WHEN HIGHER SCORES YIELD LESS FEEDBACK: SCORE SATURATION IN SELF-REWARDING LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-rewarding language models can improve through an iterative feedback loop in which the model generates multiple responses, evaluates them using a discrete scoring rubric, and trains on the resulting preferences. A known failure mode is *score saturation*, where generated responses increasingly receive scores near the top of the rubric, producing more ties and fewer informative comparisons. We develop a theoretical analysis of when and how preference optimization contributes to saturation and reshapes the feedback available to subsequent rounds. We first show that the preference optimization update direction inherently contains a component that increases the probability of the highest score, while other components of the preference gradient can either reinforce or counteract this effect. We further decompose score changes into effects arising from changes in the model's response distribution and changes in its evaluation behavior, and derive conditions under which these effects accumulate to produce score saturation. As saturation progresses, the preference comparisons increasingly reflect the gradient structure associated with increasing top-score probability, even as informative comparisons become increasingly rare. Consequently, maintaining a fixed number of them requires a sampling budget that grows inversely with the remaining probability of a non-top score, and asymptotically exponentially in the accumulated gains. Experiments on self-rewarding language models validate these theoretical results.

1 INTRODUCTION

Large language models (LLMs) are widely used both to evaluate model-generated responses (Zheng et al., 2023; Liu et al., 2023) and to provide preference feedback for model training (Bai et al., 2022; Lee et al., 2024). LLM-based evaluation can take the form of pairwise comparison, rubric-based scoring, or aggregation across repeated evaluations or multiple evaluators (Zheng et al., 2023; Verga et al., 2024; Wang et al., 2024b). In a common training paradigm with score-based rubrics (Yuan et al., 2024), candidates are assigned discrete scores, and the highest- and lowest-scoring responses in each sampled group are selected as a preference pair (i.e., *max–min selection*) for direct preference optimization (DPO) (Rafailov et al., 2023). Self-rewarding language models close this loop by using the same model to generate responses, evaluate them, and learn from the resulting preferences (Yuan et al., 2024).

However, iterative self-generated feedback can lead to *score saturation*: model-generated responses increasingly receive the highest rubric scores, while ties become more frequent (Wu et al., 2025). Related approaches seek to improve the reliability and effectiveness of such feedback, through synthetic evaluator training (Wang et al., 2024b), meta-judgment supervision (Wu et al., 2025), cross-round consistency regularization (Wang et al., 2025), and temporal self-rewarding, which constructs preference pairs across model iterations (Wang et al., 2026). These studies provide important empirical observations and practical interventions, but do not systematically characterize the mechanisms driving score saturation under iterative preference optimization. Recent theoretical work has begun to analyze iterative self-rewarding and alignment collapse (Fu et al., 2026; Gauthier et al., 2026), but focuses on broader feedback dynamics and stylized settings rather than the evolution of discrete rubric scores in score-based preference optimization. At the same time, score saturation directly affects the feedback available for subsequent training. Under max–min selection, groups in which

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

all candidates receive the same score yield no non-tied preference pair (Yuan et al., 2024). As scores concentrate at the top of the rubric, such groups become increasingly likely, reducing the supply of usable preference pairs. Yet when and how saturation arises from optimization dynamics, and how it quantitatively affects the supply of usable preferences, remain theoretically uncharacterized.

In this paper, we develop a theoretical analysis of iterative score-based DPO with max–min selection, addressing three fundamental questions:

How does score-based preference optimization contribute to the emergence of score saturation?
How do changes in response generation and evaluation contribute to its dynamics?
How does increasing score saturation affect the availability of usable preference pairs?

Crucially, we do not assume Bradley–Terry realizability, allowing us to study discrete rubric-based feedback directly without imposing a parametric model on the underlying preference structure. We distinguish the *actor*, which generates responses, from the *judge*, which assigns discrete scores, so that one framework covers both self-judge and fixed-judge settings. We find that the round-start DPO descent direction inherently contains a component that increases the probability of the top score, while the remaining preference gradients may either reinforce or oppose this component.

This local mechanism at round initialization does not by itself determine the outcome of a multi-iterate training round. We therefore track the realized change in top-score log odds and decompose it into an *actor gain*, caused by changes in the generated response distribution under the current judge, and a *judge gain*, caused by changes in how the updated responses are evaluated. We derive sufficient conditions under which these gains are positive and show that their accumulation characterizes entry into the score-saturation regime. We then close the feedback loop by analyzing how the resulting score distribution determines the next round's preference supply. Near top-score saturation, candidate groups increasingly consist of top-scoring responses, while the remaining accepted pairs increasingly follow the same top-versus-lower-score structure underlying the top-score component above. As a result, for a fixed candidate-group size, the probability of obtaining any non-tied pair vanishes in proportion to the remaining non-top-score probability. Maintaining a fixed expected preference supply therefore requires a sampling budget that grows inversely with this remaining probability and, asymptotically, exponentially with cumulative gains. More related works are provided in Appx. A. Our contributions are as follows.

- *Top-score reinforcement analysis* (Sec. 3). We establish an allocation-based decomposition of the round-start DPO descent direction, identify its component toward the top-score region, and give sufficient conditions for directional alignment.
- *Actor-judge score dynamics* (Sec. 4). We exactly decompose round-level changes in top-score log odds into actor and judge gains, derive sufficient conditions for positive gains, and characterize score saturation through their cumulative effect.
- *Preference supply analysis* (Sec. 5). We characterize preference allocation and acceptance probabilities near saturation, connecting cumulative gains to the sampling budget required to maintain an expected supply of non-tied preferences.
- *Empirical validation* (Sec. 6). Experiments with LLMs provide empirical support for the mechanisms connecting preference optimization, score saturation, and preference supply.

2 Problem Setup

We study iterative LLM-as-a-judge systems that construct preference pairs from discrete response scores. Following the common training paradigm of self-rewarding LLMs (Yuan et al., 2024), the model is updated using DPO. We distinguish two roles in such systems: the **actor**, parameterized by θ, which generates responses and is optimized using preference data, and the **judge**, parameterized by ϕ, which evaluates the generated responses and induces the preference pairs. We parameterize the two roles separately to disentangle their contributions to the system dynamics. This is an analytical distinction rather than an assumption that they must be separate models. We consider two representative judge configurations for iterative self-improvement. In the self-judge setting, the actor and judge share the same parameter at the beginning of each round, $\theta_t = \phi_t$, with role-specific prompting. In the fixed-judge setting, the actor evolves while the judge remains unchanged, $\phi_t = \phi_0$.

At round t, given a prompt x, the actor generates a response y according to $\pi_{\theta_t}(y|x)$. The judge assigns a *score* $S_{\phi_t}(x, y)$ based on x, y using the judge prompt and scoring rubric (the full prompt is in Appendix C.1). Assume the score range is $\{1, ..., L\}$, where L denotes the highest score. Define

$$p_\ell(\phi, \theta; x) = \mathbb{E}_{Y\sim\pi_\theta(.|x)}\left[\mathbb{P}\left(S_\phi\left(x, Y\right) = \ell\right)\right] \tag{1}$$

as the expected probability that a response generated by the actor θ for prompt x is assigned score ℓ by the judge ϕ. We abbreviate $p_{t,\ell} := p_\ell(\phi_t, \theta_t; x)$, the score distribution at the start of round t, and $\tilde{p}_{t,\ell} := p_\ell(\phi_t, \theta_{t+1}; x)$, the distribution after the actor update but before the judge changes; holding the judge fixed across the actor update is what later lets us separate actor-induced from judge-induced score changes. We also define $\alpha_{t,\ell} = \frac{p_{t,\ell}}{1-p_{t,L}} = \frac{p_{t,\ell}}{\sum_{s<L} p_{t,s}}$ for $\ell < L$, the fraction of the residual non-top-score probability assigned to score ℓ by the judge. Unless stated otherwise, our analysis is conditional on a fixed x, which is suppressed in the notation.

Training setting. At the beginning of round t, for the input prompt x, a group of $N \geq 2$ candidates $y_1, ..., y_N$ is sampled independently from $\pi_{\theta_t}(.|x)$ and scored by ϕ_t. The group is *accepted* if its scores are not all equal, and discarded otherwise. From an accepted group, max–min selection forms one preference pair: the highest- and lowest-scoring responses, denoted y^+ and y^-, with corresponding scores $S^+ > S^-$; ties at S^+ or S^- are broken uniformly at random among the candidates attaining them. A group therefore yields training signal only when the judge separates at least two of its candidates, so the score distribution controls not only which pairs are formed but how many. The reference policy is set to $\pi_t = \pi_{\theta_t}$, and both the reference policy and the constructed preference data remain fixed throughout the within-round optimization. Let the prompt-conditional distribution of selected pairs (y^+, y^-) under judge ϕ_t be $\Gamma_t^{(N)}$. The model is updated by minimizing

$$\mathcal{L}_t^{\text{DPO}}(\theta) = \mathbb{E}_{(y^+,y^-)\sim\Gamma_t^{(N)}}[\ell_{\log}\left(\beta\Delta_t\left(\theta; y^+, y^-\right)\right)], \quad \ell_{\log}(z) := \log(1 + e^{-z}), \tag{2}$$

where $\Delta_t(\theta; y^+, y^-) = \log\frac{\pi_\theta(y^+|x)}{\pi_t(y^+|x)} - \log\frac{\pi_\theta(y^-|x)}{\pi_t(y^-|x)}$ is the DPO margin, abbreviated $\Delta_t(\theta)$, and $\beta > 0$ is the DPO hyperparameter. Based on Eq. (2), we denote the model update direction at θ as

$$G_t(\theta) := -\nabla_\theta\mathcal{L}_t^{\text{DPO}}(\theta). \tag{3}$$

Score-conditioned quantities. We index the accepted preference data by the scores its pairs carry. A pair *has scores* (m, ℓ), $m > \ell$, if the judge assigned exactly m to y^+ and ℓ to y^-. Let the *preference allocation* for score pair (m, ℓ) be $\omega_{t,m,\ell} := \mathbb{P}(S^+ = m, S^- = \ell | S^+ > S^-)$, the share of accepted pairs carrying those two scores. Denote by π_t^ℓ and $\tilde{\pi}_t^\ell$ the pre- and post-update policy distributions conditioned on score ℓ under the same judge ϕ_t at round t, so that

$$\pi_t^\ell(y) = \frac{\pi_t(y)\mathbb{P}(S_{\phi_t}(x, y) = \ell)}{p_{t,\ell}}, \quad \tilde{\pi}_t^\ell(y) = \frac{\pi_{t+1}(y)\mathbb{P}(S_{\phi_t}(x, y) = \ell)}{\tilde{p}_{t,\ell}}, \tag{4}$$

and let $\Gamma_{t,m,\ell}^{(N)} := \Gamma_t^{(N)}(\cdot \mid S^+ = m, S^- = \ell)$ be the selected-pair distribution conditioned on that score pair.

Proposition 1 (Decomposition of the accepted preference pair distribution)**.**

$$\Gamma_{t,m,\ell}^{(N)} = \pi_t^m \otimes \pi_t^\ell, \quad \Gamma_t^{(N)} = \sum_{m>\ell} \omega_{t,m,\ell}(\pi_t^m \otimes \pi_t^\ell), \ \otimes \text{ denotes the distribution product.} \tag{5}$$

Prop. 1 separates the composition of accepted pairs from their score-conditioned response distributions. Within a score pair, the selected responses are independent draws from $\pi_t^m \otimes \pi_t^\ell$, which does not depend on N; the group size affects $\Gamma_t^{(N)}$ only through the preference allocation $\omega_{t,m,\ell}$. This split is what lets us compare the direction DPO moves in against the direction that increases the top-score probability: both are weighted sums of the same score-conditioned margin gradients, and differ only in how they weight them. Accordingly, define the *conditional DPO loss* $\varepsilon_{t,m,\ell}$ and the *conditional margin* $\Delta_{t,m,\ell}(\theta)$ on pairs with scores (m, ℓ) as

$$\varepsilon_{t,m,\ell} := \mathbb{E}_{\Gamma_{t,m,\ell}^{(N)}}\left[\ell_{\log}\left(\beta\Delta_t(\theta)\right)\right] \ \text{ and } \ \mathcal{L}_t^{\text{DPO}}(\theta) = \sum_{m>\ell} \omega_{t,m,\ell}\varepsilon_{t,m,\ell}, \ \forall\, L \geq m > \ell \geq 1, \tag{6}$$

$$\Delta_{t,m,\ell}(\theta) := \mathbb{E}_{\Gamma_{t,m,\ell}^{(N)}}\left[\Delta_t(\theta; x, y^+, y^-)\right] = \mathbb{E}_{y\sim\pi_t^m}\left[\log\frac{\pi_\theta(y|x)}{\pi_{\theta_t}(y|x)}\right] - \mathbb{E}_{y\sim\pi_t^\ell}\left[\log\frac{\pi_\theta(y|x)}{\pi_{\theta_t}(y|x)}\right], \tag{7}$$

where Eq. (7) follows from Eq. (5). Thus, $\Delta_{t,m,\ell}(\theta)$ measures, over accepted pairs receiving scores m, ℓ from judge ϕ_t, the expected log-likelihood margin that π_θ places on score-m responses over score-ℓ responses, relative to the round-t reference policy π_t; it vanishes at $\theta = \theta_t$ by construction.

Definition 1 (Score saturation). *For $\delta \in (0, 1)$, round t is δ-*score saturated *if $p_{t,L} \geq 1 - \delta$.*

Score saturation describes concentration in the actor–judge score distribution. It does not, by itself, imply improved response quality or deteriorating judgment accuracy; it says only that the judge rarely distinguishes among the actor's samples. Next, we investigate how such LLM-as-a-judge systems evolve over recursive training dynamics: how does score-based preference optimization influence $p_{t,L}$, when does it drive the system toward score saturation, and when does the resulting concentration reduce the preference feedback available to subsequent rounds? In particular, Sec. 3 analyzes the descent direction induced by the accepted pairs at round initialization, and shows that it contains a positive component along the direction increasing $p_{t,L}$. Sec. 4 characterizes the score change a completed round realizes, separates it into actor and judge contributions, and gives an exact cumulative-gain threshold for saturation. Sec. 5 analyzes how the resulting score distribution shapes the next round's preference data, and shows that near saturation the accepted pairs increasingly match the top-score gradient's structure while non-tied pairs become scarce, so that a fixed preference supply requires a sampling budget growing with the accumulated gain.

3 Top-score Reinforcement

DPO optimizes relative likelihood ratios on accepted response pairs, whereas score saturation concerns the probability of receiving the top score under the current judge. These are distinct quantities. We therefore first ask what the max–min preference distribution does to the model at the beginning of a round, comparing the DPO descent direction $G_t(\theta_t)$ against the gradient of the top-score log-odds, $v_t := \nabla_\theta \mathrm{logit} p_L(\phi_t, \theta)|_{\theta=\theta_t}$, where $\mathrm{logit}(z) = \log \frac{z}{1-z}$ is the log-odds function. The score change realized by a completed, potentially multi-step update is analyzed separately in Sec. 4.

Proposition 2 (Conditional margin gradient mixture). *At $\theta = \theta_t$,*

$$v_t = \sum_{\ell<L} \alpha_{t,\ell} \nabla_\theta \Delta_{t,L,\ell}(\theta_t), \quad G_t(\theta_t) = \frac{\beta}{2} \sum_{m>\ell} \omega_{t,m,\ell} \nabla_\theta \Delta_{t,m,\ell}(\theta_t). \tag{8}$$

Prop. 2 expresses both directions through the same family of conditional margin gradients $\{\nabla_\theta \Delta_{t,m,\ell}(\theta_t)\}_{m>\ell}$, differing only in how they weight it. The top-score gradient v_t uses top-versus-lower comparisons weighted by the composition of the non-top-score mass $\alpha_{t,\ell}$, whereas DPO follows the preference allocation induced by max-min selection. Since v_t draws only on pairs of the form (L, ℓ), which are among the comparisons G_t already averages over, we can ask how much of G_t necessarily reproduces the mixture defining v_t.

Theorem 1. *Let $c_t := \min_{\ell<L: p_{t,\ell}>0} \frac{\omega_{t,L,\ell}}{\alpha_{t,\ell}}$, then $c_t \in [p_{t,L}^{N-1}, 1]$, and*

$$G_t(\theta_t) = \frac{\beta}{2}(c_t v_t + R_t), \quad \textit{with} \quad R_t = R_t^L + R_t^{NL}, \tag{9}$$

$$R_t^L = \sum_{\ell<L} (\omega_{t,L,\ell} - c_t \alpha_{t,\ell}) \nabla_\theta \Delta_{t,L,\ell}(\theta_t), \quad R_t^{NL} = \sum_{\ell<m<L} \omega_{t,m,\ell} \nabla_\theta \Delta_{t,m,\ell}(\theta_t).$$

Thm. 1 extracts from the descent direction the largest multiple of v_t that the preference allocation permits. The coefficient c_t is fixed by the allocation itself, not by an arbitrary projection of G_t onto v_t: it is the largest mass extractable in the proportions $\{\alpha_{t,\ell}\}_{\ell<L}$ while leaving nonnegative weights on every remaining gradient in R_t. Whenever $p_{t,L} > 0$ and $v_t \neq 0$, the descent direction therefore contains the strictly positive component $c_t v_t$ along the direction that increases $p_{t,L}$, and it does so regardless of the residual. The residual gradients in R_t, by contrast, need not have nonnegative projections onto v_t. We call the component $c_t v_t$ the *structural pressure* that DPO exerts toward the current judge's high-score region; the realized dynamics are governed by the competition between it and the residual. We next bound that residual as the system approaches score saturation.

