WHEN SPARSE REWARD MEETS DENSE DISTILLATION: TRAINING DYNAMICS OF ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards provides a ***sparse*** post-training signal: a single binary outcome evaluates the entire rollout, and every token receives the same sequence-level advantage regardless of its individual contribution. To complement this sparse supervision, a growing family of methods adds a scalar-weighted teacher KL term to the policy-gradient objective, providing ***dense*** token-level guidance that may be unreliable at some positions. Despite the benefits of combining these signals, their interaction during optimization can destabilize joint training. To understand how this instability develops, we study the learning dynamics of hybrid reward–distillation training through a neural tangent kernel (NTK) analysis. We introduce the ***cross-signal NTK*** $K_{DR}(n)$, a token-level statistic that measures the alignment between reward and distillation gradients at position n. Through this analysis, we identify two failure modes: ❶ ***Magnitude drowning***, where the reward gradient exceeds the distillation gradient by orders of magnitude, so that even weak directional conflict can cause the distillation loss to ***rise*** despite its explicit inclusion in the training objective; and ❷ ***Localized directional conflict***, where the sequence-level advantage and the teacher's position-specific distribution induce opposing updates at the same token ($K_{DR}(n) < 0$). The severity of these effects depends on the optimization regime: the gradient-norm ratio $\kappa = \|\nabla\mathcal{L}_R\|/\|\nabla\mathcal{L}_D\|$ varies by roughly an order of magnitude across tasks, and our experiments reveal an empirical threshold beyond which naive mixing can lead to persistent training collapse. Motivated by these findings, we introduce the ***M3 family***, which combines magnitude normalization with three strategies for coordinating dense teacher supervision and sparse reward updates: a hard NTK-based mask that retains compatible teacher signals (M3-Select), a continuous relaxation of this mask (M3-Soft), and a fast–slow extragradient step that temporally separates teacher shaping from reward correction (M3-EG). Experiments across four model backbones and four benchmarks show that M3 maintains stable training dynamics and achieves superior performance in high-κ regimes where scalar-mixing baselines collapse.

1 INTRODUCTION

Reinforcement **L**earning with **V**erifiable **R**ewards (**RLVR**) has become a central approach to post-training reasoning-capable large language models (DeepSeek-AI, 2025; OpenAI, 2024). However, its supervision is ***sparse***: a single outcome reward evaluates the entire sequence, and every token receives the same sequence-level advantage regardless of its individual contribution. Process-reward methods partially address this limitation by evaluating intermediate reasoning steps, but step-level supervision does not directly distinguish the contributions of individual tokens. As the complementary, **teacher distillation** provides ***dense*** supervision with target distribution at each token position, although the teacher's guidance may be unreliable at some positions. Therefore, a growing family of hybrid methods therefore augments the policy-gradient objective with a scalar-weighted teacher KL term (Zhao et al., 2026; Agarwal et al., 2024), aiming to provide token-level guidance.

Despite their empirical success (Zhao et al., 2026; Agarwal et al., 2024), these hybrid methods can exhibit ***unstable training dynamics*** and, in some cases, ***catastrophic collapse*** (Figure 1c). The underlying difficulty is that a reliable but sparse outcome signal and a dense but imperfect proxy signal

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

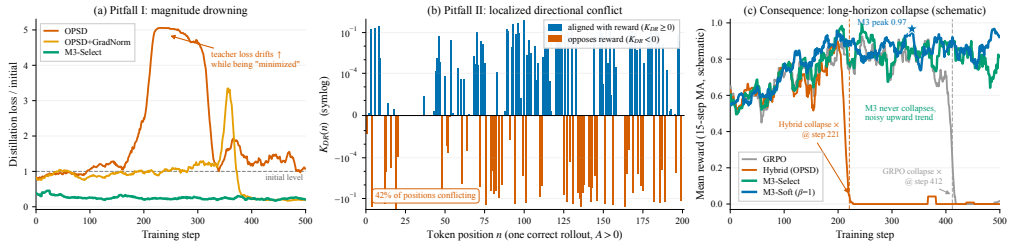


Figure 1: **Two failures of linear reward–distillation mixing** (Qwen3-1.7B, GSM8K). **(a)** Distillation loss rises under OPSD, indicating magnitude drowning. **(b)** Positive and negative $K_{DR}(n)$ interleave in a correct rollout, revealing token-local conflict hidden by aggregation. **(c)** Baselines collapse by step 500, while M3-Select and M3-Soft remain stable.

do not necessarily complement each other. Under naive scalar mixing, **they may instead compete**: ❶ one signal can overwhelm the other, ❷ or their opposing updates can cancel, progressively homogenizing the policy, eliminating reward diversity, and ultimately inducing entropy collapse.

To understand how this instability develops, we study the learning dynamics of hybrid reward–distillation training through an NTK analysis. Building on the neural tangent kernel (NTK) framework (Ren et al., 2025), we examine how reward and distillation updates interact through the model's shared parameters. We introduce the ***cross-signal NTK*** $K_{DR}(n)$, the inner product between the parameter gradients contributed by the two objectives at token position n, which measures their local alignment while accounting for the mapping from token-level residuals to parameter updates through the model's Jacobian. Our analysis identifies two failure modes: ❶ ***Magnitude drowning.*** The reward gradient can exceed the distillation gradient by orders of magnitude, as measured by the norm ratio $\kappa = \|\boldsymbol{g}_R\|/\|\boldsymbol{g}_D\|$. In this regime, even weak negative alignment can make the increase in distillation loss caused by the reward update exceed the decrease produced by the distillation update itself. Consequently, the reward or distillation loss can rise despite its explicit inclusion in the training objective (Figure 1a). ❷ ***Localized directional conflict.*** The two objectives assign updates using different information: RL broadcasts a sequence-level advantage to every token, whereas distillation uses a position-specific teacher distribution. At positions where $K_{DR}(n) < 0$ (Figure 1b), the resulting gradient contributions oppose each other. For a positive-advantage rollout, the teacher update can decrease the probability of a sampled token that the reward update seeks to reinforce. For a negative-advantage rollout, it can instead reinforce a token that the reward update seeks to suppress. Their contributions can therefore cancel in aggregate diagnostics, obscuring local conflicts. Over longer training horizons, repeated conflicting updates can reduce diversity among rollouts and diminish within-group reward variation. When all rollouts in a group receive the conflict reward, their relative advantages and policy gradient vanish. Our experiments exhibit a corresponding progression from initial reward improvement to reduced reward diversity and abrupt collapse (Figure 1c).

