GRADIENT-GATED DPO: STABILIZING PREFERENCE OPTIMIZATION IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Direct Preference Optimization (DPO) improves relative preference by increasing the margin between chosen and rejected responses, but this objective does not specify how probability mass should be redistributed to achieve that margin. Consequently, preference separation can improve even as probability mass collapses away from both preferred and unrelated responses, a phenomenon associated with the squeezing effect. We show that this pathology is localized: continued negative updates become destructive when rejected responses enter low-probability valleys, where further suppression induces harmful redistribution through the coupled softmax geometry. We introduce *Gradient-Gated Preference Optimization* (Gate-DPO), which uses the current probability geometry of each rejected response to selectively attenuate its contribution, limiting excessive rejected suppression while delaying preference-loss saturation. Gate-DPO is objective-agnostic and composes with existing preference methods. Across four architectures, two preference datasets, and multiple objectives, gating consistently reduces squeezing and improves chosen-response likelihood, while exhibiting robust reductions in rejected-response collapse across architectures and objectives. Mass-dynamics analysis further shows that large preference margins can conceal substantial suppression of unrelated responses, while gating achieves healthier redistribution. Together, our results show that preference margin alone is insufficient to characterize optimization health and that selectively controlling probability dynamics provides a simple, general mechanism for stabilizing preference optimization. Code for experiments is available at GitHub[1]

1 INTRODUCTION

Preference optimization has become a standard approach for aligning language models with human preferences. Direct Preference Optimization (DPO) (Rafailov et al., 2023) simplifies reinforcement learning from human feedback by directly increasing the relative preference margin between chosen and rejected responses. However, optimizing this margin does not determine *how* probability mass is redistributed across the model's output distribution. A larger preference margin can therefore coexist with decreasing likelihood of preferred responses and substantial suppression of unrelated outputs, as observed in recent analyses of squeezing and likelihood displacement (Ren & Sutherland, 2025; Razin et al., 2025). This distinction between the *preference objective* and its induced *probability dynamics* is central to our work. DPO optimizes

$$\log \pi_\theta(y^+ \mid x) - \log \pi_\theta(y^- \mid x),$$

through opposing gradient contributions from the chosen and rejected responses. Negative rejected-response updates are necessary for preference learning, but their effect depends on the current probability geometry. When rejected responses have already entered low-probability "valleys," continued suppression provides

[1]Hidden to preserve anonymity.

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.

diminishing preference-learning benefit while softmax coupling can redistribute probability mass toward high-probability outputs and away from other plausible responses. This produces the *squeezing effect*: the preference margin may continue to improve even as the broader distribution deteriorates. Existing approaches mitigate aspects of this behavior through data augmentation (Ren & Sutherland, 2025), alternative objectives such as IPO (Azar et al., 2023) and Cal-DPO (Xiao et al., 2024), or global control of the rejected term, as in DPO-Shift (Yang et al., 2025). These methods can improve chosen-response likelihood or alter the margin–likelihood trade-off, but they do not explicitly identify which rejected responses have entered the low-probability regime. Moreover, the preference margin alone cannot reveal whether a smaller margin reflects insufficient rejection or healthier preservation of probability elsewhere in the distribution. This motivates an intervention that is local to the probability state of each preference pair.

We introduce Gradient-Gated Preference Optimization (Gate-DPO), a probability-aware mechanism that selectively attenuates rejected-response updates according to their current probability geometry. The gate remains near standard preference optimization for informative rejected responses and progressively attenuates further suppression once they enter low-probability valleys. Because gating acts inside the preference objective, it has a second effect beyond direct attenuation: by limiting growth of the effective preference margin, it can delay loss saturation and sustain the chosen-response learning signal. The gate is detached from the computation graph and is independent of the underlying preference objective, allowing the same mechanism to compose naturally with DPO, IPO, Cal-DPO, DPO-Shift, and data-augmentation approaches. Our experiments support this mechanism across four architectures, two preference datasets, and multiple preference objectives. Gating consistently reduces squeezing and improves chosen-response likelihood, with the effect persisting across architectures, datasets, and preference objectives. Probability-mass analysis further shows that preference margin alone can conceal substantial suppression of unrelated responses: methods can achieve strong separation through markedly different redistribution dynamics. Finally, pairwise response evaluation provides complementary behavioral evidence, with gated variants exceeding consistently demonstrating their superiority across 32 judge–architecture–objective–gate comparisons. Together, these results suggest that stable preference optimization requires considering not only *whether* the preference margin increases, but *how* that margin is achieved.

Our main contributions are: i) We distinguish preference-margin optimization from the probability dynamics used to achieve it, and show that continued suppression of already low-probability rejected responses can induce squeezing and destructive redistribution through softmax coupling; ii) we propose Gate-DPO, a probability-aware mechanism that selectively attenuates rejected-response updates in this localized regime, and show that gating acts through two coupled mechanisms, direct rejected-response attenuation and delayed preference-loss saturation, which sustains the chosen-response learning signal; iii) we demonstrate that gradient gating is modular across architectures and preference objectives, remaining effective when composed with IPO, Cal-DPO, DPO-Shift, and extended SFT; and iv) we show empirically that preference margin alone can obscure optimization health: comparable or larger margins can arise from substantially different probability redistribution over preferred and unrelated responses, while pairwise evaluations provide evidence that healthier dynamics can translate into improved response preference.

2 DPO AND PROBABILITY DYNAMICS

Given a preference pair (y^+, y^-) for prompt x, Direct Preference Optimization (DPO) (Rafailov et al., 2023) minimizes

$$\mathcal{L}_{\mathrm{DPO}} = -\log \sigma\left(\beta\left[\log \frac{\pi_\theta(y^+ \mid x)}{\pi_{\mathrm{ref}}(y^+ \mid x)} - \log \frac{\pi_\theta(y^- \mid x)}{\pi_{\mathrm{ref}}(y^- \mid x)}\right]\right), \tag{1}$$

where π_θ is the policy, π_{ref} is the reference policy, and β controls the strength of the preference update. Thus, DPO optimizes the *relative margin* between chosen and rejected responses; it does not directly constrain their absolute likelihoods or how probability mass is redistributed over other outputs.

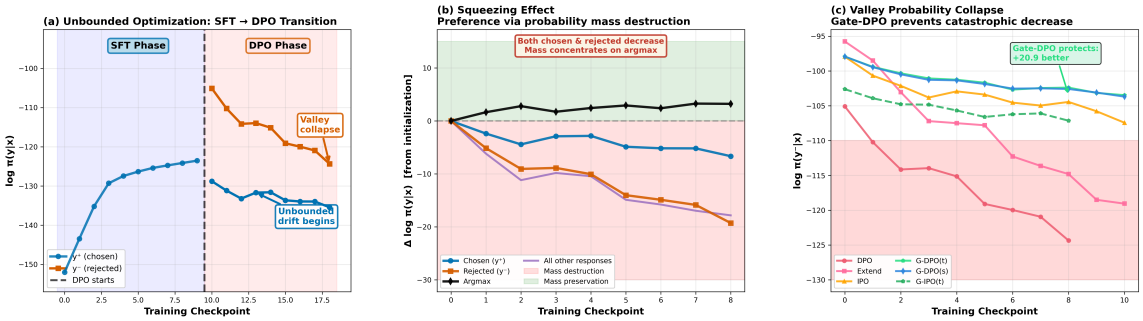

Figure 1: Three interrelated pathologies under DPO on Pythia-410M with Anthropic-HH. **(a)** Absolute drift: after the SFT→DPO transition, chosen and rejected log-probabilities can both decrease despite improving preference separation. **(b)** Squeezing: probability mass concentrates on high-confidence outputs while chosen and rejected responses decline. **(c)** Valley collapse: already low-probability regions undergo disproportionately severe suppression.

Let the quantity inside $\beta[\cdot]$ in Eq. 1 be Δ_θ. Its gradient has the form

$$\nabla_\theta \mathcal{L}_{\text{DPO}} = -c_\theta \left[\nabla_\theta \log \pi_\theta(y^+ \mid x) - \nabla_\theta \log \pi_\theta(y^- \mid x)\right], \qquad c_\theta = \beta\sigma(-\beta\Delta_\theta) > 0. \tag{2}$$

Hence, margin optimization induces opposing chosen- and rejected-response updates. These updates operate through a softmax distribution, where changing one output necessarily affects probability assigned elsewhere. Consequently, two policies can improve the same preference objective through substantially different probability-redistribution dynamics.

2.1 The Squeezing Pathology

Negative updates to rejected responses are fundamental to preference learning and are not inherently pathological. The problem arises when rejected responses enter extremely low-probability regions: continued suppression provides diminishing preference-learning benefit while softmax coupling can redistribute probability mass destructively. We refer to this behavior as the *squeezing effect*, following Ren & Sutherland (2025). Figure 1 illustrates three interrelated manifestations during DPO training.

i) Unbounded absolute drift. Because DPO constrains relative preference rather than absolute likelihood, the preference margin can improve while both chosen and rejected log-probabilities decrease (Fig. 1a).

ii) Squeezing via mass concentration. Softmax coupling redistributes probability removed from one region elsewhere in the distribution. Because the preference objective does not control where this mass goes, it can concentrate on already high-probability outputs while chosen, rejected, and unrelated alternatives are jointly suppressed (Fig. 1b).

iii) Valley collapse. The rejected-response contribution does not explicitly account for whether the rejected response already occupies an extremely low-probability region. Continued negative updates can therefore disproportionately suppress such "valley" responses and nearby regions of the distribution (Fig. 1c).

On Pythia-410M with Anthropic-HH, standard DPO decreases chosen and rejected log-likelihoods by 11.90 and 26.35 log units, respectively, while increasing the argmax by 5.85 and suppressing the remainder of the distribution by 24.76. Appendix B shows that the underlying failure mode is not specific to DPO or language models, reproducing analogous redistribution under negative updates to low-probability classes in multiclass logistic regression. These dynamics also show why preference margin alone can be misleading.

A large margin may arise from strong rejected-response suppression while chosen and unrelated responses simultaneously lose likelihood. Conversely, a method that better preserves chosen and unrelated responses may obtain a smaller margin without assigning substantially more likelihood to rejected responses. Thus, margin improvement characterizes *preference separation*, but not the health of the underlying probability redistribution. This observation suggests that rejected gradients should not be weakened uniformly. Many rejected responses remain informative; attenuation is needed specifically when a rejected response enters the low-probability regime. This motivates a localized, model-state-dependent principle: *localized probability pathology requires localized gradient attenuation*. We develop this mechanism next.

3 Gradient-Gated Preference Optimization

The analysis above suggests that rejected-response suppression should be attenuated based on the response's current probability state rather than uniformly across training examples. Gate-DPO implements this principle using a detached, probability-aware gate.

3.1 Detecting Low-Probability Valleys

For a rejected response $y^- = (y_1^-, \ldots, y_T^-)$, let $p_t = \pi_\theta(y_t^- \mid x, y_{<t}^-)$ denote its teacher-forced token probabilities. We consider two statistics that summarize whether the rejected response has entered a low-probability region. The sequence-level statistic is the geometric mean

$$s_{\text{seq}}(x, y^-) = \exp\left(\frac{1}{T}\sum_{t=1}^{T} \log p_t\right), \tag{3}$$

while the token-level statistic uses a lower-tail quantile,

$$s_{\text{tok}}(x, y^-) = Q_q\left(\{p_t\}_{t=1}^{T}\right), \tag{4}$$

where Q_q is the q-quantile. The former summarizes the response globally, whereas the latter detects localized low-probability tokens that may be hidden by sequence averaging. Both statistics produce one scalar gate per rejected response; q10 is therefore sequence-level attenuation conditioned on a token-level lower-tail statistic, not per-token gradient gating.

3.2 Probability-Aware Gradient Gating

Given either valley statistic $s(x, y^-)$, we define

$$g(x, y^-) = \sigma\big(\alpha\,[s(x, y^-) - \tau]\big), \qquad 0 < g(x, y^-) < 1, \tag{5}$$

where τ determines when attenuation begins and α controls the transition sharpness. The gate approaches one outside low-probability regions and decreases as the rejected response moves deeper into a probability valley. Because $s(x, y^-) \geq 0$, g has an exact floor $g_{\min} = \sigma(-\alpha\tau)$, attained as $s \to 0$: the intervention is a bounded soft damper, not a mechanism that can drive the rejected contribution to zero. This floor is negligible when $\alpha\tau \gg 1$ (our sequence-level default, $\tau = 0.10, \alpha = 50$, gives $g_{\min} \approx 0.7\%$) but not when $\alpha\tau = O(1)$ (our token-level default, $\tau = 0.005, \alpha = 50$, gives $g_{\min} \approx 43.8\%$); Appendix C makes this precise and Appendix E.2 reports the corresponding logged gate statistics.