Proposition 3. *Assume the norms of conditional margin gradients are bounded by D_t, then*

$$\|R_t\| \leq (1 - c_t) D_t, \quad \|\frac{2}{\beta} G_t(\theta_t) - v_t\| \leq 2(N-1)(1 - p_{t,L}) D_t. \tag{10}$$

Prop. 3 shows that score saturation reduces the residual: if $p_{t,L} \to 1$ and the conditional margin gradients are uniformly bounded, then $\|\frac{2}{\beta}G_t(\theta_t) - v_t\| \to 0$. This only bounds a difference, but does not imply directional alignment. since $\|v_t\|$ may vanish at the same time. To compare the residual against the extracted component, define $\kappa_t := \frac{(1-c_t)D_t}{c_t\|v_t\|}$.

Corollary 1. *If $\kappa_t < 1$, then $\langle v_t, G_t(\theta_t)\rangle \geq \frac{\beta c_t}{2}(1-\kappa_t)\|v_t\|^2 > 0$ and $\cos(v_t, G_t(\theta_t)) \geq \frac{1-\kappa_t}{1+\kappa_t}$.*

Cor. 1 gives the condition under which the extracted component dominates. Because κ_t compares the residual against $\|v_t\|$, what matters is the ratio, not the size of either term. Under uniformly bounded D_t, $\kappa_t \to 0$ whenever $1 - p_{t,L} = o(\|v_t\|)$, and $G_t(\theta_t)$ gets asymptotically aligned with v_t. Score saturation drives $1 - p_{t,L} \to 0$ but says nothing about $\|v_t\|$, and so does not imply alignment.

Why the top score. Thm. 1 produces a positive component toward score L. Analogous component is not guaranteed for any interior score, which is why the system has the pressure drifting toward the top of the rubric rather than toward an arbitrary score level. For any score r with $0 < p_{t,r} < 1$, define $v_t^{(r)} := \nabla_\theta \mathrm{logit} p_r(\phi_t, \theta)|_{\theta=\theta_t}$. The same calculation as in Eq. (8) gives

$$v_t^{(r)} = \sum_{\ell<r} \alpha_{t,\ell}^{(r)} \nabla_\theta \Delta_{t,r,\ell}(\theta_t) - \sum_{m>r} \alpha_{t,m}^{(r)} \nabla_\theta \Delta_{t,m,r}(\theta_t), \quad \alpha_{t,s}^{(r)} := \frac{p_{t,s}}{1-p_{t,r}},\ \forall s \in [L]/\{r\}. \tag{11}$$

For $r < L$, $v_t^{(r)}$ can carry both positive and negative coefficients, because score r can appear as either the preferred or the rejected score in an accepted pair; the preference allocation therefore need not contain a nonnegative submixture proportional to $v_t^{(r)}$. The top score is the exception: under max–min selection it can only appear on the preferred side.

The results in this section primarily concern the descent direction at the single point θ_t. Whether a completed round realizes that direction as an actual increase in $p_{t,L}$ is studied in Sec. 4.

4 Actor-Judge Framework

A completed DPO round may involve multiple optimization iterates, denoted as $\theta_{t,0} = \theta_t \to \ldots \theta_{t,k} \to \ldots \theta_{t,n_t} = \theta_{t+1}$, on the same preference data and reference policy. Along this trajectory, the general descent direction at $\theta_{t,k}$ is

$$G_t(\theta_{t,k}) = \beta \sum_{m>\ell} \omega_{t,m,\ell} \mathbb{E}_{\Gamma_{t,m,\ell}^{(N)}} \left[\sigma\left(-\beta\Delta_t\left(\theta_{t,k}\right)\right) \nabla_\theta \Delta_t(\theta_{t,k})\right]. \tag{12}$$

At round initialization $\theta_{t,0} = \theta_t$, all reference-relative margins vanish, $\Delta_t(\theta_t) = 0$, so the logistic weights satisfy $\sigma\left(-\beta\Delta_t\left(\theta_t\right)\right) = 0.5$. At later iterates, these weights and the margin gradients change, even though the preference allocation remains fixed. The extensions to within-round iterates and more discussions are provided in Appx. D. In this section, we characterize the score change between the round's endpoints and distinguish its actor and judge contributions.

4.1 Actor-Judge Decomposition

Definition 2. *Define* $\mathrm{logit}(z) = \log(\frac{z}{1-z})$. *The **actor gain** is* $\mathcal{A}_t = \mathrm{logit}(\tilde{p}_{t,L}) - \mathrm{logit}(p_{t,L})$ *and the **judge gain** is* $\mathcal{J}_t = \mathrm{logit}(p_{t+1,L}) - \mathrm{logit}(\tilde{p}_{t,L})$.

The definition follows an actor-first evaluation path. $\mathcal{A}_t$ measures the score change induced by updating the actor from θ_t to θ_{t+1} while holding the judge fixed at ϕ_t, and $\mathcal{J}_t$ measures the additional change induced by updating the judge from ϕ_t to ϕ_{t+1} while holding the post-update actor fixed. The two therefore separate a score change that comes from *the actor adapting to the current scoring rule* from one that comes from *the same response distribution being evaluated differently*. Together they decompose a completed round exactly, $\mathrm{logit}(p_{t+1,L}) - \mathrm{logit}(p_{t,L}) = \mathcal{A}_t + \mathcal{J}_t$, without requiring that the two roles be optimized separately.

The actor gain is the round-level counterpart of the local quantity studied in Sec. 3. For small $d_t = \theta_{t+1} - \theta_t$, a Taylor expansion gives $\mathcal{A}_t = \langle v_t, d_t\rangle + o(\|d_t\|)$; in particular, for a single gradient-descent step $d_t = \eta G_t(\theta_t)$, $\mathcal{A}_t = \frac{\eta\beta}{2}\left(c_t\|v_t\|^2 + \langle v_t, R_t\rangle\right) + o(\|d_t\|)$, so the decomposition

in Eq. (9) directly determines the actor gain for a sufficiently small single step. More results of relationship between round-level actor gains and local geometry are shown in Appx. D.4. For a general completed update, the endpoint analysis below characterizes the realized gain without assuming that the initial direction governs the entire trajectory. Define the cumulative gain $\Lambda_T = \sum_{t=0}^{T-1}(\mathcal{A}_t + \mathcal{J}_t)$.

Corollary 2. *The model is δ-score saturated at round T if and only if* $\Lambda_T \geq \log \frac{1-\delta}{\delta} - \mathrm{logit}(p_{0,L})$.

Cor. 2 gives an exact cumulative-gain threshold for entering the score-saturation regime. Actor and judge gains may offset one another, and gains that are positive but summable need not produce asymptotic saturation. We next give sufficient conditions for each to be positive.

4.2 Actor gain

For scores $m > \ell$, define the between-score probability change $r_{t,m,\ell} := \log \frac{\tilde{p}_{t,m}/\tilde{p}_{t,\ell}}{p_{t,m}/p_{t,\ell}}$, which measures how the actor update changes the relative probability of receiving score m versus score ℓ under the fixed judge. By Def. 2,

$$\mathcal{A}_t = -\log \sum_{\ell<L} \alpha_{t,\ell} \exp\left(-r_{t,L,\ell}\right), \quad \alpha_{t,\ell} := \frac{p_{t,\ell}}{1-p_{t,L}}. \tag{13}$$

The actor gain is therefore determined by the collection of top-versus-lower between-score probability changes $\{r_{t,L,\ell}\}_{\ell<L}$. In particular, $\min_{\ell<L} r_{t,L,\ell} \leq \mathcal{A}_t \leq \max_{\ell<L} r_{t,L,\ell}$, so $r_{t,L,\ell} > 0$ for every $\ell < L$ is sufficient for a positive actor gain; although it is not necessary, since probability changes at different scores can compensate in the aggregate.

To connect these between-score probability changes to the margins DPO optimizes, define $K_{t,s} := D_{\mathrm{KL}}(\pi_t^s \| \tilde{\pi}_t^s), s \in \{1, .., L\}$ which measures the redistribution of probability within score level s induced by the actor update.

Proposition 4. *For any two scores $m > \ell$, $r_{t,m,\ell} = \Delta_{t,m,\ell}(\theta_{t+1}) + K_{t,m} - K_{t,\ell}$.*

The conditional margin $\Delta_{t,m,\ell}(\theta_{t+1})$ measures how strongly the update favors responses with score m over those with score ℓ. The term $K_{t,m} - K_{t,\ell}$, by contrast, captures asymmetric redistribution *within* the two score-conditioned response distributions. Improving a DPO preference margin therefore does not by itself imply an increase in the probability ratio between two score levels.

When is the actor gain positive? For top-versus-lower score pairs, Prop. 4 combined with Eq. (13) separates two factors: how strongly DPO fits the top-versus-lower score pairs, and how much within-score redistribution reinforces or offsets that improvement. We relate the first factor to the DPO objective. Recall that $\varepsilon_{t,m,\ell}(\theta)$ is the conditional DPO loss on pairs with scores (m, ℓ). By convexity of the logistic loss and Eq. (6), $\Delta_{t,m,\ell}(\theta) \geq \frac{1}{\beta}\Phi(\varepsilon_{t,m,\ell}(\theta))$, $\Phi(z) = -\log(e^z - 1)$. Since $\mathcal{L}_t^{\mathrm{DPO}}(\theta) = \sum_{m>\ell} \omega_{t,m,\ell}\varepsilon_{t,m,\ell}(\theta)$, every score pair with $\omega_{t,m,\ell} > 0$ satisfies $\varepsilon_{t,m,\ell}(\theta) \leq \frac{\mathcal{L}_t^{\mathrm{DPO}}(\theta)}{\omega_{t,m,\ell}}$, and because Φ is monotonically decreasing, $\Delta_{t,m,\ell}(\theta) \geq \frac{1}{\beta}\Phi\left(\frac{\mathcal{L}_t^{\mathrm{DPO}}(\theta)}{\omega_{t,m,\ell}}\right)$. A small global DPO loss therefore controls a particular conditional margin tightly only when the corresponding score pair receives sufficient preference allocation. This gives a lower bound on the actor gain.

Theorem 2. *Actor gain $\mathcal{A}_t$ can be lower bounded:*

$$\mathcal{A}_t \geq K_{t,L} - \log \sum_{\ell<L,\ p_{t,\ell}>0} \alpha_{t,\ell} \exp\left[K_{t,\ell} - \frac{1}{\beta}\Phi\left(\frac{\mathcal{L}_t^{DPO}(\theta_{t+1})}{\omega_{t,L,\ell}}\right)\right].$$

Moreover, $\mathcal{A}_t > 0$ if the following holds for every $\ell < L$ and $p_{t,\ell} > 0$

$$\frac{1}{\beta}\Phi\left(\frac{\mathcal{L}_t^{DPO}(\theta_{t+1})}{\omega_{t,L,\ell}}\right) > K_{t,\ell} - K_{t,L}.$$

Thm. 2 shows when preference fitting suffices to produce a positive actor gain. The condition reflects a competition between the learned top-versus-lower margins and within-score redistribution: a smaller DPO loss and a larger preference allocation $\omega_{t,L,\ell}$ tighten the control on the relevant conditional margins, while $K_{t,\ell} - K_{t,L}$ captures redistribution that may offset it.

The preference allocation $\omega_{t,m,\ell}$ therefore plays a complementary role across our analysis. By Thm. 1 it determines how the conditional margin gradients contribute to the round-start descent direction; after multi-step optimization, it determines how strongly global preference fitting constrains the margins relevant to the top-score probability.

Note that $\mathcal{A}_t > 0$ means only that the updated model is more likely to receive the top score under the *fixed* judge ϕ_t. In the self-judge setting the shared update can additionally alter the scoring rule itself, which gives rise to the judge gain.

4.3 JUDGE GAIN

In the fixed judge setting, $\mathcal{J}_t = 0$. In the self-judge setting, judge gain is not produced by a separate judge-training objective. It arises because the completed model update changes the parameters used for scoring in the next round. Thus, we evaluate judge sensitivity while holding the post-update actor distribution fixed. At round t, for a given prompt x, let

$$u_t^J := \nabla_\phi \mathbb{E}_{\pi_{t+1}} \left[\mathbb{P} \left(S_\phi \left(x, Y \right) = L \right) \right] |_{\phi=\phi_t}. \tag{14}$$

When is the judge gain positive? u_t^J describes the parameter direction to which the judge's top-score probability is locally sensitive; it allows us to determine the sign of judge gain as follows.

Theorem 3. *At round t, for a given prompt x, suppose $\mathbb{E}_{\pi_{t+1}} \left[\mathbb{P} \left(S_\phi \left(x, Y \right) = L \right) \right]$ is twice differentiable with Hessian norm bounded by L_t^J. For $\triangle\phi_t = \phi_{t+1} - \phi_t = \theta_{t+1} - \theta_t = \triangle\theta_t$,*

$$\text{if } \langle u_t^J, \triangle\phi_t \rangle > \frac{L_t^J}{2} \|\triangle\phi_t\|^2, \ \mathcal{J}_t > 0; \quad \text{if } \langle u_t^J, \triangle\phi_t \rangle < -\frac{L_t^J}{2} \|\triangle\phi_t\|^2, \ \mathcal{J}_t < 0. \tag{15}$$

When u_t^J and $\triangle\phi_t = \triangle\theta_t$ are nonzero, define $\rho_t := \cos(u_t^J, \triangle\theta_t) = \frac{\langle u_t^J, \triangle\theta_t \rangle}{\|u_t^J\| \|\triangle\theta_t\|}$. Thm. 3 gives a sufficient condition $\rho_t > \frac{L_t^J \|\triangle\theta_t\|}{2\|u_t^J\|}$ for $\mathcal{J}_t > 0$. Parameter sharing in self-judge setting therefore permits an additional score-change contribution but does not fix its sign. Whether the judge reinforces or offsets actor-induced concentration depends on the alignment of the update with judge sensitivity, and on curvature bounded by L_t^J. For a single step update, $\triangle\theta_t = \eta G_t(\theta_t)$, the condition reduces to an alignment criterion between judge sensitivity and the round-start DPO descent direction.

5 PREFERENCE SUPPLY

The preceding analysis conditions on the preference data constructed at a given round and studies how optimizing these preferences changes the score distribution. In an iterative system, however, the updated actor and judge determine the score distribution from which the next round's preferences are constructed. The score changes characterized in Sec. 4 therefore feed back into both the composition and the availability of future preference data. We quantify this effect through *preference supply* V_t, defined as the number of accepted non-tied pairs at each round t.

Suppose B_t candidate groups are generated independently at round t, with $N \geq 2$ candidates per group. Under max-min selection, a candidate group is accepted exactly when its candidate scores are not all equal. Therefore, the acceptance probability of a candidate group, denoted as $a_{N,t}$, satisfies $a_{N,t} = 1 - \sum_{\ell=1}^{L} p_{t,\ell}^N$ and hence $\mathbb{E}[V_t] = B_t a_{N,t}$.

Since $a_{N,t} \geq 1 - (\max_{\ell \in [L]} p_{t,\ell})^{N-1} \sum_{\ell=1}^{L} p_{t,\ell} = 1 - (\max_{\ell \in [L]} p_{t,\ell})^{N-1}$, a score distribution that stays uniformly away from concentration on any single score (i.e., $\max_{\ell \in [L]} p_{t,\ell} \leq 1 - \zeta$ for some $\zeta > 0$) gives $a_{N,t} \geq 1 - (1-\zeta)^{N-1} > 0$. Under a fixed candidate-group budget B_t, the preference supply depletion—meaning $\mathbb{E}[V_t] \to 0$—therefore cannot occur as long as the score distribution remains uniformly non-concentrated. Note also that depletion is not specific to the highest score: concentration on *any* single score makes non-tied candidate groups increasingly rare.

In this section, we focus on top-score saturation because it is the concentration pattern the optimization of the previous sections actually produces. It also has a structural feature the other cases lack: whenever the top score appears in an accepted pair it can only appear on the preferred side, and this one-sided structure lets the remaining preference allocation near saturation be related directly to the gradient mixture defining v_t.

Proposition 5. *For fixed $N \geq 2$ and sufficiently small δ, under δ-score saturation, $a_{N,t} = N(1 - p_{t,L}) + \mathcal{O}(\delta^2) = \mathcal{O}(\delta)$, and $\omega_{t,L,\ell} = \alpha_{t,\ell} + \mathcal{O}(\delta)$ for every $\ell < L$.*

Prop. 5 establishes the connection between the accepted preference data and the top-score gradient. Conditional on acceptance, the preference allocation of score pair (L, ℓ) satisfies $\omega_{t,L,\ell} \approx \alpha_{t,\ell}$. By Eq. (8), the same $\alpha_{t,\ell}$ weights the conditional margin gradient in the decomposition of v_t. Hence, near saturation, the distribution of accepted score-pair types increasingly matches the coefficient structure of the top-score gradient. Together with the vanishing allocation on non-top score pairs, this makes $G_t(\theta_t)$ approach $\frac{\beta}{2} v_t$ in absolute norm while directional alignment additionally requires the relative residual to vanish as discussed in Sec. 3. At the same time, the probability of obtaining non-tied pairs vanishes at rate $N(1 - p_{t,L})$. Hence, the preference data become better matched, at the allocation level, to the mixture defining v_t, while the preference supply decreases. We next connect the supply reduction to cumulative dynamics and candidate generation budget.

