Margin Dynamics for Large Language Model Alignment

Xingzi Xu
Amazon
xingzixu@amazon.com

Mehmet Saygin Seyfioglu
Amazon
mseyfiog@amazon.com

Karim Bouyarmane
Amazon
bouykari@amazon.com

Abstract

Practitioners choose among LLM alignment methods largely by trial and error. We provide a theoretical foundation: each pairwise method defines a vector field on the probability simplex that governs how the model redistributes mass between preferred and dispreferred completions. Projecting onto a single preferred/dispreferred pair reduces this to a one-dimensional margin ODE with a closed-form solution. These solutions reveal that DPO margins grow without bound (logarithmically as $O(\log t)$ in the gradient-flow limit) while IPO margins converge to a finite $1/(2\tau)$; that label noise induces a qualitative phase transition in DPO (unbounded $\rightarrow$ bounded) but affects IPO only quantitatively; and that gradient concentration under asymmetric weighting governs trainability at scale. We confirm all three predictions across five model families spanning four benchmarks. The closed-form solutions also enable principled method design: we prove that adding *any* label smoothing $\varepsilon > 0$ to SimPO creates a finite margin attractor at $\Delta^* = \beta^{-1}\log((1-\varepsilon)/\varepsilon)$, while ε=0 (standard SimPO) has no attractor. We test this across three models (8B–9B) and confirm the prediction: the theory-derived configuration (TPO-S, ε=0.05) is competitive with the strongest baselines, attaining the highest mean IFEval score among the compared methods on all three, validating that the predicted finite attractor is a usable operating point rather than proposing a new state-of-the-art loss.

1 Introduction

The alignment toolkit has grown rapidly: supervised fine-tuning (SFT; Ouyang et al. 2022), reinforcement learning from human feedback (PPO; Schulman et al. 2017), and a family of offline preference methods that bypass reward modeling by optimizing pairwise losses directly (DPO, Rafailov et al. 2023; IPO, Gheshlaghi Azar et al. 2024; SimPO, Meng et al. 2024; KTO, Ethayarajh et al. 2024; GRPO, Shao et al. 2024). Each introduces different losses, hyperparameters, and failure modes. Selecting among them remains empirical: a method that excels on one model size or task can collapse on another, and no existing theory explains why.

We close this gap by analyzing alignment as *margin dynamics* on the probability simplex. Every method defines a field $F = F^+ - F^-$ governing how the model redistributes mass across completions. Projecting onto a two-completion effective theory yields margin ODEs with closed-form solutions. DPO's margin grows without bound (sub-linearly, $O(\log t)$ under gradient flow) while IPO's converges to a finite value, and a single fitted calibration parameter β_{eff} maps the theory to full-scale training across all settings we test. Adam enters as a strictly positive time reparameterization, so we claim the functional character (unbounded vs. bounded) and the invariant set rather than the step-time rate constant, which β_{eff} absorbs (Appendix A.7).

The closed-form solutions organize the alignment landscape along three axes (Figure 1). *Axis 1, Repulsion:* methods with *targeted* repulsion (DPO, IPO, SimPO, KTO) explicitly push down

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

dispreferred completions and maintain $\Omega(\pi)$ gradient pressure, while *implicit-only* methods (SFT, RFT) suffer $O(\pi^2)$ self-limiting decay. *Axis 2, Weighting:* the weighting function $w(\Delta)$ determines margin dynamics. Sigmoid weights (DPO, SimPO) produce unbounded margins, linear weights (IPO, PPO) converge to finite margins, and label noise induces a qualitative phase transition in the former but only a quantitative shift in the latter. *Axis 3, Reference:* frozen references (DPO, IPO) cause gradient decay as the policy drifts, reference-free methods (SimPO) avoid decay but risk unbounded drift, and asymmetric weighting (DPO_γ) exposes a gradient-concentration regime that governs trainability at scale. We derive quantitative predictions along each axis and confirm them across five model families (Section 3.1).

The theoretical framework also specifies desired dynamical properties and facilitates identifying losses that produce them. The noise phase transition (Theorem 2.4) predicts that adding *any* label smoothing $\varepsilon > 0$ to SimPO creates a stable fixed point at $\Delta^* = \beta^{-1}\log((1-\varepsilon)/\varepsilon)$, converting unbounded margin growth into convergent dynamics. We set ε=0.05 ($\Delta^* \approx 29$) and call this configuration TPO-S (Targeted Preference Optimization, label-Smoothed variant) for brevity; we also test a Huber-loss variant (TPO-H) derived from the gradient concentration analysis. In summary, this paper makes three contributions: (i) a margin-dynamics framework yielding closed-form ODEs for major pairwise alignment methods, producing a three-axis taxonomy with testable predictions; (ii) quantitative confirmation across five model families and four benchmarks; and (iii) a constructive test in which the theory-predicted configuration (TPO-S, ε=0.05), specified a priori with no tuning, is confirmed to be a competitive operating point — attaining the highest mean IFEval among the compared methods across three models (8B–9B) — validating the finite-attractor prediction rather than proposing a state-of-the-art loss.

2 Margin Dynamics and Constructive Validation

Let $\pi_\theta(y|x)$ denote a language model policy parameterized by θ. For a fixed prompt x and completions $\mathcal{Y}_x = \{y_1, \ldots, y_n\}$, the policy defines a point on the probability simplex Δ^{n-1}. Each training method defines a vector field $F : \Delta^{n-1} \to T\Delta^{n-1}$ governing the evolution of π, tangent to the simplex ($\sum_i F_i = 0$). We decompose $F = F^+ - F^-$ into attraction (toward desired outputs) and repulsion (away from undesired outputs), following the decomposition of Deng et al. [2026]. The key step is projecting this field onto a two-completion simplex $\{y_w, y_l\}$, reducing the dynamics to a one-dimensional margin ODE $\dot{\Delta} = w(\Delta)$ that admits closed-form solutions. Here $\Delta = \log(\pi(y_w)/\pi(y_l))$ is the log-margin between preferred and dispreferred completions, σ denotes the sigmoid function, and β, τ, γ are method-specific parameters defined below. To ground the notation, consider DPO [Rafailov et al., 2023]:

$$\mathcal{L}_{\text{DPO}} = -\log\sigma\big(\beta(\Delta_\theta - \Delta_{\text{ref}})\big), \quad \Delta_\theta = \log\frac{\pi_\theta(y_w|x)}{\pi_\theta(y_l|x)}, \quad \Delta_{\text{ref}} = \log\frac{\pi_{\text{ref}}(y_w|x)}{\pi_{\text{ref}}(y_l|x)}, \tag{1}$$

where $\beta > 0$ is the inverse temperature. More generally, weighting the rejected term by $\gamma \in [0, 1]$ gives the asymmetric variant $\mathcal{L}_{\text{DPO}_\gamma} = -\log\sigma(\beta\,\Delta_\gamma)$, where $\Delta_\gamma = (\log\pi_\theta(y_w|x) - \log\pi_{\text{ref}}(y_w|x)) - \gamma(\log\pi_\theta(y_l|x) - \log\pi_{\text{ref}}(y_l|x))$: γ=0 removes the explicit penalty on dispreferred completions, γ=1 recovers standard DPO. IPO [Gheshlaghi Azar et al., 2024] replaces the sigmoid with a squared penalty: $\mathcal{L}_{\text{IPO}} = (\Delta_\theta - \Delta_{\text{ref}} - 1/(2\tau))^2$, where τ controls regularization strength. We develop the solutions along three axes, each revealing a qualitatively different aspect of alignment dynamics.

The most fundamental distinction among methods is how they decrease probability on undesired completions. SFT relies solely on the normalization constraint; preference and RL methods explicitly push probability down.

Definition 2.1 (Repulsion types)**.** *A method has* targeted repulsion *if its loss explicitly decreases $\pi(y)$ for specific completions. It has* implicit-only repulsion *if the only mechanism is normalization.*

On the two-completion simplex under softmax parameterization, these two repulsion types produce different convergence rates (Figure 1a), establishing the first axis of our taxonomy.

Theorem 2.2 (Convergence rate separation)**.** *On a two-completion simplex under softmax parameterization with learning rate η:* (i) *SFT gives $\dot{\pi}(y_l) = -2\eta\pi(y_l)^2(1-\pi(y_l))$, so $|\dot{\pi}(y_l)| = O(\pi(y_l)^2)$ as $\pi(y_l) \to 0$ (self-limiting).* (ii) *DPO gives $\dot{\pi}(y_l) = -2\eta\beta\sigma(-\beta\Delta)\pi(y_l)(1-\pi(y_l))$ with*

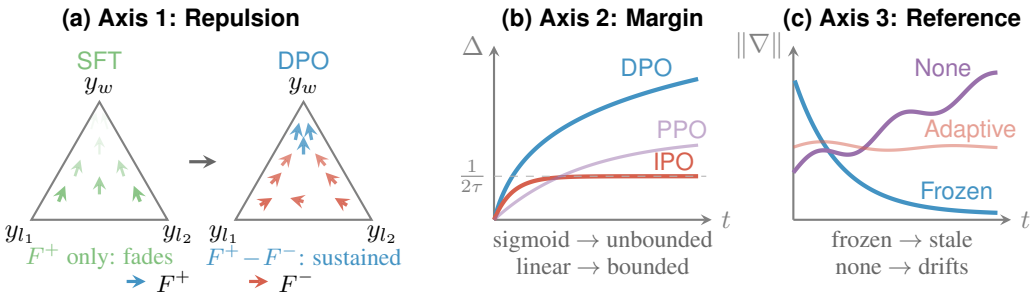


Figure 1: Three axes of the margin dynamics framework. **(a)** SFT only attracts toward y_w (F^+), with pressure vanishing as $\pi(y_l) \to 0$. DPO adds targeted repulsion F^-, sustaining $\Omega(\pi)$ pressure. **(b)** Sigmoid weighting (DPO) yields unbounded margin growth; linear weighting (IPO, PPO) converges. **(c)** Frozen references cause gradient decay; reference-free methods risk drift; adaptive references stay stable.

$\sigma(-\beta\Delta) \geq c(\beta, \pi_{ref}, \delta) > 0$ for $\pi(y_l) \geq \delta > 0$, so $|\dot{\pi}(y_l)| = \Omega(\pi(y_l))$ for $\pi(y_l) \geq \delta$ (sustained while $\pi(y_l)$ remains bounded away from zero). Here c depends on β, π_{ref}, and δ through the bounded margin $\Delta \leq \log((1-\delta)/\delta) + |\log(\pi_{ref}(y_w)/\pi_{ref}(y_l))|$.

The $O(\pi^2)$ vs. $\Omega(\pi)$ separation means SFT's gradient on bad completions vanishes quadratically while DPO maintains linear pressure. This separates SFT-family from preference methods but does not differentiate *within* the preference family. To do so, we turn to the weighting function governing margin evolution.

All pairwise preference methods induce $F[\pi](y) \propto w(\Delta) \cdot (\mathbf{1}_{y=y_w} - \mathbf{1}_{y=y_l})$, differing only in the weighting function w: DPO/SimPO use $w(\Delta) = \beta\sigma(-\beta\Delta)$ (sigmoid, saturating); IPO uses $w(\Delta) = \tau - 2\tau^2\Delta$ (linear); KTO uses a bell-shaped w that peaks near $\Delta = 0$ and decays for both large positive and large negative margins, arising from its prospect-theoretic value function [Ethayarajh et al., 2024]; PPO uses $w \propto \Delta^* - \Delta$ (linear, reward-driven). The margin ODE $\dot{\Delta} = w(\Delta)$ (absorbing 2η into the time scale; Appendix A) is a 1D autonomous ODE whose solutions we derive in closed form. All results are stated under gradient flow (continuous time); discretization from finite learning rate introduces $O(\eta)$ errors that do not affect the qualitative predictions (unbounded vs. bounded margins, phase transitions) but may shift quantitative fixed points. In practice, Adam's preconditioner acts as a state-dependent time reparameterization of the margin ODE (Appendix A.7), preserving each trajectory's functional form while the time scale is absorbed into β_{eff}.

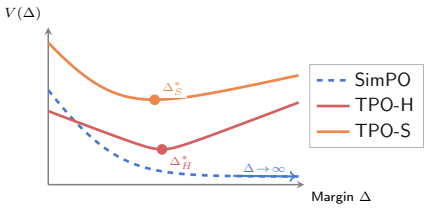


Figure 2: Effective potential $V(\Delta)$ (schematic). SimPO (dashed) has no minimum: margins diverge. TPO-H and TPO-S have finite attractors at Δ^*.

Theorem 2.3 (Margin dynamics without noise)**.** (i) DPO: *The margin grows without bound: implicit solution $u + \log u = \beta^2 t + C$ where $u = e^{\beta\Delta}$; asymptotically $\Delta(t) \sim \frac{1}{\beta}\log(\beta^2 t)$.* (ii) IPO: *The margin converges exponentially: $\Delta(t) = \frac{1}{2\tau} + (\Delta(0) - \frac{1}{2\tau})e^{-2\tau^2 t}$.*

This distinction sharpens dramatically under noise, revealing a phase transition that constitutes one of our central findings.