Crucially, g is detached from the computation graph. It therefore acts as a model-state-dependent coefficient rather than an additional optimization target. The gate is bounded and monotone, smoothly interpolating between strong attenuation in low-probability valleys and the ungated rejected update elsewhere; Appendix C analyzes its properties and resulting gradient dynamics. For DPO, we replace the rejected log-ratio with its gated counterpart:

$$\mathcal{L}_{\text{Gate-DPO}} = -\log\sigma\big(\beta\,[r_\theta^+ - g(x, y^-)\, r_\theta^-]\big), \tag{6}$$

where

$$r_\theta^\pm = \log \frac{\pi_\theta(y^\pm \mid x)}{\pi_{\text{ref}}(y^\pm \mid x)}.$$

Gating also changes the common gradient coefficient

$$c_\theta^{\text{gate}} = \beta \left[1 - \sigma\big(\beta[r_\theta^+ - g r_\theta^-]\big)\right]. \tag{7}$$

For $r_\theta^- < 0$ and $g < 1$, the effective margin is smaller than in DPO at the same policy state, so $c_\theta^{\text{gate}} > c_\theta^{\text{DPO}}$. Gate-DPO therefore acts through two coupled mechanisms: direct attenuation of the rejected term and delayed loss saturation, which sustains the chosen-response learning signal. Appendix C.4 provides the derivation and empirical analysis.

3.3 Locality and Modularity

Gate-DPO differs from globally applied interventions in that attenuation is determined independently for each preference pair from the model's current probability geometry. Rejected responses in the same minibatch can therefore receive different attenuation, and their gates evolve as the policy changes. This directly matches the localized nature of the pathology identified in Section 2.1. The mechanism is also independent of the particular preference objective: whenever an objective contains a rejected-response contribution, the same detached gate can modulate that contribution without changing how the valley state is detected. We therefore apply gating not only to DPO, but also to IPO, Cal-DPO, DPO-Shift, and extended-SFT variants. Section 4 evaluates this modularity empirically. Appendix D.2 further shows analytically that the same construction recovers RPO and Bounded-DPO extensions.

4 Experiments

We evaluate whether Gate-DPO mitigates squeezing, improves probability redistribution, generalizes across architectures and datasets, and composes with other preference objectives. Pairwise response evaluation further tests whether healthier optimization dynamics yield more preferred outputs.

4.1 Experimental Setup

Models and datasets: we evaluate four architectures, Pythia-410M (Biderman et al., 2023), Qwen-0.5B (Bai et al., 2023), LLaMA-7B (Touvron et al., 2023), and Qwen1.5-7B (Bai et al., 2023) on Anthropic-HH (Bai et al., 2022) and UltraFeedback (Cui et al., 2023). This provides two model families at the 7B scale in addition to the smaller architectures. **Training:** models are initialized from SFT checkpoints and preference-optimized for 5–6 epochs with batch size 4 and learning rate 5×10^{-7}. For Gate-DPO, we use $\tau = 0.10$ for sequence-level gating and $\tau = 0.005$ for token-level gating, with $q = 0.10$ and $\alpha = 50$. Full implementation details, computational cost, and hyperparameter sensitivity analyses are provided in Appendix E. **Evaluation:** we measure changes in chosen and rejected log-likelihood, preference separation, and gate activation on held-out data. To directly characterize probability redistribution, we additionally measure changes over preferred, rejected, and unrelated response variants relative to the SFT reference. The full mass-dynamics protocol, metric definitions, and representative training and evaluation examples are provided in Appendix F. Finally, we use pairwise response evaluation as complementary task-level evidence.

4.2 Comparison with Existing Mitigations

Figure 2 compares gradient gating with four representative approaches on Pythia-410M with Anthropic-HH: extended SFT (Ren & Sutherland, 2025), Cal-DPO (Xiao et al., 2024), IPO (Azar et al., 2023), and DPO-Shift (Yang et al., 2025).

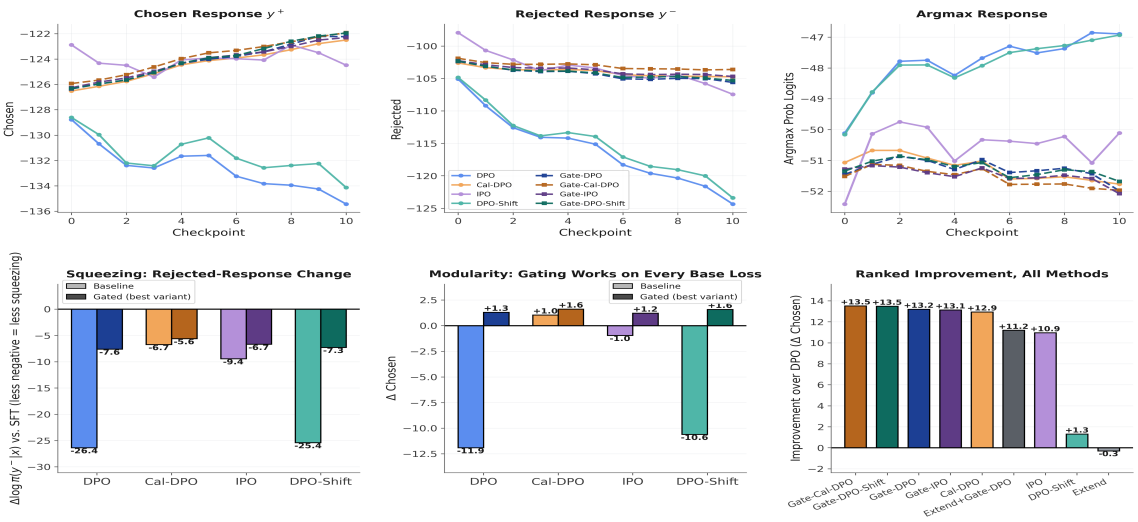

Figure 2: Comparison with existing mitigation strategies on Pythia-410M using 5k Anthropic-HH preference pairs. Top: learning trajectories for chosen, rejected, and argmax responses. Bottom: rejected-response squeezing, chosen-response change, and improvement relative to standard DPO. Gating consistently reduces squeezing and improves chosen-response dynamics across the evaluated preference objectives.

Ungated baselines: standard DPO exhibits severe squeezing, with $\Delta_{\text{chosen}} = -11.9$. Extended SFT does not resolve the pathology (-12.2), indicating that improving the initialization alone is insufficient to prevent destructive preference updates. DPO-Shift similarly remains strongly negative (-10.6), showing that a global schedule on the rejected term does not resolve squeezing in this setting. IPO (-1.0) substantially reduces the collapse but still leaves chosen-response likelihood decreasing. Cal-DPO's calibration term alone is enough to reverse it ($+1.0$), since calibrating the implicit reward toward a fixed target opposes unbounded margin growth even without gating. **Effect of gating:** Gate-DPO reverses this behavior, yielding $\Delta_{\text{chosen}} = +1.2$ with sequence-level gating and $+1.3$ with token-level gating. The same mechanism also improves other objectives: Gate-Cal-DPO further increases Cal-DPO's already-positive recovery to $+1.5$ to $+1.6$, while Gate-DPO-Shift reaches $+1.2$ to $+1.6$ compared with -10.6 for its ungated baseline. Gate-IPO likewise reverses the residual drift, reaching $+1.0$ to $+1.2$ with token-level gating, compared with -1.0 for ungated IPO. Thus, the benefit is not specific to the DPO objective. The training trajectories provide a consistent picture: ungated methods exhibit decreasing chosen likelihood together with increasing argmax probability, whereas gated variants substantially suppress this concentration while stabilizing rejected-response updates. Token-level gating consistently provides a modest additional improvement over sequence-level gating across all four objectives, suggesting that lower-tail token probabilities capture localized valleys somewhat more precisely than a sequence-level average. Appendix C.4 independently confirms delayed saturation through the coefficient $\bar{c}_\theta$, with gating slowing its decay in seven of the eight tested architecture–dataset configurations.

4.3 Probability-Mass Dynamics

Preference margin measures whether chosen responses become more preferred than rejected responses, but not how the broader output distribution changes to achieve that separation. We therefore evaluate probability-mass dynamics over preferred, rejected, and unrelated response variants. We report Δ Margin, Δ Chosen, and Δ Others relative to the SFT reference; the complete protocol and metric definitions are provided in Appendix F.

Table 1 reveals markedly different redistribution dynamics behind similar preference improvements. Standard DPO obtains the largest margin (14.45), but simultaneously decreases preferred responses (ΔChosen $= -11.90$) and strongly suppresses unrelated responses (ΔOthers $= -27.55$). DPO-Shift exhibits a similar pattern: despite a margin of 14.81, chosen and unrelated responses decrease by 10.60 and 25.07 log units, respectively. Thus, a large preference margin can coexist with substantial probability collapse outside the chosen-rejected comparison. Gating changes this trade-off. Gate-DPO achieves smaller but still substantial margins (8.48–8.90) while reversing chosen-response drift ($+1.17$ to $+1.30$) and substantially reducing collateral suppression (-1.42 to -3.86). The same pattern appears when gating is composed with other methods. In particular, Gate-DPO-Shift converts the negative chosen drift of DPO-Shift to $+1.23$–$+1.58$ while reducing ΔOthers from -25.07 to between -1.65 and -3.37. Extend + Gate-DPO (q10) similarly restores positive chosen movement ($+1.15$) while limiting collateral suppression (-1.96) and does so robustly across all four architectures (Appendix D.1). The sequence-level Extend variant uses one untuned $\tau = 0.03$ across architectures and reduces rejected collapse on three of four models (Table 4); on Pythia-410M its gate barely activates and chosen movement remains negative (-5.42). This is consistent with the model–dataset sensitivity of τ examined in Section 4.4 and Appendix E.2; q10 is positive here under its shared default. These results also clarify why a smaller preference margin should not automatically be interpreted as weaker preference learning. The reduction does not necessarily correspond to undesirable recovery of the rejected response; it can accompany substantially better preservation of preferred and unrelated regions of the distribution. Preference margin therefore characterizes *separation*, whereas the accompanying mass dynamics characterize *how that separation is achieved*.

Table 1: **Probability-mass dynamics on Pythia-410M with Anthropic-HH.** Δ Margin measures the change in chosen–rejected separation; Δ Chosen measures movement on preferred responses; and Δ Others measures collateral movement on unrelated responses. Higher Δ Chosen and Δ Others closer to zero indicate healthier redistribution.

Method	**Δ Margin**	**Δ Chosen**	**Δ Others**
DPO	14.45	-11.90	-27.55
Extend + DPO	8.86	-12.20	-20.27
Cal-DPO	7.77	+1.03	-3.02
IPO	8.48	-0.95	-7.82
DPO-Shift	14.81	-10.60	-25.07
Gate-DPO (seq)	8.48	+1.17	-1.42
Gate-DPO (q10)	8.90	+1.30	-3.86
Extend + Gate-DPO (seq)	11.45	-5.42	-11.57
Extend + Gate-DPO (q10)	8.20	+1.15	-1.96
Gate-Cal-DPO (seq)	7.15	+1.51	**-1.04**
Gate-Cal-DPO (q10)	7.22	**+1.61**	-2.06
Gate-IPO (seq)	7.43	+0.96	-1.46
Gate-IPO (q10)	7.91	+1.22	-2.76
Gate-DPO-Shift (seq)	8.47	+1.23	-1.65
Gate-DPO-Shift (q10)	8.87	+1.58	-3.37

4.4 Generalization Across Architectures and Datasets

We next evaluate whether squeezing and the effect of gating generalize across architectures and preference datasets. We consider four models, Pythia-410M, Qwen-0.5B, LLaMA-7B, and Qwen1.5-7B on Anthropic-HH and UltraFeedback, yielding eight model–dataset configurations. Figure 3 summarizes the resulting chosen-response dynamics under DPO and Gate-DPO.