Theorem 4. *For fixed $N \geq 2$, at round t, i) Assume $0 < p_{0,L} < 1$, then $\mathbb{E}[V_t] \leq \frac{NB_t}{1 + \frac{p_{0,L}}{1 - p_{0,L}} e^{\Lambda_t}}$. ii) Let $B_t^*(M) := \min\{B \in \mathbb{N} : Ba_{N,t} \geq M\}, M > 0$ be the minimum candidate-group budget required to obtain at least M accepted preference pairs in expectation. If the cumulative gain $\Lambda_t \to \infty$, then $\frac{B_t^*(M)}{\exp(\Lambda_t)} \to \frac{M p_{0,L}}{N(1 - p_{0,L})}$. iii) If the model is δ-score saturated, then $B_t^*(M) \geq \frac{M}{N\delta}$.*

Thm. 4 quantifies the preference supply bottleneck. Under a fixed budget B_t, the upper bound of expected number of accepted preference pairs decreases as cumulative gains increase. Conversely, maintaining a fixed expected preference supply requires a candidate-group budget that grows asymptotically as $B_t^*(M) \asymp \exp(\Lambda_t)$, or inversely with the remaining non-top-score probability.

Our analysis now closes the loop. Because max–min selection places the top score only on the preferred side, the accepted pairs give the round-start update a positive component along the direction that raises the top-score probability; as the resulting actor and judge gains accumulate toward score saturation, the changed score distribution both reshapes the surviving pairs and reduces the preference supply available to subsequent rounds.

6 EXPERIMENT

We use Llama-3-8B-Instruct (Grattafiori et al., 2024) with the OASST corpus (Köpf et al., 2023): of its 3466 ranked groups, 531 are held out for fixed human-ranked evaluation and the remaining 2935 drive iterative training. Evaluator SFT first adapts the model to the rubric-based judging task and the 1–5 scoring format, giving θ_0; we then train 8 rounds using DPO under both the self-judge and the fixed-judge ($\phi_t = \theta_0$) settings of Sec. 2, with $N = 7$ candidates per prompt. The rubric, hyperparameters, probe prompts, and all estimation procedures are in Appx. C.

Round-start top-score reinforcement (Sec. 3). We examine the local geometry at θ_0 on four prompts x_0–x_3. Because $N = 7$ samples per prompt make the induced estimates highly variable, we draw 128 responses per prompt from θ_0, scored by the same judge $\phi_0 = \theta_0$. Consistent with Thm. 1, the extracted component $c_t v_t$ is nonzero for all four prompts and the residual projection $\langle v_t, R_t \rangle$ is positive throughout, so the residual reinforces rather than offsets it (Fig. 1A); all four descent directions therefore satisfy $\cos(v_t, G_t) \in [0.72, 0.92] > 0$ (Fig. 1C). The guaranteed component need not dominate: for several prompts the residual contribution exceeds $c_t \|v_t\|^2$, which is the sense in which Thm. 1 identifies a structural pressure rather than an equivalent optimization objective. The relative-residual condition $\kappa_t < 1$ of Cor. 1 holds for only one prompt (Fig. 1B), confirming that it is sufficient but not necessary; that prompt, x_2, also has the largest cosine and c_t and the smallest $\langle v_t, R_t \rangle$, so stronger top-score components with weaker residual interference accompany stronger alignment. The discrepancy $\| \frac{2}{\beta} G_t - v_t \|$ remains substantial (Fig. 1D), consistent with our distinction between directional alignment and absolute approximation.

Score saturation (Sec. 4). Self-judging enters the high-score saturation regime substantially earlier than fixed-judging (Fig. 1E), and not through uniformly stronger actor adaptation: after the first update both settings have actor gains near 0.5, whereas self-judging additionally exhibits a large judge gain $\mathcal{J}_0 \approx 3.4$ (Fig. 1G). Later gains of both kinds can turn negative, producing the non-monotone dynamics in Fig. 1E. Early concentration under self-judging is therefore driven by judge-induced reassessment of the post-update response distribution as well as by actor adaptation. Saturation is

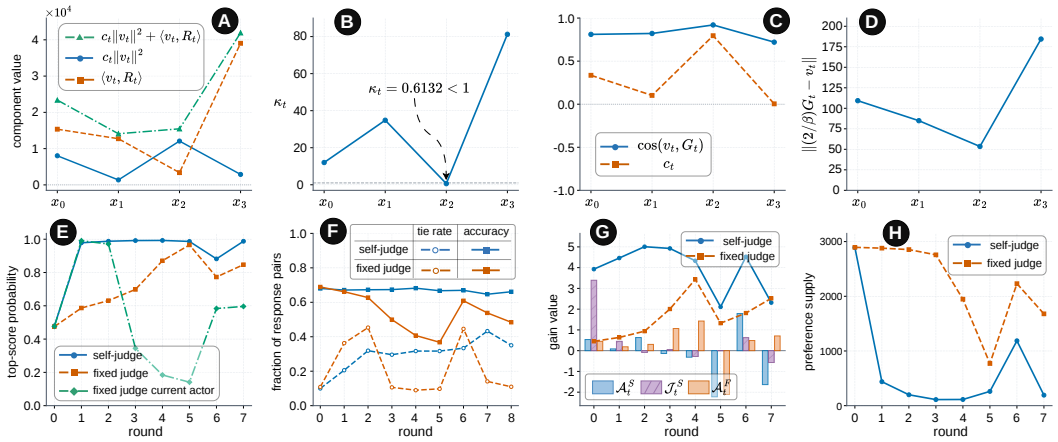


Figure 1: (A): $v_t G_t$ and its components values obtained from Thm. 1 for 4 prompts. (B): κ_t in Cor. 1 for 4 prompts. (C): c_t and the cosine reflecting the alignment between the G_t and the direction improving the top-score probability for 4 prompts. (D): The norm of the difference between two direction vectors for 4 prompts. (E): Top-score probabilities for 3 settings. *fixed judge current actor* means using the model after each round update of the fixed judge setting to score the responses generated by itself in the same round. (F): The accuracy and tie rate of models before the update from different rounds evaluated on a fixed evaluation set. (G): The actor/judge gains (bar) and Λ_{t+1} (line). The superscripts in $\mathcal{A}_t^S, \mathcal{A}_t^F, \mathcal{J}_t^S$ indicate whether the setting is self-judge (S) or fixed judge (F). $\mathcal{J}_t^F = 0$ for any t. (H): The preference supply in each round for two settings.

also judge-relative and on-policy in self-rewarding. In the fixed judge setting, scoring the current actor's responses with the actor itself (*fixed judge current actor*, Fig. 1E) yields a markedly different trajectory from that under the frozen judge, showing that high scores under the training rule need not translate into similarly high self-scores. Evaluation on the fixed human-response set (Fig. 1F) provides a complementary diagnostic. While the self-judge trajectory exhibits an overall increase in tie rate, the current models from the fixed-judge trajectory, when repurposed as judges, show a strongly non-monotonic pattern on the same responses. Therefore, the concentration is specific to the deployed feedback loop, not a model-wide tendency to score high or tie.

Reduction in preference supply (Sec. 5). Under self-judging, the preference supply falls from $V_t = 2892$ to 111 before partially recovering as the score distribution becomes less concentrated (Fig. 1H). Across all rounds, self-judging yields markedly less preference data than fixed-judging, consistent with the predicted feedback tension: concentration of the score distribution reduces the available preference data, while subsequent dispersion can restore it. Fig. 2 further shows that, under a fixed candidate-generation budget, the observed preference supply is bounded by the remaining non-top-score mass, consistent with Thm. 4(i): $\mathbb{E}[V_t] \leq NB_t\big(1 + \frac{p_{0,L}}{1-p_{0,L}} e^{\Lambda_t}\big)^{-1}$. Thus, as cumulative gains drive the system toward top-score saturation, the attainable preference supply decreases even when the candidate-generation budget is held fixed.

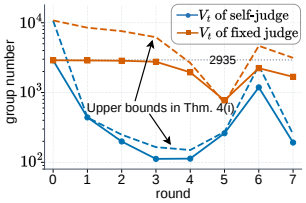


Figure 2: Preference supply V_t under fixed budget $B_t = 2935$ (solid), against the upper bounds of Thm. 4(i) (dashed), for the self-judge and fixed-judge settings.

Additional results. We further validate the generation budget and allocation matching. Due to space limit, more details are provided in Appx. C.2 and C.3. The sampling-budget diagnostics are consistent with the bounds and near-saturation scaling in Thm. 4(ii)(iii) (Fig. 3B,C). Large positive cumulative gains require substantially larger budgets, although this relation need not be globally monotone at small finite gains (Fig. 3A). For preference allocation, using the four prompts $x_0 - x_3$, two prompts with estimable non-top mass exhibit smaller discrepancies between $\omega_{t,L,\ell}$ and $\alpha_{t,\ell}$, while the other two are unidentifiable since no non-top samples remain (Fig. 4). These results support the predicted feedback tension. As $p_{t,L} \to 1$, the remaining preferences increasingly reflect the top-score gradient structure precisely while such preferences become scarce.

More discussions about limitations and future works are presented in Appx. B.

AI USE STATEMENT

In this work, we used generative AI tools for limited assistance. The theoretical framework, mathematical claims, and proofs were developed and completed by the authors. Generative AI was used only to review the proofs and help identify potential issues. For the experiments, generative AI was used to assist with writing and editing code. Generative AI was also used to improve the clarity and readability of the manuscript. We did not use generative AI for the other required-disclosure tasks, or the remaining required-disclosure tasks were not applicable to this work. We have reviewed all AI-assisted work before incorporating them into the final work. We take responsibility for the final content of this work, including all text, claims and other artifacts produced with the aid of generative AI.

REFERENCES

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Shi Fu, Yingjie Wang, Shengchao Hu, Peng Wang, and Dacheng Tao. Why self-rewarding works: Theoretical guarantees for iterative alignment of language models. *arXiv preprint arXiv:2601.22513*, 2026.

Etienne Gauthier, Francis Bach, and Michael I Jordan. Explaining and preventing alignment collapse in iterative rlhf. *arXiv preprint arXiv:2605.04266*, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Lin Gui, Cristina Gârbacea, and Victor Veitch. Bonbon alignment for large language models and the sweetness of best-of-n sampling. *Advances in Neural Information Processing Systems*, 37: 2851–2885, 2024.

Seungone Kim, Jay Shin, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Ryan Shin, Sungdong Kim, James Thorne, Minjoon Seo, et al. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations*, volume 2024, pp. 29927–29962, 2024a.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4334–4353, 2024b.

Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. *Advances in neural information processing systems*, 36:47669–47681, 2023.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797, 2025.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. Rlaif vs. rlhf: scaling reinforcement learning from human feedback with ai feedback. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, 2024.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language*

Processing, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

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

Arjun Panickssery, Samuel R Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

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

Aadirupa Saha, Aniket Wagde, and Branislav Kveton. Llm-as-a-judge on a budget. In Emtiyaz Khan, Yingzhen Li, Arno Solin, and Aaditya Ramdas (eds.), *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 4375–4383. PMLR, 02–05 May 2026. URL https://proceedings.mlr.press/v300/saha26a.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Avi Singh, John D Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J Liu, James Harrison, Jaehoon Lee, Kelvin Xu, et al. Beyond human data: Scaling self-training for problem-solving with language models. *arXiv preprint arXiv:2312.06585*, 2023.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*, 2024.

Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Ávila Pires, Michal Valko, Yong Cheng, et al. Understanding the performance gap between online and offline alignment algorithms. *arXiv preprint arXiv:2405.08448*, 2024.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9440–9450, 2024a.

Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. Self-taught evaluators. *arXiv preprint arXiv:2408.02666*, 2024b.

Yidong Wang, Xin Wang, Cunxiang Wang, Junfeng Fang, Qiufeng Wang, Jianing Chu, Xuran Meng, Shu-Xun Yang, Andrew Zhuoer Feng, Libo Qin, Wei Ye, and Shikun Zhang. Temporal self-rewarding language models: Decoupling chosen-rejected via past-future. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=8G0XzmQRG9.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 13484–13508, 2023.

Zhaoyang Wang, Weilei He, Zhiyuan Liang, Xuchao Zhang, Chetan Bansal, Ying Wei, Weitong Zhang, and Huaxiu Yao. Cream: Consistency regularized self-rewarding language models. In *International Conference on Learning Representations*, volume 2025, pp. 57017–57039, 2025.

Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason E Weston, and Sainbayar Sukhbaatar. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11548–11565, 2025.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024. URL https://openreview.net/forum?id=w3oJXMAQx2.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. Self-rewarding language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57905–57923. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/yuan24d.html.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges. In *International Conference on Learning Representations*, volume 2025, pp. 51257–51296, 2025.

A Related Work

Self-generated feedback and iterative self-improvement. Learning from human feedback combines demonstrations and preference judgments to adapt language models to user instructions, as exemplified by InstructGPT (Ouyang et al., 2022). Constitutional AI and RLAIF extend this approach by using AI-generated feedback for alignment, reducing reliance on human preference annotations (Bai et al., 2022; Lee et al., 2024).

Self-generated data also support improvement through mechanisms other than self-judging. Self-Instruct generates and filters instruction-following examples (Wang et al., 2023), whereas STaR iteratively trains on generated rationales associated with correct answers (Zelikman et al., 2022). ReST$^{\mathrm{EM}}$ applies an expectation-maximization-inspired self-training procedure to problem-solving tasks using correctness feedback (Singh et al., 2023). These approaches therefore differ in the

source of supervision used to select or compare generated responses, rather than constituting a single feedback-generation mechanism.

The closest algorithmic setting is self-rewarding language models, which use the model as a rubric-based judge and construct preference pairs for iterative DPO (Yuan et al., 2024). Subsequent work modifies different parts of this loop. Meta-Rewarding trains judgment capabilities through meta-judgments (Wu et al., 2025), while CREAM regularizes self-generated preferences using consistency across iterations (Wang et al., 2025). Temporal Self-Rewarding constructs chosen and rejected responses across model iterations to address the diminishing contrast associated with their synchronized improvement (Wang et al., 2026). Our work complements these improvement-oriented approaches by analyzing the feedback-generation mechanism of discrete-score, max–min DPO itself.

Rubric-Based judges and evaluation reliability. LLM-as-a-judge research studies both the construction of evaluators and their agreement with human assessments. MT-Bench and Chatbot Arena examine the capabilities and biases of strong LLM judges (Zheng et al., 2023), while G-Eval uses structured, chain-of-thought-based evaluation to improve agreement with human judgments (Liu et al., 2023). Prometheus develops fine-grained evaluation conditioned on user-provided rubrics (Kim et al., 2024a), and Prometheus 2 supports both direct assessment and pairwise ranking (Kim et al., 2024b). JudgeLM studies scalable judge fine-tuning and techniques for reducing several evaluation biases (Zhu et al., 2025). Self-Taught Evaluators uses synthetic data to iteratively improve evaluation without human preference annotations (Wang et al., 2024b).

Reliability studies identify position bias in pairwise evaluation (Wang et al., 2024a), as well as familiarity bias, skewed rating distributions, anchoring effects, and inconsistency across samples or prompts (Stureborg et al., 2024). Self-preference introduces a further concern when evaluators recognize and favor their own generations (Panickssery et al., 2024). RewardBench (Lambert et al., 2025) provides structured preferred–rejected comparisons for assessing reward models across chat, reasoning, and safety settings. Our object is distinct from whether an evaluator ranks a fixed collection of responses correctly. We analyze the on-policy score distribution jointly induced by the current actor and judge, and how its concentration affects preference supply under max-min selection. A high top-score probability need not imply inaccurate judgments, since distinct, genuinely strong responses can share the same highest rubric score and consequently fail to produce a non-tied training pair.

Coupled alignment and likelihood dynamics. Recent theory explicitly studies evolving feedback mechanisms. Gauthier et al. (2026) analyze iterative RLHF through coupled policy and reward-model dynamics, explaining how policy updates can influence future reward-model errors and proposing a foresighted update to mitigate alignment collapse. Fu et al. (2026) provide finite-sample guarantees for iterative self-rewarding under explicit realizability, confidence, and margin assumptions in a likelihood-based self-reward formulation. These works establish that the theoretical literature includes both failure mechanisms and conditions for successful iterative alignment; the distinction of our work is therefore not simply that the evaluator or feedback distribution can change. Instead, we study a different measurable outcome that concentration at a discrete rubric's highest score and the resulting depletion of accepted comparisons.

Importantly, improving a relative preference margin does not generally imply increasing the likelihood of the preferred response. Smaug (Pal et al., 2024) identifies such likelihood decreases and proposes DPO-Positive to address them. Razin et al. (2025) analyze likelihood displacement, in which probability can move away from both responses in a preference pair, and relate this behavior to the representations of the compared responses. Tang et al. (2024) further distinguish accurate classification of offline preference pairs from successful generation and examine the role of on-policy sampling in the performance gap between online and offline alignment. Our analysis preserves this distinction between preference fitting and probability change. The resulting contribution concerns the probability of a score-defined class of responses, rather than monotonic likelihood improvement for individual preferred responses.

Preference data construction and sampling efficiency. The generation and selection of preference data are central to iterative alignment. Xiong et al. (2024) study offline, online, and hybrid preference learning under a KL constraint, connecting theoretical guarantees to iterative preference optimization. BoNBoN analyzes best-of-n sampling and trains models to approximate its response distribution, reducing reliance on repeated best-of-n selection at inference time (Gui et al., 2024).

Related sampling issues arise in group-based reinforcement learning. Shao et al. (2024) introduce Group Relative Policy Optimization (GRPO), which estimates a baseline from rewards within a group of sampled responses. DAPO (Yu et al., 2026) explicitly notes that groups with identical binary correctness rewards provide no relative policy-gradient signal, and uses dynamic sampling to discard all-correct or all-incorrect groups while continuing to sample until a batch of effective prompts is obtained. Thus, the loss of within-group contrast and the practical use of additional sampling are already recognized phenomena.