The severity of these failure modes depends on the optimization regime. A hybrid run may initially appear stable, with reward increasing even as the distillation loss rises. Across architectures and tasks, κ varies by roughly an order of magnitude, and our experiments reveal an empirical threshold beyond which naive mixing becomes prone to persistent collapse. These findings motivate examining teacher supervision at two levels. At the *run level*, κ indicates the degree of magnitude imbalance and helps distinguish settings where standard scalar mixing remains effective from those where it collapses. At the *token level*, the sign of $K_{DR}(n)$ distinguishes locally compatible teacher updates from conflicting ones. Guided by this diagnosis, we introduce the **M3** family, which combines magnitude normalization with three strategies for coordinating teacher supervision and reward updates: **M3-Select** applies a hard NTK-based mask, retaining teacher supervision only at positions where $K_{DR}(n) \geq 0$; **M3-Soft** replaces this mask with a continuous, temperature-controlled gate; and **M3-EG** temporally separates teacher shaping from reward correction through a fast–slow extragradient step. We make the following contributions:

- We develop an NTK-based framework that characterizes the interaction between RL and distillation through a single per-token statistic, the cross-signal NTK $K_{DR}(n)$. This framework identifies two failure modes of linear mixing and explains why weight-space rebalancing methods.
- We show that κ remains stable throughout a run but varies by roughly an order of magnitude across tasks. A critical threshold separates a stable regime, in which naive mixing is effective,

from a catastrophic regime where it collapses. This yields a simple probe-batch rule for selecting the mixing strategy before full training.

- In high-κ regimes, the M3 family remains stable over horizons at which scalar-mixing baselines collapse. M3-Select is the most robust variant in the highest-κ settings, while M3-Soft recovers catastrophic cases with order-of-magnitude gains. Combined with weight averaging, which removes the gate-variance pathology identified in our analysis, M3-Soft matches or exceeds the strongest baseline across all tested architecture–dataset pairs.

2 Related Work

❶ ***LLM Reasoning via RL and Distillation.*** DeepSeek-R1 (DeepSeek-AI, 2025) spurred RL reasoning (GRPO (Shao et al., 2024), DAPO (Yu et al., 2025), Dr. GRPO (Liu et al., 2025)); OPSD (Zhao et al., 2026) and MiniLLM (Gu et al., 2024) make distillation on-policy. Hybrids differ in coupling: SDPO (Hübotter et al., 2026) self-distills from a reprompt, RLSD (Yang et al., 2026) keeps the teacher as magnitude-only reweighting, and HDPO (Ding, 2026) and DPKD (Li et al., 2024) interpolate losses. Unlike these methods, we study the dynamics of the coupling, identifying when dense teacher supervision conflicts with or is drowned by sparse reward updates.

❷ ***Multi-Objective Gradients and NTK.*** PCGrad (Yu et al., 2020), CAGrad (Liu et al., 2021), MGDA (Sener & Koltun, 2018), and NashMTL (Navon et al., 2022) operate on aggregate task gradients; GradNorm (Chen et al., 2018), UW (Kendall et al., 2018), and DWA (Liu et al., 2019) balance objectives in loss-weight space. Such global operations do not directly resolve conflicts that alternate across tokens, and we show that GradNorm becomes ineffective at $\kappa \gg 1$ (Proposition 5). Prior work applies NTK to gradient conflict and imbalance (Ren et al., 2025; Qin et al., 2025); our cross-signal NTK specializes this perspective to hybrid RL–distillation and links token-level interaction to drowning-induced collapse, complementing known RL failures such as reward hacking (Skalse et al., 2022), entropy collapse (Yu et al., 2025), and length bias (Liu et al., 2025).

3 Preliminaries and Theoretical Analysis

Notation. In this paper, scalars use italic symbols, vectors bold lowercase or Greek symbols (e.g., $\boldsymbol{g}$, $\boldsymbol{\delta}$, $\boldsymbol{\theta}$), and matrices bold uppercase symbols (e.g., $\mathbf{J}$, $\mathbf{K}$). Calligraphic symbols denote sets and losses; $\mathbb{R}$ denotes real numbers. The model has d_{par} parameters, collected in $\boldsymbol{\theta} \in \mathbb{R}^{d_{\text{par}}}$. At position n, $\boldsymbol{p}_S^n$ is the student's token distribution and $p_S^n(v) = [\boldsymbol{p}_S^n]_v$ the probability of token v. We distinguish kernel matrix $\mathbf{K}(n, m)$ from the scalar interaction score $K_{DR}(n)$ and use φ_n for alignment angles.

3.1 Problem Setup and NTK Preliminaries

We train student model $\pi_{\boldsymbol{\theta}}$ using teacher predicts and response-level rewards. For a prompt–answer pair $(x, y^*) \sim \mathcal{D}$, the student generates a response $\hat{y}$ of length N; $\hat{y}_{<n}$ is its prefix before position n.

❶ ***On-Policy Self-Distillation (OPSD).*** OPSD (Zhao et al., 2026) learns from responses sampled by the student itself. A frozen teacher receives the ground-truth answer as additional context, giving $\boldsymbol{p}_T^n = \pi_{\boldsymbol{\theta}_0}(\cdot \mid x, y^*, \hat{y}_{<n})$, while $\boldsymbol{p}_S^n = \pi_{\boldsymbol{\theta}}(\cdot \mid x, \hat{y}_{<n})$. Their distributions are compared at each position, with each contribution capped at τ:

$$\mathcal{L}_D(\boldsymbol{\theta}) = \mathbb{E}_{(x,y^*)\sim\mathcal{D}}\, \mathbb{E}_{\hat{y}\sim\pi_{\boldsymbol{\theta}}(\cdot\,|\,x)} \left[\frac{1}{N} \sum_{n=1}^{N} \min\Big(D_\lambda\big(\boldsymbol{p}_T^n \,\big\|\, \boldsymbol{p}_S^n\big),\ \tau \Big) \right], \tag{1}$$

Here, $D_\lambda(\boldsymbol{p}\|\boldsymbol{q}) = \lambda \mathrm{KL}(\boldsymbol{p}\|\boldsymbol{m}) + (1-\lambda)\mathrm{KL}(\boldsymbol{q}\|\boldsymbol{m})$ is the generalized Jensen–Shannon divergence, with $\boldsymbol{m} = \lambda\boldsymbol{p} + (1-\lambda)\boldsymbol{q}$. We use $\lambda = 1/2$, giving both distributions equal weight. The teacher thus provides dense and position-specific supervision; KL limits are given in Appendix AG.1.