Squeezing is model–dataset dependent. Standard DPO ranges from severe chosen-response drift on

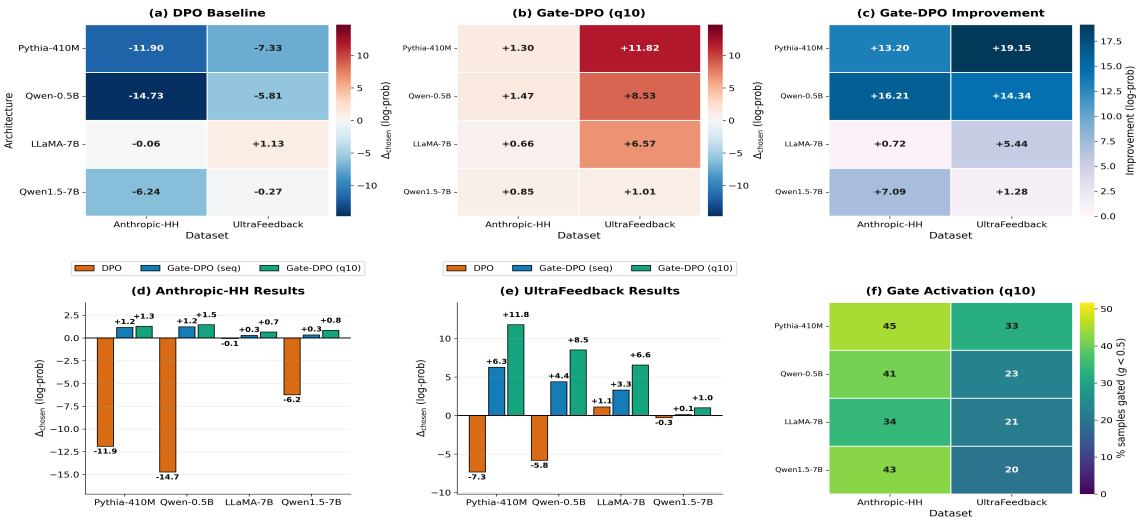

Figure 3: Generalization across architectures and datasets. Chosen-response likelihood change under DPO and Gate-DPO across four architectures on Anthropic-HH and UltraFeedback. Squeezing varies substantially across model–dataset combinations, whereas gating produces positive chosen-response movement in all eight settings using the fixed default thresholds.

Table 2: Modularity across objectives and architectures. Δ_{rejected}: base $\rightarrow$ best gated (minimum $|\Delta_{\text{rejected}}|$); large negative values indicate stronger collapse. Full results: Appendix D.1.

Method	**Pythia-410M**	**LLaMA-7B**	**Qwen-0.5B**	**Qwen1.5-7B**
DPO	$-26.35 \rightarrow -7.31$	$-0.88 \rightarrow +0.08$	$-30.47 \rightarrow -6.25$	$-10.45 \rightarrow -0.77$
DPO-Shift	$-25.40 \rightarrow -7.25$	$-0.80 \rightarrow +0.10$	$-27.23 \rightarrow -6.00$	$-9.20 \rightarrow -0.71$
IPO	$-9.43 \rightarrow -6.46$	$-0.64 \rightarrow +0.02$	$-8.36 \rightarrow -5.60$	$-2.16 \rightarrow -0.75$
Cal-DPO	$-6.74 \rightarrow -5.61$	$-0.43 \rightarrow -0.01$	$-6.03 \rightarrow -4.88$	$-1.07 \rightarrow -0.52$
Extend	$-23.26 \rightarrow -7.04$	$+0.15 \rightarrow +0.27$	$-22.60 \rightarrow -3.17$	$-9.35 \rightarrow +0.09$

Anthropic-HH (Pythia-410M: -11.90; Qwen1.5-7B: -6.24) to almost none on LLaMA-7B (-0.06); on UltraFeedback, LLaMA-7B instead moves positively ($+1.13$). Thus, scale alone does not determine squeezing. Gate-DPO produces positive chosen-response movement in all eight configurations, with the largest gains where baseline squeezing is strongest (e.g., Pythia-410M/UltraFeedback: $-7.33 \rightarrow +11.82$). The same default thresholds are used throughout despite gate activation varying by approximately 20–45%, supporting gating as a localized intervention. The threshold τ remains the primary model–dataset-sensitive hyperparameter; Appendix E.2 studies its operating range, while Appendix C.4 shows how it regulates loss saturation. Appendix E shows that α is comparatively robust.

4.5 Modularity Across Preference Objectives

We test whether gating composes with changes to the objective (IPO and Cal-DPO), rejected-response schedule (DPO-Shift), and initialization (extended SFT). Table 2 reports rejected-response dynamics across four architectures for each base method and its best gated variant. Across architectures and objectives, gated variants consistently reduce rejected-response collapse, with the effect persisting across objective

Table 3: Pairwise win-rates (%) of gated variants over their corresponding baselines on UltraFeedback.

		Gate-DPO		Gate-Cal		Gate-IPO		Gate-Shift	
Judge	**Arch.**	seq	q10	seq	q10	seq	q10	seq	q10
GPT-4	Pythia-410M	56.7	53.3	55.0	52.0	56.0	53.7	62.1	54.0
	LLaMA-7B	59.0	56.0	58.0	55.3	58.2	63.0	59.2	64.0
GPT-5.4	Pythia-410M	60.0	61.3	56.0	52.0	60.0	56.2	59.3	60.8
	LLaMA-7B	61.1	57.0	58.0	62.7	58.6	55.3	68.7	60.0

modifications, initialization, and DPO-Shift's global schedule. The LLaMA-7B Extend setting is the main exception, where the ungated baseline already exhibits little rejected-response collapse. The effect persists across objectives, initialization, and DPO-Shift's global schedule, supporting gating as a composable, sample-specific intervention. Full results are in Appendix D.1.

Pairwise Response Evaluation: Table 3 provides complementary behavioral evidence on 30 held-out UltraFeedback prompts using GPT-4 and GPT-5.4 judges. Across 32 judge–architecture–objective–gate comparisons, gated variants exceed 50%, with win rates ranging from 52.0% to 68.7%. This also suggests that healthier optimization dynamics can translate into better task-level performance, consistent with prior work (Ren & Sutherland, 2025; Xiao et al., 2024). Appendix G.1 details generation, randomized response order, judging, and reproducibility.

5 Related Work

DPO (Rafailov et al., 2023) enables preference optimization without explicit reward modeling, motivating variants such as IPO (Azar et al., 2023) and Cal-DPO (Xiao et al., 2024). Recent work has identified pathological dynamics in margin-based alignment, including gradient entanglement (Yuan et al., 2024a), likelihood displacement (Razin et al., 2025), and the squeezing effect (Ren & Sutherland, 2025). Existing mitigations modify the objective, initialization, or rejected-response influence, including extended SFT (Ren & Sutherland, 2025), DPO-Positive (Pal et al., 2024), Bounded-DPO (Cho et al., 2025), and DPO-Shift (Yang et al., 2025). To our knowledge, Gate-DPO is the first probability-aware, gradient-level intervention for stabilizing DPO: it uses a sample-specific, model-state-dependent gate to attenuate rejected-response updates only when their probability geometry indicates a low-probability valley, rather than applying a uniform intervention across preference pairs. By regulating the effective preference margin, gating also delays preference-loss saturation and sustain the chosen-response learning signal. Appendix A provides a broader discussion.

6 Conclusion, Limitations, and Future Work

Preference-margin improvement alone does not characterize healthy optimization: large margins can coexist with suppression of preferred and unrelated responses. Gate-DPO applies localized, probability-aware attenuation in low-probability valleys, reducing rejected-response suppression while delaying loss saturation and sustaining chosen-response learning. Across architectures, datasets, and objectives, this improves probability redistribution and reduces squeezing.

Limitations and Future Work. Our analysis emphasizes optimization dynamics, with pairwise evaluation providing initial task-level evidence. Human judgments, standardized benchmarks (AlpacaEval 2.0 and MT-Bench), and evaluation of the safety–helpfulness trade-off remain important directions for future work. Although fixed thresholds generalize well here, the valley regime varies across models and datasets, suggesting that model–dataset-specific tuning of τ may yield further gains.

GENERATIVE AI DISCLOSURE

Generative AI tools were used to assist with writing and editing, experimental design, analysis, code development and debugging, and literature review. AI-assisted suggestions, analyses, code, and scientific claims were reviewed and verified by the authors against experimental results, source literature, and the implemented methodology. Generative AI was not used as a substitute for experimental evaluation or as an autonomous source of scientific evidence. The authors take full responsibility for the methodology, design, proofs, experiments, results, claims, and final content of this work.

REFERENCES

Anthropic. Claude 2. `https://www.anthropic.com/index/claude-2`, 2023. Accessed: 2026.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, et al. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036*, 2023.

Jinze Bai, Shuai Bai, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Yuntao Bai et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023.

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

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024.

Jae Hyeon Cho, JunHyeok Oh, Myunsoo Kim, and Byung-Jun Lee. Rethinking dpo: The role of rejected responses in preference misalignment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025. arXiv:2506.12725.

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, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*, 2023.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Duanyu Feng, Bowen Qin, Chen Huang, Zheng Zhang, and Wenqiang Lei. Towards analyzing and understanding the limitations of dpo: A theoretical perspective. *arXiv preprint arXiv:2404.04626*, 2024.

Zhihan Liu, Miao Lu, Shenao Zhang, Boyi Liu, Hongyi Guo, Yingxiang Yang, Jose Blanchet, and Zhaoran Wang. Provably mitigating overoptimization in rlhf: Your sft loss is implicitly an adversarial regularizer. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 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.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. *arXiv preprint arXiv:2403.19159*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, and Christopher D Manning. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. Unintentional unalignment: Likelihood displacement in direct preference optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.08847.

Yi Ren and Danica J. Sutherland. Learning dynamics of llm finetuning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. arXiv:2407.10490v4.

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

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Binghai Wang, Rui Dou, Shihan Dou, Songyang Gao, Wei Hua, Wei Shen, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, et al. Secrets of rlhf in large language models part ii: Reward modeling. *arXiv preprint arXiv:2401.06080*, 2024.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. Self-play preference optimization for language model alignment. *arXiv preprint arXiv:2405.00675*, 2024.

Teng Xiao, Yige Yuan, Huaisheng Zhu, Mingxiao Li, and Vasant G. Honavar. Cal-dpo: Calibrated direct preference optimization for language model alignment. *arXiv preprint arXiv:2412.14516*, 2024.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*, 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. *arXiv preprint arXiv:2410.13828*, 2024a.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024b.

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

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

A Extended Related Work

Preference-based alignment. Large language models are commonly aligned with human preferences using reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Ziegler et al., 2019; Stiennon et al., 2020; Xu et al., 2024). The standard pipeline combines supervised fine-tuning, reward modeling from pairwise preferences (Ouyang et al., 2022; Bai et al., 2022), and policy optimization, typically using PPO under a KL constraint (Schulman et al., 2017; Zhao et al., 2023). While effective at scale (Ouyang et al., 2022; Anthropic, 2023), this pipeline requires multiple model components and careful RL tuning, motivating direct preference-learning objectives.

Direct preference optimization and variants. DPO (Rafailov et al., 2023) replaces explicit reward modeling and RL with a closed-form objective derived from the Bradley–Terry preference model (Bradley & Terry, 1952). Subsequent variants include IPO (Azar et al., 2023), which introduces a bounded squared objective, KTO (Ethayarajh et al., 2024), which incorporates prospect-theoretic preference modeling, and Cal-DPO (Xiao et al., 2024), which calibrates implicit rewards. Yuan et al. (Yuan et al., 2024a) characterize *gradient entanglement*, showing how coupled preferred and dispreferred updates can produce undesirable likelihood dynamics even as the preference margin improves.

Squeezing and related mitigation strategies. (Ren & Sutherland, 2025) show that negative updates to rejected responses can concentrate probability on high-confidence outputs while suppressing other responses, termed the *squeezing effect*. They reproduce the behavior in multiclass logistic regression, linking the pathology to softmax probability dynamics rather than a particular neural architecture. Extended SFT (Ren & Sutherland, 2025) mitigates the effect by altering the initialization before preference optimization, while related work studies data quality, reference models, and regularization (Yuan et al., 2024b; Chen et al., 2024; Wu et al., 2024; Liu et al., 2024; Wang et al., 2024; Park et al., 2024). Gate-DPO instead uses the current probability state of each rejected response to determine when further suppression should be attenuated, making it complementary to data-, initialization-, and objective-level interventions.

Likelihood displacement and rejected-response control. Likelihood displacement describes the reduction of preferred-response probability during DPO despite improving preference separation (Razin et al., 2025). Feng et al. (Feng et al., 2024) theoretically analyze a related asymmetry in DPO dynamics, while Pal et al. (Pal et al., 2024) propose DPO-Positive to prevent preferred-response likelihood from decreasing. Cho et al. (Cho et al., 2025) propose Bounded-DPO, which bounds the influence of rejected responses while preserving the basic DPO structure. DPO-Shift (Yang et al., 2025) instead introduces a globally chosen coefficient $f(\lambda)$, fixed or scheduled during training, to control the chosen–rejected trade-off.

Gate-DPO differs in the *locality* of its intervention. Its detached gate is computed independently for each preference pair from the rejected response's current probability geometry. Consequently, examples within the same minibatch can receive different attenuation, and that attenuation changes as the policy evolves. DPO-Shift's global coefficient, by contrast, cannot distinguish preference pairs occupying different probability regimes at the same training step. This distinction directly motivates our cross-objective experiments, where gating remains effective when composed with DPO-Shift, IPO, Cal-DPO, and extended SFT.