B Discussion

Our theoretical analysis and experiments examine how score-based preference optimization changes both the score distribution and the preference feedback available to subsequent rounds in the self-rewarding LLMs. We distinguish an allocation-based mechanism at round initialization, the actor and judge gains realized by a completed update, and the resulting changes in preference supply. These distinctions also clarify the scope of the present results and suggest several possible future directions for extending the analysis toward more predictive and reliable self-rewarding systems.

B.1 Limitations

Round-start local geometry and completed-round dynamics. The top-score reinforcement analysis in Sec. 3 isolates an allocation-determined component of the expected DPO descent direction at round initialization, when all reference-relative margins vanish and the logistic weights are identical. As shown in Appx. D, this component need not dominate the residual, and its presence does not imply that every subsequent optimization iterate increases the top-score probability. The actor-judge framework in Sec. 4 already accommodates multi-step updates. Its exact endpoint decomposition and actor-gain characterization do not require the initialization decomposition to remain unchanged throughout training.

Quantities and prompt-conditional analysis. The formulation incorporates finite candidate groups and stochastic scoring, but the analytical results concern population score probabilities, preference allocations, expected descent directions, and expected preference supply. We do not establish non-asymptotic guarantees for their empirical estimates or for the discrepancy between optimization on sampled preferences and the corresponding population objective. The independent-candidate-set comparisons and bootstrap diagnostics in experiments provide empirical evidence and uncertainty assessments under the evaluated settings, rather than finite-sample guarantees for the complete iterative procedure. Moreover, our theory is mainly conditional on a fixed prompt. The endpoint identities can be evaluated prompt by prompt for a realized shared-parameter update, but they do not by themselves characterize interactions among prompt-specific gradients or changes in the representation of prompts within accepted training data.

Score saturation versus externally assessed quality. Score saturation describes concentration in the on-policy actor–judge score distribution. It does not, by itself, identify the quality of generated responses, the correctness of accepted preferences, or the judge's agreement with an external evaluation criterion. The fixed human-response evaluation in Fig. 1F examines a different object from the on-policy score probabilities in Fig. 1E, and the reported trajectories illustrate that these diagnostics need not evolve together. These experiments support the distinction between score concentration and judgment performance, but they do not provide a theoretical account of their joint evolution. Thus, the present results diagnose how non-tied preference supply can diminish; they do not establish that reducing score concentration necessarily improves response quality or judgment accuracy.

Different analytical roles of actor and judge gains. The actor and judge gains have different connections to the training objective. DPO directly optimizes relative likelihood margins on the constructed preferences, allowing the actor gain to be related to conditional margins, preference allocation, and within-score redistribution. In the self-judge setting, the judge gain instead arises because the updated parameters are subsequently used for scoring under role-specific prompting; the analyzed procedure contains no separate judge-training objective that constrains this change. Accordingly, Thm. 3 characterizes judge gain through the alignment of the completed update with judge sensitivity and a curvature bound. This analytical asymmetry reflects the different roles of the two quantities, rather than an assumption that judge-induced changes are weaker. The current

formulation does not specify additional structure linking generation and scoring behavior that would determine this alignment from the DPO objective alone.

B.2 Future directions

Trajectory-level analysis and predictive conditions for saturation. In Appx. D, we extend the relationship between DPO descent directions and the top-score gradient throughout within-round optimization. As shown in Appx. D, two effects require explicit treatment. First, the logistic factors become pair-dependent and reweight individual margin gradients, including gradients within the same score pair. Second, the current policy changes while the constructed preference data and reference policy remain fixed. Tracking these effects together with the residual geometry could connect the initialization mechanism more directly to the accumulated actor gain. Retaining conditional-loss information, rather than controlling every relevant margin through the global DPO loss, could also yield sharper sufficient conditions when preference allocation is uneven.

Combining such trajectory-level estimates with judge dynamics would enable more predictive conditions for sustained concentration or non-saturation. The objective would be to identify properties of the model and training procedure that force Λ_t to diverge, keep it below a specified saturation threshold, or permit repeated entry into and exit from the saturation regime. This requires more than establishing positive gains at individual rounds. Positive but summable gains need not produce asymptotic saturation. An end-to-end analysis should also account for the feedback from diminishing preference supply to subsequent optimization, including updates based on very small accepted batches or additional sampling triggered by insufficient data. Such results could distinguish conditions for reaching saturation from conditions for remaining there.

Statistically reliable estimation across prompts and rounds. As mentioned in Appx. B.1, finite-sample extensions could develop state-dependent uncertainty bounds for score probabilities, preference allocation, conditional gradients, and actor-judge gains. These bounds should reflect the scarcity of non-top-score observations rather than assume uniform estimation accuracy throughout training. An important practical distinction is between a score level having zero population probability and simply not appearing in a finite candidate set. Sequential sampling procedures could use this distinction to allocate additional measurements where the available data cannot support the relevant conditional estimates.

Based on the analysis in Sec. 5, a complementary direction is to strengthen the expected-supply characterization with high-probability guarantees under adaptive budgets and evolving policies. Extending the analysis to a prompt distribution would additionally require tracking prompt-dependent acceptance and interactions among gradients under shared-parameter training. The quantities being aggregated should be specified explicitly: averaging prompt-conditional log-odds gains is generally different from computing gains from prompt-averaged score probabilities. An analysis that preserves this distinction could connect prompt-level mechanisms to population-level measurements while characterizing how rejection changes the composition of subsequent training data.

Interventions for reliable and computationally efficient feedback. The framework suggests several targets for intervention, including quality-supervised judge updates, constraints on undesirable changes in scoring behavior, and adaptive preference construction. These possibilities should be evaluated through their effects on actor gain, judge gain, preference supply, and externally assessed performance, rather than through score dispersion alone. Fixing the judge removes the judge-gain contribution but does not prevent actor-induced concentration. Moreover, suppressing top-score concentration is not sufficient to prevent preference depletion if probability instead concentrates on another score level. An effective intervention should therefore consider the entire score distribution while preserving the correctness of the distinctions it introduces.

Recovering accepted-pair counts should also be distinguished from improving feedback quality. Increasing the generation budget can compensate for scarce non-tied groups without correcting unreliable judgments or restoring meaningful scoring resolution. For fixed N, Thm. 4 implies that the minimum total number of generated candidates required to maintain a fixed expected preference supply satisfies $NB_t^*(M) \sim \frac{M}{1-p_{t,L}}$ as $p_{t,L} \to 1$. Thus, the inverse-non-top-mass cost remains when the budget is measured in generated candidates rather than candidate groups. Jointly adapting group size or changing the candidate distribution requires a separate analysis of the resulting allocation, acceptance probabilities, and generation and scoring costs. The objective is not to prevent

high scores, but to preserve reliable comparative feedback as generation improves. Progress should therefore be assessed jointly through externally evaluated actor and judge performance, preference correctness, and the computational cost required to sustain learning.

C Experiment Supplements

C.1 Llama-3-8B experiment details and additional experiments

The scoring rubric used in judge model is the following.

> Review the user's instruction and the assistant response using an additive 5-point quality scale.
> Add 1 point if the response is relevant and contains at least some useful information.
> Add a second point if it addresses a substantial part of the instruction, even if incomplete.
> Add a third point if it answers the main request in a useful way.
> Add a fourth point if it is direct, clear, well organized, and helpful, with only minor issues.
> Add a fifth point if it is expertly tailored to the instruction, insightful, concise, and free of distracting material.
> User instruction: {instruction}
> Assistant response: <response> {response} </response>
> Briefly justify the total score in no more than 100 words, then end with exactly this format: Score: <integer from 1 to 5>

Four prompts used in related experiments are as following.

> x_0: Can you tell me if these two pytorch operations are equivalent: "torch.bmm(attention_slice, value)" and "einsum('b i j, b j d → b i d', attention_slice, value)"
>
> x_1: Provide me an example about how to use the pythagorean theorem
>
> x_2: What is the air-speed velocity of an unladen swallow?
>
> x_3: Write a greentext story.
> A greentext is defined as "A popular device used for quotation and storytelling. They began on chan imageboards, where quoting text places a > in front of it, turning it green. This effect can also be done manually. Before long, this feature was put to innovative use. When quoting another post, greentext is frequently used for summary, criticism, and ridicule. Its other major use is storytelling, in which a lengthy story is told through fragmentary statements in greentext."
> An example of a greentext story is, "
> >be me
> >23 days ago
> >driving drunk as fuck, run a red light, ram into a fucking cop car at 20 mph
> >cars are not too damaged
> >drunk mind is scared as shit
> >unable think of a reason, stab myself with my pocket knife before the cop comes over
> >hide the knife in the glove box
> >say "I have been stabbed officer"
> >cop gives me a ride to the nearest hospital
> Mfw crisis avoided."

Specific setting of DPO training. We use 3466 ranked groups from OASST dataset Köpf et al. (2023). 531 groups are held out for fixed human-ranked evaluation and the remaining 2935 prompts are used in the iterative experiments.

Before preference optimization, we perform evaluator fine-tuning to obtain a common initial checkpoint θ_0. The evaluator is trained for 2 epochs with learning rate 10^-4, maximum sequence length 2048. During SFT, we first convert the human ranks in OASST into pointwise supervision for rubric-based judging. For a prompt with n ranked responses, we sort the responses from best to worst and map position $i \in \{0, ..., n-1\}$ to a 5-point target score using $s_i = \mathrm{clip}(5 - \mathrm{round}(\frac{4\mathrm{i}}{\mathrm{n}-1}))$. Each response is then converted into an evaluator-format SFT example whose input contains the user instruction, the response to be evaluated, and the 5-point scoring rubric. The target output follows the

judge format and ends with $Score : s_i$. 8896 examples are obtained from the 2935 groups. 7562 of these are used for the SFT training set, and the remaining examples serve as the evaluation set. We train Llama-3-8B-Instruct on the 7562 examples. The resulting evaluator-adapted checkpoint is denoted θ_0 and initializes both the self-judge and fixed-judge settings.

We run 8 DPO rounds for both self-judge and fixed judge settings. At round t, the current model samples 7 responses for each of the 2935 training prompts. In actor side, the temperature is 0.7, top-p is 0.9, and at most 640 newly generated tokens. To reduce sampling variability in the diagnostic judge evaluation (Saha et al., 2026), each response is evaluated 5 times using the fixed scoring rubric, with judge temperature 0.6, top-p 0.9, and at most 160 generated tokens, and we get 5 scores for each response. We average the valid sampled scores of each candidate before pair construction. The practical training pipeline emits at most one pair per prompt. It requires at least two scoreable candidates and a raw mean-score range of at least 0.01. A chosen response is selected among candidates lying within 0.25 of the maximum mean score using length control, while the rejected response is selected from the minimum-score candidates. Identical selected texts are discarded. Each round trains for 2 epochs with $\beta = 0.1$, learning rate $\eta = 3 \times 10^{-5}$, maximum sequence length 2048, batch size 32. The reference policy and constructed preference dataset are fixed during the within-round optimization.

C.1.1 Supplement of round-start top-score reinforcement experiment

For each prompt $x \in \{x_0, .., x_3\}$, $y_1, ..., y_{128}$ are sampled from $\pi_{\theta_0}(y|x)$ and scored by the same judge $\phi_0 = \theta_0$ with the scoring rubric mentioned above. The corresponding scores are denoted as $s_1, ..., s_{128}$.

For $\ell \in [1, 2, 3, 4, 5]$, $p_{0,\ell}$ is estimated by $\frac{n_\ell}{128}$ where n_ℓ is the number of responses with score ℓ, and therefore, $\alpha_{0,\ell} = \frac{p_{0,\ell}}{1-p_{0,5}}$. Define $P_0[u : v] = \sum_{r=u}^{v} p_{0,r}$. It's easy to check that for $N = 7$ independent samples from the empirical score distribution $\{p_{0,\ell}\}_{\ell=1}^{5}$, the probability that the maximum score is m and the minimum score is $\ell < m$ is

$$\gamma_{0,m,\ell} = P_0[\ell : m]^7 - P_0[\ell + 1 : m]^7 - P_0[\ell : m - 1]^7 + P_0[\ell + 1 : m - 1]^7. \tag{16}$$

As described in the main paper, the max-min selection rejects all-equal-score groups and breaks endpoint ties uniformly over candidate indices. Based on Eq. 16, the preference allocation is estimated by

$$\omega_{0,m,\ell} = \frac{\gamma_{0,m,\ell}}{1 - \sum_{\ell=1}^{5} p_{0,\ell}^7} = \frac{\gamma_{0,m,\ell}}{\sum_{m>\ell} \gamma_{0,m,\ell}},$$

satisfying $\sum_{m>\ell} \omega_{0,m,\ell} = 1$.

We estimate the conditional mean of the response log-probability gradient as

$$g_{0,i} = \nabla_\theta \log \pi_\theta(y_i|x)|_{\theta=\theta_0}, \quad \mu_{0,\ell} = \frac{1}{n_\ell} \sum_{i:s_i=\ell} g_{0,i}.$$

Therefore, the estimated conditional margin gradient is

$$\nabla_\theta \Delta_{0,m,\ell}(\theta_0) = \mu_{0,m} - \mu_{0,\ell}.$$

By Proposition 2, the estimated top-score log-odds gradient and descent direction are

$$v_0 = \sum_{\ell:\, n_\ell>0,\ \ell<5} \alpha_{0,\ell} \nabla_\theta \Delta_{0,m,\ell}(\theta_0) = \mu_{0,5} - \sum_{\ell:\, n_\ell>0,\ \ell<5} \alpha_{0,\ell} \mu_{0,\ell},$$

$$G_0 = \frac{\beta}{2} \sum_{m>\ell} \omega_{0,m,\ell} \nabla_\theta \Delta_{0,m,\ell}(\theta_0).$$

This estimator reconstructs the expected descent direction from allocation weights and conditional gradient means. Moreover, we compute the largest top-score submixture whose lower-score composition matches $\alpha_{0,\ell}$:

$$c_0 = \min_{\ell<5, n_\ell>0} \frac{\omega_{0,5,\ell}}{\alpha_{0,\ell}}.$$

The residual is then

$$R_0 = \frac{2}{\beta} G_0 - c_0 v_0.$$

Equivalently, it's easy to estimate R_t^N and R_t^{NL}. We estimate the maximum conditional gradient by

$$D_0 = \max_{m>\ell} \|\nabla_\theta \Delta_{0,m,\ell}(\theta_0)\|$$

and κ_0 in Corollary 1 by its definition $\kappa_0 = \frac{(1-c_0)D_0}{c_0\|v_0\|}$.

C.1.2 Sampling Estimation Settings

There are two sampling estimation methods used in our experiments. The first one is detailed in Appx. C.1.1 with 128 independent responses. The other one is used in estimating the $p_{t,L}$ in Fig. 1E and its related quantities like gains. We use 2935 prompts and their corresponding responses to get the Monte-Carlo estimate.

Why two sampling estimation methods in experiments. The realized actor and judge gains are endpoint quantities of the actual $N = 7$ training system. Once the trained checkpoints (models) are fixed, they are estimated broadly over 2935 prompts from their endpoint top-score probabilities. In contrast, the round-start mechanism depends explicitly on the N-dependent preference allocation and score-conditioned gradients, so we densely sample each fixed prompt to estimate these quantities while retaining the training-time group size $N = 7$. The 128 response sets A and B used in Sec. 6, C.2 and C.3 improve the Monte-Carlo sampling accuracy and do not change max-min group size. Therefore, the set size of A and B and the max-min group size N play different roles. Increasing the former only reduces Monte-Carlo sampling error in estimating a fixed actor-judge state, while changing the latter changes the preference allocation $\omega_{t,m,\ell}$ and therefore defines a different DPO update direction.

C.1.3 Supplement of Score Saturation and Preference supply experiments

How to calculate tie rate and accuracy in Fig. 1(F). In Fig. 1(F), we evaluate models at each round as the judge on a fixed set of 531 human-ranked response groups from the same dataset OASST Köpf et al. (2023), which is separate from the 2935 prompts in the iterative self-rewarding experiments. A single data point in the OASST dataset has the structure like the following:

"prompt": "Hello", "responses": [{"response": "Hello, my name is Open Assistant. How may I help you today?", "rank": 0}, {"response": "Hello to you as well", "rank": 1}, {"response": "Hi", "rank": 2}]

and lower rank values corresponds to the responses within the same response set that are judged by humans to be better. For each prompt, we construct gold preference pairs from the fixed set responses with strictly different human ranks. Specifically, we list all the valid pair combinations in which two responses have different human ranks for every prompt as the gold pairs.

Next, the model scores every response five times, and we average the scores to obtain a response-level score R. For a human-preferred gold pair (y^+, y^-), we count the comparison as correct if $R(y^+) > R(y^-)$, incorrect if $R(y^+) < R(y^-)$, and tied if $R(y^+) = R(y^-)$. We report the **tie rate**, the fraction of gold pairs receiving equal scores, and **accuracy** which is calculated by assigning scores of 1, 0, and 0.5 to correct, incorrect, and tied comparisons, respectively, summing all the scores and dividing the gold pair number. Thus, the reported accuracy jointly reflects ranking correctness and comparison coverage, rather than accuracy conditioned only on non-tied pairs.

In the following experiments, we select models at round $t \in \{0, 1, 6\}$ to cover qualitatively different score regimes. By Fig. 1E, θ_0 provides the initial unsaturated baseline, θ_1 corresponds to the early highly score-saturated state, and θ_6 captures a later state in which the top-score probability $p_{t,L}$ has substantially decreased again.

