(y_l \mid x) = g_l^\top \Delta\theta + o(\|\Delta\theta\|). \tag{13}$$

Since π_{ref} is fixed, the reference-relative offset δ_θ of Eq. 6 changes by the same expressions up to the positive factor $1/|y|$, so it suffices to track the log-probabilities. The target dynamics is

$$\Delta \log \pi_\theta(y_w \mid x) > 0 \qquad \text{and} \qquad \Delta \log \pi_\theta(y_l \mid x) < 0. \tag{14}$$

DPO step. Differentiating $\mathcal{L}_{\text{DPO}} = -\log \sigma(m_\theta)$ with $m_\theta = \beta\big(\log \pi_\theta(y_w \mid x) - \log \pi_\theta(y_l \mid x)\big)$ gives $\nabla_\theta \mathcal{L}_{\text{DPO}} = -\beta w\,(g_w - g_l)$, where $w = \sigma(-m_\theta) \in (0, 1)$. The descent step $\Delta\theta_{\text{DPO}} = \eta\beta w\,(g_w - g_l)$ substituted into Eq. 13 yields

$$\Delta \log \pi_\theta(y_w \mid x) = \eta\,\beta\,w\,(G_w - \rho), \qquad \Delta \log \pi_\theta(y_l \mid x) = \eta\,\beta\,w\,(\rho - G_l). \tag{15}$$

As $\eta\beta w > 0$, the chosen log-probability rises iff $\rho < G_w$ and the rejected falls iff $\rho < G_l$, so Eq. 14 holds iff

$$\rho \;<\; \min(G_w, G_l). \tag{16}$$

When $\rho \geq G_w$ the chosen log-probability decreases, yet the margin still grows, since $\Delta \log \pi_\theta(y_w) - \Delta \log \pi_\theta(y_l) = \eta\beta w\,(G_w + G_l - 2\rho) \geq 0$ by Cauchy–Schwarz.

ARB step. Consider a step at which both hinges in Eq. 8 are active. An inactive penalty contributes no gradient and leaves Eq. 15 unchanged. The bounds are detached from the computation graph, and with π_{ref} fixed $\nabla_\theta \delta_\theta(x, y) = \frac{1}{|y|}\nabla_\theta \log \pi_\theta(y \mid x)$, so the active penalty $\mathcal{P} = \lambda_w(F_w - \delta_\theta(x, y_w)) + \lambda_l(\delta_\theta(x, y_l) - C_l)$ has gradient $\nabla_\theta \mathcal{P} = -\frac{\lambda_w}{|y_w|} g_w + \frac{\lambda_l}{|y_l|} g_l$. Descending along $\mathcal{L}_{\text{DPO}} + \mathcal{P}$ gives

$$\Delta\theta_{\text{ARB}} = \eta\big[(a + b_w)\,g_w - (a + b_l)\,g_l\big], \qquad a = \beta w > 0, \;\; b_w = \tfrac{\lambda_w}{|y_w|} \geq 0, \;\; b_l = \tfrac{\lambda_l}{|y_l|} \geq 0, \tag{17}$$

with DPO recovered at $b_w = b_l = 0$. Substituting into Eq. 13,

$$\Delta \log \pi_\theta(y_w \mid x) = \eta\big[(a+b_w)G_w - (a+b_l)\rho\big], \qquad \Delta \log \pi_\theta(y_l \mid x) = \eta\big[(a+b_w)\rho - (a+b_l)G_l\big]. \tag{18}$$

Dividing by $\eta(a + b_l) > 0$ and writing $t = \frac{a+b_w}{a+b_l}$, Eq. 14 becomes

$$t\,G_w > \rho \qquad \text{and} \qquad t\,\rho < G_l. \tag{19}$$

DPO is fixed at $t = 1$, whereas $b_w, b_l \geq 0$ realize any $t \in (0, \infty)$, taking $b_l = 0$ sweeps $t \in [1, \infty)$ and taking $b_w = 0$ sweeps $t \in (0, 1]$, since $a > 0$.