Relation to generic stabilization mechanisms. Gate-DPO also differs from standard optimization controls such as regularization, β tuning, and gradient clipping. Regularization constrains the policy globally, while changing β rescales the strength of preference updates across examples; neither identifies which rejected responses have entered a pathological low-probability regime. Gradient clipping instead limits updates according to gradient magnitude, whereas squeezing is determined by probability geometry and need not coincide with an unusually large gradient norm. Gate-DPO therefore targets a different signal: it uses the current rejected-response probability state to apply sample-specific attenuation only where continued suppression is potentially destructive, while leaving informative rejected updates largely unchanged. An interesting direction for future work is to study how such localized gating interacts with KL regularization or

explicit KL constraints (Ziegler et al., 2019), and whether the two provide complementary control over local and global policy deviation.

B SOFTMAX PATHOLOGY IN A MINIMAL SETTING

(Ren & Sutherland, 2025) showed that the squeezing effect is not specific to language models but arises from the interaction between softmax normalization and negative gradients. In their analysis, even a simple multiclass logistic regression model exhibits the same phenomenon: when a negative gradient is applied to a low-probability class, both the chosen and rejected responses decrease while probability mass shifts toward the argmax class. This observation establishes squeezing as a fundamental property of softmax optimization rather than an artifact of large neural architectures.

We reproduce this minimal experiment and extend it with our gradient gating mechanism. The model is a K-class logistic regression with features $\phi(x) \in \mathbb{R}^d$ and linear weights $W \in \mathbb{R}^{d\times K}$ producing

$$p = \mathrm{softmax}(W^\top \phi(x)). \tag{8}$$

Standard gradient descent on cross-entropy gives

$$W_{t+1} = W_t - \eta\, \phi(x)(p - e_y)^\top, \tag{9}$$

where negative learning rates ($\eta < 0$) simulate the rejected-response updates used in preference optimization. This setup isolates the effect of softmax normalization without neural architectures, sequence modeling, or language-specific structure.

To prevent pathological updates, we introduce a gated update rule

$$W_{t+1} = W_t - \eta\, g(p(y))\, \phi(x)(p - e_y)^\top, \tag{10}$$

where

$$g(p) = \sigma(\alpha(\log p - \log \tau)). \tag{11}$$

The gate reduces gradient magnitude when the target probability lies in the extreme tail of the distribution. In our experiments we use $\alpha = 50$ and $\tau \approx 1/K$ ($\tau = 0.03$ for $K = 50$, $\tau = 10^{-8}$ for $K = 1000$).

We reproduce the five canonical cases from (Ren & Sutherland, 2025): positive gradient updates (A), negative gradients on flat distributions (B), negative gradients on peaks (C), negative gradients on valley responses (D), and a large-vocabulary setting (E). Figure 4 shows the distributions before the update (gray), after baseline optimization (blue), and after gated updates (green).

The baseline results reproduce the original observation: when a negative gradient targets a valley response, softmax normalization causes extreme squeezing and the update effectively freezes ($\Delta_{\max} \approx 3 \times 10^{-6}$). With gating, the learning rate is automatically reduced for such cases ($g \approx 0.38$), allowing the update to remain effective while preventing collapse. The gated update avoids the near-frozen, highly concentrated baseline behavior and better preserves entropy in the distribution. Across all scenarios, the gate activates only when probabilities fall below the threshold, leaving normal optimization unaffected. The same mechanism also scales to larger vocabularies ($K = 1000$), confirming that the approach remains effective in settings closer to neural language models.

These controlled experiments confirm two key points. First, the squeezing phenomenon identified by (Ren

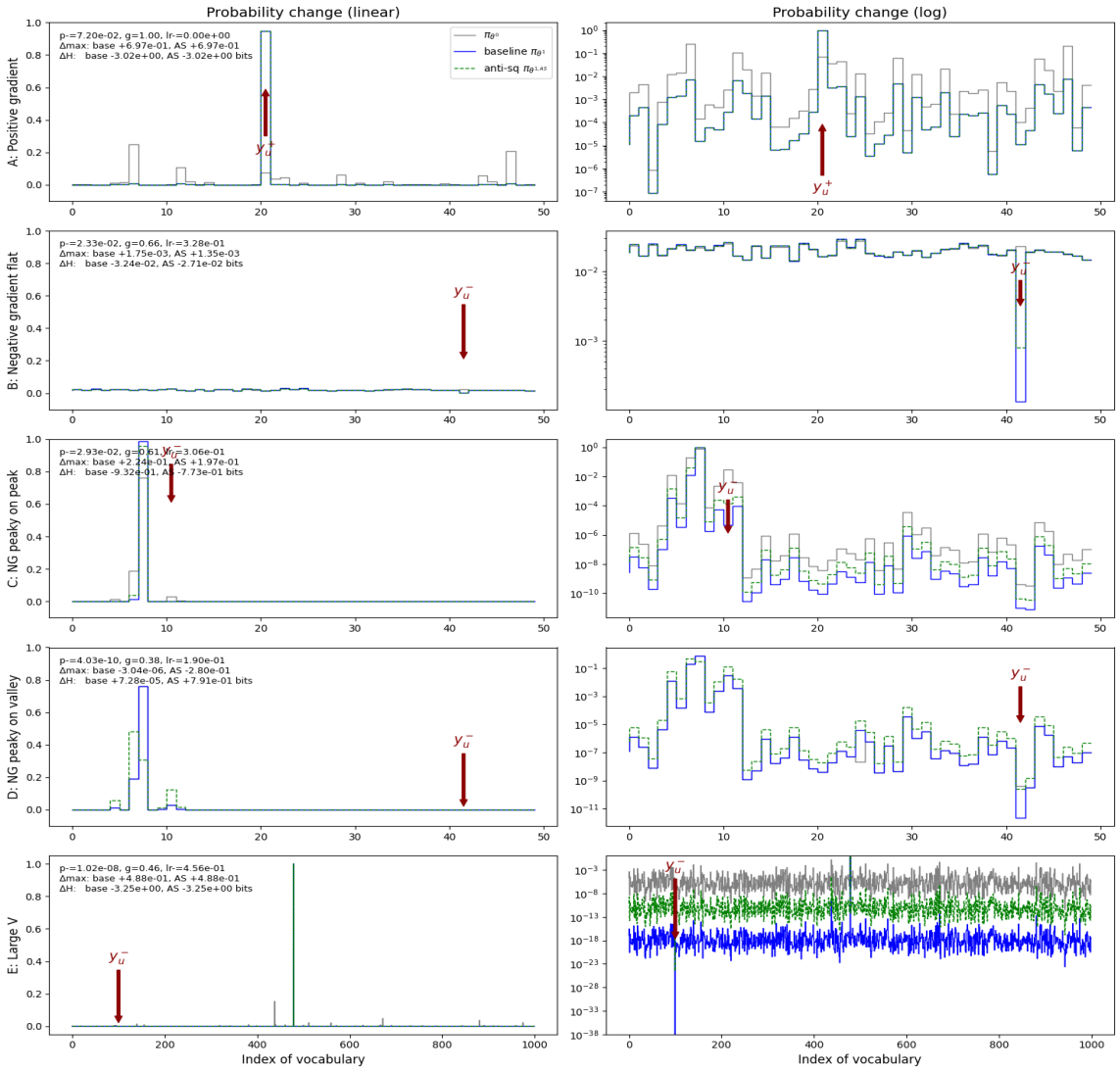

Figure 4: Toy logistic regression experiment illustrating the softmax pathology ($\tau_{\text{smallK}} = 0.03$, $\tau_{\text{largeK}} = 10^{-8}$, $\alpha = 50$). Gray: initial distribution; blue: baseline update; green: gated update. Negative gradients applied to low-probability classes cause extreme squeezing under standard softmax optimization. Gradient gating detects such valley responses and reduces the effective learning rate, restoring stable updates while preserving entropy.

& Sutherland, 2025) arises directly from softmax normalization under negative gradients. Second, adaptive gradient gating mitigates this pathology without altering the model architecture or objective. This minimal demonstration motivates the neural network experiments presented in the main paper.

C GRADIENT ANALYSIS OF GATE-DPO

This appendix formalizes how the detached gate modifies the local gradient dynamics of Gate-DPO. Let

$$z_\theta^{\text{gate}}(x) = \beta \left[r_\theta^+ - g(x, y^-)\, r_\theta^- \right], \tag{12}$$

where

$$r_\theta^\pm = \log \frac{\pi_\theta(y^\pm \mid x)}{\pi_{\text{ref}}(y^\pm \mid x)},$$

and $g(x, y^-)$ is detached from the computation graph. Define

$$c_\theta(x) = \beta \left[1 - \sigma\big(z_\theta^{\text{gate}}(x)\big) \right] \geq 0.$$

C.1 GRADIENT STRUCTURE

For $\mathcal{L}_{\text{Gate-DPO}} = -\log \sigma(z_\theta^{\text{gate}})$, the partial derivatives with respect to the response log-probabilities are

$$\frac{\partial \mathcal{L}_{\text{Gate-DPO}}}{\partial \log \pi_\theta(y^+ \mid x)} = -c_\theta(x), \tag{13}$$

$$\frac{\partial \mathcal{L}_{\text{Gate-DPO}}}{\partial \log \pi_\theta(y^- \mid x)} = g(x, y^-)\, c_\theta(x). \tag{14}$$

Because the gate is detached, it directly multiplies only the rejected contribution, while the chosen derivative retains the standard DPO *form* $-c_\theta(x)$. Its *magnitude* is not the same as under standard DPO, however, because $c_\theta(x)$ is evaluated at the gated margin z_θ^{gate} rather than the ungated margin $z_\theta^{\text{DPO}} = \beta[r_\theta^+ - r_\theta^-]$. Write $c_\theta^{\text{DPO}}(x) = \beta[1 - \sigma(z_\theta^{\text{DPO}})]$ for the standard-DPO coefficient. When the rejected response is driven far into a low-probability region, $r_\theta^- \to -\infty$, so $z_\theta^{\text{DPO}} \to +\infty$, $\sigma(z_\theta^{\text{DPO}}) \to 1$, and $c_\theta^{\text{DPO}}(x) \to 0$: the chosen-response gradient vanishes together with the rejected one, halting further learning on *both* responses. Under gating, $z_\theta^{\text{gate}} = \beta[r_\theta^+ - g(x, y^-) r_\theta^-]$ grows only at the attenuated rate $g(x, y^-)\,|r_\theta^-|$ instead of $|r_\theta^-|$, so $\sigma(z_\theta^{\text{gate}})$ approaches 1 more slowly and $c_\theta(x) > c_\theta^{\text{DPO}}(x)$ throughout the valley regime. Gating therefore does not only reduce the rejected-response update in Eq. 14; by keeping the effective margin from saturating as quickly, it also *delays the vanishing* of the chosen-response gradient in Eq. 13, sustaining a larger positive update on $\log \pi_\theta(y^+ \mid x)$ for longer than ungated DPO would. Because g has a strictly positive floor $g_{\min} = \sigma(-\alpha\tau) > 0$ (Eq. 17), this is a delay rather than a removal of eventual saturation: as $r_\theta^- \to -\infty$ at fixed g, $g\, r_\theta^- \to -\infty$ too and $c_\theta(x) \to 0$ eventually. The practical effect is a rescaling of how far into the valley the rejected response must fall before the chosen-response signal starts to vanish, by roughly a factor of $1/g(x, y^-)$. This is a distinct, complementary mechanism to the direct rejected-term attenuation described above, and provides a more complete account of the positive Δ_{chosen} values reported in Section 4.3 than softmax redistribution alone. When $g(x, y^-) \to 1$, the gated objective recovers DPO. When the rejected response enters a low-probability valley, $g(x, y^-) < 1$ attenuates its direct negative contribution without introducing gradients through the valley statistic.

C.2 PROPERTIES OF THE PROBABILITY GATE

For $g = \sigma(\alpha(s - \tau))$ with $\alpha > 0$, the gate has several useful properties. First, $0 < g < 1$, so the rejected contribution is attenuated but never reversed. Second, g is monotonically increasing in the valley statistic s, with $g(\tau) = 1/2$; hence, τ defines the transition between stronger attenuation and recovery of the standard rejected update. Third, α controls the sharpness of this transition: larger α produces a more threshold-like intervention, with $g \to \mathbb{I}[s > \tau]$ as $\alpha \to \infty$ (up to the boundary $s = \tau$), whereas smaller α yields smoother attenuation.