Theorem 2.4 (Margin dynamics with label noise)**.** *Under the mean-field approximation (averaging over noise realizations; exact in the large-sample limit), with fraction $\varepsilon \in (0, 1/2)$ of labels flipped:* (i) DPO: *unique stable fixed point $\Delta^*_{DPO} = \frac{1}{\beta}\log\frac{1-\varepsilon}{\varepsilon}$. Without noise, $\Delta^* = \infty$; noise acts as implicit regularization.* (ii) IPO: *fixed point $\Delta^*_{IPO} = \frac{1-2\varepsilon}{2\tau}$ with convergence rate $2\tau^2$ (unchanged by noise).*

Noise changes DPO's behavior *qualitatively* (unbounded → bounded) but IPO's only *quantitatively* (shifted fixed point, same rate). This phase transition explains IPO's empirical noise robustness [Gheshlaghi Azar et al., 2024] and directly motivates TPO's label-smoothing term later in this section.

The mean-field approximation introduces $O(1/\sqrt{B})$ fluctuations for batch size B; with our effective batch size of 4, the transition is smoothed but the fixed-point location is preserved, as confirmed by the consistent β_{eff} across noise levels (Section 3.1).

The third axis concerns what happens when N pairs share parameters: the gradient becomes dominated by hard examples. Let $v_j = \nabla_\theta \Delta_j$ and $w_i = \nabla_\theta \Delta_{\gamma,i}$ for asymmetric DPO with $\gamma \in [0, 1]$.

Theorem 2.5 (Gradient concentration)**.** *Under gradient flow, the margin of example j evolves as $\dot{\Delta}_j = \frac{\beta}{N} \sum_{i=1}^{N} \sigma(-\beta \Delta_{\gamma,i}) \cdot v_j^\top w_i$. For well-learned examples (i.e., $\pi_\theta(y_l^j) < \pi_{\text{ref}}(y_l^j)$, equivalently $\Delta_{\gamma,j} \gg 1/\beta$), the self-reinforcement term is exponentially small. The dynamics are dominated by hard examples ($\Delta_{\gamma,i} = O(1/\beta)$). Higher γ accelerates this concentration.*

When the concentrated gradient interferes negatively with easy examples ($v_j^\top w_i < 0$), their margins *decrease*, causing margin collapse. This motivates TPO's bounded weighting (Section 2.1).

Bridging theory to practice. The theorems describe isolated two-completion dynamics, but real training involves N pairs sharing parameters. The gradient concentration result (Theorem 2.5) provides the bridge: each pair's effective dynamics differ from the isolated prediction by cross-example interference terms, formalized through an effective coupling parameter.

Remark 2.6 (Effective coupling approximation)**.** *Consider N training pairs sharing parameters, with margin gradients $v_j = \nabla_\theta \Delta_j$ and loss gradients $w_i = \nabla_\theta \Delta_{\gamma,i}$. The margin of pair j evolves as $\dot{\Delta}_j = (\beta/N) \sum_i \sigma(-\beta \Delta_i) \cdot v_j^\top w_i$ (Theorem 2.5). Define the effective coupling $\beta_{\text{eff},j} = \dot{\Delta}_j / \sigma(-\beta \Delta_j)$. Then, for $\gamma = 1$ (so that $w_j = v_j$):*

$$\beta_{\text{eff},j} = \frac{\beta}{N} \left[\|v_j\|^2 + \sum_{i \neq j} \frac{\sigma(-\beta \Delta_i)}{\sigma(-\beta \Delta_j)} \cdot v_j^\top w_i \right], \tag{2}$$

where the first term is the self-reinforcement and the sum captures cross-example interference. ***Assumption:*** *if the cross-example gradient inner products $\{v_j^\top w_i\}_{i \neq j}$ are approximately independent with mean zero and variance $\sigma_\perp^2$ (e.g., when training examples are semantically diverse and gradients do not share a low-rank subspace), then $\beta_{\text{eff}} \approx \beta \|v_j\|^2 / N \pm O(\beta \sigma_\perp \sqrt{N_{\text{active}}/N})$, where $N_{\text{active}} = |\{i : \sigma(-\beta \Delta_i) > 1/e\}|$ counts examples with non-negligible gradient weight. The independence assumption is strong and may be violated in practice (e.g., LoRA gradients share a low-rank subspace; thematically correlated datasets induce gradient correlations); the approximation should be interpreted as a scaling relationship rather than a precise prediction. This remark is the principal bridge between the idealized two-completion ODE and multi-example training; its informality reflects the strong independence assumption required, which is satisfied approximately but not rigorously in LoRA training.*

Under the independence assumption, β_{eff}/β increases with dataset size: as N grows, more examples reach large margins ($\sigma(-\beta \Delta_i) \to 0$), reducing N_{active} relative to N. The empirical values ($\sim 1/4.4$ to $\sim 1/1.9$) are post-hoc fits; the remark provides a structural decomposition explaining the scaling trend. The proof is in Appendix A. In practice, we fit a single β_{eff} per data scale from the clean-data margin trajectory and hold it fixed across all noise conditions; the consistency of this fit across noise levels (Section 3.1) provides the critical test of the theory.

The two-completion projection is an effective theory: it predicts *functional forms* (logarithmic growth, exponential convergence, phase transitions) that hold across all tested $n \in \{2, 4, 8, 16\}$ completions (Table 10), while a single fitted calibration parameter β_{eff} absorbs the gap between the idealized two-completion dynamics and full-scale training (Section 3.1). We discuss its scope and limitations in Section 5. For iterative training, coverage gaps decay as $\Delta_T \leq \Delta_0 (1 - p_{\min})^{KT}$ (Theorem A.9, Appendix A). Additional results (PPO dynamics, GRPO gradient vanishing, reference staleness, optimal adaptive reference, and diversity selection) are in Appendix A.

Table 1 organizes the landscape along the three axes, revealing two families: *margin-driven* methods (DPO, IPO, SimPO, KTO) and *reward-driven* methods (PPO, GRPO). PPO bridges both: its KL penalty induces linear margin dynamics like IPO (Proposition A.7). Table 2 maps each theoretical result to its experimental test. Having identified the failure modes along each axis, we derive a loss family that addresses all three simultaneously.

Method	Axis 1: Repulsion	Axis 2: Weighting	Axis 3: Reference
SFT / RFT / STaR	Implicit only	—	—
PPO / RLHF	Targeted	Linear (non-sat.)	Frozen π_{ref} (KL)
GRPO	Targeted (imbal.)	Advantage-weighted	None
DPO	Targeted	Sigmoid (saturating)	Frozen π_{ref}
IPO	Targeted	Linear (non-sat.)	Frozen π_{ref}
SimPO	Targeted	Sigmoid (saturating)	None
KTO	Targeted	Bell-shaped (robust)	Implicit (KL term)
TPO-H	Targeted	Huber (bounded)	Reference-free
TPO-S	Targeted	Sigmoid + smoothing	Reference-free

Table 1: Three-axis taxonomy of LLM training methods.

2.1 From Theory to Loss Design

The framework is not only explanatory but constructive: specify desired dynamical properties of $w(\Delta)$, then identify losses that produce them. We require five properties: **D1** a finite attractor ($w(\Delta^*) = 0$); **D2** convergence rate bounded away from zero ($|w|$ bounded below by a positive constant far from Δ^*); **D3** stability ($w'(\Delta^*) < 0$); **D4** bounded gradient ($|w| \leq c$), preventing the concentration of Theorem 2.5; and **D5** asymmetry (stronger push below the target). SimPO's $w(\Delta) = \beta\sigma(-\Delta)$ is strictly positive (**D1** fails); IPO's $w(\Delta) = \tau - 2\tau^2\Delta$ satisfies **D1**–**D3** but grows without bound (**D4** fails).

We present two complementary instantiations; the "TPO" prefix denotes *targeted preference optimization*: both losses are derived from desiderata D1–D5 but via different weighting functions (Huber vs. label-smoothed sigmoid), not a shared functional form. ***TPO-H*** (Huber variant) uses the clamped linear weighting $w^*(\Delta) = \text{clamp}(\Delta^* - \Delta, -c, c)$, satisfying D1–D4 (but not D5). Integrating yields the Huber loss $\mathcal{L}_{\text{TPO-H}}$ (Appendix B). ***TPO-S*** (label-smoothed sigmoid) exploits Theorem 2.4 directly: adding $\varepsilon > 0$ to SimPO yields $w(\Delta) = (1-\varepsilon)\sigma(-\Delta) - \varepsilon\sigma(\Delta)$, with a unique zero at $\Delta^* = \beta^{-1}\log((1-\varepsilon)/\varepsilon)$:

$$\mathcal{L}_{\text{TPO-S}} = (1-\varepsilon)\big(-\log\sigma(\delta)\big) + \varepsilon\big(-\log\sigma(-\delta)\big), \quad \delta = \beta(s_w - s_l - \mu), \tag{3}$$

where s_w, s_l are length-normalized log-probabilities (as in SimPO), with β=0.1, μ=0.5, ε=0.05. (We use μ for SimPO's target-margin parameter to avoid conflict with the asymmetric weighting $\gamma \in [0,1]$ in DPO_γ.) TPO-S satisfies all five desiderata and preserves SimPO's sigmoid curriculum while the label-smoothing term creates a restoring force at Δ^* (Figure 2). The desiderata identify *when* and *why* these loss designs work, and predict the exact attractor location as a closed-form function of ε. While cDPO [Mitchell, 2023] also uses label smoothing, it retains a frozen reference and treats ε as a tuning knob; TPO-S is reference-free, includes a target margin $\mu > 0$, and derives ε from the phase transition (Theorem 2.4), giving a closed-form relationship between ε and the attractor location. The empirical gap between TPO-S and cDPO (2–6 points on IFEval, Table 3) reflects the reference axis (Axis 3), not ε itself. Llama 2's preference margin loss [Touvron et al., 2023] also injects a margin term into the DPO sigmoid, using a per-example offset $m(s_w, s_l)$ derived from Likert-scale annotations. However, such a shift leaves $w(\Delta) = \beta\sigma(-\beta(\Delta - m/\beta))$ strictly positive, so the margin still grows without bound, and only the initial condition is shifted. TPO-S's ε changes the *sign* of w at Δ^*, a qualitatively different modification that produces the finite attractor.

What the theory adds over grid search. A practitioner grid-searching ε over SimPO would eventually find a good value. The theory predicts a qualitative phase transition: *any* $\varepsilon > 0$ creates a finite attractor while ε=0 does not, with the exact attractor location $\Delta^* = \beta^{-1}\log((1-\varepsilon)/\varepsilon)$ as a closed-form function of ε. Practitioners can directly control the attractor: ε=0.01 gives $\Delta^* \approx 46$ (weaker regularization), ε=0.1 gives $\Delta^* \approx 22$ (stronger). We set ε=0.05 ($\Delta^* \approx 29$) as a moderate default without per-model tuning. A sweep over $\varepsilon \in \{0.01, \ldots, 0.2\}$ confirms the predicted monotonic ordering (Appendix E.3).

The desiderata paradox. Both TPO-H and TPO-S satisfy D1–D4; TPO-S additionally satisfies D5. The performance gap is small (0.4–0.6 points on IFEval), confirming that the finite-attractor property (D1–D3) is the primary driver. TPO-S's sigmoid provides exponentially small but nonzero gradients everywhere, giving it a slight edge over TPO-H's hard clipping. Full derivation in Appendix B.

Result	Prediction	Experiment	Confirmed?
Main predictions (closed-form margin ODEs)			
Thm. 2.3	DPO $\Delta \sim \log t$; IPO $\Delta \to 1/(2\tau)$	Fig. 3	✓ functional form
Thm. 2.4	Noise: DPO phase transition, IPO rate-preserving	Fig. 4	✓ 8% error
Thm. 2.2	$\gamma = 0$ vs $\gamma > 0$ sharp separation	Fig. 5	✓ qualitative
Supporting predictions (Appendix)			
Thm. 2.5	High $\gamma \to$ collapse	Fig. 5	✓ non-monotonic
Thm. A.9	Exponential gap decay	App. C	✓ quantitative
Prop. A.3	GRPO fails at low p	App. C	✓ qualitative
Prop. A.4	Frozen ref constrains drift	App. C	✓ 82× gap

Table 2: Theory-to-experiment mapping. Main predictions test the closed-form margin ODEs; supporting predictions are in the appendix. A single calibration parameter β_{eff} maps the two-completion theory to full-scale training.