Feasibility. We determine when some admissible t satisfies Eq. 19. If $\rho \leq 0$, both inequalities hold for every $t > 0$, as $tG_w > 0 \geq \rho$ and $t\rho \leq 0 < G_l$. If $\rho > 0$, they read $t > \rho/G_w$ and $t < G_l/\rho$, and such a t exists iff $\rho/G_w < G_l/\rho$, i.e. $\rho^2 < G_w G_l$. Both cases reduce to

$$\rho \;<\; \sqrt{G_w G_l}, \tag{20}$$

which by Cauchy–Schwarz fails only when g_w, g_l are positively collinear. Comparing with Eq. 16, DPO succeeds on $\rho < \min(G_w, G_l)$ and ARB on the strictly larger set $\rho < \sqrt{G_w G_l}$. The difference

$$\min(G_w, G_l) \;\leq\; \rho \;<\; \sqrt{G_w G_l} \tag{21}$$

is nonempty whenever $G_w \neq G_l$. In this regime the binding inequality in Eq. 19 is $t > \rho/G_w > 1$, met by increasing b_w, i.e. strengthening the chosen-side floor weight λ_w.

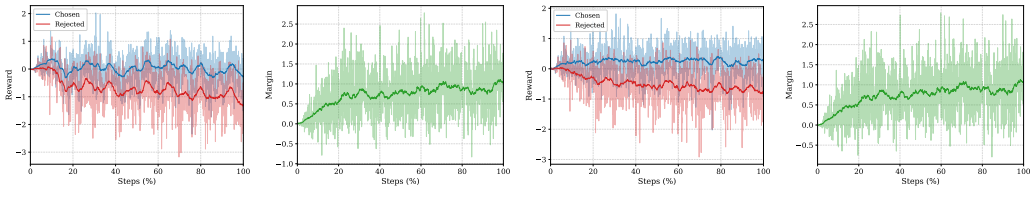
B PSEUDOCODE

Algorithm 1 Adaptive Reward Bounds (ARB)

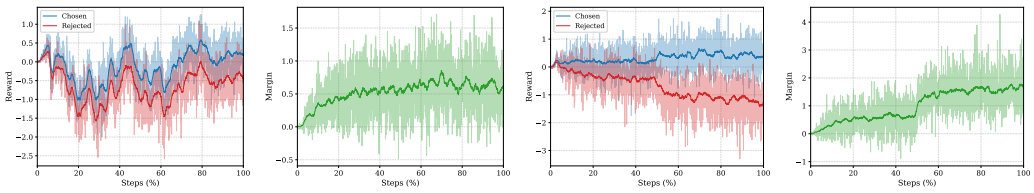
Require: policy π_θ, frozen reference π_{ref}, temperature β, weights λ_w, λ_l, warmup T_0, window size W, slide period S, biases $\delta_w^{\text{bias}}, \delta_l^{\text{bias}} \geq 0$

init queues $Q_w, Q_l \leftarrow \varnothing$; bounds $F_w, C_l \leftarrow \perp$; READY $\leftarrow$ FALSE; counter $c \leftarrow 0$
for optimizer step $t = 1, 2, \ldots$ **do**
  sample batch $\mathcal{B}_t = \{(x, y_w, y_l)\}$
  compute $\log \pi_\theta$ and $\log \pi_{\text{ref}}$ for y_w, y_l
  $\mathcal{L}_{\text{DPO}} \leftarrow -\frac{1}{|\mathcal{B}_t|} \sum \log \sigma\big(\beta s_w - \beta s_l\big), \quad s_\cdot = \log \frac{\pi_\theta(y_\cdot|x)}{\pi_{\text{ref}}(y_\cdot|x)}$
  $\bar{\delta}_w^{(t)} \leftarrow \frac{1}{|\mathcal{B}_t|} \sum \frac{1}{|y_w|} \big(\log \pi_\theta(y_w \mid x) - \log \pi_{\text{ref}}(y_w \mid x)\big)$
  $\bar{\delta}_l^{(t)} \leftarrow \frac{1}{|\mathcal{B}_t|} \sum \frac{1}{|y_l|} \big(\log \pi_\theta(y_l \mid x) - \log \pi_{\text{ref}}(y_l \mid x)\big)$
  if READY **and** $t \geq T_0$ **then**
    $\mathcal{P}^{(t)} \leftarrow \lambda_w \operatorname{ReLU}\big(F_w - \bar{\delta}_w^{(t)}\big) + \lambda_l \operatorname{ReLU}\big(\bar{\delta}_l^{(t)} - C_l\big)$
  else
    $\mathcal{P}^{(t)} \leftarrow 0$
  end if
  $\theta \leftarrow \theta - \eta \nabla_\theta \big(\mathcal{L}_{\text{DPO}} + \mathcal{P}^{(t)}\big)$
  // update the sliding-window bound state under stop-gradient
  append $\text{sg}[\bar{\delta}_w^{(t)}]$ to Q_w, append $\text{sg}[\bar{\delta}_l^{(t)}]$ to Q_l
  if $|Q_w| > W$ **then**
    pop the oldest entry from Q_w and Q_l
  end if
  if not READY **and** $|Q_w| \geq W$ **then**
    $F_w \leftarrow \text{median}(Q_w) + \delta_w^{\text{bias}}$; $C_l \leftarrow \text{median}(Q_l) - \delta_l^{\text{bias}}$; READY $\leftarrow$ TRUE; $c \leftarrow 0$
  else if READY **then**
    $c \leftarrow c + 1$
    if $c \geq S$ **then**
      $F_w \leftarrow \text{median}(Q_w) + \delta_w^{\text{bias}}$; $C_l \leftarrow \text{median}(Q_l) - \delta_l^{\text{bias}}$; $c \leftarrow 0$
    end if
  end if
end for