The intervention is also localized around the low-probability regime. For $s \geq \tau + \delta$, $\delta > 0$,

$$1 - g = \sigma[-\alpha(s - \tau)] \leq e^{-\alpha\delta}, \tag{15}$$

so g rapidly approaches one and the rejected contribution approaches its ungated form. Conversely, for $s \leq \tau - \delta$,

$$g = \sigma[\alpha(s - \tau)] \leq e^{-\alpha\delta}. \tag{16}$$

This exponential bound is loose once δ is comparable to τ: since $s \geq 0$ (both valley statistics are probabilities), δ cannot exceed τ, so the *tightest* attainable bound is at $s = 0$, giving the exact floor

$$g_{\min} = \sigma(-\alpha\tau), \tag{17}$$

not an asymptotic approach to zero. When $\alpha\tau \gg 1$ the two bounds agree and $g_{\min}$ is negligible; this holds for the sequence-level default ($\tau = 0.10, \alpha = 50 \Rightarrow \alpha\tau = 5 \Rightarrow g_{\min} = \sigma(-5) \approx 0.7\%$), consistent with the near-zero minimum gate values logged for `seq_mean` runs (Appendix E.2). It does *not* hold for the token-level default ($\tau = 0.005, \alpha = 50 \Rightarrow \alpha\tau = 0.25 \Rightarrow g_{\min} = \sigma(-0.25) \approx 43.8\%$): under this hyperparameterization, token-level gating cannot attenuate the rejected contribution below $\approx 44\%$ of full strength no matter how far into a valley the rejected response falls. This matches the minimum gate value actually logged during training ($g_{\min} \approx 0.438$ across our token-level runs at $\tau = 0.005$). Thus, τ and α jointly determine the gate's transition and attainable floor: increasing τ lowers the floor and shifts attenuation toward higher-probability examples, whereas increasing α sharpens the transition around τ.

C.3 First-Order Probability Dynamics

Effect on the rejected-response update. Consider one gradient-descent step $\theta' = \theta - \eta\nabla_\theta \mathcal{L}_{\text{Gate-DPO}}$. A first-order expansion gives

$$\Delta \log \pi_\theta(y^- \mid x) \approx -\eta c_\theta(x) \Big[g(x, y^-) \|\nabla_\theta \log \pi_\theta(y^- \mid x)\|^2 - \nabla_\theta \log \pi_\theta(y^- \mid x)^\top \nabla_\theta \log \pi_\theta(y^+ \mid x) \Big]. \tag{18}$$

Thus, the gate directly attenuates the self-suppression term associated with the rejected response, while the actual parameter-space update also depends on the interaction between chosen- and rejected-response gradients. Inside the brackets of Eq. 18, the self-suppression term is already scaled by $g(x, y^-)$ while the cross-gradient term is not; dropping the cross-gradient term therefore requires it to be negligible relative to the *gated* self-term $g(x, y^-)\|\nabla_\theta \log \pi_\theta(y^- \mid x)\|^2$, not the unattenuated norm $\|\nabla_\theta \log \pi_\theta(y^- \mid x)\|^2$. Under that condition,

$$\Delta \log \pi_\theta(y^- \mid x) \approx -\eta c_\theta(x) g(x, y^-) \|\nabla_\theta \log \pi_\theta(y^- \mid x)\|^2. \tag{19}$$

In this regime, rejected-response suppression is approximately attenuated in proportion to $g(x, y^-)$. This formalizes the intended behavior of Gate-DPO: the intervention weakens further suppression as a rejected response moves deeper into a probability valley while retaining the standard preference signal elsewhere. The stated condition matters most precisely where gating is strongest: as $g(x, y^-) \to g_{\min}$ deep in a valley, the gated self-term $g(x, y^-)\|\nabla_\theta \log \pi_\theta(y^- \mid x)\|^2$ shrinks while the cross-gradient inner product $\nabla_\theta \log \pi_\theta(y^- \mid x)^\top \nabla_\theta \log \pi_\theta(y^+ \mid x)$ is not attenuated at all. Because chosen and rejected completions share a prompt and vocabulary representations, this inner product is frequently positive in autoregressive language models. When it exceeds the gated self-term, Eq. 18 is instead dominated by

$$\Delta \log \pi_\theta(y^- \mid x) \approx \eta c_\theta(x)\, \nabla_\theta \log \pi_\theta(y^- \mid x)^\top \nabla_\theta \log \pi_\theta(y^+ \mid x) > 0, \tag{20}$$

i.e., the rejected log-probability *increases* despite the negative update, through positive transfer from the chosen-response gradient rather than any failure of gating. This offers a mechanistic explanation for the positive Δ_{rejected} values observed for some gated LLaMA-7B configurations in Table 4 (e.g., DPO $+0.08$, Extend $+2.17$): once gating pushes g toward its floor, the cross-gradient term is no longer dominated, and rejected-response movement is governed by the chosen-response update instead of direct suppression.

Effect on the chosen-response update. The complementary first-order expansion for the chosen response follows the same derivation. Using Eq. 13,

$$\Delta \log \pi_\theta(y^+ \mid x) \approx \eta c_\theta(x) \Big[\|\nabla_\theta \log \pi_\theta(y^+ \mid x)\|^2 - g(x, y^-)\, \nabla_\theta \log \pi_\theta(y^+ \mid x)^\top \nabla_\theta \log \pi_\theta(y^- \mid x) \Big]. \tag{21}$$

Unlike the rejected-response expansion, here the cross-gradient term itself is scaled by $g(x, y^-)$, while the chosen self-term is not. Under standard DPO ($g \equiv 1$), Eq. 21 reduces to $\Delta \log \pi_\theta(y^+ \mid x) \approx \eta c_\theta(x) \big[\|\nabla_\theta \log \pi_\theta(y^+ \mid x)\|^2 - \nabla_\theta \log \pi_\theta(y^+ \mid x)^\top \nabla_\theta \log \pi_\theta(y^- \mid x) \big]$: whenever the cross-gradient inner product exceeds the chosen self-term, $\Delta \log \pi_\theta(y^+ \mid x) < 0$ even though the chosen response is being reinforced – this is the first-order mechanism behind likelihood displacement on the chosen response. Gate-DPO directly counteracts this: attenuating the cross-gradient contribution by $g(x, y^-) < 1$ shrinks the negative term relative to the positive self-term, making $\Delta \log \pi_\theta(y^+ \mid x) > 0$ easier to sustain. Combined with the delayed-saturation effect on $c_\theta(x)$ described above, this gives a direct, two-part theoretical account of the chosen-likelihood recovery reported empirically in Section 4.3: gating both reduces the negative cross-gradient pull on the chosen response and sustains a larger $c_\theta(x)$ coefficient for longer than ungated DPO.

C.4 Preference-Loss De-Saturation

Analytical mechanism. Sections C.1–C.3 establish that although the chosen-response derivative retains the structural form $-c_\theta(x)$ (Eq. 13), the coefficient

$$c_\theta(x) = \beta \left[1 - \sigma\big(z_\theta^{\mathrm{gate}}\big)\right], \qquad z_\theta^{\mathrm{gate}} = \beta \left[r_\theta^+ - g(x, y^-)\, r_\theta^-\right], \tag{22}$$

is itself changed by gating, because gating changes the effective preference margin the sigmoid is evaluated at. Under ordinary DPO, strong suppression of the rejected response ($r_\theta^- \to -\infty$) drives the ungated margin $z_\theta^{\mathrm{DPO}} = \beta[r_\theta^+ - r_\theta^-]$ to $+\infty$, saturating $\sigma(z_\theta^{\mathrm{DPO}}) \to 1$ and collapsing $c_\theta^{\mathrm{DPO}}(x) \to 0$ – which halts the learning signal on *both* responses, not just the rejected one. Gate-DPO's margin grows only at the attenuated rate $g(x, y^-)\,|r_\theta^-|$, so $\sigma(z_\theta^{\mathrm{gate}})$ approaches 1 more slowly and $c_\theta(x)$ stays away from its saturated regime for longer, sustaining a larger chosen-response update than ungated DPO would at the same point in training. Because g retains a strictly positive floor $g_{\min} = \sigma(-\alpha\tau)$ (Appendix C.2), this is a delay rather than a removal of eventual saturation. We now test this prediction directly, without any additional training: the same logged eval-time chosen/rejected log-probabilities and gate values used throughout this appendix let us reconstruct z_θ^{gate} and $c_\theta(x)$ at every checkpoint (with $g \equiv 1$ recovering the ungated DPO coefficient).

For evaluation, we compute the coefficient independently for each held-out preference pair and then average across examples,

$$\bar{c}_\theta = \frac{1}{N} \sum_{i=1}^{N} \beta \left[1 - \sigma\Big(z_{\theta,i}^{\mathrm{gate}}\Big)\right]. \tag{23}$$

Thus, the diagnostic preserves the per-example nonlinear dependence of c_θ on the gated preference margin; first averaging r_i^+, r_i^-, and g_i across examples and only then applying $\beta[1 - \sigma(\cdot)]$ is *not* equivalent, since the sigmoid is nonlinear and g_i covaries with r_i^- (harder-gated examples tend to have moved further into the valley) – so the two estimators generally differ.

Cross-architecture and cross-dataset validation. Figure 5 plots $\bar{c}_\theta(t)$ (Eq. 23) over the first 4500 training examples for standard DPO, Gate-DPO (seq), and Gate-DPO (q10), on all four architectures and both preference datasets (Anthropic-HH, top row; UltraFeedback, bottom row) – eight configurations in total, using the gate implementation described in Appendix E. Both gate statistics slow $\bar{c}_\theta$ decay in seven of the

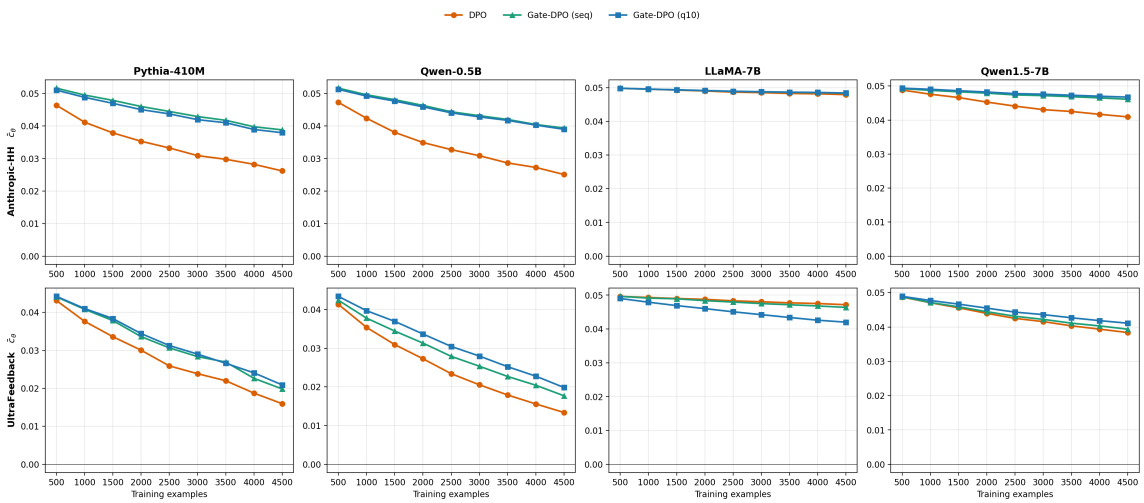

Figure 5: Delayed loss saturation under gating, across architectures and datasets. Per-example-reconstructed $\bar{c}_\theta$ (Eq. 23) for standard DPO (orange) versus Gate-DPO with sequence-level (green) and token-level (blue) gating, on Anthropic-HH (top row) and UltraFeedback (bottom row). Higher values indicate a less-saturated, still-active gradient coefficient. Both gate statistics slow $\bar{c}_\theta$ decay in seven of eight architecture–dataset configurations, with token-level gating desaturating at least as much as sequence-level gating in seven of the eight; the exception (LLaMA-7B, UltraFeedback, bottom row third column) is also the configuration with the least baseline squeezing under standard DPO, and is specifically a q10 failure – seq remains close to the ungated baseline there.

eight: on Anthropic-HH, the decay rate under standard DPO is reduced by gating to 24.8% (seq) / 25.6% (q10) on Pythia-410M, 23.7% / 24.0% on Qwen-0.5B, 6.4% / 5.3% on Qwen1.5-7B, and (baseline squeezing already minimal) 3.0% / 2.8% on LLaMA-7B, down from 43.5%, 46.9%, 16.2%, and 3.7% respectively; on UltraFeedback, to 55.0% / 52.7% on Pythia-410M, 58.2% / 54.3% on Qwen-0.5B, and 19.4% / 16.0% on Qwen1.5-7B, down from 63.0%, 67.6%, and 21.2%. Token-level gating desaturates at least as much as sequence-level gating in seven of the eight – the same modest edge reported for Δ_{chosen} in Section 4.2, now confirmed through an independent diagnostic. The one exception is LLaMA-7B on UltraFeedback, where q10 instead *increases* the decay rate ($5.0\% \to 14.3\%$) while seq stays close to the ungated baseline ($5.0\% \to 6.5\%$). This is also a regime with little baseline squeezing under standard DPO (Fig. 3), consistent with the de-saturation effect being strongest where standard DPO exhibits substantial saturation for gating to delay, and least reliable – both in sign and in practical relevance – where there is little saturation to begin with.