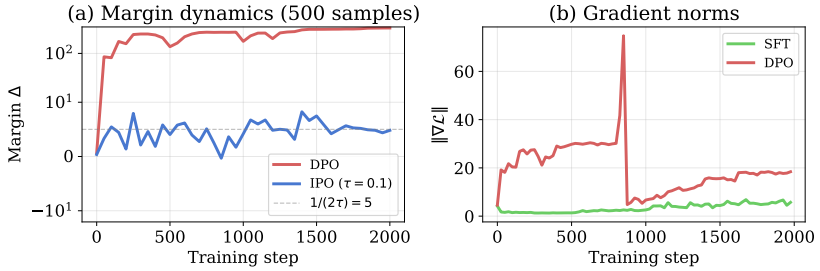


Figure 3: Margin and gradient dynamics over 2000 training steps (Theorem 2.3). Left: DPO's margin grows without bound, consistent with $u + \log u = at + c$ ($R^2 = 0.85$; the remaining variance reflects discrete-time effects and cross-example interference not captured by the isolated ODE). Right: DPO's gradient norm is $10\times$ SFT's, confirming that targeted repulsion sustains $\Omega(\pi)$ gradient pressure (Definition 2.1).

The four theorems and the fitted β_{eff} approximation yield concrete, falsifiable predictions: specific functional forms, a noise-induced phase transition, and a concentration-driven collapse pattern, all testable with a single fitted calibration parameter per scale.

3 Experiments

We validate the framework's predictions and test TPO across three models, two data regimes (on-policy and off-policy), and both preference accuracy and generation quality. We evaluate on MATH [Hendrycks et al., 2021] (transition regime: 37% base solve rate), GSM8K [Cobbe et al., 2021] (high-coverage: >70%), and AIME 2024 [Mathematical Association of America, 2024] (low-coverage: 0%), spanning the coverage spectrum of Theorem A.9. Primary model: Qwen3-8B with LoRA [Hu et al., 2022] (r=64). Baselines: SFT, RFT, DPO (β=0.1), IPO (τ=0.1), SimPO, KTO, cDPO, ORPO, GRPO; all use default hyperparameters from their respective papers unless noted. Dynamical predictions track margin Δ over training steps; we fit a single β_{eff} per data scale and use it across all noise conditions without refitting.

3.1 Prospective Predictions

We test three predictions derived from the margin ODEs. The first concerns the functional form of margin trajectories (Theorem 2.3): DPO's sigmoid weighting predicts unbounded growth as $O(\log t)$, while IPO's linear weighting predicts exponential convergence to $1/(2\tau)$. Figure 3 confirms this separation: DPO reaches 1540 at step 2000 while IPO oscillates near the predicted fixed point of 5. The separation replicates on full MATH (DPO 465 vs. IPO ∼0), GSM8K (257 vs. 8.5), and MBPP [Austin et al., 2021] (160 vs. 1.1). A τ-sweep validates the IPO fixed point quantitatively across $\tau \in \{0.01, \ldots, 0.2\}$ (Appendix E.2).

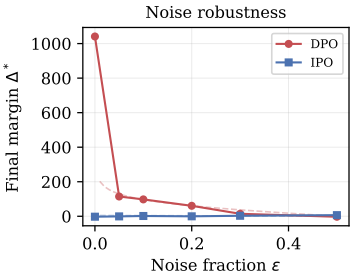


Figure 4: Noise phase transition (Theorem 2.4). DPO's final margin (y-axis) plotted against label-noise fraction ε (x-axis) collapses from unbounded growth at ε=0 to a finite fixed point $\Delta^* = \beta_{\text{eff}}^{-1} \log((1-\varepsilon)/\varepsilon)$ for $\varepsilon > 0$ (dashed: theory, solid: empirical). IPO's fixed point shifts with ε but the convergence rate is preserved.

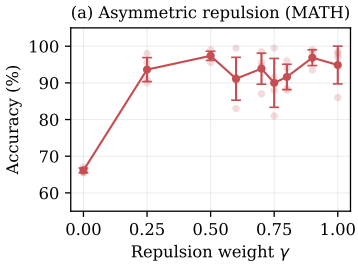


Figure 5: Repulsion strength (Theorem 2.2). γ=0 plateaus at 66%; γ=0.5 achieves 97%; γ=0.7–0.8 shows a collapse valley (Theorem 2.5).

This distinction sharpens under noise, providing the second and most striking test. Theorem 2.4 predicts that label noise ε should bound DPO's margin at $\Delta^* = \beta^{-1} \log \frac{1-\varepsilon}{\varepsilon}$ while leaving IPO's convergence rate $2\tau^2$ unchanged. Figure 4 confirms: DPO's margin decreases monotonically with noise ($1041 \to 114 \to 97 \to 64$ for $\varepsilon \in \{0, 0.05, 0.1, 0.2\}$), while IPO remains stable. We fit $\beta_{\text{eff}} = \beta/4.4$ once from the clean-data trajectory and hold it constant across all noise conditions; consistency across noise levels ($\beta/4.5$ and $\beta/4.4$ from ε=0.1 and 0.2) confirms the log-formula. The ratio β_{eff}/β varies from $\sim 1/4.4$ (500 samples) to $\sim 1/1.9$ (full scale); predicting β_{eff} a priori remains open. The qualitative predictions (functional forms, phase transitions, ordering of attractors) are parameter-free; only the quantitative mapping to specific margin values requires β_{eff}.

The third axis tests repulsion strength (Theorem 2.2): any $\gamma > 0$ should maintain $\Omega(\pi)$ repulsion while γ=0 reverts to $O(\pi^2)$. Figure 5 confirms: γ=0 plateaus at 66% while γ=0.25 reaches near-ceiling held-out accuracy (∼99%). Higher γ triggers gradient concentration (Theorem 2.5), producing a non-monotonic pattern; measured on the 50 tracked hard examples of Appendix E.5 (mean over seeds), γ=0.5 achieves 97% while γ=0.75 degrades to 87% (and γ=0.25 recovers to 91% on this harder tracked set). The pattern is absent on GSM8K (100% for all $\gamma > 0$), confirming collapse requires hard examples. Standard DPO (γ=1.0) shows the widest variance (82–98% over 3 seeds). Cross-model validation on Mistral-7B confirms monotonic degradation with γ (98% $\to$ 88% $\to$ 83% for $\gamma = 0.25, 0.75, 1.0$).

All three predictions replicate on GSM8K and MBPP (Appendix E.5–E.1), confirming that the dynamics are governed by $w(\Delta)$, not by architecture or domain specifics.

3.2 Theory-Predicted Configurations across Model Scales

Having confirmed the framework's descriptive predictions, we test its constructive power: does the theory-predicted configuration match or exceed existing methods? We evaluate on DeepSeek-R1-0528-Qwen3-8B (8B, 9K on-policy pairs), Qwen3.5-9B (9B, 5.4K pairs), and Ministral-3-8B (8B, 2.8K pairs). All methods use LoRA (r=64); TPO-S uses β=0.1, μ=0.5, ε=0.05 (predicted by Theorem 2.4); TPO-H uses β=2.0, μ=1.0. No hyperparameter tuning was performed on ε; the theory predicts that *any* $\varepsilon > 0$ creates a finite attractor. We report IFEval (objective instruction-following) and UltraFeedback preference accuracy.

The theory-derived TPO-S attains the highest mean IFEval score among the compared methods across all three models (Table 3), with the effect significant on MI-8B (+2.6 over SimPO, p=0.013) and within noise on DS-8B (p=0.36); we read this as confirming the predicted attractor is a competitive operating point, not as a state-of-the-art claim. Method ranking aligns with the taxonomy: reference-free methods with finite attractors (TPO-S, TPO-H, SimPO) occupy the top three positions on every model, reference-dependent methods form a middle tier, and ORPO [Hong et al., 2024] falls far

[2] p-values (TPO-S vs. SimPO, Welch's t-test, 5 seeds): MI p=0.013, QW p=0.047, DS p=0.36.

	DS-R1-8B	QW3.5-9B	MI-8B
TPO-S	**66.8 ± 0.3**	**61.4 ± 0.4**	**76.5 ± 0.3**
TPO-H	66.4 ± 0.3	60.8 ± 0.3	*76.1 ± 0.9*
SimPO	*66.5 ± 0.6*	*60.9 ± 0.2*	73.9 ± 1.4
DPO	65.1 ± 0.2	60.6 ± 0.8	71.9 ± 0.6
KTO	64.2 ± 0.8	60.4 ± 0.5	72.2 ± 1.0
cDPO	64.1 ± 0.7	59.8 ± 0.4	70.6 ± 0.5
IPO	64.2 ± 0.7	59.7 ± 0.6	70.3 ± 1.4
ORPO	52.6 ± 1.7	51.2 ± 7.5	49.5 ± 6.2
SFT	63.1	59.8	70.3

Table 3: IFEval instruction-following accuracy (%) on UltraFeedback-trained models (mean ± std, 5 seeds). IFEval is an objective benchmark with no judge model. The theory-derived TPO-S attains the highest mean among the compared methods on all three, confirming the predicted attractor is a competitive operating point. TPO-S adds negligible wall-clock overhead over SimPO: a single extra $\log \sigma(-\delta)$ evaluation per example (the ε-weighted label-smoothing term in Eq. 3).[2]

	On-policy (Math)			Off-policy (UltraFeedback)		
	DS-8B	QW-9B	MI-8B	DS-8B	QW-9B	MI-8B
TPO-S	**73.6**±.5	**72.0**±1.9	59.4±1.1	**55.2**±1.9	**59.4**±1.9	*58.2*±2.2
SimPO	*72.4*±1.9	*69.4*±2.4	58.6±6.2	*51.2*±.8	*54.6*±2.4	**60.2**±.8
TPO-H	71.4±2.6	68.0±1.9	**59.8**±1.9	47.8±.8	53.4±4.2	51.2±1.1
KTO	50.2±1.3	64.2±.8	*59.6*±1.1	46.8±1.1	52.0±2.0	52.8±1.5
DPO	47.0±2.0	61.0±2.3	55.6±8.8	45.4±.5	47.0±4.8	55.2±2.8
cDPO	46.6±1.5	59.0±.7	57.2±1.5	45.2±.4	45.6±.9	46.0±1.2
IPO	43.8±.4	55.4±1.5	50.8±.4	43.4±.5	42.8±1.3	45.2±.8
ORPO	42.5±.0	53.8±3.6	46.5±.0	42.4±.8	41.8±.4	43.0±.0

Table 4: Preference accuracy (%, mean ± std, 5 seeds) on on-policy math data and off-policy UltraFeedback [Cui et al., 2024] (60K pairs). TPO-S leads or ties on most settings. The reference-free vs. reference-dependent gap exceeds 24 points on DS-R1-8B.

behind. GRPO and PPO are online RL methods analyzed theoretically (Proposition A.3); we exclude them from Table 3 to keep the comparison within the offline-pairwise setting.

Ablation: reference vs. attractor. Table 3 implicitly contains a 2×2 ablation over reference model (present/absent) and finite attractor ($\varepsilon>0$ / $\varepsilon=0$), with DPO, cDPO, SimPO, and TPO-S filling the four cells. Removing the reference contributes $+1.2$ points on average (SimPO $-$ DPO: $+1.4$, $+0.3$, $+2.0$), while adding the attractor on top of a reference-free loss contributes $+1.1$ (TPO-S $-$ SimPO: $+0.3$, $+0.5$, $+2.6$). Adding the attractor to a reference-based loss *hurts*: cDPO $-$ DPO averages -1.0, negative on every model—when the frozen reference already bounds margins, additional clipping reduces useful gradient signal. The two axes are complementary: TPO-S combines both, and neither alone matches its gains on MI-8B ($+4.6$ over DPO).

Table 4 shows two robust patterns and one suggestive one. Robustly: TPO-S leads or ties on most settings, and reference-free methods consistently outscore reference-dependent ones (gap exceeding 24 points on DS-8B on-policy, Proposition A.4). Suggestively, and within the seed variance of this table, TPO-S retains its edge off-policy on DS-8B and QW-9B, consistent with the finite attractor providing regularization under distribution mismatch; we do not over-read the per-setting margins here given the seed-to-seed spread. Method rankings are preserved under the standard SFT $\rightarrow$ preference pipeline (Appendix D).

Additional experiments (method rankings, diversity/coverage, n-completion validation, ablations) are in Appendix C. On mathematical reasoning, TPO reaches 98.7% on MATH vs. 93.8% for DPO (Appendix C.6), with the gain concentrated in the transition regime (Theorem 2.5); on GSM8K both reach 100%; on AIME both score $\sim$12% ($p_t \approx 0$, Theorem A.9).