❷ ***Group Relative Policy Optimization (GRPO).*** GRPO (Shao et al., 2024), a critic-free variant of PPO (Schulman et al., 2017), samples G responses per prompt and assigns each a verifiable reward $r^{(i)} = r(\hat{y}^{(i)}, y^*) \in \{0, 1\}$. Its advantage $A_i = (r^{(i)} - \bar{r})/(\sigma_r + \epsilon)$ compares that reward with the group mean $\bar{r}$ and standard deviation σ_r, where $\epsilon > 0$ prevents division by zero:

$$\mathcal{L}_R(\boldsymbol{\theta}) = -\mathbb{E}_{(x,y^*)\sim\mathcal{D}}\, \mathbb{E}_{\{\hat{y}^{(i)}\}_{i=1}^{G}\sim\pi_{\boldsymbol{\theta}}(\cdot\,|\,x)} \left[\frac{1}{G} \sum_{i=1}^{G} A_i\, \frac{1}{N_i} \sum_{n=1}^{N_i} \log \pi_{\boldsymbol{\theta}}\Big(\hat{y}_n^{(i)} \,\Big|\, x, \hat{y}_{<n}^{(i)} \Big) \right], \tag{2}$$

where $N_i = |\hat{y}^{(i)}|$. Positive advantages encourage sampled responses and negative advantages discourage them. The same advantage weights every token, without identifying which tokens caused success or failure. Gradients hold the sampled responses and advantages fixed.

❸ ***NTK in Learning Dynamics.*** Following (Ren et al., 2025), let $\boldsymbol{z}^n \in \mathbb{R}^{|\mathcal{V}|}$ be the policy's logits, the scores converted by softmax into probabilities over vocabulary $\mathcal{V}$. The Jacobian $\mathbf{J}^n = \nabla_{\boldsymbol{\theta}} \boldsymbol{z}^n \in \mathbb{R}^{d_{\text{par}} \times |\mathcal{V}|}$ describes their dependence on the parameters. For $\mathcal{L} = N^{-1} \sum_n \ell^n$, the chain rule gives

$$\nabla_{\boldsymbol{\theta}} \mathcal{L} = \frac{1}{N} \sum_n \mathbf{J}^n \boldsymbol{\delta}^n, \qquad \boldsymbol{\delta}^n = \nabla_{\boldsymbol{z}^n} \ell^n, \qquad \mathbf{K}(n,m) = (\mathbf{J}^n)^\top \mathbf{J}^m. \tag{3}$$

The residual $\boldsymbol{\delta}^n$ describes how the loss changes with each logit. The empirical NTK $\mathbf{K}(n,m)$ (Jacot et al., 2018) couples positions through their shared parameters: an update driven by position m can change predictions at n. Following Ren et al. (2025), we use the Action–Kernel–Gradient (AKG) decomposition to analyze these changes.

3.2 Hybrid Update Dynamics and Loss Interactions

In OPSD, we assume the hybrid loss between policy update and distillation loss is $\mathcal{L}_H = (1-\alpha)\mathcal{L}_R + \alpha \mathcal{L}_D$, where $\alpha \in [0,1]$ weights distillation. A step of size η gives $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}_H$. Expanding the logits to first order and summing over T steps yields:

$$\boldsymbol{z}_T^n = \boldsymbol{z}_0^n + \sum_{t=0}^{T-1} \Delta \boldsymbol{z}_t^n, \qquad \Delta \boldsymbol{z}_t^n = -\frac{\eta}{N} \sum_{m=1}^{N} \mathbf{K}_t(n,m) \Big[(1-\alpha)\, \boldsymbol{\delta}_R^m + \alpha\, \boldsymbol{\delta}_D^m \Big] + O(\eta^2), \tag{4}$$

where $\boldsymbol{z}_t^n = \boldsymbol{z}^n(\boldsymbol{\theta}_t)$ and $\Delta \boldsymbol{z}_t^n = \boldsymbol{z}_{t+1}^n - \boldsymbol{z}_t^n$. At each step, the kernel maps the combined residual at every position m to a change in the logits at n. The residuals are evaluated at step t:

$$\boldsymbol{\delta}_R^m = -A(\boldsymbol{e}_{\hat{y}_m} - \boldsymbol{p}_S^m), \qquad \boldsymbol{\delta}_D^m = \boldsymbol{p}_S^m - \boldsymbol{p}_T^m. \tag{5}$$

The one-hot vector $\boldsymbol{e}_{\hat{y}_m}$ selects the sampled token. The reward residual encourages or discourages this token according to A; the teacher residual compares the full distributions. We use a local forward-KL model of distillation: near agreement, the unclipped JSD gradient is $\lambda(1-\lambda)\boldsymbol{\delta}_D^m + O(\|\boldsymbol{p}_S^m - \boldsymbol{p}_T^m\|^2)$, with the leading constant absorbed into the teacher scale. Clipped tokens contribute zero gradient (Appendix AG.1).

First-Order Loss Changes. With $\boldsymbol{g}_R = \nabla_{\boldsymbol{\theta}} \mathcal{L}_R$ and $\boldsymbol{g}_D = \nabla_{\boldsymbol{\theta}} \mathcal{L}_D$, the same update gives

$$\Delta \mathcal{L}_R \approx -\eta \Big[(1-\alpha) \|\boldsymbol{g}_R\|^2 + \alpha \langle \boldsymbol{g}_R, \boldsymbol{g}_D \rangle \Big], \qquad \Delta \mathcal{L}_D \approx -\eta \Big[\alpha \|\boldsymbol{g}_D\|^2 + (1-\alpha) \langle \boldsymbol{g}_D, \boldsymbol{g}_R \rangle \Big], \tag{6}$$

Each squared-gradient term describes an objective's own decrease; the inner product describes the other update's effect. Positive alignment helps both objectives, while negative alignment opposes their progress. A loss increases only when this opposing contribution exceeds its own decrease within the first-order approximation.