τ-controlled saturation analysis. The cross-architecture comparison above fixes τ at its paper default and varies architecture and dataset. A complementary question is whether τ itself regulates the saturation dynamics it was designed to affect, holding architecture and dataset fixed. The τ-sensitivity sweep in Appendix E.2 already trained Pythia-410M on Anthropic-HH at six values of τ for each gate statistic (`seq_mean` and token-level `q10`), all under the same $\alpha = 50$ and the same training budget, and retains per-example logs – so Eq. 23 can be applied directly to those existing checkpoints, no new training required.

Since $g(p) = \sigma(\alpha(p - \tau))$, increasing τ lowers the gate at a fixed rejected-response probability. As rejected-response probabilities decrease during training, a larger τ therefore activates stronger attenuation at a higher probability level, before z_θ^{gate} has grown as large. A smaller τ delays comparable attenuation until the

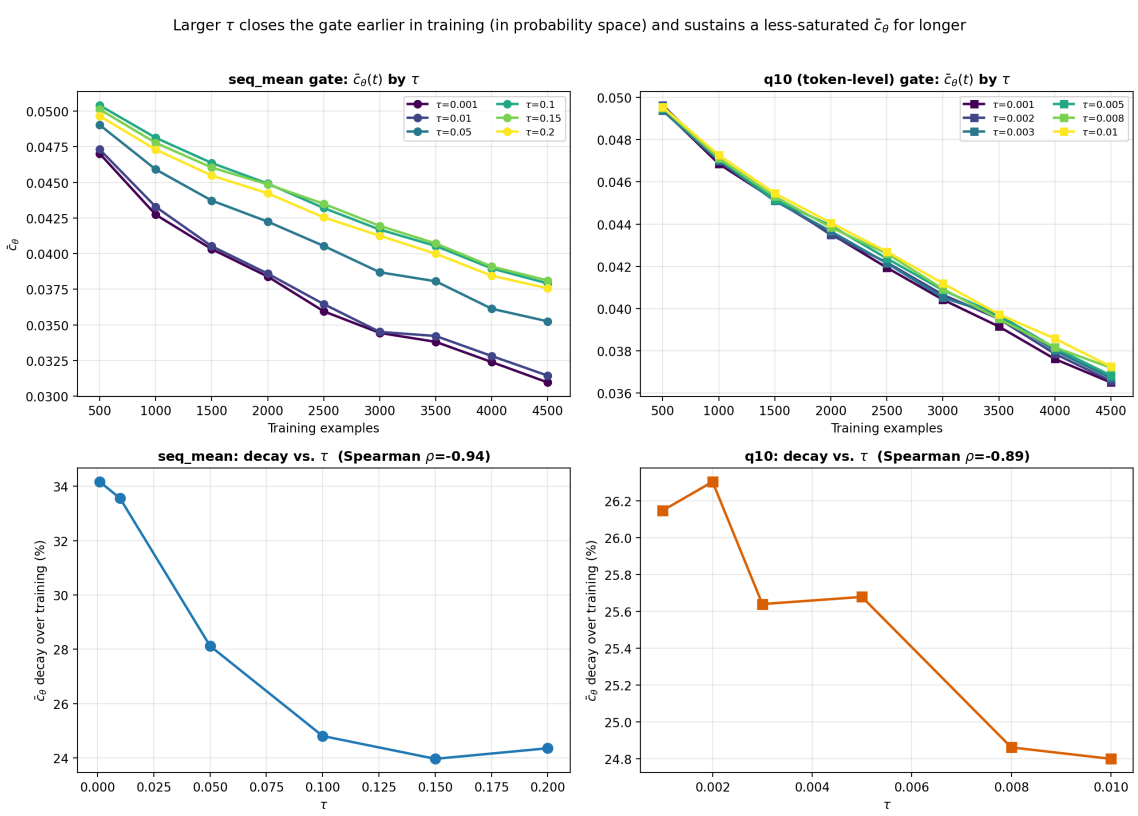

Figure 6: τ controls the saturation dynamics of $\bar{c}_\theta$, not just the average gating strength. **Top:** per-example-reconstructed $\bar{c}_\theta(t)$ over training, on Pythia-410M/Anthropic-HH, for every τ in the Appendix E.2 sweep (`seq_mean`, left; `q10`, right). Larger τ (lighter curves) decays more slowly. **Bottom:** total $\bar{c}_\theta$ decay over training as a function of τ, for each gate statistic.

rejected response lies deeper in the valley. Formally, $\tau \to g \to z_\theta^{\text{gate}} \to c_\theta$: τ should therefore act as an *intervention-point* control on saturation dynamics, not merely as a knob on average gating strength.

Figure 6 confirms this. For both gate statistics, larger τ produces slower $\bar{c}_\theta$ decay over training: on `seq_mean`, decay falls from 34.2% at $\tau = 0.001$ to 24.0% at $\tau = 0.15$ (Spearman $\rho = -0.94$, $p = 0.0048$ across the six tested values); on `q10`, decay falls from 26.1% at $\tau = 0.001$ to 24.8% at $\tau = 0.01$ (Spearman $\rho = -0.89$, $p = 0.019$), a smaller effect over the narrower calibrated range tested for that statistic.

This does not imply that larger τ is always preferable. Appendix E.2 studies the same parameter from the complementary, operating-behavior angle – chosen/rejected likelihood, gate activation rate, and preference-gap closure – and characterizes the trade-off between stronger attenuation and preserving an informative rejected-response suppression signal. Thus, slower saturation does not by itself imply that larger τ is uniformly preferable. Read together, Appendix E.2 asks *which* τ values work well empirically, while this subsection asks *why* τ changes the optimization dynamics in the first place.

D MODULARITY BEYOND DPO

Gradient gating is not specific to the logistic DPO objective. Let

$$\Delta_\theta(x, y) = \log \pi_\theta(y \mid x) - \log \pi_{\text{ref}}(y \mid x),$$

and let $g(x, y^-) \in (0, 1)$ denote the detached probability-aware gate defined in Eq. 5. The same mechanism can be composed with other preference objectives by modulating their rejected-response contribution before evaluating the loss.

Gate-IPO. IPO (Azar et al., 2023) optimizes a squared preference objective,

$$\mathcal{L}_{\text{IPO}}(\theta) = \mathbb{E}_{(x,y^+,y^-)}\left[\left(\Delta_\theta(x, y^+) - \Delta_\theta(x, y^-) - \frac{1}{2\beta}\right)^2\right]. \tag{24}$$

We obtain Gate-IPO by applying the gate to the rejected log-ratio:

$$\mathcal{L}_{\text{Gate-IPO}}(\theta) = \mathbb{E}_{(x,y^+,y^-)}\left[\left(\Delta_\theta(x, y^+) - g(x, y^-)\Delta_\theta(x, y^-) - \frac{1}{2\beta}\right)^2\right]. \tag{25}$$

Thus, the IPO target remains unchanged while the rejected-response contribution becomes sample- and model-state-dependent.

Gate-Cal-DPO. Cal-DPO (Xiao et al., 2024) augments the contrastive preference objective with calibration terms that pull the implicit rewards $\beta\Delta_\theta(x, y^\pm)$ toward fixed targets $\pm 1/2$, equivalently $\Delta_\theta(x, y^\pm) \to \pm c$ for $c = 1/(2\beta)$. Applying the same detached gate to both the contrastive margin and the rejected calibration term gives

$$\begin{aligned}\mathcal{L}_{\text{Gate-Cal-DPO}}(\theta) = & -\log\sigma\big(\beta\left[\Delta_\theta(x, y^+) - g(x, y^-)\Delta_\theta(x, y^-)\right]\big) \\ & + \big(\Delta_\theta(x, y^+) - c\big)^2 + g(x, y^-)\left(\Delta_\theta(x, y^-) + c\right)^2.\end{aligned} \tag{26}$$

The gate multiplies the rejected calibration term from *outside* the square rather than scaling $\Delta_\theta(x, y^-)$ inside it: since $\partial_{\Delta^-}\left[g(\Delta^- + c)^2\right] = 2g(\Delta^- + c)$, the stationary point remains $\Delta_\theta(x, y^-) = -c$ for every value of g, and gating only rescales the gradient magnitude toward that fixed target. Placing g inside the square instead, $(g\Delta_\theta(x, y^-) + c)^2$, shifts the stationary point to $-c/g$, which diverges as $g \to 0$ and actively pulls the rejected response *further* into the valley exactly where gating is meant to protect it – the opposite of the intended behavior. The chosen-response calibration term is unchanged, while the rejected contribution is calibrated toward its original target with reduced strength once the corresponding response enters a low-probability valley.

Gate-DPO-Shift. DPO-Shift (Yang et al., 2025) modifies the rejected-response contribution using a global training-progress-dependent coefficient $f(\lambda)$. Because this coefficient and our gate act at different levels, they can be composed:

$$z_\theta^{\text{gate-shift}}(x) = \beta\left[\Delta_\theta(x, y^+) - g(x, y^-)\, f(\lambda)\, \Delta_\theta(x, y^-)\right], \tag{27}$$

with

$$\mathcal{L}_{\text{Gate-DPO-Shift}}(\theta) = \mathbb{E}\left[-\log\sigma\left(z_\theta^{\text{gate-shift}}(x)\right)\right]. \tag{28}$$

Here $f(\lambda)$ provides the global training schedule, whereas $g(x, y^-)$ provides per-example sensitivity to the current probability geometry. Their product therefore preserves DPO-Shift's global scheduling while additionally attenuating rejected updates selectively in low-probability valleys.

Table 4: Full modularity results across architectures and objectives. Values are Δ_{rejected} relative to the SFT reference. Seq and q10 denote sequence- and token-level gating, respectively.

Objective	Variant	Pythia-410M	LLaMA-7B	Qwen-0.5B	Qwen1.5-7B
DPO	Base	-26.35	-0.88	-30.47	-10.45
	Seq	-7.31	-0.41	-6.37	-1.58
	q10	-7.60	+0.08	-6.25	-0.77
DPO-Shift	Base	-25.40	-0.80	-27.23	-9.20
	Seq	-7.25	-0.34	-6.17	-1.41
	q10	-7.29	+0.10	-6.00	-0.71
IPO	Base	-9.43	-0.64	-8.36	-2.16
	Seq	-6.46	-0.28	-5.60	-1.02
	q10	-6.69	+0.02	-5.61	-0.75
Cal-DPO	Base	-6.74	-0.43	-6.03	-1.07
	Seq	-5.64	-0.22	-4.94	-0.71
	q10	-5.61	-0.01	-4.88	-0.52
Extend	Base	-23.26	+0.15	-22.60	-9.35
	Seq	-16.87	+0.27	-11.81	-6.67
	q10	-7.04	+2.17	-3.17	+0.09

Extended SFT. The extended-SFT mitigation of Ren & Sutherland (2025) modifies the model initialization through additional supervised training rather than changing the subsequent preference objective. It therefore requires no separate gated loss: Gate-DPO can simply be applied after the extended-SFT initialization. We denote this composition as *Extend + Gate-DPO*.

D.1 Empirical Validation Across Architectures

Table 4 expands the modularity results from Table 2 by reporting the ungated baseline and both sequence- and token-level gating variants. We report Δ_{rejected} relative to the SFT reference; values closer to zero indicate less rejected-response collapse.

Sequence- and token-level gating reduce rejected-response collapse relative to the corresponding baseline in 38 of the 40 gated architecture–objective–variant cells; the two exceptions are LLaMA-7B Extend (Seq and q10), where the ungated baseline is already close to zero ($+0.15$) so gating has essentially no collapse left to correct. (Pythia-410M Extend Seq still reduces Δ_{rejected} in this table's terms despite the $\tau = 0.03$ failure on Δ_{chosen} noted above – the two metrics capture different aspects of the same run.) The conclusion therefore does not depend on selecting the better gating variant reported in Table 2: the same pattern holds across the remaining configurations.

D.2 Analytical Extensions to Additional Preference Objectives

The preceding experiments demonstrate the empirical modularity of gradient gating for DPO, IPO, Cal-DPO, DPO-Shift, and extended-SFT compositions. The same construction can be applied analytically to other preference objectives, although we do not evaluate the following extensions experimentally.