C.2 Generation budget theory validation

In our setting, $B_t = 2935$ for any t. Fig. 2 compares historical pair counts under a fixed budget of $B_t = 2935$ candidate groups with the estimated non-top-mass envelope motivated by Thm. 4(i). The

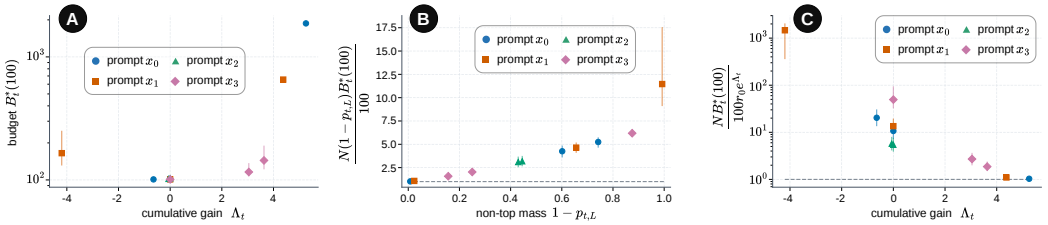


Figure 3: (A): Fix $M = 100$, the cumulative gains and budgets $B_t^*(100)$ of four prompts at θ_i, $i \in \{0, 1, 6\}$. (B): Non-top masses and $\frac{N(1-p_{t,L})B_t^*(100)}{100}$ of four prompts at θ_i, $i \in \{0, 1, 6\}$. (C): Cumulative gains and $\frac{NB_t^*(100)}{100r_0e^{\Lambda_t}}$ of four prompts at θ_i, $i \in \{0, 1, 6\}$. $r_0 = \frac{p_{0,L}}{1-p_{0,L}}$. The lines denote 95% candidate-bootstrap intervals which are obtained from 1000 independent resamples of the 128 samples set within each fixed prompt-model state.

pair counts come from the practical data builder, whereas the theorem bounds expected preference supply under max-min selection.

Similar to Sec. 6, based on the same four prompts, we use 128 responses sampled independently from $\theta_0/\theta_1/\theta_6$ for each prompt. From the score frequencies, we can estimate $p_{t,\ell}$. Based on this, we estimate $a_{N,t} = 1 - \sum_\ell p_{t,\ell}^N$ and $B_t^*(100) = \lceil \frac{100}{a_{N,t}} \rceil (N = 7)$ as the required candidate-group budget to obtain 100 accepted pairs in expectation.

Fig. 3C examines the approximation in Thm. 4(ii) through

$$\frac{NB_t^*(100)}{100r_0e^{\Lambda_t}}, \quad r_0 = \frac{p_{0,L}}{1-p_{0,L}}.$$

The displayed states with the largest positive cumulative gains have normalized budgets closest to one, whereas states farther from top-score saturation show larger deviations. This pattern is consistent with the predicted near-saturation approximation in Thm. 4(ii). Moreover, as shown in Fig. 3A, the high-positive-gain state requires a larger budget $B_t^*(M)$ than its initial state, consistent with the budget increase predicted as cumulative gains drive the system toward top-score saturation.

However, this relationship need not be globally monotone at finite gains, since the budget B_t^* depends on concentration of the entire score distribution, while Λ_t measures only changes in top-score odds. This distinction is shown for x_1 in Fig. 3A: the state with $\Lambda_t \approx -4$ requires a larger estimated budget than the state near $\Lambda_t \approx 0$, despite having a substantially smaller top-score cumulative gain. Such behavior is consistent with the theory, because concentration on a non-top score can also suppress $a_{N,t} = 1 - \sum_\ell p_{t,\ell}^N$ and thereby increase $B_t^*(M)$.

To examine Thm. 4(iii), Fig. 3B displays $\frac{N(1-p_{t,L})B_t^*(100)}{100}$. Notice that by Def. 1, if $\frac{N(1-p_{t,L})B_t^*(100)}{100} > 1$, then $B_t^*(100) \geq \frac{100}{N(1-p_{t,L})} \geq \frac{100}{N\delta}$ which satisfying Thm. 4(iii). All displayed finite point estimates are at least one, and the smallest non-top-mass states lie closest to 1. This is consistent with the lower bound.

C.3 Validation of allocation matching

We use two independent candidate sets A and B for each prompt-model state. For each set, they are generated following the same procedure described in Sec. 6. In this experiment, set A is used to estimate the score distribution and hence the non-top composition $\alpha_{t,\ell}$, while set B is used independently to estimate the realized preference allocation $\omega_{t,m,\ell}$ under strict max-min selection. The estimation calculations are the same with the estimation in Sec. 6 which is detailed in Sec. C.1.1.

Why two sets. To avoid spurious agreement caused by shared sampling noise, we estimate the two quantities from independent candidate sets. If both $\alpha_{t,\ell}$ and $\omega_{t,L,\ell}$ are computed from the same finite samples, fluctuations in the empirical score distribution affect both estimates in correlated ways. For example, over-sampling a particular lower-score level can simultaneously increase its estimated non-top-score composition and its induced preference allocation, making the two quantities appear closer even when the underlying population quantities differ. We therefore estimate these

Prompt	$t = 0$	$t = 1$	$t = 6$
x_0	0.5821 [0.4454, 0.7217]	NA	0.5822 [0.4805, 0.6901]
x_1	0.4341 [0.3354, 0.5567]	0.2200 [0, 0.6758]	0.7464 [0.4840, 1.0000]
x_2	0.0969 [0.0041, 0.2983]	NA	0.2392 [0.0948, 0.4084]
x_3	0.5363 [0.4683, 0.7011]	0.0299 [0, 0.2219]	0.3098 [0.1463, 0.5421]

Table 1: Cross-set allocation discrepancy D_a for $N = 7$. Brackets denote 95% candidate-bootstrap intervals which are obtained from 1000 independent resamples of the 128 sets A and B candidates within each fixed prompt-model state. These intervals can measure the certainty of corresponding discrepancy. NA indicates that the conditional preference allocation is not estimable because the independent set B contains no non-top-score candidates; it is not interpreted as zero discrepancy.

two quantities from two independent sets. This cross-set design reduces such shared-noise effects, so the matching between $\alpha_{t,\ell}$ and $\omega_{t,L,\ell}$ provides a more informative empirical test of the allocation matching predicted by Prop. 5.

Top- and non-top-score candidate counts in the independent sets A and B for each prompt and model are shown in Table 2.

Fig. 4A-D compare the non-top-score composition $\alpha^A_{t,\ell}$ estimated from set A with the preference allocation $\omega^B_{t,L,\ell}$ estimated from set B. To summarize the full score-pair discrepancy, including accepted pairs whose two scores are both below $L = 5$, we define

$$D_a = \frac{1}{2}\left[\sum_{\ell<L} |\omega^B_{t,L,\ell} - \alpha^A_{t,\ell}| + \sum_{\ell<m<L} \omega^B_{t,m,\ell}\right].$$

This quantity is the total-variation distance between the observed preference allocation in set B and the target allocation that places mass $\alpha^A_{t,\ell}$ on each top-versus-lower score pair (L, ℓ) and zero mass on non-top-versus-non-top pairs. Hence, $D_a = 0$ requires both $\omega^B_{t,L,\ell} = \alpha^A_{t,\ell}$ for all $\ell < L$ and vanishing preference allocation on pairs with $m < L$. The values of D_a are shown in Table 1.

For the highly score-saturated model θ_1, preference allocation remains estimable for two of the four prompts $x_0 - x_3$ because set B contains no non-top-score candidates as shown in Table 2. For x_1 and x_3, Both prompts exhibit smaller discrepancy D_a than at θ_0 and θ_6. For x_1, D_a decreases from 0.4341 to 0.2200, although the estimate is uncertain because only 3 and 5 non-top candidates are observed in sets A and B, respectively. x_3 provides the clearest case, its discrepancy D_a decreases from 0.5363 at θ_0 to 0.0299 at θ_1, before increasing to 0.3098 at θ_6. The orange circles and squares in Fig. 4D show that at θ_1, for x_3, the dominant lower score masses in $\alpha^A_{t,\ell}$ and $\omega^B_{t,L,\ell}$ nearly coincide, while the non-top-pair allocation is negligible. In Fig. 4B, compared to x_3, x_2 shows a larger redistribution across neighboring lower-score levels.

These results validate the allocation matching in Prop. 5, as $p_{t,L} \to 1$ (approximately satisfied at θ_1), $\omega_{t,L,\ell} \approx \alpha_{t,\ell}$ which means the preference allocation approximates the coefficient structure defining the top-score gradient. For x_0 and x_2, however, the preference allocation at θ_1 cannot be estimated since non-top responses disappear. These cases can be regarded as unidentified preference allocations due to the finite sampling statistical errors.

D Top-Score Reinforcement Beyond Round Initialization

We extend the top-score reinforcement analysis in Section 3 to non-initial within-round iterates and explore the relationship between it and round-level actor gains. The analysis is highly structurally consistent with Section 3, the difference lies in **the additional perturbation induced by the current policy $\pi_{\theta_{t,k}}$ on longer coincides with the fixed round-start policy** π_t which serves as the DPO reference. This perturbation results in two coupled effects:

- A shift between the current and training score-conditioned distributions.
- Nonuniform logistic reweighting caused by nonzero DPO margins.

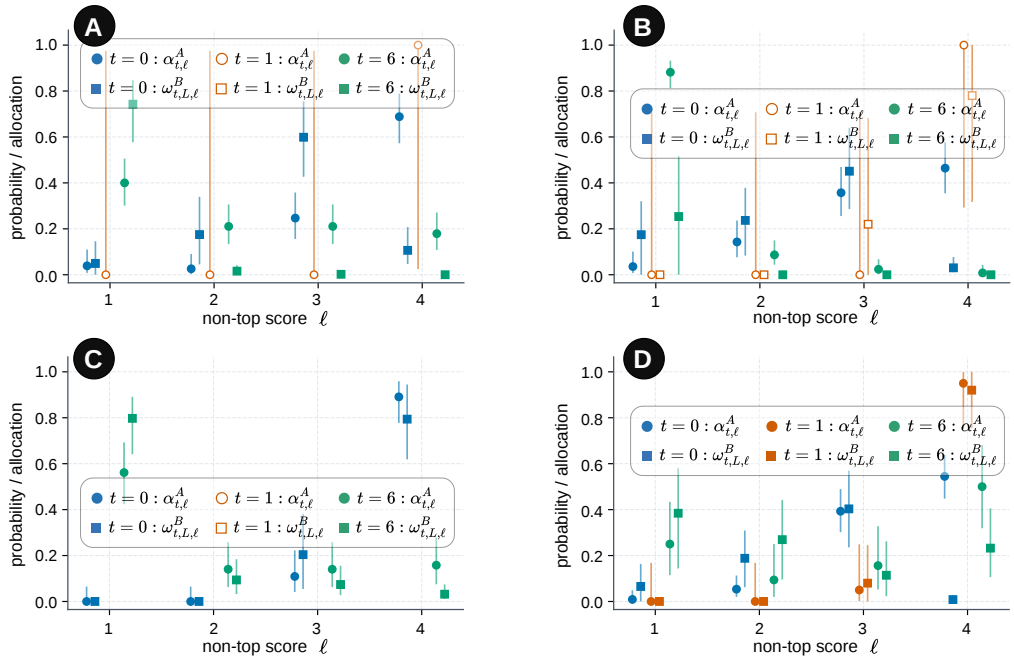


Figure 4: Comparisons between estimated non-top-score composition $\alpha_{t,\ell}$ and preference allocation $\omega_{t,L,\ell}$ among 4. (A)-(D) correspond to the four prompts $x_0 - x_3$, evaluated at checkpoints $t \in \{0, 1, 6\}$ with $N = 7$ candidates per group. Colors denote checkpoints, error bars are 95% candidate-bootstrap intervals. Hollow markers denote estimating with fewer than 8 observed non-top candidates. Quantities with an undefined denominator are omitted rather than set to 0.

Prompt	model	Set A		Set B	
		Top	Non-top	Top	Non-top
x_0	θ_0	51	77	39	89
x_0	θ_1	127	1	128	0
x_0	θ_6	33	95	23	105
x_1	θ_0	44	84	34	94
x_1	θ_1	125	3	123	5
x_1	θ_6	1	127	3	125
x_2	θ_0	73	55	66	62
x_2	θ_1	128	0	128	0
x_2	θ_6	71	57	70	58
x_3	θ_0	16	112	18	110
x_3	θ_1	108	20	109	19
x_3	θ_6	96	32	97	31

Table 2: Top- and non-top-score candidate counts in the independent sets A and B. Each set contains 128 candidates for every prompt-model state.

The round-start analysis in Section 3 is therefore recovered as the zero-drift special case, while the results below quantify how this structure persists under controlled within-round policy drift. As in the main analysis, all statements concern the population objective conditional on a fixed prompt x. The judge ϕ_t, reference policy π_t, and training distribution $\Gamma_t^{(N)}$ remain fixed throughout the round, and hence so does the preference allocation $\omega_{t,m,\ell}$. We assume that the policies have a common positive response support, and differentiation can be interchanged with the relevant expectations, and that $0 < p_{t,L} < 1$. All sums involving score-conditioned distributions are restricted to positive-probability score levels, which remain unchanged within the round under the common-support assumption.

D.1 Changes in the Descent Direction

For an iterate $\theta_{t,k}$, define the pairwise logistic weight and margin gradient:

$$q_{t,k}(y^+, y^-) := \sigma\big(-\beta\Delta_t(\theta_{t,k}; y^+, y^-)\big),$$
$$h_{t,k}(y^+, y^-) := \nabla_\theta\Delta_t(\theta_{t,k}; y^+, y^-) = \nabla_\theta \log \pi_{\theta_{t,k}}(y^+ \mid x) - \nabla_\theta \log \pi_{\theta_{t,k}}(y^- \mid x).$$

The reference policy contributes no derivative because it is fixed. Eq. (12) becomes

$$G_t(\theta_{t,k}) = \beta \sum_{m>\ell} \omega_{t,m,\ell} \mathbb{E}_{\Gamma^{(N)}_{t,m,\ell}}[q_{t,k} h_{t,k}]. \tag{17}$$

The conditional margin gradient in the main paper is still evaluated under the original training distribution:

$$\nabla_\theta \Delta_{t,m,\ell}(\theta_{t,k}) = \mathbb{E}_{\Gamma^{(N)}_{t,m,\ell}}[h_{t,k}].$$

Writing $\bar{q}_{t,k,m,\ell} := \mathbb{E}_{\Gamma^{(N)}_{t,m,\ell}}[q_{t,k}]$, gives the exact identity

$$\mathbb{E}_{\Gamma^{(N)}_{t,m,\ell}}[q_{t,k} h_{t,k}] = \bar{q}_{t,k,m,\ell} \nabla_\theta \Delta_{t,m,\ell}(\theta_{t,k}) + \mathrm{Cov}_{\Gamma^{(N)}_{t,m,\ell}}(q_{t,k}, h_{t,k}), \tag{18}$$

where $\mathrm{Cov}_P(q, h) := \mathbb{E}_P[(q - \mathbb{E}_P q)(h - \mathbb{E}_P h)]$. Thus, **logistic weighting changes both the relative contributions of different score pairs and the weighting of response pairs within each score pair**. Replacing the pairwise logistic weight by its conditional mean is exact only when the corresponding covariance vanishes. Moreover, $\bar{q}_{t,k,m,\ell}$ generally differs from the logistic weight evaluated at the conditional mean margin.

At initialization of one round, $q_{t,0} \equiv 1/2$, so all covariance terms vanish and Proposition 2 is recovered. At later iterates in the same round, the pairwise margin gradients may themselves change with the parameters. Importantly, even though the training distribution remains fixed, the score-conditioned response distributions defining the current top-score gradient evolve with the actor.

D.2 An Exact Decomposition at Non-initial Iterates

Similar to $p_{t,L}$ and $\alpha_{t,\ell}$ in the main paper, define the current score probabilities and non-top-score proportions under the fixed judge:

$$p_{t,k,s} := p_s(\phi_t, \theta_{t,k}), \qquad \alpha_{t,k,\ell} := \frac{p_{t,k,\ell}}{1 - p_{t,k,L}}, \quad \ell < L.$$

The corresponding score-conditioned distributions at $\theta_{t,k}$ are

$$\pi^s_{t,k}(y) := \frac{\pi_{\theta_{t,k}}(y \mid x)\mathbb{P}(S_{\phi_t}(x, y) = s)}{p_{t,k,s}}, \qquad Q_{t,k,m,\ell} := \pi^m_{t,k} \otimes \pi^\ell_{t,k}.$$

These distributions describe the current actor ($\theta_{t,k}$) under the fixed judge, and they do not replace the fixed training distribution. Let $v_{t,k} := \nabla_\theta \operatorname{logit} p_L(\phi_t, \theta)|_{\theta=\theta_{t,k}}$. Similar to the decomposition of v_t in Proposition 2, the current top-score gradient at $\theta_{t,k}$ retains a nonnegative mixture representation:

$$v_{t,k} = \sum_{\ell<L} \alpha_{t,k,\ell} \mathbb{E}_{Q_{t,k,L,\ell}}[h_{t,k}]. \tag{19}$$

In fact, $\nabla_\theta \log p_s(\phi_t, \theta)|_{\theta=\theta_{t,k}} = \mathbb{E}_{\pi^s_{t,k}}[\nabla_\theta \log \pi_{\theta_{t,k}}(Y \mid x)]$. Applying this identity to top-versus-lower comparisons and using $\sum_{\ell<L} p_{t,k,\ell} = 1 - p_{t,k,L}$ establishes Eq. (19).