3.3 Token-Level Decomposition of Gradient Interactions

Cross-Position Interactions. The overall inner product can hide local conflicts. Each term pairs the teacher signal at n with the reward signal at m through the shared kernel. The sum includes same-position and cross-position interactions, whose positive and negative contributions can cancel. Expanding both gradients gives that:

$$\langle \boldsymbol{g}_D, \boldsymbol{g}_R \rangle = \frac{1}{N^2} \sum_{n,m} \underbrace{(\boldsymbol{\delta}_D^n)^\top}_{\text{teacher residual}} \underbrace{\mathbf{K}(n,m)}_{\text{kernel coupling}} \underbrace{\boldsymbol{\delta}_R^m}_{\text{reward residual}}. \tag{7}$$

Gradient Magnitude and Alignment. Expanding the squared norms in the same way gives the magnitude ratio when $\|\boldsymbol{g}_D\| > 0$:

$$\kappa^2 = \frac{\|\boldsymbol{g}_R\|^2}{\|\boldsymbol{g}_D\|^2} = \frac{\sum_{n,m} (\boldsymbol{\delta}_R^n)^\top \mathbf{K}(n,m)\, \boldsymbol{\delta}_R^m}{\sum_{n,m} (\boldsymbol{\delta}_D^n)^\top \mathbf{K}(n,m)\, \boldsymbol{\delta}_D^m}, \tag{8}$$

the numerator measures reward-gradient strength and the denominator teacher-gradient strength; $\kappa \gg 1$ indicates strong imbalance. For nonzero gradients, the normalized inner product $\Phi_{DR} =$

$\langle \boldsymbol{g}_D, \boldsymbol{g}_R \rangle / (\|\boldsymbol{g}_D\| \|\boldsymbol{g}_R\|) \in [-1, 1]$ measures direction independently of magnitude. Positive values indicate alignment and negative values interference. These quantities arise from one gradient Gram matrix: its diagonal contains squared norms and its off-diagonal contains the cross inner product (Appendix AG.2).

3.4 Token-Level Conflict and Magnitude Drowning

We first locate conflict at individual positions, then examine how gradient imbalance amplifies its effect on the teacher loss.

Definition 1 (Cross-Signal Token-Level NTK). *Using the local residual model of Section 3.2, define*

$$K_{DR}(n) = \langle \mathbf{J}^n \boldsymbol{\delta}_D^n, \mathbf{J}^n \boldsymbol{\delta}_R^n \rangle, \qquad \boldsymbol{\delta}_D^n = \boldsymbol{p}_S^n - \boldsymbol{p}_T^n, \qquad \boldsymbol{\delta}_R^n = -A \left(\boldsymbol{e}_{\hat{y}_n} - \boldsymbol{p}_S^n \right). \tag{9}$$

where $A = (r - \bar{r})/(\sigma_r + \epsilon)$. This scalar compares the teacher and reward parameter gradients contributed by the same position.

Positive, negative, and zero scores define the sets Ω_+, Ω_-, and Ω_0. A zero score also includes vanishing gradients. $K_{DR}(n)$ is the $m = n$ summand in Eq. 7, before the common factor N^{-2}; it does not include cross-position interactions.

❶ ***Token-Level Conflict.*** On a positive-advantage response, a negative $K_{DR}(n)$ means that the teacher's same-position contribution lowers the sampled token's log-probability. The full update need not do so, since reward and other-position contributions also matter. Corollary 1 derives this local effect. The conflict rate $C_{\mathrm{NTK}} = |\{n : K_{DR}(n) < 0\}|/N$ is approximately 40% in our measurements, an empirical observation rather than a consequence of the definition.

❷ ***Magnitude Drowning.*** Write $\cos\varphi = \Phi_{DR}$. Equation 6 gives $\Delta\mathcal{L}_D \approx -\eta \|\boldsymbol{g}_D\|^2 [\alpha + (1 - \alpha)\kappa \cos\varphi]$. For nonzero gradients and $0 < \alpha < 1$, its first-order sign condition is $\Delta\mathcal{L}_D > 0 \iff \cos\varphi < -\frac{\alpha}{(1-\alpha)\kappa}$. For fixed α, the threshold approaches zero from below as $O(\kappa^{-1})$: when the reward gradient is large, even weak negative alignment can outweigh the teacher's own descent. Imbalance alone is insufficient; negative alignment must also satisfy the threshold.

As the student approaches the teacher distribution, $\boldsymbol{\delta}_D^n$ shrinks, while the reward residual need not shrink at the same rate. The kernel maps both into parameter space, so their magnitudes and directions jointly determine κ. Our experiments identify magnitude imbalance as the dominant failure mode, with ratios reaching 1.3×10^4. Ratios are approximately stable across the tested LoRA ranks 8–64 but vary substantially across tasks (Proposition 7). Extended scaling analysis and projection-baseline limitations appear in Appendices AG.2 and M.

3.5 Reward-Signal Degeneration and Training Collapse

Repeated teacher contributions at conflicting positions may suppress successful response patterns. If the resulting responses become less diverse in reward, GRPO receives less information for distinguishing them. This is a possible training mechanism, not a direct consequence of the local gradient identity. The final step is exact: if all G sampled responses receive the same reward, then $r^{(i)} = \bar{r}$, $\sigma_r = 0$, and every advantage $A_i = 0$. All reward residuals vanish, so that group contributes no reward gradient, $\boldsymbol{g}_R^{(x)} = 0$. Only the teacher term can contribute to its hybrid update. However, one such group does not establish collapse: it may contain all successes or all failures, and later sampling or updates from other prompts may restore reward variation. Persistent failure requires poor responses to remain dominant without recovery of a useful reward signal. The abrupt drops in Figure 1c are consistent with this mechanism; Section 5.2 provides empirical tests.

Remark 1 (Empirical Gate-Selection Threshold). *Our configurations show a transition near $\kappa^* \approx 5 \times 10^3$: hard masking is most useful at high imbalance, while soft gating often retains more useful supervision at lower imbalance. This empirical guideline is not a universal collapse threshold; it depends on the compatible-token fraction, conflict strength, and gate-estimation error. Section 4 introduces the gating methods.*

4 Methodology

M3 turns the preceding analysis into two coupled decisions: how strongly each signal enters the update, and where teacher guidance is admitted. We first normalize token residuals, then allocate

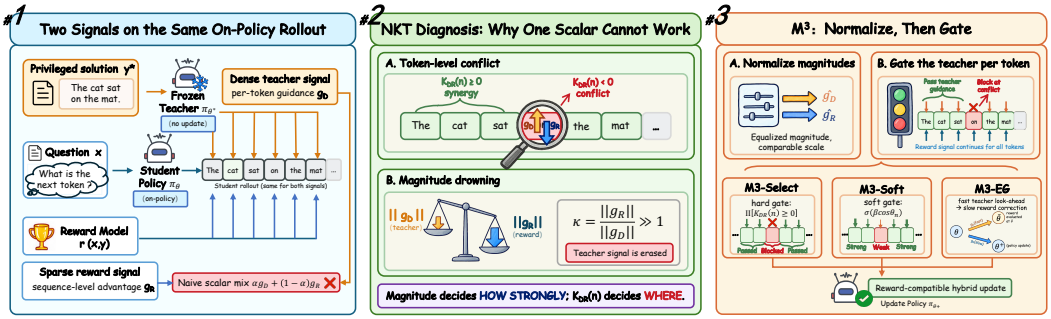


Figure 2: **From interaction analysis to hybrid updates.** Shared parameters couple dense teacher and sparse reward supervision. M3 controls residual scale and allocates teacher influence by token-level compatibility; its extragradient extension separates teacher shaping from reward correction.

teacher weight using local compatibility (Figure 2). M3-Select and M3-Soft implement this allocation with hard and continuous gates. M3-EG extends the same principle to update timing, letting the teacher shape where the reward direction is evaluated.