C HH-RLHF



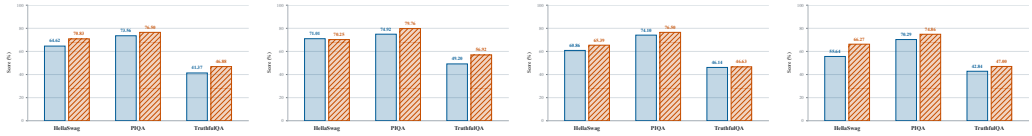
(a) Llama3.1-8B Instruct



(b) Qwen3-8B Instruct

Figure 8: Reward dynamics under DPO (left) and ARB (right) on HH-RLHF.

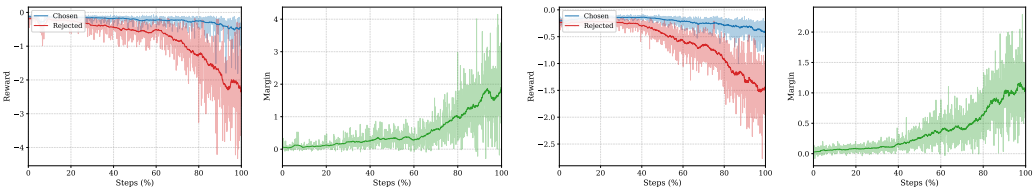
D ADDITIONAL METRICS



(a) Llama3.1-8B Base (b) Llama3.1-8B Instruct (c) Qwen3-8B Base (d) Qwen3-8B Instruct

Figure 9: Downstream capability across HellaSwag (raw), PIQA (raw) and TruthfulQA (MC1) on DPO (blue) and ARB (orange).

E REFERENCE-FREE POLICY



(a) Llama3.1-8B Instruct (b) Qwen3-8B Instruct

Figure 10: ARB Reward dynamics under reference-free condition.

F SETTING

Table 2: Hyperparameters for the four models.

Model	W	S	δ_w^{bias}	λ_w	δ_l^{bias}	λ_l
Llama3.1-8B Base	100	10	0.1	0.5	0.3	0.2
Llama3.1-8B Instruct	100	10	0.3	0.4	0.1	0.2
Qwen3-8B Base	100	10	0.3	0.3	0.1	0.2
Qwen3-8B Instruct	100	10	0.1	0.5	0.1	0.2

Tab. 2 lists the hyperparameters specific to the objective. ARB introduces the two directional biases of Eqs. 4–5 and the two penalty weights of Eq. 8, and the two parameters that govern the sliding window, namely its length W and the period S at which the bounds are recomputed. We report the weights λ_w and λ_l. All runs are implemented in OpenRLHF (Hu et al., 2025). We train Llama3.1-8B series and Qwen3-8B series on UltraFeedback (Cui et al., 2023) and HH-RLHF (Bai et al., 2022).