Unlike Proposition 2, Eq. (19) uses the current distributions $Q_{t,k,L,\ell}$, whereas Eq. (17) uses the original training distributions $\Gamma^{(N)}_{t,m,\ell}$. To compare the two directions in a common representation, define

$$\rho_{t,k,s}(y) := \frac{\pi^s_{t,k}(y)}{\pi^s_t(y)} = \frac{\pi_{\theta_{t,k}}(y \mid x)}{\pi_t(y \mid x)} \frac{p_{t,s}}{p_{t,k,s}},$$

and

$$Z_{t,k} := \sum_{m>\ell} \omega_{t,m,\ell} \mathbb{E}_{\Gamma^{(N)}_{t,m,\ell}}[q_{t,k}] \in (0, 1).$$

Further define

$$\lambda_{t,k,m,\ell}(y^+, y^-) := \frac{\omega_{t,m,\ell} q_{t,k}(y^+, y^-)}{Z_{t,k} \rho_{t,k,m}(y^+) \rho_{t,k,\ell}(y^-)}. \tag{20}$$

By Proposition 1 and a change of measure,

$$\frac{G_t(\theta_{t,k})}{\beta Z_{t,k}} = \sum_{m>\ell} \mathbb{E}_{Q_{t,k,m,\ell}}[\lambda_{t,k,m,\ell} h_{t,k}], \quad \sum_{m>\ell} \mathbb{E}_{Q_{t,k,m,\ell}}[\lambda_{t,k,m,\ell}] = 1. \tag{21}$$

The weights $\lambda_{t,k,m,\ell}$ account for both logistic reweighting and changes in score-conditioned distributions; they are not new preference allocations.

Proposition 6 (Non-initial top-score decomposition)**.** *Define*

$$c^\star_{t,k} := \min_{\ell<L: p_{t,\ell}>0} \inf_{(y^+,y^-)\sim Q_{t,k,L,\ell}} \frac{\lambda_{t,k,L,\ell}(y^+, y^-)}{\alpha_{t,k,\ell}}. \tag{22}$$

Then $0 \le c^\star_{t,k} \le 1$*, and*

$$G_t(\theta_{t,k}) = \beta Z_{t,k} \left(c^\star_{t,k} v_{t,k} + R^\star_{t,k} \right), \tag{23}$$

where

$$R^\star_{t,k} = \sum_{\ell<L} \mathbb{E}_{Q_{t,k,L,\ell}} \left[\left(\lambda_{t,k,L,\ell} - c^\star_{t,k} \alpha_{t,k,\ell} \right) h_{t,k} \right] + \sum_{\ell<m<L} \mathbb{E}_{Q_{t,k,m,\ell}}[\lambda_{t,k,m,\ell} h_{t,k}]. \tag{24}$$

All remaining weights in Eq. (24) are nonnegative. The coefficient $c^\star_{t,k}$ is the largest extractable multiple of the current top-score mixture under the pointwise non-negativity constraint in this representation. If, additionally, $\|h_{t,k}(y^+, y^-)\| \le B_{t,k}$ almost surely under all positive-probability score-pair distributions, then

$$\|R^\star_{t,k}\| \le (1 - c^\star_{t,k}) B_{t,k}, \quad \left\| \frac{G_t(\theta_{t,k})}{\beta Z_{t,k}} - v_{t,k} \right\| \le 2(1 - c^\star_{t,k}) B_{t,k}. \tag{25}$$

Proof. By Eq. (22), for every positive-probability $\ell < L$,

$$\lambda_{t,k,L,\ell} - c^\star_{t,k} \alpha_{t,k,\ell} \ge 0$$

almost surely. Integrating, summing over ℓ, and applying Eq. (21) gives

$$c^\star_{t,k} \le \sum_{\ell<L} \mathbb{E}_{Q_{t,k,L,\ell}}[\lambda_{t,k,L,\ell}] \le 1.$$

Similar to Thm. 1, subtracting $c^\star_{t,k} v_{t,k}$ by Eq. (19) from Eq. (21) proves Eq. (23)–(24). Any larger coefficient makes a remaining weight negative on a set of positive measure in at least one top-versus-lower score pair, proving maximality within this representation.

The total remaining weight in Eq. (24) is

$$\sum_{m>\ell} \mathbb{E}_{Q_{t,k,m,\ell}}[\lambda_{t,k,m,\ell}] - c^\star_{t,k} \sum_{\ell<L} \alpha_{t,k,\ell} = 1 - c^\star_{t,k}.$$

Nonnegativity and the pairwise gradient bound yield the first norm bound. Eq. (19) also gives $\|v_{t,k}\| \le B_{t,k}$, so

$$\left\| \frac{G_t(\theta_{t,k})}{\beta Z_{t,k}} - v_{t,k} \right\| \le (1 - c^\star_{t,k}) \|v_{t,k}\| + \|R^\star_{t,k}\|,$$

which proves the second bound. □

At initialization of round t,

$$\rho_{t,0,s} \equiv 1, \quad Z_{t,0} = \tfrac{1}{2}, \quad \lambda_{t,0,m,\ell} \equiv \omega_{t,m,\ell},$$

and hence

$$c^\star_{t,0} = c_t, \qquad v_{t,0} = v_t, \qquad R^\star_{t,0} = R_t,$$

recovering Theorem 1 exactly. Round initialization is special because nonuniform logistic reweighting and the mismatch between training and current distributions disappaer simultaneously. At this

point, the remaining weights are constant within each score pair, so the weaker conditional margin gradient bound in Proposition 3 suffices. For $\theta_{t,k}$ with $k \geq 1$ (non-initialization), Eq. (25) uses a pairwise bound to control the additional nonuniform weighting.

The one-sided role of the top score is unchanged. It can appear only on the preferred side of accepted pairs. **What changes is the extractable mixture coefficient, which is no longer determined by preference allocation alone.** Positive individual logistic weights do not by themselves provide a uniform positive lower bound on $c^{\star}_{t,k}$.

D.3 Quantitative Persistence Under Bounded Within-Round Policy Drift

Corollary 3 (Quantitative persistence)**.** *Suppose*

$$\left|\log \frac{\pi_{\theta_{t,k}}(y \mid x)}{\pi_t(y \mid x)}\right| \leq \xi_{t,k} \tag{26}$$

on the common response support, then

$$c^{\star}_{t,k} \geq c_t e^{-(4+2\beta)\xi_{t,k}} \geq p_{t,L}^{N-1} e^{-(4+2\beta)\xi_{t,k}} > 0. \tag{27}$$

Under the preceding pairwise gradient bound ($B_{t,k}$ in Proposition 6),

$$\begin{aligned}\left\| \frac{G_t(\theta_{t,k})}{\beta Z_{t,k}} - v_{t,k} \right\| &\leq 2\left(1 - c_t e^{-(4+2\beta)\xi_{t,k}}\right) B_{t,k} \\ &\leq 2\left[(N-1)(1-p_{t,L}) + (4+2\beta)\xi_{t,k}\right] B_{t,k}.\end{aligned} \tag{28}$$

Proof. Write $u_{t,k}(y) := \log \frac{\pi_{\theta_{t,k}}(y|x)}{\pi_t(y|x)}$. For every positive-probability score s,

$$\frac{p_{t,k,s}}{p_{t,s}} = \mathbb{E}_{\pi_t^s}[e^{u_{t,k}(Y)}] \in [e^{-\xi_{t,k}}, e^{\xi_{t,k}}].$$

The same bounds hold for $(1 - p_{t,k,L})/(1 - p_{t,L})$. Since $\Delta_t(\theta_{t,k}; y^+, y^-) = u_{t,k}(y^+) - u_{t,k}(y^-)$, we have

$$q_{t,k},\ Z_{t,k} \in [\sigma(-2\beta\xi_{t,k}), \sigma(2\beta\xi_{t,k})], \quad \frac{q_{t,k}}{Z_{t,k}} \geq e^{-2\beta\xi_{t,k}}.$$

For a top-versus-lower score pair, Eq. (20) gives

$$\frac{\lambda_{t,k,L,\ell}}{\alpha_{t,k,\ell}} = \frac{\omega_{t,L,\ell}}{\alpha_{t,\ell}} \frac{q_{t,k}}{Z_{t,k}} \frac{p_{t,k,L}(1-p_{t,k,L})}{p_{t,L}(1-p_{t,L})} e^{-u_{t,k}(y^+) - u_{t,k}(y^-)}.$$

The last two factors are each bounded below by $e^{-2\xi_{t,k}}$. Taking $\inf$ and then minimizing over ℓ proves the first inequality in Eq. (27). The second follows from Theorem 1. Finally,

$$1 - c_t e^{-(4+2\beta)\xi_{t,k}} \leq (1 - c_t) + (4 + 2\beta)\xi_{t,k},$$

and

$$1 - c_t \leq 1 - p_{t,L}^{N-1} \leq (N-1)(1-p_{t,L}).$$

Combining these inequalities with Eq. (25) proves Eq. (28). □

Eq. (28) separates the residual control in Proposition 3 into two contributions as following.

- **Round-start score saturation.** Same as in Proposition 3, the allocation mismatch $1 - c_t$ at round initialization is controlled by round-start concentration $(N-1)(1-p_{t,L})$.
- **Within-round policy drift.** $(4 + 2\beta)\xi_{t,k}$ controls the additional within-round mismatch induced by the policy drift ($c^{*}_{t,k} \geq c_t e^{-(4+2\beta)\xi_{t,k}}$) consists of two terms. $2\beta\xi_{t,k}$ controls nonuniform logistic reweighting, while the remaining $4\xi_{t,k}$ controls the change from the training score-conditioned distributions to their current counterparts.

Notice that concentration enters through $p_{t,L}$, rather than $p_{t,k,L}$, because the preference allocation was fixed at round initialization. Moreover, Eq. (26) is a sufficient stability condition, not a consequence of decreasing the DPO loss.

For directional alignment, suppose $c^\star_{t,k} > 0$ and $v_{t,k} \neq 0$, and define

$$\kappa^\star_{t,k} := \frac{(1 - c^\star_{t,k})B_{t,k}}{c^\star_{t,k}\|v_{t,k}\|}.$$

If $\kappa^\star_{t,k} < 1$, Eq. (23)–(25) and Cauchy-Schwarz Inequality imply

$$\begin{aligned} \langle v_{t,k}, G_t(\theta_{t,k})\rangle &\geq \beta Z_{t,k} c^\star_{t,k}(1 - \kappa^\star_{t,k})\|v_{t,k}\|^2 > 0, \\ \cos(v_{t,k}, G_t(\theta_{t,k})) &\geq \frac{1 - \kappa^\star_{t,k}}{1 + \kappa^\star_{t,k}}. \end{aligned} \tag{29}$$

Since $\nabla_\theta p_L = p_{t,k,L}(1 - p_{t,k,L})v_{t,k}$ at the current iterate $\theta_{t,k}$, this condition also guarantees a positive first-order change in top-score probability along the DPO descent direction.

As at initialization, an absolute difference bound does not imply directional alignment when $\|v_{t,k}\|$ also becomes small. Moreover, the current gradient scale is $\beta Z_{t,k}$, rather than $\beta/2$, and

$$\|G_t(\theta_{t,k})\| \leq \beta Z_{t,k} B_{t,k}.$$

Therefore, even a well-aligned normalized direction can produce a weak update when $Z_{t,k}$ is small.

D.4 Implications for Whole-Round Actor Gain

Similar to Section 4, the preceding results concern the current descent direction, to connect them to the change realized by a completed round, let

$$f_t(\theta) := \operatorname{logit} p_L(\phi_t, \theta), \qquad d_{t,k} := \theta_{t,k+1} - \theta_{t,k}.$$

Assume that ∇f_t is $L_{f,t}$-Lipschitz on the line segments joining consecutive iterates. Since $\mathcal{A}_t = f_t(\theta_{t,n_t}) - f_t(\theta_{t,0})$ by its definition, summation over the iterates in the same round gives

$$\mathcal{A}_t \geq \sum_{k=0}^{n_t-1} \left[\langle v_{t,k}, d_{t,k}\rangle - \frac{L_{f,t}}{2}\|d_{t,k}\|^2\right]. \tag{30}$$

If we further consider the gradient descent, $d_{t,k} = \eta_{t,k} G_t(\theta_{t,k})$, $\eta_{t,k} > 0$, Eq. (23)–(25) further yield

$$\mathcal{A}_t \geq \sum_{k=0}^{n_t-1} \eta_{t,k}\beta Z_{t,k}\left[c^\star_{t,k}\|v_{t,k}\|^2 - (1 - c^\star_{t,k})B_{t,k}\|v_{t,k}\|\right] - \frac{L_{f,t}}{2}\sum_{k=0}^{n_t-1}\eta^2_{t,k}\|G_t(\theta_{t,k})\|^2. \tag{31}$$

To verify this bound, substitute the decomposition into Eq. (30) and use $\langle v_{t,k}, R^\star_{t,k}\rangle \geq -(1 - c^\star_{t,k})B_{t,k}\|v_{t,k}\|$. A positive RHS of Eq. (31) is sufficient for positive whole-round actor gain. In particular, *strict directional alignment at every iterate, together with sufficiently small step sizes, guarantees positive actor gain*. This condition is not necessary, however, because gains at different iterates can compensate over the completed round.

No judge gain analysis. Since the judge remains fixed at every within-round $\theta_{t,k}$, these results concern only $\mathcal{A}_t$ in both the self-judge and fixed judge settings. The additional endpoint change induced by updating the judge is still captured separately by $\mathcal{J}_t$.

Overall, non-initial top-score reinforcement depends jointly on mixture persistence, residual geometry, and the actual optimization trajectory. The endpoint actor-judge decomposition in Section 4 remains applicable without requiring these sufficient conditions to hold throughout the round.

E Proofs

In this section, we present the derivations of the equations and the proofs of the propositions, lemmas, corollaries, and theorems in the order in which they appear in the main paper. **Proofs for other equations, lemmas, theorems, etc., in the appendix are provided where they are proposed.**

Derivation of Eq. (4). $\pi^\ell_t(y) = \frac{\pi_t(y)\mathbb{P}(S_{\phi_t}(x,y)=\ell)}{p_{t,\ell}}$, $\tilde{\pi}^\ell_t(y) = \frac{\pi_{t+1}(y)\mathbb{P}(S_{\phi_t}(x,y)=\ell)}{\tilde{p}_{t,\ell}}$.

Proof. For every score ℓ with $p_{t,\ell} > 0$, Bayes' rule gives

$$\mathbb{P}(Y = y \mid S_{\phi_t}(x, Y) = \ell) = \frac{\mathbb{P}(Y = y, S_{\phi_t}(x, Y) = \ell)}{\mathbb{P}(S_{\phi_t}(x, Y) = \ell)} = \frac{\pi_t(y)\mathbb{P}(S_{\phi_t}(x, y) = \ell)}{p_{t,\ell}} = \pi_t^\ell(y).$$

Applying the same argument to $Y \sim \pi_{t+1}(\cdot \mid x)$, while retaining the judge ϕ_t, yields

$$\tilde{\pi}_t^\ell(y) = \frac{\pi_{t+1}(y)\mathbb{P}(S_{\phi_t}(x, y) = \ell)}{\tilde{p}_{t,\ell}}.$$

□

Derivation of Eq. (5) in Proposition 1.

$$\Gamma_{t,m,\ell}^{(N)} = \pi_t^m \otimes \pi_t^\ell,\ \Gamma_t^{(N)} = \sum_{m>\ell} \omega_{t,m,\ell}(\pi_t^m \otimes \pi_t^\ell),\ \otimes \text{ denotes the distribution product.}$$

Proof. Let S_i be the score assigned to candidate Y_i. Independent response sampling and conditionally independent scoring imply that (Y_i, S_i), $i = 1, \ldots, N$, are independent and identically distributed. Fix a positive-probability score vector $(s_1, \ldots, s_N)$. Conditional on this vector, the responses remain independent and satisfy

$$Y_i \mid (S_1 = s_1, \ldots, S_N = s_N) \sim \pi_t^{s_i}.$$

Suppose that $\max_i s_i = m$ and $\min_i s_i = \ell$, where $m > \ell$. Any selected maximum-score index i and minimum-score index j are distinct and satisfy $s_i = m$ and $s_j = \ell$. Thus, for measurable response sets A, B,

$$\mathbb{P}(Y_i \in A, Y_j \in B \mid S_1 = s_1, \ldots, S_N = s_N) = \pi_t^m(A)\pi_t^\ell(B).$$

Tie-breaking uses only the scores and auxiliary randomness. Averaging over the selected indices therefore preserves this product:

$$\mathbb{P}(Y^+ \in A, Y^- \in B \mid S_1 = s_1, \ldots, S_N = s_N) = \pi_t^m(A)\pi_t^\ell(B).$$

The right-hand side is identical for every score vector with maximum m and minimum ℓ. Averaging over such vectors gives

$$\Gamma_{t,m,\ell}^{(N)} = \pi_t^m \otimes \pi_t^\ell.$$

Finally, conditioning on the score pair of an accepted group yields

$$\Gamma_t^{(N)} = \sum_{m>\ell} \mathbb{P}(S^+ = m, S^- = \ell \mid S^+ > S^-)\Gamma_{t,m,\ell}^{(N)} = \sum_{m>\ell} \omega_{t,m,\ell}(\pi_t^m \otimes \pi_t^\ell).$$

□

Derivation of Eq. (6). $\varepsilon_{t,m,\ell} := \mathbb{E}_{\Gamma_{t,m,\ell}^{(N)}}[\ell_{log}(\beta\Delta_t(\theta))]$ and $L_t^{DPO}(\theta) = \sum_{m>\ell} \omega_{t,m,\ell}\varepsilon_{t,m,\ell},\ \forall\, L \geq m > \ell \geq 1.$

Proof. Writing explicitly the dependence on θ, and $\varepsilon_{t,m,\ell}(\theta) = \mathbb{E}_{\Gamma_{t,m,\ell}^{(N)}}[\ell_{\log}(\beta\Delta_t(\theta))]$. By conditioning on the score pair of the accepted preference,

$$\begin{aligned} L_t^{DPO}(\theta) &= \mathbb{E}_{\Gamma_t^{(N)}}[\ell_{\log}(\beta\Delta_t(\theta))] \\ &= \sum_{m>\ell} \omega_{t,m,\ell}\mathbb{E}_{\Gamma_{t,m,\ell}^{(N)}}[\ell_{\log}(\beta\Delta_t(\theta))] \\ &= \sum_{m>\ell} \omega_{t,m,\ell}\varepsilon_{t,m,\ell}(\theta). \end{aligned}$$

Each conditional loss is nonnegative because $\ell_{\log}(z) = \log(1 + e^{-z}) \geq 0$. □

Derivation of Eq. (7). $\Delta_{t,m,\ell}(\theta) := \mathbb{E}_{\Gamma^{(N)}_{t,m,\ell}}[\Delta_t(\theta; x, y^+, y^-)] = \mathbb{E}_{y\sim\pi_t^m}\left[\log \frac{\pi_\theta(y|x)}{\pi_{\theta_t}(y|x)}\right] - \mathbb{E}_{y\sim\pi_t^\ell}\left[\log \frac{\pi_\theta(y|x)}{\pi_{\theta_t}(y|x)}\right]$.