Label-smoothing sensitivity. Theorem 2.4 predicts that any $\varepsilon > 0$ creates a finite attractor at $\Delta^*=\beta^{-1}\log((1-\varepsilon)/\varepsilon)$, with higher ε yielding a lower equilibrium. A sweep on DS-R1-8B (UltraFeedback, 60K pairs) confirms: final margins decrease monotonically—$1605 \rightarrow 1264 \rightarrow 832 \rightarrow$

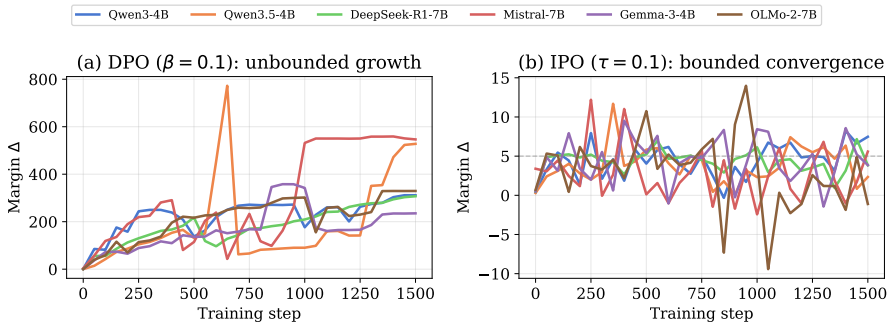


Figure 6: Margin dynamics replicated across five model families (Qwen, Gemma, OLMo, Mistral, DeepSeek). DPO's margin grows without bound and IPO's converges near the predicted fixed point $1/(2\tau) = 5$ (dashed line) on every architecture. Only the growth rate varies across families.

$455 \to 341$ for $\varepsilon \in \{0.01, 0.03, 0.05, 0.1, 0.2\}$—while SimPO ($\varepsilon$=0) diverges. Observed margins are 25–35$\times$ the closed-form Δ^* (e.g., 832 vs. 29 at ε=0.05), reflecting β_{eff} renormalization (Remark 2.6). Downstream accuracy is flat across ε (49–52%), confirming that any $\varepsilon > 0$ suffices (Appendix E.3).

Cross-Architecture Universality The theory operates on the probability simplex, shared by all autoregressive models with softmax outputs, so its predictions should be architecture-agnostic. We replicate Prediction 1 on six additional models spanning five families. In every case, DPO's margin grows to 204–629 while IPO's remains bounded near $1/(2\tau) = 5$ (Figure 6). Growth rates vary 3$\times$ across families, reflecting differences in β_{eff}, but the functional form is invariant: logarithmic for DPO, exponential convergence for IPO, despite different tokenizers, vocabularies, and parameter counts.

4 Related Work

Alignment methods and unified frameworks. We analyze RLHF/PPO [Ouyang et al., 2022, Schulman et al., 2017], DPO [Rafailov et al., 2023], IPO [Gheshlaghi Azar et al., 2024], SimPO [Meng et al., 2024], KTO [Ethayarajh et al., 2024], GRPO [Shao et al., 2024], and others as instances of a single dynamical mechanism. Deng et al. [2026] introduce the F^+/F^- decomposition for image generation on continuous velocity fields; we borrow this decomposition but work on a discrete simplex with 1D margin ODEs. Several frameworks characterize *what* each method optimizes—f-PO [Han et al., 2025] via f-divergences, RPO [Sun et al., 2025] via a common objective, DCR [Ruiz Luyten and van der Schaar, 2025] via gradient flow—while we characterize *how* the optimization unfolds. Llama 2's preference margin loss [Touvron et al., 2023] shifts the DPO sigmoid by a per-example offset derived from Likert annotations; in our taxonomy this leaves $w(\Delta)$ strictly positive (no attractor). The noise phase transition of Theorem 2.4 exemplifies the difference: f-PO can show that DPO and IPO minimize different divergences, but cannot predict that DPO's margin grows logarithmically while IPO's converges exponentially, or that noise creates a phase transition in one but not the other.

Failure modes and iterative training. Concurrent work identifies individual DPO pathologies—gradient imbalance [Yan et al., 2025, Ma et al., 2025], likelihood displacement [Razin et al., 2024], noise fragility [Chowdhury et al., 2024], GRPO zero-gradient collapse [Zhang et al., 2025]—which our framework unifies as instances of a single cross-method pattern governed by the three axes. DAPO [Yu et al., 2025] removes the reference model entirely (reference-free, bounded convergence in our taxonomy). Iterative methods [Yuan et al., 2024, Zelikman et al., 2022, Guo et al., 2024] refresh the training distribution; our coverage theorem (Theorem A.9) formalizes when they succeed ($p_t > 0$) and stall ($p_t \to 0$).

5 Conclusion

We showed that major pairwise LLM alignment methods are instances of a single dynamical mechanism: a vector field on the probability simplex whose margin ODE admits closed-form solutions. The three main predictions (margin trajectories, noise-induced phase transitions, and repulsion strength) are confirmed across five model families. We validated the framework's constructive prediction that *any* $\varepsilon > 0$ label smoothing creates a finite attractor for SimPO across three models; the theory-derived TPO-S (ε=0.05) is a competitive operating point, attaining the highest mean IFEval among the compared methods on all three, which confirms the predicted attractor is usable rather than establishing a new state-of-the-art loss.

Limitations. The two-completion simplex is the central modeling choice: qualitative predictions hold at $n \in \{2, 4, 8, 16\}$ (Table 10), but quantitative accuracy requires β_{eff}, fitted per experiment (2.3× variation across scales). The margin is a sufficient state variable only in the leakage-free / independent-gradient regime: when the chosen and rejected gradients are strongly aligned ($\langle g_w, g_l \rangle > \|g_w\|^2$, with $g = \nabla_\theta \log \pi$), mass escaping to unsampled completions can decouple the margin from $\pi(y_w)$ (likelihood displacement, Razin et al., 2024), so the margin captures relative preference dynamics and predicting downstream quality from it alone is out of scope—consistent with the weak margin–accuracy correlation ($\rho \approx 0.13$) on AIME. TPO-S's gain over SimPO is modest (1.2 pts on DS 8B); on MI 8B KTO leads on chat. The γ-sweep collapse valley shows high seed variance (82–98%). Experiments use LoRA (r=64) and Adam; Adam acts as a time reparameterization absorbed into β_{eff} (Appendix A.7), and the residual R^2=0.85 (Figure 3) reflects discrete-time and cross-example effects.

References

Kavosh Asadi, Xingzi Xu, Julien Han, Ege Beyazit, Idan Pipano, Dominique Perrault-Joncas, Shoham Sabach, Mohammad Ghavamzadeh, and Karim Bouyarmane. C2-DPO: Constrained controlled direct preference optimization. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=7h5Ho9t5NL.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. Provably robust DPO: Aligning language models with noisy feedback. In *International Conference on Machine Learning*, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems, 2021.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with scaled ai feedback, 2024. URL https://arxiv.org/abs/2310.01377.

Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting, 2026.

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

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling, 2025.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 4447–4455. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, Johan Ferret, and Mathieu Blondel. Direct language model alignment from online ai feedback, 2024. URL https://arxiv.org/abs/2402.04792.

Jiaqi Han, Mingjian Jiang, Yuxuan Song, Stefano Ermon, and Minkai Xu. f-po: Generalizing preference optimization with f-divergence minimization. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 1144–1152. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/han25a.html.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Advances in Neural Information Processing Systems*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Qinwei Ma, Jingzhe Shi, Can Jin, Jenq-Neng Hwang, Serge Belongie, and Lei Li. Gradient imbalance in direct preference optimization. *arXiv preprint arXiv:2502.20847*, 2025.

Mathematical Association of America. American Invitational Mathematics Examination (AIME), 2024. URL https://artofproblemsolving.com/wiki/index.php/2024_AIME_I.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=3Tzcot1LKb.

Eric Mitchell. A note on DPO with noisy preferences and relationship to IPO. https://ericmitchell.ai/cdpo.pdf, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. Unintentional unalignment: Likelihood displacement in direct preference optimization. *Advances in Neural Information Processing Systems*, 2024.

Max Ruiz Luyten and Mihaela van der Schaar. The reasoning-creativity trade-off: Toward creativity-driven problem solving. *arXiv preprint arXiv:2601.00747*, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Shengyang Sun, Yian Zhang, Alexander Bukharin, David Mosallanezhad, Jiaqi Zeng, Soumye Singhal, Gerald Shen, Adithya Renduchintala, Tugrul Konuk, Yi Dong, et al. Reward-aware preference optimization: A unified mathematical framework for model alignment. *arXiv preprint arXiv:2502.00203*, 2025.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL `https://arxiv.org/abs/2307.09288`.

Yuzi Yan, Yibo Miao, Jialian Li, Yipin Zhang, Jian Xie, Zhijie Deng, and Dong Yan. 3D-properties: Identifying challenges in DPO and charting a path forward. In *International Conference on Learning Representations*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models, 2024.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, 2022.

Zhi Zhang, Zhen Han, Costas Mavromatis, Qi Zhu, Yunyi Zhang, Sheng Guan, Dingmin Wang, Xiong Zhou, Shuai Wang, Soji Adeshina, et al. Train less, learn more: Adaptive efficient rollout optimization for group-based reinforcement learning. *arXiv preprint arXiv:2602.14338*, 2025.

A Proofs

Remark A.1 (Extension beyond two completions)**.** *The two-completion analysis is exact at $n = 2$ (no leakage by construction). As $n \to \infty$ with all completions labeled, repelling every incorrect completion concentrates mass on correct ones, and the analysis again becomes tight, though this limit is unrealistic in practice. The approximation error is largest at intermediate n, where repelling a subset of incorrect completions redistributes mass to unsampled ones, the displacement phenomenon of Asadi et al. [2026]. This makes the two-completion simplex a* conservative *approximation: it predicts the advantage of targeted repulsion in the leakage-free regime, while the actual advantage with $n > 2$ is strictly larger. We validate empirically that the qualitative predictions (unbounded DPO margins, bounded IPO margins) hold at $n \in \{2, 4, 8, 16\}$ (Appendix C.8), though quantitative accuracy at $n > 2$ with the same β_{eff} remains unvalidated. Diversity-aware selection (Proposition A.6) mitigates leakage by covering the incorrect-completion space, consistent with the +29% GRPO gain at $K = 4$.*

Notation and conventions. All results are stated on a two-completion simplex $\{y_w, y_l\}$ under softmax parameterization $\pi(y_i) = e^{\theta_i}/(e^{\theta_w} + e^{\theta_l})$. We assume gradient flow (continuous time) with full-batch gradients; SGD noise from mini-batching is not analyzed but does not affect the qualitative predictions.

Remark A.2 (Connection to generative modeling)**.** *The vector field F on the probability simplex is the same mathematical object studied in continuous normalizing flows [Lipman et al., 2023] and flow matching [Geng et al., 2025]. In those settings, a learned velocity field transports a source distribution to a target along a probability path; here, the loss function* induces *the velocity field on the margin coordinate $\Delta = \log(\pi(y_w)/\pi(y_l))$, yielding the 1D autonomous ODEs we solve in closed form. The noise fixed point of Theorem 2.4–where the forward and reverse "flows" balance–is the discrete analog of score equilibrium in diffusion models [Ho et al., 2020, Song et al., 2021]: label-smoothed TPO (Section 2.1) exploits this by engineering a stable equilibrium at a prescribed margin, converting the noise phase transition from a failure mode into a design principle.*

For IPO, we analyze the reference-free variant $\mathcal{L} = (\tau\Delta_\theta - 1/2)^2$ where $\Delta_\theta = \log(\pi(y_w)/\pi(y_l))$; with a frozen reference, the dynamics are identical up to a constant shift Δ_{ref} in the fixed point. For DPO, the full margin is $\tilde{\Delta} = \Delta_\theta - \Delta_{\text{ref}}$ where $\Delta_{\text{ref}} = \log(\pi_{\text{ref}}(y_w)/\pi_{\text{ref}}(y_l))$ is constant under a frozen reference; the ODE $d\tilde{\Delta}/dt = \beta\sigma(-\beta\tilde{\Delta})$ is autonomous in $\tilde{\Delta}$, so the reference shifts the initial condition but does not change the functional form of the solution. The key identity is the softmax Jacobian:

$$\dot{\pi}(y_l) = \pi(y_l)(1 - \pi(y_l))\left(\dot{\theta}_l - \dot{\theta}_w\right). \tag{4}$$

For any loss $\mathcal{L}$, gradient descent gives $\dot{\theta}_y = -\eta\frac{\partial\mathcal{L}}{\partial\theta_y}$. Since DPO and IPO losses depend on $\theta_w - \theta_l$, both logits receive equal and opposite gradients, yielding $\dot{\theta}_w - \dot{\theta}_l = 2\eta \cdot w(\Delta)$ where $w(\Delta)$ is the method-specific weighting. We absorb 2η into the time scale, writing $\frac{d\Delta}{dt} = w(\Delta)$.

A.1 Axis 1: Convergence Rate Separation (Theorem 2.2)

Proof sketch. Compute the softmax gradient for each loss, substitute into the Jacobian (4), and compare the rate at which $|\dot{\pi}(y_l)|$ vanishes as $\pi(y_l) \to 0$.