4.1 Controlling Signal Scale

In a global mixture, comparable weighted gradient norms require $\alpha \approx \kappa/(1+\kappa)$; the coefficient must compensate for scale before it can express a preference between the signals. M3 instead controls scale at the residual level, before conversion into a parameter update. Using the residual convention of Section 3.2, define

$$\widehat{\boldsymbol{\delta}}_q^n = \frac{\boldsymbol{\delta}_q^n}{\|\boldsymbol{\delta}_q^n\| + \epsilon}, \qquad \boldsymbol{h}_q^n = \mathbf{J}^n \widehat{\boldsymbol{\delta}}_q^n, \qquad q \in \{D, R\}. \tag{10}$$

Here $\epsilon > 0$ stabilizes division and keeps zero residuals zero. Normalization reduces the dependence of mixing on raw residual magnitudes. Its positive scaling preserves the sign of $K_{DR}(n)$, but does not equalize parameter-gradient norms after multiplication by $\mathbf{J}^n$. It therefore controls residual scale while leaving the directional coordination problem to the gate. The teacher budget acts on these rescaled contributions, while a common update scale controls their overall size. This separates the intended allocation of supervision from raw magnitude disparity that constrains naive mixing.

4.2 Allocating Teacher Influence Across Positions

With a token-dependent teacher weight $\alpha_n \in [0, \alpha_{\max}]$ and $\alpha_{\max} < 1$, the common update:

$$\begin{aligned} \boldsymbol{h}_H^n &= \underbrace{\alpha_n \boldsymbol{h}_D^n}_{\text{teacher contribution}} + \underbrace{(1-\alpha_n)\boldsymbol{h}_R^n}_{\text{reward contribution}}, \\ \boldsymbol{g}_H &= \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{h}_H^n, \qquad \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \bar{s}\, \boldsymbol{g}_H. \end{aligned} \tag{11}$$

The shared exponential-moving-average scale $\bar{s}$ sets the update magnitude; α_n controls its local composition. Teacher and reward weights vary together: rejecting the teacher restores reward weight to one. The remaining choice is how compatibility determines α_n.

M3-Select: retain compatible teacher contributions. The hard gate directly implements the sign criterion:

$$\alpha_n^* = \alpha_{\max} \mathbb{I}\{K_{DR}(n) \geq 0\}. \tag{12}$$

This removes negative teacher projections onto the same-position reward direction. Since normalization preserves signs, the resulting contribution satisfies

$$\langle \boldsymbol{h}_H^n, \boldsymbol{h}_R^n \rangle = (1-\alpha_n^*)\|\boldsymbol{h}_R^n\|^2 + \alpha_n^* \langle \boldsymbol{h}_D^n, \boldsymbol{h}_R^n \rangle \geq (1-\alpha_{\max})\|\boldsymbol{h}_R^n\|^2. \tag{13}$$

Rejected positions retain the full local reward contribution. This is a same-position guarantee: cross-position interactions in the aggregate update remain. A zero score is admitted by convention and may simply reflect a vanishing signal. The gate ceiling controls how much teacher influence an admitted position receives; the score's sign controls whether it receives that influence at all. These two choices need not be tied to one global loss weight.

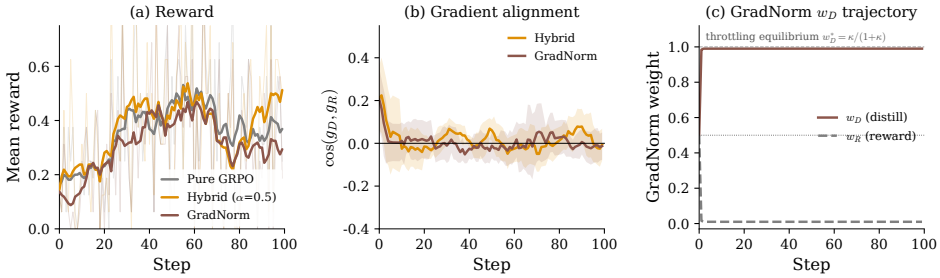


Figure 3: GradNorm challenge on InternLM2.5-1.8B GSM8K: training reward (left), gradient cosine (center), and learned distillation weight (right).

M3-Soft: vary teacher influence continuously. Hard decisions can change abruptly near zero alignment. M3-Soft smooths this transition, retaining partial supervision when compatibility is weak or uncertain:

$$\alpha_n = \alpha_{\max}\sigma(\beta \cos\varphi_n), \quad \cos\varphi_n = \frac{K_{DR}(n)}{\|\boldsymbol{g}_D^n\|\|\boldsymbol{g}_R^n\| + \varepsilon}. \tag{14}$$

Here σ is the logistic sigmoid, $\beta \geq 0$ controls sharpness, and $\varepsilon > 0$ stabilizes the score. Larger β approaches hard selection away from zero; at zero, the weight remains $\alpha_{\max}/2$. Soft gating also retains some negatively aligned teacher contributions, trading strict local exclusion for smoother supervision. It therefore does not inherit Eq. 13 (Appendix AG.4). At $\beta = 0$, all positions receive the same teacher weight $\alpha_{\max}/2$; increasing sharpness progressively makes allocation depend on compatibility. Both variants thus share the same normalized update, with their distinction confined to the teacher-allocation rule.