Proof. By Proposition 1, the conditional selected-pair law is $\pi_t^m \otimes \pi_t^\ell$. Hence

$$\begin{aligned}\Delta_{t,m,\ell}(\theta) &= \mathbb{E}_{\Gamma^{(N)}_{t,m,\ell}}\left[\log \frac{\pi_\theta(Y^+ \mid x)}{\pi_t(Y^+ \mid x)} - \log \frac{\pi_\theta(Y^- \mid x)}{\pi_t(Y^- \mid x)}\right] \\ &= \mathbb{E}_{\pi_t^m}\left[\log \frac{\pi_\theta(Y \mid x)}{\pi_t(Y \mid x)}\right] - \mathbb{E}_{\pi_t^\ell}\left[\log \frac{\pi_\theta(Y \mid x)}{\pi_t(Y \mid x)}\right].\end{aligned}$$

The conditional distributions and the reference policy do not depend on the within-round optimization variable θ. Therefore,

$$\nabla_\theta \Delta_{t,m,\ell}(\theta) = \mathbb{E}_{\pi_t^m}[\nabla_\theta \log \pi_\theta(Y \mid x)] - \mathbb{E}_{\pi_t^\ell}[\nabla_\theta \log \pi_\theta(Y \mid x)].$$

□

Proof of Proposition 2. At $\theta = \theta_t$,

$$v_t = \sum_{\ell<L} \alpha_{t,\ell} \nabla_\theta \Delta_{t,L,\ell}(\theta_t),\ G_t(\theta_t) = \frac{\beta}{2} \sum_{m>\ell} \omega_{t,m,\ell} \nabla_\theta \Delta_{t,m,\ell}(\theta_t).$$

Proof. The judge parameters ϕ_t are held fixed when differentiating with respect to the actor parameters, including in the self-judge setting. For every positive-probability score s, Eq. (1) and differentiation under the expectation imply

$$\begin{aligned}\nabla_\theta p_s(\phi_t, \theta; x)|_{\theta=\theta_t} &= \sum_y \mathbb{P}(S_{\phi_t}(x, y) = s)\ \nabla_\theta \pi_\theta(y \mid x)|_{\theta=\theta_t} \\ &= p_{t,s}\, \mathbb{E}_{\pi_t^s}\left[\nabla_\theta \log \pi_\theta(Y \mid x)|_{\theta=\theta_t}\right].\end{aligned}$$

Consequently,

$$\nabla_\theta \log p_s(\phi_t, \theta; x)|_{\theta=\theta_t} = \mathbb{E}_{\pi_t^s}\left[\nabla_\theta \log \pi_\theta(Y \mid x)|_{\theta=\theta_t}\right].$$

Using $1 - p_L = \sum_{\ell<L} p_\ell$, we obtain

$$\begin{aligned}v_t &= \nabla_\theta \left(\log p_L - \log \sum_{\ell<L} p_\ell\right)\Bigg|_{\theta=\theta_t} \\ &= \sum_{\ell<L} \alpha_{t,\ell} \left(\mathbb{E}_{\pi_t^L}\left[\nabla_\theta \log \pi_\theta(Y \mid x)|_{\theta=\theta_t}\right] - \mathbb{E}_{\pi_t^\ell}\left[\nabla_\theta \log \pi_\theta(Y \mid x)|_{\theta=\theta_t}\right]\right) \\ &= \sum_{\ell<L} \alpha_{t,\ell} \nabla_\theta \Delta_{t,L,\ell}(\theta_t).\end{aligned}$$

At initialization, $\Delta_t(\theta_t; y^+, y^-) = 0$ for every preference pair. Hence $\sigma(-\beta\Delta_t(\theta_t)) = 1/2$, and

$$\begin{aligned}G_t(\theta_t) &= \frac{\beta}{2}\mathbb{E}_{\Gamma_t^{(N)}}[\nabla_\theta \Delta_t(\theta_t)] = \frac{\beta}{2} \sum_{m>\ell} \omega_{t,m,\ell} \mathbb{E}_{\Gamma^{(N)}_{t,m,\ell}}[\nabla_\theta \Delta_t(\theta_t)] \\ &= \frac{\beta}{2} \sum_{m>\ell} \omega_{t,m,\ell} \nabla_\theta \Delta_{t,m,\ell}(\theta_t).\end{aligned}$$

□

Proof of Theorem 1. Let $c_t := \min_{\ell<L: p_{t,\ell}>0} \frac{\omega_{t,L,\ell}}{\alpha_{t,\ell}}$, then $c_t \in [p_{t,L}^{N-1}, 1]$, and

$$G_t(\theta_t) = \frac{\beta}{2}(c_t v_t + R_t),\ R_t = R_t^L + R_t^{NL},$$

$$R_t^L = \sum_{\ell<L} (\omega_{t,L,\ell} - c_t \alpha_{t,\ell}) \nabla_\theta \Delta_{t,L,\ell}(\theta_t),\ R_t^{NL} = \sum_{\ell<m<L} \omega_{t,m,\ell} \nabla_\theta \Delta_{t,m,\ell}(\theta_t).$$

Proof. Notice that $a_{N,t}$ is the probability that a candidate group is accepted. Acceptance requires at least one non-top score. Hence the union bound gives

$$a_{N,t} \le N(1 - p_{t,L}).$$

Fix $\ell < L$ with $p_{t,\ell} > 0$. The event that exactly one candidate receives score ℓ and the remaining $N-1$ candidates receive score L has probability $N p_{t,\ell} p_{t,L}^{N-1}$. This event necessarily produces score pair (L, ℓ), so

$$\omega_{t,L,\ell} \ge \frac{N p_{t,\ell} p_{t,L}^{N-1}}{a_{N,t}}.$$

Dividing by $\alpha_{t,\ell} = p_{t,\ell}/(1 - p_{t,L})$ yields

$$\frac{\omega_{t,L,\ell}}{\alpha_{t,\ell}} \ge \frac{N(1 - p_{t,L}) p_{t,L}^{N-1}}{a_{N,t}} \ge p_{t,L}^{N-1}.$$

Taking the minimum over ℓ proves $c_t \ge p_{t,L}^{N-1}$. For the upper bound, a minimum is no greater than a weighted average:

$$c_t \le \sum_{\ell < L:\, p_{t,\ell} > 0} \alpha_{t,\ell} \frac{\omega_{t,L,\ell}}{\alpha_{t,\ell}} = \sum_{\ell < L} \omega_{t,L,\ell} \le 1.$$

By the definition of c_t,

$$\omega_{t,L,\ell} - c_t \alpha_{t,\ell} \ge 0$$

for every active lower score. If $c > c_t$, this inequality fails for an index attaining the minimum. Thus c_t is the largest coefficient satisfying all of these allocation-level constraints.

Using Proposition 2 and adding and subtracting $c_t \alpha_{t,\ell}$ in each top-versus-lower coefficient,

$$\begin{aligned}\frac{2}{\beta} G_t(\theta_t) = c_t \sum_{\ell < L} \alpha_{t,\ell} \nabla_\theta \Delta_{t,L,\ell}(\theta_t) + \sum_{\ell < L} (\omega_{t,L,\ell} - c_t \alpha_{t,\ell}) \nabla_\theta \Delta_{t,L,\ell}(\theta_t) \\ + \sum_{\ell < m < L} \omega_{t,m,\ell} \nabla_\theta \Delta_{t,m,\ell}(\theta_t) = c_t v_t + R_t^L + R_t^{NL}.\end{aligned}$$

Finally, $0 < p_{t,L} < 1$ implies $c_t > 0$, but the vector $c_t v_t$ is nonzero only when $v_t \neq 0$. In that case,

$$\langle \nabla_\theta p_L(\phi_t, \theta_t; x), c_t v_t \rangle = c_t p_{t,L}(1 - p_{t,L}) \|v_t\|^2 > 0.$$

□

Derivation of Eq. (11). $v_t^{(r)} = \sum_{\ell < r} \alpha_{t,\ell}^{(r)} \nabla_\theta \Delta_{t,r,\ell}(\theta_t) - \sum_{m > r} \alpha_{t,m}^{(r)} \nabla_\theta \Delta_{t,m,r}(\theta_t)$, $\alpha_{t,s}^{(r)} := \frac{p_{t,s}}{1 - p_{t,r}}$, $\forall s \in [L]/\{r\}$.

Proof. The same calculation as in Proposition 2 gives

$$v_t^{(r)} = \sum_{s \neq r} \alpha_{t,s}^{(r)} \left(\mathbb{E}_{\pi_t^r} \left[\nabla_\theta \log \pi_\theta(Y \mid x)|_{\theta=\theta_t} \right] - \mathbb{E}_{\pi_t^s} \left[\nabla_\theta \log \pi_\theta(Y \mid x)|_{\theta=\theta_t} \right] \right).$$

For $s < r$, the difference is $\nabla_\theta \Delta_{t,r,s}(\theta_t)$. For $s > r$, it is $-\nabla_\theta \Delta_{t,s,r}(\theta_t)$. Thus

$$v_t^{(r)} = \sum_{\ell < r} \alpha_{t,\ell}^{(r)} \nabla_\theta \Delta_{t,r,\ell}(\theta_t) - \sum_{m > r} \alpha_{t,m}^{(r)} \nabla_\theta \Delta_{t,m,r}(\theta_t).$$

□

Positive component is not guaranteed for $r < L$. To show that a nonnegative submixture proportional to $v_t^{(r)}$ is not guaranteed for $r < L$, consider two responses y_r, y_L with deterministic scores r, L, and let

$$\pi_\theta(y_L \mid x) = \sigma(\theta), \qquad \pi_\theta(y_r \mid x) = 1 - \sigma(\theta).$$

Every accepted group produces score pair (L, r), so $\omega_{t,L,r} = 1$. Moreover,

$$\Delta_{t,L,r}(\theta) = \theta - \theta_t, \qquad \nabla_\theta \Delta_{t,L,r}(\theta_t) = 1, \qquad G_t(\theta_t) = \frac{\beta}{2}.$$

However,

$$v_t^{(r)} = \nabla_\theta \operatorname{logit}(1 - \sigma(\theta)) = -1.$$

No nonnegative submixture of the only available preference gradient can equal a strictly positive multiple of $v_t^{(r)}$. Moreover, the claim is a non-guarantee, not a universal impossibility. For $r = 1$, only negative-sign terms occur in Eq. (11). If all higher scores have zero probability, a lower rubric score can also be maximal on the active support.

Proof of Proposition 3. Assume the norms of conditional margin gradients are bounded by D_t, then

$$\|R_t\| \le (1 - c_t)D_t, \ \|\frac{2}{\beta}G_t(\theta_t) - v_t\| \le 2(N-1)(1 - p_{t,L})D_t.$$

Proof. All coefficients in R_t are nonnegative by Theorem 1. Their total mass is

$$\sum_{\ell<L}(\omega_{t,L,\ell} - c_t\alpha_{t,\ell}) + \sum_{\ell<m<L}\omega_{t,m,\ell} = \sum_{m>\ell}\omega_{t,m,\ell} - c_t\sum_{\ell<L}\alpha_{t,\ell} = 1 - c_t.$$

The triangle inequality therefore gives

$$\|R_t\| \le (1 - c_t)D_t.$$

By Proposition 2, v_t is a convex combination of conditional margin gradients, so $\|v_t\| \le D_t$. Using Theorem 1,

$$\frac{2}{\beta}G_t(\theta_t) - v_t = (c_t - 1)v_t + R_t.$$

Hence

$$\left\|\frac{2}{\beta}G_t(\theta_t) - v_t\right\| \le (1 - c_t)\|v_t\| + \|R_t\| \le 2(1 - c_t)D_t.$$

Finally,

$$1 - c_t \le 1 - p_{t,L}^{N-1} = (1 - p_{t,L})\sum_{j=0}^{N-2} p_{t,L}^j \le (N-1)(1 - p_{t,L}).$$

Combining these inequalities proves

$$\left\|\frac{2}{\beta}G_t(\theta_t) - v_t\right\| \le 2(N-1)(1 - p_{t,L})D_t.$$

For fixed N, uniformly bounded D_t, and $p_{t,L} \to 1$, the right-hand side converges to zero. □

Proof of Corollary 1. If $\kappa_t < 1$, then $\langle v_t, G_t(\theta_t)\rangle \ge \frac{\beta c_t}{2}(1 - \kappa_t)\|v_t\|^2 > 0$ and $\cos(v_t, G_t(\theta_t)) \ge \frac{1-\kappa_t}{1+\kappa_t}$.

Proof. Assume $c_t > 0$, $v_t \ne 0$, and $\kappa_t < 1$. Proposition 3 gives

$$\|R_t\| \le (1 - c_t)D_t = \kappa_t c_t\|v_t\|.$$

By Theorem 1 and the Cauchy–Schwarz inequality,

$$\langle v_t, G_t(\theta_t)\rangle = \frac{\beta}{2}\left(c_t\|v_t\|^2 + \langle v_t, R_t\rangle\right) \ge \frac{\beta}{2}\left(c_t\|v_t\|^2 - \|v_t\|\|R_t\|\right)$$
$$\ge \frac{\beta c_t}{2}(1 - \kappa_t)\|v_t\|^2 > 0.$$

In particular, $G_t(\theta_t) \ne 0$. Also,

$$\|G_t(\theta_t)\| \le \frac{\beta}{2}(c_t\|v_t\| + \|R_t\|) \le \frac{\beta c_t}{2}(1 + \kappa_t)\|v_t\|.$$

Therefore,

$$\cos(v_t, G_t(\theta_t)) = \frac{\langle v_t, G_t(\theta_t)\rangle}{\|v_t\|\|G_t(\theta_t)\|} \ge \frac{1 - \kappa_t}{1 + \kappa_t}.$$

□

Derivation of Eq. (12). $G_t(\theta_{t,k}) = \beta \sum_{m>\ell} \omega_{t,m,\ell} \mathbb{E}_{\Gamma^{(N)}_{t,m,\ell}} \left[\sigma\left(-\beta\Delta_t\left(\theta_{t,k}\right)\right) \nabla_\theta \Delta_t(\theta_{t,k})\right]$.

Proof. The gradient identity derived from Eq. (3) holds at every within-round iterate:

$$G_t(\theta_{t,k}) = \beta \mathbb{E}_{\Gamma^{(N)}_t} \left[\sigma(-\beta\Delta_t(\theta_{t,k}))\nabla_\theta \Delta_t(\theta_{t,k})\right].$$

Conditioning on the score pair and using the fixed round-start preference allocation gives

$$G_t(\theta_{t,k}) = \beta \sum_{m>\ell} \omega_{t,m,\ell} \mathbb{E}_{\Gamma^{(N)}_{t,m,\ell}} \left[\sigma(-\beta\Delta_t(\theta_{t,k}))\nabla_\theta \Delta_t(\theta_{t,k})\right].$$

At initialization the sigmoid factor is identically $1/2$. At later iterates it generally varies across individual pairs and remains inside the conditional expectation. □

Proof of Corollary 2. The model is $\delta-$score saturated at round T if and only if $\Lambda_T \geq \log \frac{1-\delta}{\delta} - \text{logit}(p_{0,L})$.

Proof. Summing the exact actor-judge decomposition gives

$$\Lambda_T = \sum_{t=0}^{T-1} \left[\text{logit}(p_{t+1,L}) - \text{logit}(p_{t,L})\right] = \text{logit}(p_{T,L}) - \text{logit}(p_{0,L}).$$

The logit function is strictly increasing on $(0,1)$ because $\frac{d}{dp}\text{logit}(p) = \frac{1}{p(1-p)} > 0$. Consequently,

$$p_{T,L} \geq 1-\delta \iff \text{logit}(p_{T,L}) \geq \log\frac{1-\delta}{\delta} \iff \Lambda_T \geq \log\frac{1-\delta}{\delta} - \text{logit}(p_{0,L}).$$

□

Derivation of Eq. (13). $\mathcal{A}_t = -\log \sum_{\ell<L} \alpha_{t,\ell} \exp(-r_{t,L,\ell}),\ \alpha_{t,\ell} := \frac{p_{t,\ell}}{1-p_{t,L}}$.