Proof. Part (i): SFT. The cross-entropy loss $\mathcal{L} = -\log\pi(y_w)$ gives gradients $\dot{\theta}_y = \eta(p^*(y) - \pi(y))$. With $p^*(y_w) = 1$, $p^*(y_l) = 0$:

$$\dot{\theta}_w = \eta\pi(y_l), \qquad \dot{\theta}_l = -\eta\pi(y_l). \tag{5}$$

Thus $\dot{\theta}_l - \dot{\theta}_w = -2\eta\pi(y_l)$. By (4):

$$\dot{\pi}(y_l) = -2\eta\pi(y_l)^2(1 - \pi(y_l)) = O(\pi(y_l)^2) \text{ as } \pi(y_l) \to 0. \tag{6}$$

Part (ii): DPO. The loss $\mathcal{L} = -\log\sigma(\beta\tilde{\Delta})$ with $\tilde{\Delta} = (\theta_w - \theta_l) - \log(\pi_{\text{ref}}(y_w)/\pi_{\text{ref}}(y_l))$ gives:

$$\frac{\partial\mathcal{L}}{\partial\theta_w} = -\beta\sigma(-\beta\tilde{\Delta}), \qquad \frac{\partial\mathcal{L}}{\partial\theta_l} = \beta\sigma(-\beta\tilde{\Delta}). \tag{7}$$

So $\dot{\theta}_l - \dot{\theta}_w = -2\eta\beta\sigma(-\beta\Delta)$, and by (4):

$$\dot{\pi}(y_l) = -2\eta\beta\sigma(-\beta\Delta)\pi(y_l)(1 - \pi(y_l)). \tag{8}$$

When $\pi(y_l) \geq \delta > 0$, the margin Δ is bounded above, so $\sigma(-\beta\Delta) \geq c(\beta, \pi_{\text{ref}}, \delta) > 0$ where c depends on β, π_{ref}, and δ through the bounded margin. Therefore $|\dot{\pi}(y_l)| = \Omega(\pi(y_l))$: sustained. □

A.2 Axis 2: Margin Dynamics (Theorems 2.3–2.4)

Theorem 2.3 (without noise). *Proof sketch.* Solve the 1D ODE $d\Delta/dt = w(\Delta)$ for each weighting function. DPO's sigmoid yields a nonlinear ODE solvable via substitution; IPO's linear weighting gives a standard linear ODE.

Proof. (i) DPO. The ODE $\frac{d\Delta}{dt} = \frac{\beta}{1+e^{\beta\Delta}}$ becomes, under $u = e^{\beta\Delta}$:

$$\frac{du}{dt} = \frac{\beta^2 u}{1+u}. \tag{9}$$

Separating variables and integrating: $\log u + u = \beta^2 t + C$ where $C = \beta\Delta(0) + e^{\beta\Delta(0)}$. For large t, u dominates $\log u$, so $u(t) \sim \beta^2 t$ and $\Delta(t) \sim \frac{1}{\beta}\log(\beta^2 t)$: unbounded but logarithmic. (This asymptotic is valid for $\beta^2 t \gg 1$; at early times the implicit solution should be used directly.)

(ii) IPO. The ODE $\frac{d\Delta}{dt} = \tau - 2\tau^2\Delta$ is linear with fixed point $\Delta^* = 1/(2\tau)$. Writing $\delta = \Delta - \Delta^*$:

$$\frac{d\delta}{dt} = -2\tau^2\delta \implies \Delta(t) = \frac{1}{2\tau} + \left(\Delta(0) - \frac{1}{2\tau}\right) e^{-2\tau^2 t}. \tag{10}$$

Exponential convergence to $1/(2\tau)$ with rate $2\tau^2$. □

Theorem 2.4 (with label noise). *Proof sketch.* Average the ODE over correct and flipped pairs, treating the noise as i.i.d. across examples. This is a mean-field approximation: we replace the per-example noise realization with its expectation, which becomes exact in the large-sample limit (law of large numbers applied to the batch gradient). For finite batches, the approximation introduces fluctuations of order $O(1/\sqrt{B})$ where B is the batch size. For DPO, the opposing sigmoid terms create a stable fixed point. For IPO, the linear structure preserves the convergence rate.

Proof. (i) DPO. The averaged ODE is:

$$\frac{d\Delta}{dt} = (1-\varepsilon)\beta\sigma(-\beta\Delta) - \varepsilon\beta\sigma(\beta\Delta). \tag{11}$$

Setting to zero and using $\sigma(-x) = e^{-x}\sigma(x)$: $(1-\varepsilon)e^{-\beta\Delta^*} = \varepsilon$, giving $\Delta^* = \frac{1}{\beta}\log\frac{1-\varepsilon}{\varepsilon}$. Stability: differentiating the full averaged field, $f'(\Delta) = -(1-\varepsilon)\beta^2\sigma(\beta\Delta)\sigma(-\beta\Delta) - \varepsilon\beta^2\sigma(\beta\Delta)\sigma(-\beta\Delta) = -\beta^2\sigma(\beta\Delta)\sigma(-\beta\Delta) < 0$ everywhere.

(ii) IPO. A flipped pair has effective margin $-\Delta$, giving loss $(\tau\Delta + 1/2)^2$. Averaging:

$$\frac{d\Delta}{dt} = -2\tau^2\Delta + \tau(1-2\varepsilon). \tag{12}$$

Fixed point $\Delta^* = (1-2\varepsilon)/(2\tau)$, rate $2\tau^2$ unchanged. Noise shifts the target but not the speed. □

Proposition A.3 (GRPO gradient vanishes with sparse rewards)**.** *For GRPO with binary rewards, group size K, success probability p: $\mathbb{E}[\|\nabla\mathcal{L}\|^2] \propto p(1-p)\cdot G^2/K$, maximized at $p = 1/2$ and vanishing as $p \to 0$ or $p \to 1$. The zero-gradient probability is $(1-p)^K + p^K$.*

Proof. The advantage $A_i = R_i - \bar{R}$ gives $\sum A_i^2 = k(K-k)/K$. Since $\mathbb{E}[k(K-k)] = Kp(1-p)(K-1)$:

$$\mathbb{E}\left[\frac{1}{K^2}\sum A_i^2\right] = \frac{(K-1)p(1-p)}{K^2} \sim \frac{p(1-p)}{K}. \tag{13}$$

Maximized at $p = 1/2$, vanishing as $p \to 0$ or 1. Zero-gradient probability: $(1-p)^K + p^K$. □

A.3 Axis 3: Reference Strategy (Proposition A.4, Proposition A.5)

Proposition A.4 (Gradient decay under staleness)**.** *For any method with frozen π_{ref} whose gradient includes $\sigma(-\beta\Delta)$, the per-example gradient norm is $\|\nabla\mathcal{L}\| = \beta\cdot\sigma(-\beta\Delta)\cdot\|\nabla\log\pi(y_w) - \nabla\log\pi(y_l)\|$. As training progresses, Δ grows on correctly-learned examples, driving $\sigma(-\beta\Delta)\to 0$. Reference-free methods (SimPO, ORPO) avoid this decay but still experience sigmoid-based gradient decay from growing margins; the staleness proposition isolates the* additional *decay caused by the frozen reference.*

Proof. The DPO gradient is $\nabla_\theta \mathcal{L} = -\beta\sigma(-\beta\Delta)[\nabla_\theta \log \pi_\theta(y_w|x) - \nabla_\theta \log \pi_\theta(y_l|x)]$. Taking norms:

$$\|\nabla_\theta \mathcal{L}\| = \beta \cdot \sigma(-\beta\Delta) \cdot \|\nabla_\theta \log \pi_\theta(y_w|x) - \nabla_\theta \log \pi_\theta(y_l|x)\|. \tag{14}$$

Since $\Delta = \Delta_\theta - \Delta_{\text{ref}}$ and Δ_θ grows on correctly-learned examples (as established by Theorem 2.3 for DPO), $\sigma(-\beta\Delta) \to 0$. The log-probability gradient term is bounded for finite-parameter models, so the sigmoid decay dominates. For SimPO, $\Delta = \Delta_\theta - \mu$ with fixed μ, so no *reference-staleness* decay occurs. However, SimPO still experiences sigmoid-based gradient decay from growing margins ($\sigma(-\beta\Delta_\theta) \to 0$ as Δ_θ grows). The distinction is that DPO has *both* margin-based decay and reference-staleness decay, while SimPO has only the former. □

Proposition A.5 (Optimal adaptive reference)**.** *The reference minimizing $\lambda_1 \mathrm{KL}(\pi_{ref}\|\pi_\theta) + \lambda_2 \mathrm{KL}(\pi_{ref}\|\pi_0)$ is $\pi^*_{ref}(y) \propto \pi_0(y)^{1-\alpha}\pi_\theta(y)^\alpha$ with $\alpha = \lambda_1/(\lambda_1+\lambda_2)$. At $\alpha = 0$: frozen (DPO). At $\alpha = 1$: self-referencing (SimPO-like).*

Proof. The objective $J(\pi_{\text{ref}}) = \lambda_1 \mathrm{KL}(\pi_{\text{ref}}\|\pi_\theta) + \lambda_2 \mathrm{KL}(\pi_{\text{ref}}\|\pi_0)$ is strictly convex. Setting the Lagrangian derivative to zero:

$$(\lambda_1 + \lambda_2) \log \pi_{\text{ref}}(y) = \lambda_1 \log \pi_\theta(y) + \lambda_2 \log \pi_0(y) + \text{const}. \tag{15}$$

Exponentiating: $\pi^*_{\text{ref}}(y) \propto \pi_\theta(y)^\alpha \cdot \pi_0(y)^{1-\alpha}$ with $\alpha = \lambda_1/(\lambda_1+\lambda_2)$.

Note: We minimize $\mathrm{KL}(\pi_{\text{ref}}\|\pi_\theta)$ (forward KL) rather than $\mathrm{KL}(\pi_\theta\|\pi_{\text{ref}})$ (staleness) because the forward direction yields a closed form. Both vanish at $\pi_{\text{ref}} = \pi_\theta$. □

A.4 Diversity and PPO (Propositions A.6–A.7)

Proposition A.6 (Diversity improves field estimation)**.** *For a weighted estimate $\hat{V} = \sum w_i v_i$ with independent v_i, $\mathbb{E}[v_i] = V$, $\mathrm{Var}(v_i) = \sigma^2 I$, and $\sum w_i = 1$: MSE$(\hat{V}) = \sigma^2/d_{eff}$ where $d_{eff} = (\sum w_i^2)^{-1} \le N$, with equality iff weights are uniform.*

Proof. For $\hat{V} = \sum w_i v_i$ with $\sum w_i = 1$ and independent v_i with $\mathbb{E}[v_i] = V$, $\mathrm{Var}(v_i) = \sigma^2 I$:

$$\mathrm{MSE}(\hat{V}) = \mathrm{Var}(\hat{V}) = \sigma^2 \sum w_i^2 = \sigma^2/d_{\text{eff}}. \tag{16}$$

By Cauchy–Schwarz, $1 = (\sum w_i)^2 \le N \sum w_i^2$, so $d_{\text{eff}} \le N$ with equality iff $w_i = 1/N$. □

Proposition A.7 (PPO-KL margin dynamics)**.** *Under PPO with KL penalty β (not the clipped PPO variant, whose non-smooth objective does not admit a closed-form ODE): $\frac{d\Delta}{dt} = \eta[(R(y_w) - R(y_l)) - \beta(\Delta - \Delta_{ref})]$, a linear ODE with fixed point $\Delta^* = (R(y_w) - R(y_l))/\beta + \Delta_{ref}$ and rate $\eta\beta$.*

Proof. The KL-regularized gradient gives $\dot{\theta}_y = \eta[R(y) - \beta \log(\pi(y)/\pi_{\text{ref}}(y)) + \text{const}]$. For two completions:

$$\frac{d\Delta}{dt} = \eta[(R(y_w) - R(y_l)) - \beta(\Delta - \Delta_{\text{ref}})], \tag{17}$$

a linear ODE with fixed point $\Delta^* = (R(y_w) - R(y_l))/\beta + \Delta_{\text{ref}}$ and rate $\eta\beta$. □

A.5 Coverage Expansion (Theorem A.9)

The following result is a straightforward application of coupon-collector-style reasoning. Its contribution is not the proof technique but the formalization of when iterative alignment methods stall ($p_t \to 0$) vs. succeed ($p_t > 0$), connecting the coverage gap to the margin dynamics framework: methods with targeted repulsion maintain $p_t > 0$ on solvable problems, while SFT-only methods see p_t decay.

Definition A.8 (Coverage)**.** *For reward threshold τ, the coverage of policy π is* $\mathrm{Cov}(\pi, \tau) = \mathbb{P}_x\big[\exists\, y \sim \pi(\cdot|x) : R(x, y) \ge \tau\big]$.

Theorem A.9 (Iterative coverage expansion)**.** *Let $\Delta_0 = 1 - \mathrm{Cov}(\pi_0, \tau)$ be the initial coverage gap. After T iterations with K samples per prompt and discovery probabilities $\{p_t\}$, assuming (a) covered prompts remain covered across iterations (no catastrophic forgetting) and (b) $p_t > 0$ for uncovered prompts:*

$$\mathbb{E}[\mathrm{Cov}(\pi^{(T)}, \tau)] \geq 1 - \Delta_0 \prod_{t=0}^{T-1} (1 - p_t)^K. \tag{18}$$

When $p_t \geq p_{\min} > 0$, coverage gaps decay exponentially: $\Delta_T \leq \Delta_0 (1 - p_{\min})^{KT}$.