4.3 Coordinating the Signals in Time

The synchronous variants combine directions evaluated at the same parameters. M3-EG instead uses the gated teacher field $\boldsymbol{m}_D(\boldsymbol{\theta}) = N^{-1}\sum_n \alpha_n^* \boldsymbol{h}_D^n$ to construct a temporary point, then evaluates the normalized reward field $\boldsymbol{h}_R(\boldsymbol{\theta}) = N^{-1}\sum_n \boldsymbol{h}_R^n$ there:

$$\begin{aligned} \widetilde{\boldsymbol{\theta}} &= \boldsymbol{\theta} - \eta_{\text{in}}\boldsymbol{m}_D(\boldsymbol{\theta}), && \text{teacher look-ahead,} \\ \boldsymbol{\theta}^+ &= \boldsymbol{\theta} - \eta_{\text{out}}\boldsymbol{h}_R(\widetilde{\boldsymbol{\theta}}), && \text{reward correction.} \end{aligned} \tag{15}$$

The reward correction is applied from the original parameters, after restoring them; no gradient is propagated through the temporary step. Thus teacher guidance changes the reward evaluation point rather than entering the committed update additively. Appendix E.1 gives the procedure and the conditions on the inner displacement and outer step for local reward descent.

4.4 Training Rule and Scope

Each iteration samples student responses, obtains teacher distributions and verifier advantages, and forms the two residual fields. Their compatibility scores determine the gates; normalized, gated contributions then define either the synchronous direction or the extragradient step (Algorithms 2 and 1). Gates and normalization factors specify update coefficients and are held fixed when applying the direction. Clipped teacher terms and zero-advantage reward terms contribute zero; normalization does not recreate missing supervision.

The local projection bound directly supports the hard selection rule. Aggregate stationarity and conditional variance bounds require a more restrictive population model: smooth lower-bounded reward loss, orthogonal position subspaces, matched teacher and mean reward norms, deterministic conditional teacher directions, and gates fixed before fresh reward noise. These conditions are not enforced by residual normalization. Theorem 1 and Proposition 2 state the resulting guarantees; the experiments assess behavior beyond that model.

Table 1: Cell reports accuracy (top) and its percentage-point change from GRPO (bottom; ↑ / ↓); †Hard masking is vacuous for single-token ARC.

Method	Qwen3-1.7B			Qwen2.5-1.5B			InternLM2.5-1.8B			Llama-3.2-1B		
	GSM8K	SVAMP	ARC	GSM8K	SVAMP	ARC	GSM8K	SVAMP	ARC	GSM8K	SVAMP	ARC
Reference: pure RL												
GRPO	0.696	0.940	0.732	0.726	0.825	0.689	0.404	0.670	0.599	0.552	0.730	0.533
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Teacher-augmented baselines												
Hybrid (OPSD)	0.786	0.932	0.796	0.717	0.857	0.693	0.492	0.608	0.604	0.563	0.722	0.544
	↑+9.0	↓-0.8	↑+6.4	↓-0.9	↑+3.2	↑+0.4	↑+8.8	↓-6.2	↑+0.5	↑+1.1	↓-0.8	↑+1.1
OPSD+GradNorm	0.765	0.917	0.752	0.735	0.837	0.695	0.459	0.687	0.590	0.547	0.717	0.370
	↑+6.9	↓-2.3	↑+2.0	↑+0.9	↑+1.2	↑+0.6	↑+5.5	↑+1.7	↓-0.9	↓-0.5	↓-1.3	↓-16.3
RLSD	0.8317	0.923	0.748	0.725	0.860	0.684	0.471	0.737	0.602	**0.568**	0.648	0.545
	↑+13.6	↓-1.7	↑+1.6	↓-0.1	↑+3.5	↓-0.5	↑+6.7	↑+6.7	↑+0.3	↑+1.6	↓-8.2	↑+1.2
SDPO	0.587	0.567	0.742	0.085	0.612	0.424	0.316	0.432	0.586	0.220	0.483	0.432
	↓-10.9	↓-37.3	↑+1.0	↓-64.1	↓-21.3	↓-26.5	↓-8.8	↓-23.8	↓-1.3	↓-33.2	↓-24.7	↓-10.1
***Ours**: boundary-gated mixing (M3)*												
M3-Select†	0.803	0.943	0.275	0.459	0.585	0.571	0.394	0.538	0.599	0.516	0.673	0.459
	↑+10.7	↑+0.3	↓-45.7	↓-26.7	↓-24.0	↓-11.8	↓-1.0	↓-13.2	0.0	↓-3.6	↓-5.7	↓-7.4
M3-Soft	0.8324	0.945	**0.807**	**0.752**	**0.905**	**0.736**	**0.501**	**0.777**	0.608	**0.568**	0.750	**0.556**
	↑+13.6	↑+0.5	↑+7.5	↑+2.6	↑+8.0	↑+4.7	↑+9.7	↑+10.7	↑+0.9	↑+1.6	↑+2.0	↑+2.3
M3-EG	**0.8446**	**0.952**	0.731	0.749	0.868	0.692	0.440	0.468	**0.6135**	0.552	**0.758**	0.534
	↑+14.9	↑+1.2	↓-0.1	↑+2.3	↑+4.3	↑+0.3	↑+3.6	↓-20.2	↑+1.5	0.0	↑+2.8	↑+0.1

M3-Soft ≥ best non-M3 baseline in **12/12** cells (11 strict wins, 1 exact tie).

Table 2: Long-horizon reward on GSM8K after 500 steps (last-50 mean). ↓ denotes collapse.

	Reference and adaptive baselines			***Ours**: boundary-gated mixing*	
Architecture	**GRPO**	**Hybrid**	**GradNorm**	**M3-Select**	**M3-Soft**
Qwen3-1.7B	↓0.003	↓0.000	↓0.000	0.809	**0.866**
Qwen3-0.6B	↓0.003	↓0.000	↓0.053	0.644	**0.673**
Llama-1B	↓0.004	↓0.003	↓0.328	**0.497**	0.485

5 Experiments

5.1 Setup

Our main evaluation covers Qwen3-1.7B, Qwen2.5-1.5B, InternLM2.5-1.8B, and Llama-3.2-1B on GSM8K, SVAMP, and ARC-Challenge, with Qwen3-0.6B and MATH added in supplementary long-horizon runs. We define $\bar{\kappa}$ as the mean of κ_t over RL-active steps of the corresponding naive-Hybrid run. Full experimental details are provided in Appendix A.

5.2 Does Magnitude Drowning Actually Occur?

Figure 1a shows loss inversion under scalar mixing. With $\bar{\kappa}$ ranging from 1.4×10^3 to 1.3×10^4, the reward update overwhelms distillation; panel c shows the resulting long-horizon collapse. We next test whether global norm balancing fixes it.

A controlled challenge. GradNorm trails both GRPO and naive Hybrid in training reward (Figure 3, left), despite a near-zero aggregate gradient cosine (center). Its distillation weight rapidly approaches $w_D^* = \kappa/(1+\kappa) \approx 1$ (right), leaving the effective reward weight at only $1 - w_D^* \approx 1/\kappa$. GradNorm therefore achieves global norm balance only by nearly turning off RL; because the same weight is applied to every token, it also cannot separate compatible from conflicting teacher updates. This motivates M3's position-dependent gate; further diagnostics are reported in Appendix B.4.