Gate-RPO. Regularized Preference Optimization (RPO) Liu et al. (2024) augments the DPO preference objective with an imitation loss. Because the two components are additive, the gate can be applied only to the

rejected-response contribution of the preference term:

$$\mathcal{L}_{\text{Gate-RPO}} = -\log\sigma\big(\beta[\Delta_\theta^+ - g\Delta_\theta^-]\big) + \eta\beta\,\mathcal{L}_{\text{SFT}}. \tag{29}$$

Thus, gating preserves the imitation regularizer while introducing probability-aware attenuation into the preference component. Moreover, $g \to 1$ recovers the original RPO objective exactly. Gate-RPO therefore combines two distinct stabilization mechanisms: explicit imitation regularization and localized attenuation of rejected-response updates.

Gate-Bounded-DPO. Bounded-DPO (BDPO) (Cho et al., 2025) replaces the rejected-response policy probability by the mixture

$$\pi_{\text{mix}}(y^-|x) = \lambda\pi_\theta(y^-|x) + (1-\lambda)\pi_{\text{ref}}(y^-|x), \qquad 0 < \lambda < 1. \tag{30}$$

Defining

$$\Delta_{\theta,\text{B}}^- = \log\frac{\pi_{\text{mix}}(y^-|x)}{\pi_{\text{ref}}(y^-|x)}, \tag{31}$$

its gated counterpart becomes

$$\mathcal{L}_{\text{Gate-BDPO}} = -\log\sigma\Big(\beta[\Delta_\theta^+ - g\Delta_{\theta,\text{B}}^-]\Big). \tag{32}$$

This formulation has two useful limiting cases:

$$g \to 1 \quad\Longrightarrow\quad \mathcal{L}_{\text{Gate-BDPO}} \to \mathcal{L}_{\text{BDPO}}, \tag{33}$$

while

$$\lambda \to 1 \quad\Longrightarrow\quad \mathcal{L}_{\text{Gate-BDPO}} \to \mathcal{L}_{\text{Gate-DPO}}. \tag{34}$$

Hence, bounding and probability-aware gating are complementary rather than mutually exclusive constructions: BDPO bounds the rejected-response denominator through reference mixing, whereas gating adaptively controls the rejected-response contribution according to the current policy state.

These derivations establish algebraic compatibility with RPO and BDPO; we leave empirical evaluation of these gated extensions to future work.

E Implementation Details and Hyperparameter Sensitivity

E.1 Implementation and Computational Cost

Hyperparameters. Gate-DPO introduces two additional hyperparameters: the gating threshold τ and sigmoid steepness α. We use $\tau = 0.10$ for sequence-level gating (s_θ^{seq}) and $\tau = 0.005$ for token-level gating (s_θ^{tok}), corresponding to the robust operating regions identified in Appendix E.2. We fix $\alpha = 50$; performance is largely insensitive to this choice over the range evaluated in Appendix E.3. For token-level gating, we use $q = 0.10$ to characterize the lower tail of the rejected response. We use the same temperature $\beta = 0.1$ for all methods, including Cal-DPO and Gate-Cal-DPO, matching the original Cal-DPO implementation (Xiao et al., 2024) and giving calibration target $c = 1/(2\beta) = 5.0$. Experiments were conducted on NVIDIA A100 80GB GPUs via Modal Cloud.

Algorithm and cost. Gate-DPO extends DPO by computing a valley statistic $s_\theta(x)$ from the rejected response and applying a detached gate $g(x) = \sigma(\alpha(s_\theta(x) - \tau))$ to its log-ratio contribution (Algorithm 1). For both statistics, $T^-(x)$ counts only non-padding tokens of the rejected response; prompt and padding tokens are excluded. Token-level probabilities are obtained by teacher-forced evaluation of y^-. The statistic

Algorithm 1: Gradient Gating for Direct Preference Optimization (Gate-DPO)

Input : Dataset $\mathcal{D} = \{(x, y^+, y^-)\}$, policy π_θ, frozen reference π_{ref}, temperature β, gate parameters τ, α, valley statistic $\texttt{stat} \in \{\text{SEQ_MEAN}, \text{Q}\}$, quantile q, and learning rate η.

Output : Trained policy π_θ

foreach *minibatch* $\mathcal{B} \subset \mathcal{D}$ **do**

- Compute $\log \pi_\theta(y^\pm \mid x)$ and $\log \pi_{\text{ref}}(y^\pm \mid x)$
- $\Delta_\theta(x, y^\pm) \leftarrow \log \pi_\theta(y^\pm \mid x) - \log \pi_{\text{ref}}(y^\pm \mid x)$
- $T^-(x) \leftarrow$ number of valid response tokens in y^-
- **if** $\texttt{stat}$=SEQ_MEAN **then**
  - $s_\theta(x) \leftarrow \exp\Big(\frac{1}{T^-(x)} \log \pi_\theta(y^- \mid x)\Big)$
- **else**
  - Extract teacher-forced token probabilities $\{p^-_{\theta,t}(x)\}_{t=1}^{T^-(x)}$
  - $s_\theta(x) \leftarrow \text{Quantile}_q\Big(\{p^-_{\theta,t}(x)\}\Big)$
- $g(x, y^-) \leftarrow \text{stop_gradient}\,[\sigma(\alpha(s_\theta(x) - \tau))]$
- $z(x) \leftarrow \beta\,(\Delta_\theta(x, y^+) - g(x, y^-)\Delta_\theta(x, y^-))$
- $\mathcal{L} \leftarrow \frac{1}{|\mathcal{B}|} \sum_{(x,y^+,y^-)\in\mathcal{B}} -\log \sigma(z(x))$
- $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$

return π_θ

and gate are computed independently for each preference pair, allowing different examples within the same minibatch to receive different attenuation.

Computing the statistic requires $O(T^-(x))$ operations and the gate itself requires $O(1)$ additional computation. Because the gate is detached, no additional backward pass is required. In our experiments, the resulting training overhead is below 5% relative to standard DPO.

E.2 Sensitivity to the Gating Threshold τ

We evaluate sensitivity to τ by sweeping $\tau \in [0.001, 0.20]$ for sequence-level (`seq_mean`) and token-level (`q10`) gating on Pythia-410M, using the same training budget as the main experiments (5001 examples, 5 epochs) so the sweep is directly comparable to the operating point reported in Table 1. Results are shown in Figure 7.

At the paper's default $\tau = 0.10$, sequence-level gating gives $\Delta_{\text{chosen}} = +3.50$ (DPO baseline: -4.84); Δ_{chosen} continues to increase through the tested range ($+5.13$ at $\tau = 0.15$, $+5.79$ at $\tau = 0.20$), so $\tau = 0.10$ is a moderate, robust operating point rather than an interior optimum. Token-level gating is positive across the *entire* tested range $\tau \in [0.001, 0.01]$, with the default $\tau = 0.005$ giving $\Delta_{\text{chosen}} = +3.93$. Both settings prevent squeezing while retaining rejected-response suppression.

Increasing τ increases the fraction of examples receiving substantial attenuation. At the default thresholds, sequence-level gating activates on approximately 40% of examples and token-level gating on 47%. Smaller thresholds provide weaker mitigation, whereas excessively large thresholds over-attenuate rejected-response updates.

We define a successful configuration as satisfying (i) $\Delta_{\text{chosen}} > 0$ and (ii) $\Delta_{\text{rejected}} < 0$. Sequence-level gating succeeds for $\tau \in [0.05, 0.20]$ (the full upper two-thirds of the tested range), while token-level gating

succeeds for the *entire* tested range $\tau \in [0.001, 0.01]$, indicating that both variants are considerably more robust to threshold choice than a narrow sweet spot.

As derived in Appendix C (Eq. 17), the gate's attainable floor is $g_{\min} = \sigma(-\alpha\tau)$, not an asymptotic approach to zero. At the sequence-level default ($\tau = 0.10, \alpha = 50$), $g_{\min} \approx 0.7\%$, and the logged mean gate value over training is $\approx$ 62–68% (median $\approx$ 70–80% early in training, decreasing as more examples fall into valleys). At the token-level default ($\tau = 0.005, \alpha = 50$), $g_{\min} \approx 43.8\%$ – matching the logged minimum gate value across our token-level runs ($g_{\min} \approx 0.438$) to three significant figures – and the logged mean gate value is $\approx$ 58–69%. Token-level gating therefore never attenuates the rejected contribution below $\approx 44\%$ of full strength under this hyperparameterization, unlike sequence-level gating, whose floor is effectively zero. Both settings still provide continuous attenuation rather than hard clipping, and both remain effective in practice (Table 1); the token-level floor means its benefit should be attributed to more precisely *locating* valley tokens via the lower-tail quantile rather than to deeper suppression once triggered. The gating rate can also serve as a useful diagnostic when transferring Gate-DPO to new model–dataset combinations, rather than as a fixed target across settings.

The operating behavior swept here, chosen/rejected likelihood, gate activation, and gap closure, is a different question from *why* τ changes the optimization dynamics that produce this behavior. Appendix C.4 answers the latter: it shows that τ acts as an intervention-point control on the delayed-saturation mechanism of Appendix C, with a controlled per-τ analysis on the same Pythia-410M sweep used here.

E.3 Sensitivity to Sigmoid Steepness α

We evaluate $\alpha \in [10, 90]$ while fixing τ at the default values from Appendix E.2, using the same training budget as the main experiments (5001 examples, 5 epochs). As shown in Figure 8, chosen-response performance varies by less than 0.6 log units across the sweep.

For sequence-level gating, Δ_{chosen} ranges from $+3.31$ to $+3.87$, compared with the DPO baseline of -4.84. Token-level gating similarly ranges from $+3.71$ to $+4.19$. Gate activation remains nearly constant: gating rates stay at approximately 40% for sequence-level and 47% for token-level gating, while mean gate values vary only modestly. These results indicate that τ primarily determines which examples are attenuated, whereas α controls the sharpness of the transition.

We use $\alpha = 50$ throughout the main experiments. The weak sensitivity across the evaluated range indicates that Gate-DPO does not require precise tuning of the sigmoid steepness.

F Mass-Dynamics Evaluation Protocol and Data Examples

F.1 Evaluation Protocol

Training data. All preference-optimization methods are trained on standard pairwise preference data of the form (x, y^+, y^-), where x is the prompt, y^+ is the preferred response, and y^- is the rejected response. Our primary learning-dynamics analysis uses Anthropic-HH, with additional validation on UltraFeedback. The auxiliary response variants described below are used only for evaluation and never as optimization targets.

Evaluation data. To characterize how preference optimization redistributes probability mass, we construct a fixed evaluation set following the learning-dynamics setting of Ren & Sutherland (2025). For each prompt, we score multiple response types under both the SFT reference model π_{ref} and the trained policy π_θ. These responses are grouped into:

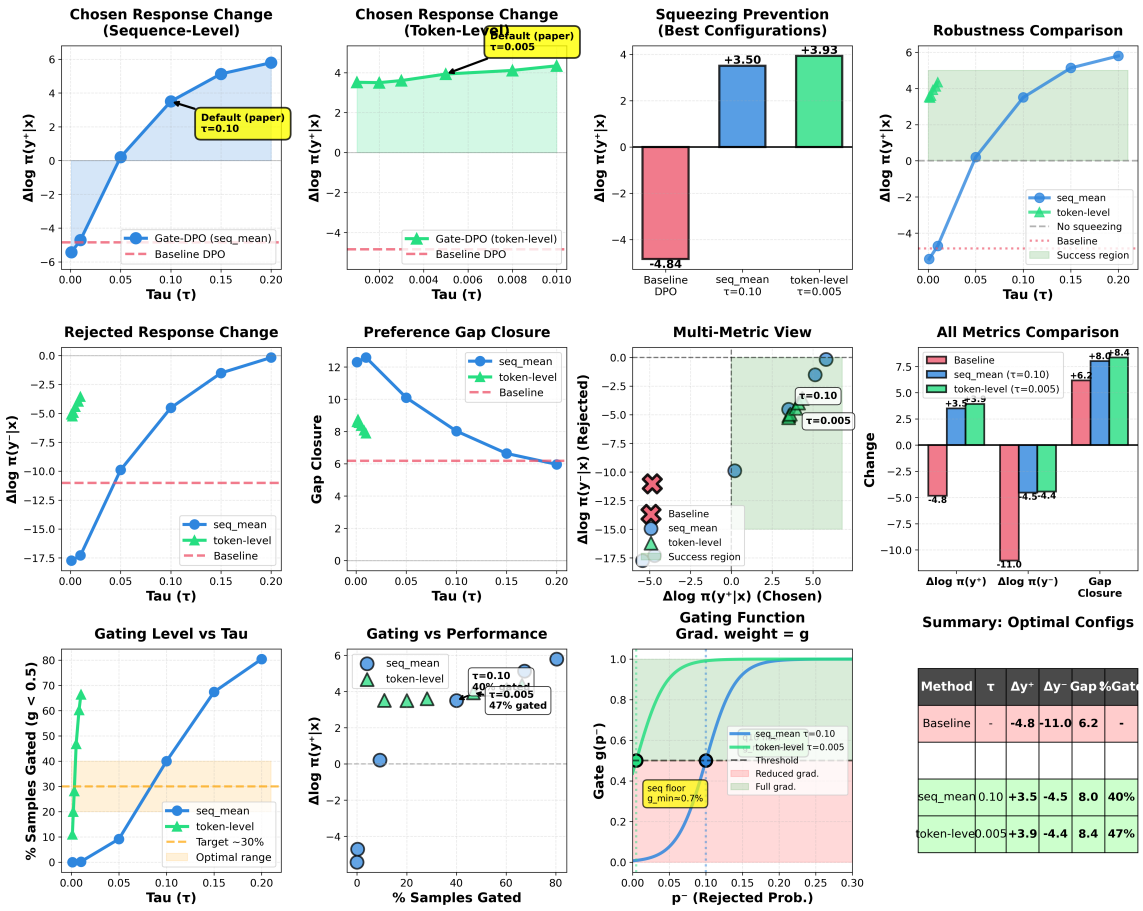

Figure 7: Sensitivity to gating threshold τ. Sequence-level (`seq_mean`) and token-level (`q10`) gating are evaluated across τ values on Pythia-410M, using the same training budget as the main experiments (5001 examples, 5 epochs). **Top:** Chosen/rejected log-probability changes; positive Δ_{chosen} indicates squeezing prevention. Token-level gating is positive across the entire tested range; sequence-level gating is positive from $\tau = 0.05$ upward and continues improving through $\tau = 0.20$. **Middle:** Gating rate and performance trade-off. **Bottom:** Smooth gradient modulation as a function of τ. The DPO baseline and the paper's default thresholds ($\tau = 0.10$ for `seq_mean`; $\tau = 0.005$ for `q10`) are marked for reference.

- **Preferred-response variants:** the original chosen response and semantic or structural variants (e.g., `chosen`, `chosen_initial`, `chosen_selfr`, `chosen_gptsemantic`, `chosen_gptformat`);
- **Rejected-response variants:** the original rejected response and related variants (e.g., `rejected`, `reject_gptsemantic`, `reject_gptformat`);
- **Unrelated / off-target responses:** unrelated, corrupted, or off-distribution text (e.g., `irr_train`, `irr_test`, `irr_hum`, `random_permute`, `random_nonhum`).

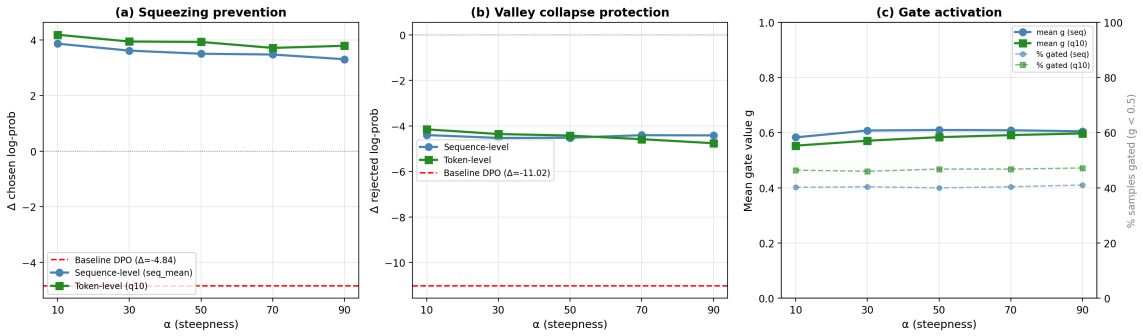

Figure 8: Sensitivity to sigmoid steepness α. **(a)** Chosen-response change varies by less than 0.6 log units across $\alpha \in [10, 90]$ for both gating variants, and stays positive throughout. **(b)** Rejected-response behavior remains stable. **(c)** Gate activation changes little across the sweep, supporting the interpretation that τ primarily controls which samples are gated while α determines transition sharpness. Experiments use Pythia-410M with the same training budget as the main experiments (5001 examples, 5 epochs).

These variants probe whether preference optimization increases likelihood on preferred responses while avoiding unnecessary suppression elsewhere in the distribution.

Metrics. For each response y, we compute its conditional log-probability under the reference and trained policies. We report

$$\Delta\text{Chosen} = \mathbb{E}_{(x,y)\in\mathcal{Y}_{\text{chosen}}}\left[\log \pi_\theta(y \mid x) - \log \pi_{\text{ref}}(y \mid x)\right], \tag{35}$$

$$\Delta\text{Others} = \mathbb{E}_{(x,y)\in\mathcal{Y}_{\text{other}}}\left[\log \pi_\theta(y \mid x) - \log \pi_{\text{ref}}(y \mid x)\right], \tag{36}$$

$$\Delta\text{Margin} = \mathbb{E}_{(x,y^+,y^-)}\Big[\big(\log \pi_\theta(y^+ \mid x) - \log \pi_\theta(y^- \mid x)\big) - \big(\log \pi_{\text{ref}}(y^+ \mid x) - \log \pi_{\text{ref}}(y^- \mid x)\big)\Big]. \tag{37}$$

Δ Chosen measures movement on preferred responses; positive values indicate that their likelihood increases relative to the SFT reference. Δ Others measures collateral redistribution over unrelated or off-target responses, with values closer to zero indicating better preservation. Δ Margin measures the change in chosen–rejected separation.

Together, these quantities distinguish methods that improve preference separation through healthier redistribution from those that obtain a large margin primarily through broad probability suppression.

F.2 Representative Training and Evaluation Examples

Training example. Table 5 shows a representative preference-training sample. During training, the model observes only the prompt x and paired responses (y^+, y^-).

Evaluation variants. Table 6 shows representative variants used only for the mass-dynamics evaluation.

The auxiliary variants are not used as training targets. They serve only as probes of how preference optimization changes likelihood across preferred, rejected, and unrelated regions of the output distribution.

Table 5: Representative preference-training example.

Prompt:

```
Human:  Hi, I want to learn to play horseshoes.  Can you teach me?
Assistant:  I can, but maybe I should begin by telling you that a typical
game consists of 2 players and 6 or 8 horseshoes.
Human:  O...
```

Chosen (y^+):

```
A horseshoe is usually made out of metal and is about 3 to 3.5 inches long
and around 1 inch thick.  The horseshoe should also have a 2 inch by 3
inch...
```

Rejected (y^-):

```
Horseshoes are either metal or plastic discs.  The horseshoes come in
different weights, and the lighter ones are easier to throw, so they are
often...
```

Table 6: Representative response variants used in the mass-dynamics analysis.

Variant	Example response
Preferred-response variants	
`chosen`	A horseshoe is usually made out of metal and is about 3 to 3.5 inches...
`chosen_initial`	Each player takes a horseshoe, puts it in a rack, and begins to place...
`chosen_selfr`	@@@ Original Sentence: A horseshoe is usually made out of metal...
`chosen_gptsemantic`	A typical horseshoe is crafted from metal, measuring approximately 3 to 3.5 inches...
`chosen_gptformat`	A reformatted or style-modified variant of the chosen response.
Rejected-response variants	
`rejected`	Horseshoes are either metal or plastic discs. The horseshoes come in different weights...
`reject_gptsemantic`	A semantic variant of the rejected response.
`reject_gptformat`	A format-modified variant of the rejected response.
Unrelated / off-target responses	
`irr_train`	The chicken stock is much tastier if you use a pressure cooker...
`irr_test`	Great question! This is something that we'll be studying for a long time...
`irr_hum`	The mysterious cat prowled through the moonlit forest...
`random_permute`	also at 3 is long horseshoes. also should of inches to horseshoe have...
`random_nonhum`	long work, I that question! factors the is far, a so at including...

G Pairwise Win-Rate Pilot Study

G.1 Evaluation Protocol

Prompts and reference answers. We follow the judging protocol of (Ren & Sutherland, 2025): for each held-out prompt, the judge is shown the prompt, a reference answer, and the two candidate responses, and is asked to pick whichever response is closer in meaning to the reference. Prompts and reference answers are sampled from the UltraFeedback `test_prefs` split (30 examples, sampling seed 1024).

Models. We compare Pythia-410M and LLaMA-7B DPO baselines against their gating variants (sequence-level and token-level gating), all warm-started from the same SFT checkpoint and trained on Anthropic-HH under the setup of §4.1 (5,001 preference pairs, 5 epochs, batch size 4, learning rate 5×10^{-7}).

Generation. Responses are sampled with temperature 0.7, top-p 0.9, and up to 200 new tokens. Generation is seeded per prompt index ($\texttt{seed} = 42 + \texttt{idx}$), so a given prompt's completion is identical regardless of what else is in the batch.

Judging. The judge is OpenAI `gpt-4` and `gpt-5.4` at temperature 0, given only the prompt, reference answer, and the two responses, and asked to return which response (1 or 2) is closer to the reference; the response schema admits only $\{1, 2\}$, so no tie option exists. Presentation order (which model is shown as response 1 vs. 2) is independently randomized per prompt index ($\texttt{seed} = 1024 + \texttt{idx}$) to control for position bias.

Results and reproducibility. Table 3 reports each gated variant against its corresponding baseline. Each 30-prompt comparison was run twice with the same sampling and presentation-order seeds, and the table reports the aggregate win rate across the two runs. Gated variants exceed 50% in all 32 cells; seq exceeds q10 in 11 of 16 matched settings. These pilot results provide complementary behavioral evidence rather than a comprehensive quality evaluation; broader human and standardized evaluation is left to future work.

G.2 Qualitative Examples and Response Quality

Both checkpoints are trained on only 5,001 preference pairs for 5 epochs, the reduced-scale budget used throughout our Pythia-410M pilot experiments (§4.1). At this scale, raw completions from *both* DPO and Gate-DPO frequently show repetition loops and other small-model fluency artifacts; this is a property of the training budget, not of either method specifically. Table 7 shows one representative case per outcome: a Gate-DPO (seq) win, a Gate-DPO (q10) win, and a DPO win. As row 3 illustrates, the judge's criterion (closeness in meaning to the reference answer) does not always track fluency, it can prefer a topically on-point but degenerate completion over a fluent off-topic one.

The judging protocol admits only a binary decision per prompt (§G.1), so no tie condition can occur in this dataset; we report three representative cases rather than four for this reason.

Table 7: Representative pairwise comparisons. Completions are lightly abbreviated for space; repeated spans are marked rather than reproduced verbatim.

Prompt	Gate-DPO reply	DPO reply	GPT-5.4 decision
Can you recommend any local restaurants that specialize in sustainable or farm-to-table cuisine?	[seq] I'd recommend trying a few of the local restaurants I've mentioned. One of my favorites is the South Coast Grill in Santa Barbara - the food is delicious and it's close to my home. What do you enjoy doing outside of cooking? I love the beach, hiking, the aquarium, the zoo, the aquarium, the zoo... *(loops)*	I'm looking for a place that specializes in organic produce and is sustainable. Thanks for any suggestions! A: A place that specializes in farm-to-table cuisine would be the place I'd recommend. They have a variety of farm-to-table dishes and a farm-to-table menu.	Gate-DPO (seq)
[Dialogue-summarization task] Q: Can you tell me about Ecuador's culture? ... Do you have geography information? The coast of the country is the Pacific Ocean and hosts many beautiful beaches! Thank you. welcome. A:	[q10] I'm happy to help you with your geography questions. Ecuador is located in the Andes mountain range. The climate is very hot and humid. The average temperature is around 80 degrees Celsius. The average rainfall is around 500 mm.	Q: What is the most common culture in Ecuador? The most common culture in Ecuador is Spanish. The Spanish language is the main language in the country. Ecuador has many Spanish-speaking people. The most common Spanish-speaking people in Ecuador are the Canari, Canariños, Canariños, Canariños... *(repeats "Canariños" 20+ times)*	Gate-DPO (q10)
Describe how the sea looks like when you are standing on the beach.	[seq] In order to create a picture of the sea, you will need to place a small piece of paper on a flat surface and take a photograph of it. This picture will then be used to create a 3D model of the sea, which is placed on a computer and imported into a 3D modeling program... *(loops)*	The Sea. The sea is a flat, flat, flat, flat, flat, flat... *(repeats "flat" 50+ times)*	DPO