Proof sketch. Induction on iterations. At each step, every uncovered prompt has at least K independent chances to discover a solution, reducing the coverage gap by a factor of $(1 - p_t)^K$.

Proof. Define $U_t = \{x : \text{no completion } y \text{ with } R(x, y) \geq \tau \text{ has been found by iteration } t\}$ as the set of uncovered prompts. Let $\Delta_t = \mathbb{P}[x \in U_t]$ denote the coverage gap.

Base case: $\Delta_0 = 1 - \mathrm{Cov}(\pi_0, \tau)$ by definition.

Inductive step: At iteration t, for each prompt $x \in U_t$, we sample K completions from $\pi^{(t)}(\cdot|x)$. Let $p_t = \inf_{x \in U_t} \mathbb{P}_{y \sim \pi^{(t)}}[R(x, y) \geq \tau \mid x]$ be the worst-case discovery probability over uncovered prompts. The probability that *all* K samples fail to produce a τ-good completion is at most $(1 - p_t)^K$ (by independence of samples). Therefore: $\mathbb{P}[x \in U_{t+1} \mid x \in U_t] \leq (1 - p_t)^K$. Taking expectations over prompts: $\Delta_{t+1} = \mathbb{P}[x \in U_{t+1}] \leq \mathbb{P}[x \in U_t] \cdot (1 - p_t)^K = \Delta_t \cdot (1 - p_t)^K$.

Telescoping: By induction, $\Delta_T \leq \Delta_0 \prod_{t=0}^{T-1} (1 - p_t)^K$, so $\mathrm{Cov}(\pi^{(T)}, \tau) = 1 - \Delta_T \geq 1 - \Delta_0 \prod_{t=0}^{T-1} (1 - p_t)^K$.

When $p_t \geq p_{\min} > 0$ for all t, the product is bounded by $(1 - p_{\min})^{KT}$, giving exponential decay of the coverage gap: $\Delta_T \leq \Delta_0 (1 - p_{\min})^{KT}$. □

A.6 Gradient Concentration (Theorem 2.5)

Proof. The asymmetric DPO loss for example i is $\ell_i = -\log \sigma(\beta \Delta_{\gamma,i})$ where $\Delta_{\gamma,i} = \log(\pi_\theta(y_w^i)/\pi_{\mathrm{ref}}(y_w^i)) - \gamma \log(\pi_\theta(y_l^i)/\pi_{\mathrm{ref}}(y_l^i))$. Since π_{ref} is frozen, $\nabla_\theta \Delta_{\gamma,i} = \nabla_\theta \log \pi(y_w^i) - \gamma \nabla_\theta \log \pi(y_l^i) =: w_i$. By the chain rule:

$$\nabla_\theta \mathcal{L} = -\frac{\beta}{N} \sum_{i=1}^{N} \sigma(-\beta \Delta_{\gamma,i}) \cdot w_i. \tag{19}$$

Under gradient flow $\dot{\theta} = -\nabla_\theta \mathcal{L}$, the margin $\Delta_j = \log \pi(y_w^j) - \log \pi(y_l^j)$ has gradient $v_j = \nabla_\theta \Delta_j$, so:

$$\dot{\Delta}_j = v_j^\top \dot{\theta} = \frac{\beta}{N} \sum_{i=1}^{N} \sigma(-\beta \Delta_{\gamma,i}) \cdot v_j^\top w_i. \tag{20}$$

This is exact in continuous time; v_j, w_i, and $\Delta_{\gamma,i}$ all depend on $\theta(t)$.

For $\Delta_{\gamma,j} \gg 1/\beta$: $\sigma(-\beta \Delta_{\gamma,j}) = (1 + e^{\beta \Delta_{\gamma,j}})^{-1} \leq e^{-\beta \Delta_{\gamma,j}}$, exponentially small. For hard examples with $\Delta_{\gamma,i} = O(1/\beta)$, $\sigma(-\beta \Delta_{\gamma,i}) = \Theta(1)$, so these dominate the sum.

For the γ-monotonicity: when $\pi_\theta(y_l^i) < \pi_{\mathrm{ref}}(y_l^i)$ (i.e., the model has successfully decreased the dispreferred probability below the reference, which we term a "learned" example), $\log(\pi_\theta(y_l^i)/\pi_{\mathrm{ref}}(y_l^i)) < 0$, so $\partial \Delta_{\gamma,i}/\partial \gamma = -\log(\pi_\theta(y_l^i)/\pi_{\mathrm{ref}}(y_l^i)) > 0$. Higher γ increases $\Delta_{\gamma,i}$, decreasing $\sigma(-\beta \Delta_{\gamma,i})$ and concentrating the gradient on hard examples faster. Early in training, this condition may not hold for all examples. □

A.7 Effective Coupling Approximation (Remark 2.6)

Proof. From Theorem 2.5, the margin of pair j evolves as:

$$\dot{\Delta}_j = \frac{\beta}{N} \sum_{i=1}^{N} \sigma(-\beta \Delta_i) \cdot v_j^\top w_i. \tag{21}$$

Isolating the self-term ($i = j$) and noting that for $\gamma = 1$, $w_j = v_j$ (as assumed in Remark 2.6; for general γ, the self-term is $v_j^\top w_j \neq \|v_j\|^2$):

$$\dot{\Delta}_j = \frac{\beta}{N}\left[\sigma(-\beta\Delta_j)\|v_j\|^2 + \sum_{i\neq j}\sigma(-\beta\Delta_i)\cdot v_j^\top w_i\right]. \tag{22}$$

Dividing by $\sigma(-\beta\Delta_j)$ to extract the effective coupling:

$$\beta_{\text{eff},j} := \frac{\dot{\Delta}_j}{\sigma(-\beta\Delta_j)} = \frac{\beta}{N}\left[\|v_j\|^2 + \sum_{i\neq j}\frac{\sigma(-\beta\Delta_i)}{\sigma(-\beta\Delta_j)}\cdot v_j^\top w_i\right]. \tag{23}$$

Cross-term analysis. The inner products $v_j^\top w_i$ for $i \neq j$ depend on the parameter-sharing structure. In the LoRA setting with rank r and model dimension d, the gradients $v_j, w_i \in \mathbb{R}^{2rd}$ (two LoRA matrices per layer). **Independence assumption (stated in Remark 2.6):** For distinct prompts with sufficient diversity, the gradient directions are approximately independent, giving $\mathbb{E}[v_j^\top w_i] \approx 0$ and $\mathrm{Var}[v_j^\top w_i] = \sigma_\perp^2$. This assumption is strong: LoRA gradients share a low-rank subspace of dimension $2rd \ll d^2$, and thematically correlated training examples (e.g., multiple math problems of the same type) will have positively correlated gradients. The $O(1/\sqrt{N})$ error scaling below depends entirely on this assumption; when it is violated, the cross-term contribution may not average out, and β_{eff} may deviate from the self-reinforcement approximation by more than the stated error. The sigmoid ratio $\sigma(-\beta\Delta_i)/\sigma(-\beta\Delta_j)$ is $O(1)$ for active examples and exponentially small for well-learned ones. Defining $N_{\text{active}} = |\{i : \sigma(-\beta\Delta_i) > 1/e\}|$, the cross-term sum has mean zero and standard deviation $O(\sigma_\perp\sqrt{N_{\text{active}}})$ under the independence assumption, giving:

$$\beta_{\text{eff},j} = \frac{\beta\|v_j\|^2}{N} \pm O\left(\frac{\beta\sigma_\perp\sqrt{N_{\text{active}}}}{N}\right). \tag{24}$$

Scaling with dataset size. As training progresses, well-learned examples have $\Delta_i \gg 1/\beta$, so $\sigma(-\beta\Delta_i) \to 0$ and they exit the active set. The number of active examples N_{active} depends on the difficulty distribution: for a dataset with fraction f of hard examples (near the decision boundary), $N_{\text{active}} \approx fN$. As N grows with fixed difficulty distribution, $\beta_{\text{eff}} \approx \beta\|v_j\|^2/N$ while the noise term scales as $\beta\sigma_\perp\sqrt{fN}/N = \beta\sigma_\perp\sqrt{f/N}$, which vanishes. This explains why β_{eff}/β approaches a stable value at large N: the self-reinforcement term dominates and the cross-example noise averages out.

Empirical consistency. The observed $\beta_{\text{eff}}/\beta \approx 1/4.4$ at 500 samples and $\approx 1/1.9$ at full scale is consistent with N_{active} decreasing relative to N as the dataset grows: with more data, a larger fraction of examples are well-separated (easy), leaving fewer active examples to dilute the self-reinforcement signal.

Adam as time reparameterization. Our experiments use AdamW, not gradient flow. Adam's update is $\theta_{t+1} = \theta_t - \eta\,\hat{m}_t/(\sqrt{\hat{v}_t} + \epsilon)$, which in continuous-time form is $\dot{\theta} = -P_t\nabla\mathcal{L}$ with $P_t = \mathrm{diag}(1/(\sqrt{\hat{v}_t} + \epsilon))$ a positive-definite preconditioner. Projecting onto the scalar margin, $\dot{\Delta} = v^\top P_t\nabla\mathcal{L} = (v^\top P_t v)\cdot w(\Delta)$, where $v^\top P_t v > 0$ is a state-dependent positive scalar. Because P_t is positive definite, $v^\top P_t v > 0$ is a strictly positive (state-dependent) scalar, so this is a *time reparameterization* of the gradient-flow ODE: multiplying $w(\Delta)$ by a positive factor $c(\Delta) = v^\top P_t v$ leaves its zero set unchanged and preserves the sign of w' at each zero (since $(cw)'(\Delta^*) = c(\Delta^*)\,w'(\Delta^*)$ when $w(\Delta^*) = 0$). Hence, under a positive but possibly *nonuniform* $c(\Delta)$, the following are invariant: the zeros of w (fixed-point locations), the sign of w' at each zero (stability), the sign of $\dot{\Delta}$ (monotone direction), and—since $c > 0$ cannot change the sign of $c\,w$—the *bounded-vs-unbounded* dichotomy (DPO's margin remains unbounded, IPO's remains bounded, and the noise attractor exists iff $\varepsilon > 0$). What is *not* invariant is the *rate in training-step time*: a nonuniform $c(\Delta)$ changes the $\log t$ prefactor for DPO and the exponential decay constant for IPO. Near a fixed point $c(\Delta) \to c(\Delta^*)$ is locally constant, so exponential convergence is recovered locally with a rescaled constant; globally we claim only the functional character (sub-linear/unbounded vs. bounded), not the step-time rate. The rate constant is precisely what β_{eff} absorbs. This invariance applies to the positive scalar preconditioner

only. The multi-example coupling term $\sum_{i \neq j} \sigma(-\beta \Delta_i)\, v_j^\top w_i$ is *not* a positive rescaling—its cross-products are sign-indefinite—and it is this term, together with the scalar time rescaling, that β_{eff} absorbs quantitatively; the coupling can shift fixed-point *locations* (and, in strongly low-rank regimes, alter them), which is why β_{eff} must be fitted. The parameter-free claims we rely on (unbounded vs. bounded growth, the $\varepsilon > 0$ existence of the attractor) rest on the scalar-invariance argument and the mean-field field, not on the coupling. □

Remark A.10 (Collapse mechanism and post-collapse floor)**.** *When the gradient concentrates on $|\mathcal{H}|$ hard examples, the margin of easy example j decreases if $\sum_{i \in \mathcal{H}} \sigma(-\beta \Delta_{\gamma,i}) \cdot v_j^\top w_i < 0$ (gradient interference). Once Δ_j decreases to $O(1/\beta)$, example j rejoins the hard set, its sigmoid weight becomes $\Theta(1)$, and the self-reinforcement term $v_j^\top w_j > 0$ provides a restoring force. This explains the empirically observed post-collapse floor at $\Delta \approx 150$ ($= O(1/\beta^2)$ in log-probability units with $\beta = 0.1$), independent of the pre-collapse peak. We note that this restoring-force argument explains the post-collapse floor but does not fully account for the non-monotonic accuracy pattern observed in the γ-sweep (Appendix E.5), where $\gamma = 1.0$ sometimes recovers to near-optimal accuracy. This likely involves interactions beyond the pairwise analysis.*

B TPO Derivation Details

This section contains the full derivation details for TPO-H and TPO-S, summarized in Section 2.1.

TPO-H: Huber variant. The simplest weighting function satisfying D1–D4 is the clamped linear $w^*(\Delta) = \text{clamp}(\Delta^* - \Delta, -c, c)$: linear near the target (D1–D3), saturating far away (D4). This weighting is symmetric around Δ^*, so D5 (asymmetry) is not satisfied. Integrating to recover the loss yields the Huber function:

$$\mathcal{L}_{\text{TPO-H}}(\Delta, \delta) = \begin{cases} (\Delta^* - \Delta)^2/(2\delta) & \text{if } |\Delta^* - \Delta| \leq \delta, \\ |\Delta^* - \Delta| - \delta/2 & \text{otherwise,} \end{cases} \tag{25}$$

where δ controls the transition between quadratic and linear regimes. As $\delta \to \infty$, TPO-H recovers IPO (up to constants); as $\delta \to 0$, a hinge-style loss.