5.3 Can Token-Level Gating Beat Global Mixing?

Accuracy across architectures. Across four architectures and three datasets (Table 1), M3-Soft matches or exceeds the strongest non-M3 baseline in all 12 cells under the per-cell best-observed protocol (11 strict wins, one tie). The largest gains occur on Qwen2.5-SVAMP (0.905, +4.5pp over RLSD), Qwen2.5-ARC (0.736, +4.1pp over OPSD+GradNorm), and InternLM-SVAMP (0.777, +4.0pp over RLSD). On the two Llama arithmetic cells, M3-Soft ties RLSD on GSM8K (0.568) and improves over GRPO on SVAMP (0.750, +2.0pp). The Qwen3 margins are tighter: 0.8324 on GSM8K (+0.07pp over RLSD) and 0.945 on SVAMP (+0.5pp over GRPO).

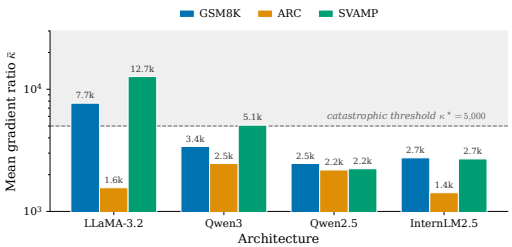


Figure 4: Mean gradient-magnitude ratio $\bar{\kappa}$ for 12 architecture–dataset pairs under naive Hybrid; dashed line: empirical threshold $\kappa^* = 5{,}000$.

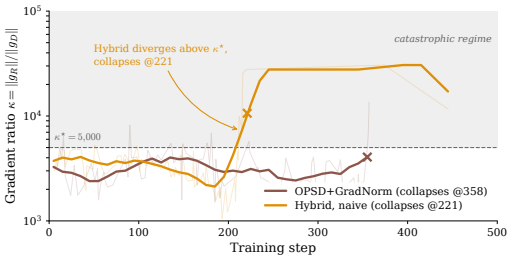


Figure 5: Runtime magnitude ratio. Hybrid and GradNorm cross $\kappa^* = 5{,}000$ before collapsing at steps 221 and 358.

Table 3: One-step sampled-token log-probability change by pre-update conflict region (30 trajectories, 8,492 tokens). Negative values indicate suppression by the update.

Region	OPSD	GRPO	Hybrid
Compatible ($K_{DR} > \epsilon$, n=851)	$+1.6 \times 10^{-3}$	$+2.7 \times 10^{-3}$	$+2.0 \times 10^{-3}$
Neutral ($\lvert K_{DR}\rvert \leq \epsilon$, n=5191)	≈ 0	≈ 0	≈ 0
Conflicting ($K_{DR} < -\epsilon$, n=2450)	$+0.3 \times 10^{-4}$	$+7.3 \times 10^{-4}$	-2.4×10^{-4}

Long-horizon stability. Table 2 compares last-50 training reward after 500 steps on GSM8K. Across all three architectures, every reference or globally balanced baseline triggers the collapse criterion, whereas M3-Select and M3-Soft remain non-collapsed. M3-Soft attains the highest final reward on both Qwen3 scales, while M3-Select is marginally higher on Llama-3.2-1B. Detailed phase-wise trajectories and collapse times for Qwen3-1.7B are reported in Table 8 of Appendix B.2.

Regime Dependence of Token-Level Gating. Across the 12 architecture–dataset cells, $\bar{\kappa}$ ranges from 1,423 to 12,705 (Figure 4). Figure 5 provides the within-run counterpart: on Qwen3-1.7B GSM8K, Hybrid crosses the empirical threshold $\kappa^* = 5{,}000$ near step 200 before collapsing at step 221, while GradNorm crosses later and collapses at step 358. In the two high-κ Llama arithmetic cells (7,690/12,705), Hybrid's reference-rate training reward falls to 0.031/0.000 on GSM8K/SVAMP, whereas M3-Soft reaches 0.647/0.822 (Appendix B.1). At lower κ, hard masking can be overly restrictive: on Qwen3-ARC ($\bar{\kappa} = 2{,}465$), M3-Select reaches 0.275 accuracy, compared with 0.807 for M3-Soft (Table 1). Thus, $\bar{\kappa}$ indicates the severity of magnitude imbalance and the appropriate gate strength, rather than directly predicting accuracy. Together, these results support κ as a regime indicator for instability and appropriate gate strength, rather than a predictor of absolute accuracy; exact per-cell $\bar{\kappa}$ values and reference-rate rewards are in Appendix B.1.

Local Consequence of Token-Level Conflict. We test the sign-specific prediction on 30 correct Qwen3-1.7B GSM8K trajectories comprising 8,492 tokens. Tokens are partitioned by the pre-update cross-signal NTK, after which we apply one OPSD, GRPO, or hybrid update. On compatible positions, the hybrid update increases the sampled-token log-probability by 2.0×10^{-3}; on conflicting positions, it decreases it by 2.4×10^{-4} (Table 3). Over the same conflicting subset, the pure OPSD and GRPO updates yield positive changes. Because the partition is computed before the update, the result tests the local sign prediction rather than defining conflict from the observed logit change.

6 Conclusion and Future Work

We study when dense teacher supervision can complement sparse verifiable rewards in reasoning-model post-training. Our NTK analysis separates their interaction into token-level compatibility, captured by the cross-signal NTK $K_{DR}(n)$, and scale imbalance, captured by κ, exposing localized directional conflict and magnitude drowning. This diagnosis leads to the M3 family, which combines magnitude normalization with hard, soft, or temporally decoupled compatibility gating. Across four model families and three reasoning benchmarks, M3-Soft matches or exceeds the strongest non-M3 baseline, while M3 variants remain stable over 500-step GSM8K training where non-gated baselines collapse. In the future, we plan to scale this application to support larger and more complex agentic scenarios such as coding and deep research.

AI USE STATEMENT

In this work, large language models (LLMs) were used for language polishing, figure design assistance, coding support, and mathematical proof assistance. Specifically, LLMs were used to improve the clarity, grammar, and readability of the manuscript, refine its stylistic quality, and suggest alternative phrasings to reduce redundancy. They also provided suggestions for figure design, visualization layouts, and graphical presentation; all final figures were created, verified, and curated by the authors using the authors' experimental data. In addition, LLMs assisted with code development, intermediate mathematical derivations, proof construction, and consistency checking. All assumptions, formal statements, derivations, proofs, code, figures, and other LLM-assisted content were independently reviewed, verified, and revised by the authors before inclusion. The research questions, core methodology, scientific contributions, experimental design, critical analyses, and final decisions were independently determined by the authors.