Proof. For every active lower score ℓ, $e^{-r_{t,L,\ell}} = \frac{p_{t,L}\tilde{p}_{t,\ell}}{\tilde{p}_{t,L}p_{t,\ell}}$. Multiplying by $\alpha_{t,\ell}$ gives

$$\alpha_{t,\ell}e^{-r_{t,L,\ell}} = \frac{p_{t,L}\tilde{p}_{t,\ell}}{(1-p_{t,L})\tilde{p}_{t,L}}.$$

The common-support condition ensures that no previously zero-probability score level acquires positive probability under the fixed judge. Thus

$$\begin{aligned}\sum_{\ell<L} \alpha_{t,\ell}e^{-r_{t,L,\ell}} &= \frac{p_{t,L}}{(1-p_{t,L})\tilde{p}_{t,L}} \sum_{\ell<L} \tilde{p}_{t,\ell} = \frac{p_{t,L}(1-\tilde{p}_{t,L})}{(1-p_{t,L})\tilde{p}_{t,L}} \\ &= \exp\left(\text{logit}(p_{t,L}) - \text{logit}(\tilde{p}_{t,L})\right) = e^{-\mathcal{A}_t}.\end{aligned}$$

Taking the negative log proves Eq. (13). □

Proof of Proposition 4. For any two scores $m > \ell$, $r_{t,m,\ell} = \Delta_{t,m,\ell}(\theta_{t+1}) + K_{t,m} - K_{t,\ell}$.

Proof. Because the same judge ϕ_t defines both conditional distributions, Eq. (4) gives, on the support of π^s_t,

$$\frac{\pi^s_t(y)}{\tilde{\pi}^s_t(y)} = \frac{\pi_t(y)}{\pi_{t+1}(y)} \frac{\tilde{p}_{t,s}}{p_{t,s}}.$$

Taking logarithms and expectation under π^s_t yields

$$K_{t,s} = -\mathbb{E}_{\pi^s_t}\left[\log\frac{\pi_{t+1}(Y)}{\pi_t(Y)}\right] + \log\frac{\tilde{p}_{t,s}}{p_{t,s}}.$$

Equivalently,

$$\mathbb{E}_{\pi^s_t}\left[\log\frac{\pi_{t+1}(Y)}{\pi_t(Y)}\right] = \log\frac{\tilde{p}_{t,s}}{p_{t,s}} - K_{t,s}.$$

Applying this identity to $s = m$ and $s = \ell$ and using Eq. (7),

$$\begin{aligned}\Delta_{t,m,\ell}(\theta_{t+1}) &= \log \frac{\tilde{p}_{t,m}}{p_{t,m}} - \log \frac{\tilde{p}_{t,\ell}}{p_{t,\ell}} - K_{t,m} + K_{t,\ell} \\ &= r_{t,m,\ell} - K_{t,m} + K_{t,\ell}.\end{aligned}$$

Rearranging proves the proposition. □

Proof of $\Delta_{t,m,\ell}(\theta) \geq \frac{1}{\beta}\Phi(\frac{L_t^{DPO}(\theta)}{\omega_{t,m,\ell}})$.

Proof. The logistic loss is convex because $\ell''_{\log}(z) = \frac{e^z}{(1+e^z)^2} > 0$. Jensen's inequality gives

$$\varepsilon_{t,m,\ell}(\theta) = \mathbb{E}_{\Gamma^{(N)}_{t,m,\ell}}[\ell_{\log}(\beta\Delta_t(\theta))] \geq \ell_{\log}\left(\beta\mathbb{E}_{\Gamma^{(N)}_{t,m,\ell}}[\Delta_t(\theta)]\right) = \ell_{\log}(\beta\Delta_{t,m,\ell}(\theta)).$$

Therefore,

$$e^{-\beta\Delta_{t,m,\ell}(\theta)} \leq e^{\varepsilon_{t,m,\ell}(\theta)} - 1,$$

which implies

$$\Delta_{t,m,\ell}(\theta) \geq \frac{1}{\beta}\Phi(\varepsilon_{t,m,\ell}(\theta)), \qquad \Phi(z) = -\log(e^z - 1).$$

By Eq. (6) and non-negativity of all conditional losses,

$$L_t^{DPO}(\theta) \geq \omega_{t,m,\ell}\varepsilon_{t,m,\ell}(\theta).$$

Thus, whenever $\omega_{t,m,\ell} > 0$,

$$\varepsilon_{t,m,\ell}(\theta) \leq \frac{L_t^{DPO}(\theta)}{\omega_{t,m,\ell}}.$$

Since $\Phi'(z) = -\frac{e^z}{e^z-1} < 0$ $(z > 0)$, we can conclude that

$$\Delta_{t,m,\ell}(\theta) \geq \frac{1}{\beta}\Phi\left(\frac{L_t^{DPO}(\theta)}{\omega_{t,m,\ell}}\right).$$

□

Proof of Theorem 2. $\mathcal{A}_t \geq -\log\sum_{\ell<L,\ p_{t,\ell}>0}\alpha_{t,\ell}\exp\left[K_{t,\ell} - \frac{1}{\beta}\Phi\left(\frac{L_t^{DPO}(\theta_{t+1})}{\omega_{t,L,\ell}}\right)\right]$. In particular, a sufficient condition for $\mathcal{A}_t > 0$ is $\sum_{\ell<L,\ p_{t,\ell}>0}\alpha_{t,\ell}\exp\left[K_{t,\ell} - K_{t,L} - \frac{1}{\beta}\Phi\left(\frac{L_t^{DPO}(\theta_{t+1})}{\omega_{t,L,\ell}}\right)\right] < 1$. A simpler sufficient condition is if

$$\frac{1}{\beta}\Phi(\frac{L_t^{DPO}(\theta_{t+1})}{\omega_{t,L,\ell}}) > K_{t,\ell} - K_{t,L} \text{ for every } \ell < L,\ p_{t,\ell} > 0.$$

Proof. For every active lower score ℓ, Proposition 4 and the conditional-margin bound imply

$$r_{t,L,\ell} \geq \frac{1}{\beta}\Phi\left(\frac{L_t^{DPO}(\theta_{t+1})}{\omega_{t,L,\ell}}\right) + K_{t,L} - K_{t,\ell}.$$

Therefore,

$$e^{-r_{t,L,\ell}} \leq \exp\left[K_{t,\ell} - K_{t,L} - \frac{1}{\beta}\Phi\left(\frac{L_t^{DPO}(\theta_{t+1})}{\omega_{t,L,\ell}}\right)\right].$$

Using Eq. (13) and monotonicity of the negative logarithm gives the stronger intermediate bound

$$\mathcal{A}_t \geq -\log\sum_{\ell<L:\ p_{t,\ell}>0}\alpha_{t,\ell}\exp\left[K_{t,\ell} - K_{t,L} - \frac{1}{\beta}\Phi\left(\frac{L_t^{DPO}(\theta_{t+1})}{\omega_{t,L,\ell}}\right)\right].$$

Factoring out $e^{-K_{t,L}}$ and using $K_{t,L} \geq 0$ yields

$$\mathcal{A}_t \geq K_{t,L} - \log \sum_{\ell<L:\, p_{t,\ell}>0} \alpha_{t,\ell} \exp\left[K_{t,\ell} - \frac{1}{\beta}\Phi\left(\frac{L_t^{DPO}(\theta_{t+1})}{\omega_{t,L,\ell}}\right)\right]$$
$$\geq -\log \sum_{\ell<L:\, p_{t,\ell}>0} \alpha_{t,\ell} \exp\left[K_{t,\ell} - \frac{1}{\beta}\Phi\left(\frac{L_t^{DPO}(\theta_{t+1})}{\omega_{t,L,\ell}}\right)\right].$$

This proves the displayed lower bound in the theorem.

If the sum in the stronger intermediate bound is less than one, its negative logarithm is positive, and hence $\mathcal{A}_t > 0$. In particular, if

$$\frac{1}{\beta}\Phi\left(\frac{L_t^{DPO}(\theta_{t+1})}{\omega_{t,L,\ell}}\right) > K_{t,\ell} - K_{t,L}$$

for every active lower score, every exponential term in that sum is strictly less than one. Since the $\alpha_{t,\ell}$ sum to one, the entire sum is strictly less than one. □

Proof of Theorem 3. At round t, for a given prompt x, suppose $\mathbb{E}_{\pi_{t+1}}\left[\mathbb{P}\left(S_\phi\left(x, Y\right) = L\right)\right]$ is twice differentiable with Hessian norm bounded by L_t^J. For $\triangle\phi_t = \phi_{t+1} - \phi_t = \theta_{t+1} - \theta_t = \triangle\theta_t$,

$$\text{if } \langle u_t^J, \triangle\phi_t\rangle > \frac{L_t^J}{2}\|\triangle\phi_t\|^2,\ \mathcal{J}_t > 0;\ \text{if } \langle u_t^J, \triangle\phi_t\rangle < -\frac{L_t^J}{2}\|\triangle\phi_t\|^2,\ \mathcal{J}_t < 0.$$

Proof. Let

$$f(\phi) = \mathbb{E}_{\pi_{t+1}}[\mathbb{P}(S_\phi(x, Y) = L)],$$

with π_{t+1} fixed. Then

$$f(\phi_t) = \tilde{p}_{t,L}, \qquad f(\phi_{t+1}) = p_{t+1,L}, \qquad \nabla f(\phi_t) = u_t^J.$$

Define

$$g(s) = f(\phi_t + s\triangle\phi_t), \qquad s \in [0, 1].$$

By the chain rule,

$$g'(0) = \langle u_t^J, \triangle\phi_t\rangle,\ g''(s) = \triangle\phi_t^\top \nabla_\phi^2 f(\phi_t + s\triangle\phi_t)\triangle\phi_t.$$

The Hessian bound along the segment implies

$$|g''(s)| \leq L_t^J\|\triangle\phi_t\|^2.$$

The one-dimensional Taylor theorem therefore gives

$$\left|p_{t+1,L} - \tilde{p}_{t,L} - \langle u_t^J, \triangle\phi_t\rangle\right| \leq \frac{L_t^J}{2}\|\triangle\phi_t\|^2.$$

Consequently,

$$p_{t+1,L} - \tilde{p}_{t,L} \geq \langle u_t^J, \triangle\phi_t\rangle - \frac{L_t^J}{2}\|\triangle\phi_t\|^2,$$

and

$$p_{t+1,L} - \tilde{p}_{t,L} \leq \langle u_t^J, \triangle\phi_t\rangle + \frac{L_t^J}{2}\|\triangle\phi_t\|^2.$$

The first condition in Eq. (15) makes the probability difference strictly positive, whereas the second makes it strictly negative. Since the logit function is strictly increasing on $(0, 1)$,

$$\text{sign}(\mathcal{J}_t) = \text{sign}(p_{t+1,L} - \tilde{p}_{t,L}),$$

which proves both claims. In the self-judge setting, parameter sharing at successive round starts gives $\triangle\phi_t = \phi_{t+1} - \phi_t = \theta_{t+1} - \theta_t = \triangle\theta_t$. □

Proof of Proposition 5. For fixed $N \geq 2$ and sufficiently small δ under $\delta-$score saturation, $a_{N,t} = N(1 - p_{t,L}) + \mathcal{O}(\delta^2) = \mathcal{O}(\delta)$, and $\omega_{t,L,\ell} = \alpha_{t,\ell} + \mathcal{O}(\delta)$ for every $\ell < L$.

Proof. Let $h := 1 - p_{t,L},\ 0 < h \le \delta$ and H be the number of non-top scores in a candidate group. Then $H \sim \mathrm{Binomial}(N, h)$. A group with $H = 1$ is necessarily accepted, whereas acceptance requires $H \ge 1$. Thus

$$Nh(1-h)^{N-1} = \mathbb{P}(H = 1) \le a_{N,t} \le \mathbb{P}(H \ge 1) \le Nh.$$

It follows that

$$0 \le Nh - a_{N,t} \le Nh[1 - (1-h)^{N-1}] \le N(N-1)h^2.$$

Consequently,

$$a_{N,t} = Nh + O(h^2) = N(1 - p_{t,L}) + O(\delta^2).$$

Also, $a_{N,t} \le N\delta$, so $a_{N,t} = O(\delta)$. In particular,

$$\frac{a_{N,t}}{N(1-p_{t,L})} \to 1 \text{ as } p_{t,L} \to 1.$$

For the allocation statement, Theorem 1 implies

$$0 \le \omega_{t,L,\ell} - c_t \alpha_{t,\ell} \le 1 - c_t,$$

because all residual coefficients are nonnegative and have total mass $1 - c_t$. Since

$$\omega_{t,L,\ell} - \alpha_{t,\ell} = (\omega_{t,L,\ell} - c_t\alpha_{t,\ell}) - (1 - c_t)\alpha_{t,\ell},$$

and both terms on the right belong to $[0, 1 - c_t]$,

$$|\omega_{t,L,\ell} - \alpha_{t,\ell}| \le 1 - c_t \le 1 - p_{t,L}^{N-1} \le (N-1)(1 - p_{t,L}) \le (N-1)\delta.$$

This proves $\omega_{t,L,\ell} = \alpha_{t,\ell} + O(\delta)$. □

Proof of Theorem 4. For fixed $N \ge 2$, at round t, i) Assume $0 < p_{0,L} < 1$, then $\mathbb{E}[V_t] \le \frac{NB_t}{1+\frac{p_{0,L}}{1-p_{0,L}}e^{\Lambda_t}}$. ii) Let $B_t^*(M) := \min\{B \in \mathbb{N} : Ba_{N,t} \ge M\}, M > 0$ be the minimum candidate-group budget required to obtain at least M accepted preference pairs in expectation. If the cumulative gain $\Lambda_t \to \infty$, then $\frac{B_t^*(M)}{\exp(\Lambda_t)} \to \frac{Mp_{0,L}}{N(1-p_{0,L})}$. iii) If the model is $\delta-$score saturated, then $B_t^*(M) \ge \frac{M}{N\delta}$.

Proof. We first prove (i) $\mathbb{E}[V_t] \le \frac{NB_t}{1+\frac{p_{0,L}}{1-p_{0,L}}e^{\Lambda_t}}$. The acceptance probability satisfies $a_{N,t} = 1 - \sum_{\ell=1}^{L} p_{t,\ell}^N \le 1 - p_{t,L}^N$. Using the geometric-sum identity,

$$1 - p_{t,L}^N = (1 - p_{t,L}) \sum_{j=0}^{N-1} p_{t,L}^j \le N(1 - p_{t,L}).$$

Therefore, $\mathbb{E}[V_t] = B_t a_{N,t} \le NB_t(1 - p_{t,L})$. The cumulative-gain identity gives

$$1 - p_{t,L} = \frac{1}{1 + \frac{p_{0,L}}{1-p_{0,L}}e^{\Lambda_t}}.$$

Substitution yields

$$\mathbb{E}[V_t] \le \frac{NB_t}{1 + \frac{p_{0,L}}{1-p_{0,L}}e^{\Lambda_t}}.$$

Next, we prove (ii). For $a_{N,t} > 0$, the constraint $Ba_{N,t} \ge M$ is equivalent to $B \ge M/a_{N,t}$. Hence $B_t^*(M) = \left\lceil \frac{M}{a_{N,t}} \right\rceil$, and

$$0 \le B_t^*(M) - \frac{M}{a_{N,t}} < 1.$$

Assume $\Lambda_t \to +\infty$. The cumulative-gain identity implies $p_{t,L} \to 1$ and

$$(1 - p_{t,L})e^{\Lambda_t} = \frac{e^{\Lambda_t}}{1 + \frac{p_{0,L}}{1-p_{0,L}}e^{\Lambda_t}} \to \frac{1 - p_{0,L}}{p_{0,L}}.$$

By Proposition 5, $\frac{a_{N,t}}{1-p_{t,L}} \to N$. Therefore,

$$a_{N,t}e^{\Lambda_t} = \frac{a_{N,t}}{1-p_{t,L}}(1-p_{t,L})e^{\Lambda_t} \to \frac{N(1-p_{0,L})}{p_{0,L}}.$$

The rounding error satisfies

$$0 \le \frac{B_t^*(M)}{e^{\Lambda_t}} - \frac{M}{a_{N,t}e^{\Lambda_t}} < e^{-\Lambda_t} \to 0.$$

Consequently,

$$\frac{B_t^*(M)}{e^{\Lambda_t}} \to \frac{Mp_{0,L}}{N(1-p_{0,L})}.$$

Since $a_{N,t} \sim N(1-p_{t,L})$ and $a_{N,t} \to 0$, the same rounding argument also gives $B_t^*(M) \sim \frac{M}{N(1-p_{t,L})}$.

Finally, we prove (iii). Under δ-score saturation, $1-p_{t,L} \le \delta$. Hence $a_{N,t} \le N(1-p_{t,L}) \le N\delta$. Every feasible budget B satisfies $Ba_{N,t} \ge M$, so

$$B \ge \frac{M}{a_{N,t}} \ge \frac{M}{N\delta}.$$

Taking the minimum over feasible integer budgets proves

$$B_t^*(M) \ge \frac{M}{N\delta}.$$

If $a_{N,t} = 0$, no finite budget is feasible. □