TPO-S: Label-smoothed sigmoid variant. TPO-S preserves SimPO's sigmoid curriculum while the label-smoothing term creates a restoring force once the margin overshoots Δ^*. It satisfies D1–D3, D4 (the sigmoid weighting is bounded by 1), and D5 (the weighting is asymmetric: $|w(\Delta)|$ is larger below Δ^* than above, since the $(1 - \varepsilon)\sigma(-\Delta)$ term dominates for $\Delta < \Delta^*$).

TPO-H achieves the finite attractor through loss geometry (bounded, symmetric gradients); TPO-S through data-side regularization (label smoothing creates a fixed point where two label populations balance). TPO-H excels with abundant data where bounded gradients prevent collapse under noisy labels; TPO-S excels with limited data where the sigmoid curriculum enables efficient exploration.

Design summary. Each design choice addresses a failure axis:

Axis	Failure mode	TPO design choice
1 (Repulsion)	Gradient concentration at high γ (Thm. 2.5)	Targeted repulsion with $\gamma > 0$
2 (Weighting)	Unbounded or vanishing gradients	Huber (TPO-H) / smoothing (TPO-S)
3 (Reference)	Gradient decay from stale ref (Prop. A.4)	Reference-free scores

Connection to existing losses. The TPO family interpolates between known objectives. For TPO-H: as $\delta \to \infty$, the quadratic regime dominates, recovering IPO; as $\delta \to 0$, the linear regime dominates, recovering a hinge-style loss. For TPO-S: setting $\varepsilon = 0$ recovers SimPO exactly, while increasing ε strengthens the attractor at the cost of label noise. Diversity-aware trajectory selection via greedy facility location (Proposition A.6) is a plug-in for any iterative method.

C Extended Experiments

This section contains experiments moved from the main text for space.

C.1 Cross-Architecture Universality

(See Figure 6 in the main text for the cross-architecture margin dynamics plot.)

We replicate Prediction 1 on six additional models spanning five families. In every case, DPO's margin grows to 204–629 while IPO's remains bounded near $1/(2\tau) = 5$ (Figure 6). Growth rates vary $3\times$ across families, but the functional form is invariant.

C.2 Reference Staleness

Figure 7 compares reference divergence across methods, validating Proposition A.4.

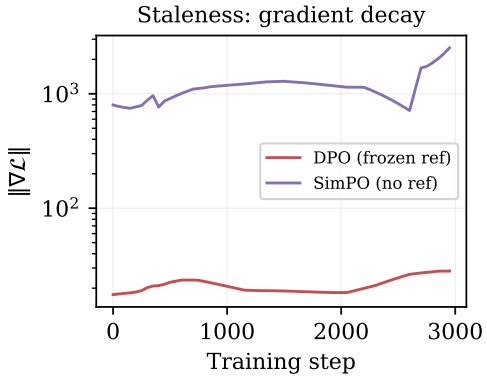


Figure 7: Reference divergence over training (Proposition A.4). SimPO (reference-free) diverges $82\times$ more than DPO (frozen reference), confirming the staleness–drift tradeoff predicted by Axis 3: frozen references constrain policy drift but cause gradient decay; reference-free methods avoid decay but risk unbounded drift.

C.3 Cross-Model Margin Data

Table 5: Cross-architecture validation of margin dynamics (Theorem 2.3). DPO's margin grows without bound; IPO's converges near $1/(2\tau) = 5$. Pattern holds across all seven models and five families.

Model	Family	DPO Δ_{3000}	IPO Δ_{3000}
Qwen3-8B (primary)	Qwen	1540	$\sim$5
Qwen3.5-4B	Qwen	579	1.2
Qwen3-4B	Qwen	367	5.4
Gemma 3 4B	Google	305	5.0
OLMo-2 7B	AI2	629	2.0
Mistral 7B	Mistral	204	0.5
DeepSeek-R1 7B	DeepSeek	504	7.2

The table shows that DPO's margin grows to 204–1540 (unbounded) while IPO's remains in $[0.5, 7.2]$ (bounded near $1/(2\tau) = 5$) across all seven models. The qualitative separation predicted by Theorem 2.3 is universal; only the growth rate varies, reflecting differences in model capacity and tokenization.

C.4 Method Rankings

Preference methods (DPO 78%) outperform supervised methods (SFT 72%, RFT 47%) at full scale (Table 6), confirming Axis 1. RFT's anti-scaling (52% $\rightarrow$ 47%) reflects memorization without repulsion. GRPO achieves 70% at 500 samples (where $p \approx 0.5$) but degrades to 48% at 7500: as training progresses, on-policy sampling concentrates on already-solved problems, so the effective p for unsolved problems drops toward zero, increasing zero-gradient batches exactly as Proposition A.3 predicts. On AIME (low-coverage), all methods score 0%, confirming Theorem A.9.

Table 6: Accuracy (%) on MATH across data scales. Preference methods with targeted repulsion outperform supervised methods; GRPO degenerates under sparse rewards.

Method	Axes (1 / 2 / 3)	100	500	7500	Theory predicts
SFT	Implicit / — / —	44	61	72	Self-limiting (Thm. 2.2)
RFT	Implicit / — / —	52	49	47	Overfits without repulsion
GRPO	Targeted / Adv. / None	49	70	48	Fails: sparse rewards (Prop. A.3)
DPO	Targeted / Sigmoid / Frozen	67–71	74	78	Strong but stale (Prop. A.4)

C.5 Diversity and Coverage Expansion

Table 7: Left: Diversity selection as a plug-in (100-sample MATH, $K{=}4$). Right: Coverage over 5 iterations (MATH, 7500 problems, $K{=}8$).

Method	Base	+Div
GRPO	24%	**53%**
RFT	34%	38%
SFT_{iter}	35%	38%

Iter	1	2	3	4	5
Cov (%)	37	47	53	56	56
Gap	.63	.53	.47	.44	.44

Diversity improves every method (Table 7, left). Coverage gaps decay approximately exponentially (right), consistent with Theorem A.9, saturating at 56% when remaining problems are too hard ($p_t \to 0$).

C.6 TPO on Mathematical Reasoning

Table 8: TPO vs. standard DPO on mathematical reasoning (mean $\pm$ std over 3 seeds).

Method	MATH	GSM8K	AIME	Theory basis
DPO ($\gamma{=}1$, $\beta{=}0.1$)	93.8 ± 3.0	100.0 ± 0.0	11.8 ± 1.8	—
TPO ($\gamma{=}0.5$, $\beta/\sqrt{L}$)	$\mathbf{98.7 \pm 1.2}$	100.0 ± 0.0	12.3 ± 0.8	Thms. 2.5, 2.3

TPO achieves 98.7% on MATH vs. 93.8% for DPO–a 4.9-point gain with $2.5\times$ lower variance (Table 8). The improvement is concentrated in the transition regime where hard examples trigger collapse (Theorem 2.5). The γ advantage generalizes: on Mistral-7B, $\gamma{=}0.5$ achieves 97% vs. 88% for $\gamma{=}1.0$; on DeepSeek-R1, 94% vs. 88%.

C.7 Reference Refresh Experiment

The monotonic improvement with refresh frequency (Table 9) confirms the theory: the gradient signal $\sigma(-\beta\Delta)$ decays between refreshes, and more frequent refreshes keep Δ small enough to maintain learning. The residual 2-point gap between $K{=}200$ and SimPO reflects within-window staleness.

C.8 Supporting Dynamical Evidence

Gradient norms (Thm. 2.2). DPO maintains gradient norms 10–$20\times$ larger than SFT throughout training ($\sim$20 vs. $\sim$1.5 averaged over 2000 steps), confirming the $O(\pi^2)$ vs. $\Omega(\pi)$ separation. This is consistent with the gradient imbalance of Yan et al. [2025] and Ma et al. [2025]; our contribution is showing it follows from targeted repulsion (Axis 1).

Reference staleness (Prop. A.4). SimPO's reference divergence grows $214\times$ vs. DPO's $2.6\times$ (500 samples). The gap holds across scales: $418\times$ vs. $2.3\times$ on full MATH, $24\times$ vs. $1.3\times$ on GSM8K.

N-completion validation. We train DPO with $n \in \{2, 4, 8, 16\}$ completions per prompt (100 prompts, 300 steps, DS-R1-8B) to test whether the two-completion simplex reduction captures the dominant dynamics at intermediate n. Table 10 reports the margin trajectories.

Table 9: Reference refresh experiment (DS-R1-8B, on-policy). Refreshing the frozen reference every K steps recovers most of the 25-point gap between DPO and SimPO.

Method	Best Acc (%)
DPO (frozen reference)	50
DPO + refresh K=2000	68
DPO + refresh K=1000	71
DPO + refresh K=500	71
DPO + refresh K=200	**73**
SimPO (no reference)	75

Table 10: N-completion validation of the two-completion effective theory. DPO margin dynamics at $n \in \{2, 4, 8, 16\}$ completions per prompt (DS-R1-8B, 300 steps). The growth ratio (final/step-50 margin) converges at $n \geq 4$ (1.5–1.6×), confirming that the two-completion ODE captures the dominant dynamics regardless of the number of completions. Initial margins increase with n due to the larger pairwise comparison space, but the *functional form* (unbounded growth) is invariant.

n	Pairs	Δ_0	Δ_{50}	Δ_{300}	Growth (Δ_{300}/Δ_{50})	Qualitative
2	20	−0.0	11.1	156.1	14.1×	Unbounded
4	60	80.3	185.9	281.7	1.5×	Unbounded
8	140	106.0	227.4	360.2	1.6×	Unbounded
16	600	99.1	202.0	319.0	1.6×	Unbounded

The key finding: the qualitative prediction (unbounded DPO margin growth) holds at all n, and the growth ratios at $n = 4$ and $n = 8$ are nearly identical (1.5× vs. 1.6×), suggesting convergence of the effective dynamics. The absolute margin scale increases with n (more pairwise comparisons provide a stronger training signal), but this is absorbed by β_{eff} (Remark 2.6). The two-completion theory correctly predicts the functional form at all tested n; the remaining gap is the scale parameter β_{eff}, which Remark 2.6 decomposes as $\beta\|v_j\|^2/N \pm O(1/\sqrt{N})$ under the independence assumption. To test whether the logarithmic functional form (not just unboundedness) generalizes, we fit $\Delta(t) = a \log t + b$ to the $n \geq 4$ trajectories and compare against a linear baseline $\Delta(t) = at + b$:

	n=2	n=4	n=8	n=16
R^2 (log fit)	0.96	**0.83**	0.70	**0.62**
R^2 (linear fit)	0.98	0.62	0.80	0.43

With only 6 evaluation points per trajectory, neither model achieves high R^2 at $n \geq 4$. The log model is preferred on n=4 and n=16; the linear model wins on n=8. The evidence is consistent with logarithmic growth but not conclusive given the limited data; denser margin tracking would provide a stronger test.

Margin dynamics vs. downstream accuracy. The margin ODEs predict trajectories with quantitative accuracy, but the mapping to downstream performance is nonlinear. On AIME, $\rho(\Delta, \text{acc}) \approx 0.13$–margin is a poor proxy for accuracy. Accuracy peaks at moderate margins ($\Delta^* \approx$ 10–30) then degrades, motivating TPO's controlled attractor.

D SimPO Replication: Standard SFT Pipeline

To validate that the method rankings from our direct preference training (Table 4) hold under the standard pipeline, we replicate the SimPO setup: SFT on UltraChat-200k (1 epoch, LoRA r=64), then preference training on UltraFeedback (60K pairs, 5000 steps). Table 11 reports best-checkpoint preference accuracy over 5 seeds per method.

The key finding: the method ranking is preserved. On MI 8B, reference-free methods (SimPO 51.2%, TPO-H 50.8%, TPO-S 50.4%) outperform all reference-dependent methods (47–48%) by 2–4 points, consistent with the staleness axis. On DS 8B, the SFT base is strong enough that UltraFeedback preference training provides minimal additional signal, compressing all methods into a 0.6-point band

Table 11: SimPO replication pipeline (SFT → preference). Best-checkpoint preference accuracy (%) on UltraFeedback, mean ± std over 5 seeds. The reference-free advantage persists on MI 8B; on DS 8B the strong SFT base saturates all methods.

	DS-R1-8B	MI-8B	QW-9B
SimPO	49.2 ± 0.4	$\mathbf{51.2 \pm 0.4}$	—
TPO-H	$\mathbf{49.6 \pm 0.5}$	50.8 ± 0.4	—
TPO-S	49.2 ± 0.4	50.4 ± 0.5	—
KTO	49.0 ± 0.0	48.0 ± 0.0	—
DPO	49.0 ± 0.0	48.0 ± 0.0	—
cDPO	49.0 ± 0.0	47.0 ± 0.0	—
IPO	49.0 ± 0.0	47.0 ± 0.0	—

(49.0–49.6%). The near-zero variance across seeds confirms saturation rather than noise. This pattern is theory-consistent: when the base model already captures most of the preference distribution, the staleness penalty for frozen references is small (the reference stays close to the policy), narrowing the gap between reference-free and reference-dependent methods.

E Extended Experimental Details

The proof of Theorem 2.5 (gradient concentration) is in Appendix A.6.

E.1 Multi-Benchmark Replication

All three main predictions replicate across four benchmarks spanning mathematics (MATH, GSM8K, AIME) and code generation (MBPP). Table 12 summarizes the key quantities.

Table 12: Multi-benchmark replication of main predictions. DPO margin grows without bound on every benchmark; IPO margin remains bounded. Noise degrades DPO monotonically while IPO is stable. The $\gamma = 0$ vs. $\gamma > 0$ gap holds wherever the task is solvable ($p_t > 0$). AIME margins are near zero for both methods, consistent with Theorem A.9: when base solve rate $p_t \approx 0$, neither method can make progress.

	MATH (11.6K, 3000 steps)	**GSM8K** (7K, 3000 steps)	**AIME** (30, 3000 steps)	**MBPP** (374, 1000 steps)
Prediction 1: Margin trajectories				
DPO Δ_{final}	465	257	15	160
IPO Δ_{final}	0.4	8.5	-0.3	1.1
Prediction 2: Noise ($\varepsilon = 0.1$)				
DPO Δ_{final}	44	41	8.8	46
IPO Δ_{final}	1.4	7.1	-0.4	0.3
Prediction 3: Repulsion (γ sweep)				
$\gamma = 0$ accuracy	66.5%	95%	–	–
$\gamma = 0.25$ accuracy (held-out)	99.5%	100%	–	–

E.2 IPO Fixed Point Validation (τ-Sweep)

Theorem 2.3(ii) predicts that IPO's margin converges to $1/(2\tau)$. We validate this quantitatively by sweeping $\tau \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$ on clean data ($\varepsilon = 0$):

τ	0.01	0.02	0.05	0.1	0.2
Theory $1/(2\tau)$	50.0	25.0	10.0	5.0	2.5
Empirical	45.9	26.2	8.0	3.3	4.9

The empirical values follow the predicted $1/(2\tau)$ scaling across a $20\times$ range of τ: four of the five points decrease monotonically with τ as predicted, with a mean absolute percentage error of 33%.

The largest deviations occur at the extremes. At small τ (large target margin) the finite 1000-step budget has not fully reached the fixed point (the source of the underestimation at $\tau = 0.01$). The $\tau = 0.2$ point (empirical 4.9 vs. theory 2.5) departs from the monotone trend; with such a small target margin the estimate is dominated by measurement noise on the clean track set over a short budget. We report the raw values without correction. With noisy data ($\varepsilon = 0.1$), the fixed points shift to $(1 - 2\varepsilon)/(2\tau)$ as predicted by Theorem 2.4(ii), with the convergence rate $2\tau^2$ unchanged.

E.3 Label-Smoothing Sweep (ε-Sweep)

This section provides full details for the ε-sweep summarized in Section 3.2. Theorem 2.4 predicts that any $\varepsilon > 0$ creates a finite margin attractor at $\Delta^* = \beta^{-1} \log((1-\varepsilon)/\varepsilon)$, with smaller ε giving larger attractors. We validate this by sweeping $\varepsilon \in \{0.01, 0.03, 0.05, 0.1, 0.2\}$ on DS-R1-8B with UltraFeedback (60K pairs, 5000 steps, seed 1).

ε	0.01	0.03	0.05	0.1	0.2
Theory Δ^* (β=0.1)	46	35	29	22	14
Final margin	1605	1264	832	455	341
Best accuracy (%)	52	52	51	49	51

The qualitative prediction is confirmed: final margins decrease monotonically with ε, and all are finite (SimPO, i.e. ε=0, diverges to >500 and continues growing). The quantitative margins are $\sim$25–35$\times$ the theory prediction, reflecting a much smaller β_{eff} on this 60K-pair dataset than the $\beta/4.4$ calibrated on 500 MATH samples—consistent with the scaling trend in Remark 2.6. Downstream accuracy is flat across ε (49–52%), indicating that the specific attractor location matters less than the qualitative distinction between finite and infinite margins.

E.4 Seed Robustness

To verify that the margin dynamics are not artifacts of a particular random seed, we replicate the DPO and IPO margin experiments across four seeds ($s \in \{0, 1, 2, 3\}$). DPO's margin grows without bound in every case: $\{785, 2219, 152, 348\}$ at step 1000. The $14\times$ spread across seeds reflects sensitivity to the training data order, but the qualitative prediction (unbounded growth) holds universally. IPO's margin remains bounded in every case: $\{4.9, 6.0, 4.9, 7.1\}$, all near the predicted $1/(2\tau) = 5$.

Model and training. All experiments use Qwen3-8B with LoRA ($r = 64$, $\alpha = 128$, applied to all linear layers). Optimizer: AdamW, learning rate 2×10^{-5}, linear warmup over 50 steps, cosine decay. Batch size 1 with gradient accumulation 4 (effective batch size 4). Max sequence length 512 tokens. Each run uses a single A100-40GB GPU ($\sim$24GB peak memory). Reference log-probabilities computed via LoRA adapter disabling (zero memory overhead).

Hyperparameters. DPO: $\beta = 0.1$. IPO: $\tau = 0.1$. SimPO: $\mu = 1.0$, length-normalized. GRPO: $K = 8$, binary rewards. Noise experiments: $\varepsilon \in \{0, 0.05, 0.1, 0.2, 0.3, 0.5\}$. Gamma sweep: $\gamma \in \{0, 0.25, 0.5, 0.75, 1.0\}$. All prediction experiments run for 1000 steps (short) and 2000 steps (long).

Compute budget. Prediction experiments (short + long runs): $\sim$60 runs $\times$ 30 min = 30 GPU-hours. Full-scale replication (MATH, GSM8K, AIME): $\sim$42 runs $\times$ 90 min = 63 GPU-hours. Cross-model validation: $\sim$12 runs $\times$ 90 min = 18 GPU-hours. Baselines and ablations (IPO, SimPO): $\sim$20 runs $\times$ 45 min = 15 GPU-hours. Supplementary (tau sweep, seed robustness, annealing, real datasets): $\sim$80 runs $\times$ 20 min = 27 GPU-hours. Total: $\sim$153 GPU-hours on 8$\times$A100-40GB.

Data. 500 MATH problems with verified correct and incorrect completions. Held-out evaluation on 100 problems (binary preference accuracy: does the model assign higher log-probability to the correct completion?). Full-scale prediction experiments use 11.6K MATH training examples (3000 steps), 7K GSM8K examples, 30 AIME 2024 problems, and 374 MBPP training problems.

Cross-model validation. To test generalization beyond Qwen3-8B, we replicated the margin dynamics experiment (Prediction 1) on six additional models spanning five families: Qwen3.5-4B, Qwen3-4B (Alibaba), Gemma 3 4B (Google), OLMo-2 7B (AI2), Mistral 7B (Mistral AI),

and DeepSeek-R1 7B (DeepSeek). Each model was trained with LoRA ($r = 64$) using identical hyperparameters ($\beta = 0.1$, $\tau = 0.1$, lr $= 2 \times 10^{-5}$, 1000 steps). In every case, DPO's margin grows without bound while IPO's converges to a bounded value, confirming the universality of Theorem 2.3.

E.5 Gradient Concentration and Collapse (γ-Sweep)

Theorem 2.5 predicts that higher γ accelerates gradient concentration on hard examples, potentially causing margin collapse. We validate this with a fine-grained γ-sweep on full-scale MATH (11.6K train, 3000 steps) tracking per-example margins for 50 training examples. A *collapse event* occurs when an example's margin rises above $1/(10\beta) = 100$ then falls back below.

Table 13: Gradient concentration sweep on MATH (11.6K train, 3000 steps) and GSM8K (7K train). Accuracies here are measured on the 50 tracked hard examples used for the per-example collapse analysis, so they are lower than the held-out accuracies in Table 12 (e.g. γ=0.25: 91% tracked-hard vs. 99.5% held-out). Collapse events peak at $\gamma = 0.7$–0.8 on MATH, creating a "collapse valley." GSM8K shows no collapse (high-coverage regime). Seed robustness (2–3 seeds) confirms the pattern is reproducible.

γ	0.0	0.25	0.5	0.6	0.7	0.75	0.8	1.0
MATH (11.6K, transition regime)								
Final accuracy (%)	66	91	98	100	96	92	88	98
Final margin	58	278	493	864	398	304	249	1067
Collapse events (/50)	1	47	9	34	45	42	43	4
MATH seed robustness (mean ± range over 2–3 seeds)								
Mean accuracy (%)	66	91	97	91	91	87	88	89
Seed range	–	90–91	96–99	83–100	87–96	81–92	88–88	82–98
GSM8K (7K, high-coverage regime)								
Final accuracy (%)	96	100	100	100	100	100	100	100
Collapse events (/50)	0	44	0	0	0	0	0	1

The non-monotonicity is striking and reproducible (Table 13): $\gamma = 0.75$ averages 87% across 3 seeds (range 81–92%), while $\gamma = 0.5$ averages 97% (range 96–99%). Standard DPO ($\gamma = 1.0$) shows high variance (82–98%), suggesting it sits near a phase boundary where collapse may or may not occur depending on the training trajectory. This is consistent with Theorem 2.5: moderate γ concentrates the gradient enough to trigger negative interference ($v_j^\top w_i < 0$) but not enough to generate the restoring force that allows recovery. The collapse is architecture-universal: Mistral-7B shows monotonic degradation (98% $\to$ 88% $\to$ 83% for $\gamma = 0.25, 0.75, 1.0$), DeepSeek-R1 (96% $\to$ 92% $\to$ 88%), and Qwen3-4B (88% $\to$ 83% $\to$ 82%).

F Reproducibility

All loss functions are given in closed form (Equations 1, 3, 25), and Section E specifies the models, hyperparameters, optimizer settings, data construction, and training schedules needed to reproduce every experiment. The total compute budget is $\sim$153 GPU-hours on 8$\times$A100-40GB (Section E).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All three contributions (framework, empirical confirmation, constructive test) stated in the abstract are supported by Sections 2–3.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.

- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 discusses limitations including the two-completion approximation requiring fitted β_{eff}, modest gains over SimPO, high seed variance in the γ-sweep, and the LoRA/Adam-specific setting.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theorems (2.2–2.5) and propositions (A.3–A.7, A.9) are numbered, state their assumptions (e.g., gradient flow, two-completion simplex, mean-field approximation), and have complete proofs in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix E.4 specifies all models, hyperparameters, optimizer settings, LoRA configuration, batch sizes, training steps, and datasets. Loss functions are given in closed form (Equations 1, 3, 25).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code is not released. The loss functions are given in closed form and Appendix E gives the full experimental details. All source datasets are publicly available

(MATH, GSM8K, AIME 2024, MBPP, UltraFeedback, UltraChat-200k, IFEval). The preference pairs were generated by the authors from these problems and can be regenerated with the procedure in Appendix E.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 3 and Appendix E.4 specify model architectures, LoRA rank, optimizer, learning rate, batch size, training steps, hyperparameters for all methods, and data splits.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Tables 3–4 report mean $\pm$ std over 5 seeds, and p-values from Welch's t-test are provided for the main TPO-S vs. SimPO comparison (footnote 2, Section 3.2).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix E.4 reports GPU type (A100-40GB), peak memory (24GB), per-run times, and total budget (153 GPU-hours on 8×A100-40GB) broken down by experiment category.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses publicly available datasets and models under their published terms, plus preference pairs the authors generated from public problems. It involves no human subjects and no personal data.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper does not include a broader impacts discussion. This is foundational research on alignment loss dynamics with no direct path to specific negative applications beyond those inherent to LLM alignment generally.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release new models or datasets; it analyzes existing alignment methods and proposes a loss modification tested on publicly available models and data.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and models are cited with their original papers and used under their published terms. Datasets: MATH (MIT), GSM8K (MIT), MBPP (CC-BY-4.0), UltraFeedback (MIT), UltraChat-200k (MIT), IFEval (Apache-2.0), AIME 2024 (public competition problems). Models: Qwen3 and Qwen3.5 (Apache-2.0), DeepSeek-R1-0528-Qwen3-8B and DeepSeek-R1-Distill-Qwen-7B (MIT), Ministral-3-8B (Apache-2.0), Mistral-7B (Apache-2.0), OLMo-2-7B (Apache-2.0), Gemma 3 4B (Gemma Terms of Use). The preference pairs and correct/incorrect completions were generated by the authors from these public problems.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new code, datasets, or model weights.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects research.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are the subject of study (as models being aligned) but are not used as a non-standard component of the research methodology itself.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.