ETHICS STATEMENT

All experiments in this work were carried out using publicly available language models and standard reasoning benchmarks, including GSM8K, MATH, SVAMP, and ARC-Challenge, in accordance with their respective licenses and terms of use. The study does not involve human or animal subjects, and we did not collect, use, or disclose any personally identifiable information or private user data.

REPRODUCIBILITY STATEMENT

All experiments use publicly available base models (Qwen3-0.6B, Qwen3-1.7B, Qwen2.5-1.5B, InternLM2.5-1.8B, and Llama-3.2-1B) and standard benchmarks (GSM8K, MATH, SVAMP, and ARC-Challenge). Full hyperparameters, training protocols, held-out split construction, and the rerun/best-of-candidates selection procedure are specified in Appendix A and Section 5.1. The cross-signal NTK diagnostic $K_{DR}(n)$ and the M3 gating rules are specified in closed form in Sections 3.4–4 and Algorithm 2; no proprietary data or infrastructure is required to reproduce the main results.

REFERENCES

Rishabh Agarwal et al. On-policy distillation of language models: Learning from self-generated mistakes. *ICLR*, 2024.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International Conference on Machine Learning*, 2018.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Ken Ding. HDPO: Hybrid distillation policy optimization via privileged self-distillation. *arXiv preprint arXiv:2603.23871*, 2026.

Yuxian Gu et al. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Yixing Li, Yuxian Gu, Li Dong, Dequan Wang, Yu Cheng, and Furu Wei. Direct preference knowledge distillation for large language models. *arXiv preprint arXiv:2406.19774*, 2024.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-Like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In *International Conference on Machine Learning*, 2022.

OpenAI. Learning to reason with LLMs. 2024.

Xiaohan Qin, Xiaoxing Wang, Ning Liao, and Junchi Yan. NTKMTL: Mitigating task imbalance in multi-task learning from neural tangent kernel perspective. *Advances in Neural Information Processing Systems*, 2025. arXiv:2510.18258.

Yi Ren, Danica J Sutherland, et al. Learning dynamics of LLM finetuning. In *International Conference on Learning Representations*, 2025.

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

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Joar Skalse, Nikolaus H R Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled RLVR. *arXiv preprint arXiv:2604.03128*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Table 4: Default experimental hyperparameters.

Component	Hyperparameter	Value
LoRA	Rank r	64
LoRA	Scaling α_{LoRA}	128
Distillation	Divergence	Generalized JSD
Distillation	JSD coefficient λ	$\frac{1}{2}$
Distillation	Token clipping τ	0.05
Reinforcement learning	Rollouts per group G	8
Hybrid baseline	Default mixing coefficient α	0.5
Optimization	Reference learning rate	5×10^{-5}

A Experimental Details

Rollout and optimization protocol. We use the on-policy OPSD construction of Section 3.1. With LoRA disabled, the frozen teacher receives a privileged prompt containing the ground-truth solution, whereas the student receives only the problem and its sampled prefix. Unless otherwise stated, experiments use the defaults in Table 4.

Evaluation metrics and regime statistics. Training reward is the last-20-step mean unless a table states otherwise, and accuracy is measured by greedy decoding on GSM8K ($n = 1319$), SVAMP ($n = 600$), and ARC ($n = 1172$). The SVAMP split excludes all prompts used by the 100- and 200-step training runs. For each architecture–dataset pair, $\bar{\kappa}$ is the mean of $\kappa_t = \|\boldsymbol{g}_R(t)\|/\|\boldsymbol{g}_D(t)\|$ over RL-active steps of the corresponding naive-Hybrid run and is used as a regime indicator. A run is marked as collapsed when reward remains below 0.05 for ten consecutive steps.

Candidate selection and post-processing. For each method–cell pair, the evaluated candidates include the reference run and, where available, seed replicates, checkpoints every 20 steps for horizons up to $T \in \{100, 200, 400\}$, and conservative-rate reruns at 1–2 $\times 10^{-5}$. M3-Soft additionally includes the evaluated $(\beta, \alpha_{\max}, T)$ sweep. SDPO uses its selected conservative-rate candidate; on InternLM-ARC, this is the 20-step pre-collapse checkpoint. Conservative-rate candidates are made available to all methods on Llama GSM8K and SVAMP. Collapsed runs are retained and reported at their measured reward and accuracy rather than excluded. SWA is applied to the Qwen3- and InternLM-GSM8K chains and to the Llama-GSM8K conservative-rate seed-42 chain using last-two-checkpoint averaging. Because candidates and checkpoints are chosen using accuracy and no independent validation split is defined, these entries are reported as *best observed* rather than validation-selected results. Seed-level robustness and SWA values are reported in Section B.1.

B Extended Experiments

B.1 Detailed Cross-Architecture Results

Accuracy. Table 5 directly compares M3-Soft with the strongest non-M3 baseline in each architecture–dataset cell. Under the per-cell best-observed protocol, M3-Soft matches or exceeds the strongest baseline in all 12 cells (11 strict wins, one tie). The largest gains occur on Qwen2.5-SVAMP (0.905, +4.5pp over RLSD), Qwen2.5-ARC (0.736, +4.1pp over OPSD+GradNorm), and InternLM-SVAMP (0.777, +4.0pp over RLSD); Llama-GSM8K is the exact tie with RLSD at 0.568.

Seed robustness and SWA. Seed replicates qualify the best-observed results in Table 5. **Wide-margin cells.** The ranking is stable on Qwen2.5-SVAMP, where the M3-Soft seed mean remains +2.4pp above RLSD, and on Llama-SVAMP, where both replicates exceed GRPO by +7.7 and +10.2pp. Qwen2.5-GSM8K also wins for both available replicates, while two of three Qwen2.5-ARC seeds exceed the strongest baseline and the third ties it. **Tight-margin cells.** The Qwen3-GSM8K replicates remain within 1.2pp below RLSD before SWA, and the InternLM-GSM8K win is likewise obtained only after averaging. SWA raises the selected Qwen3-, InternLM-, and Llama-GSM8K chains by 0.013, 0.015, and 0.004, respectively, yielding accuracies of 0.8324, 0.501, and 0.568; it is not uniformly beneficial, decreasing the Llama seed-43 chain from 0.562 to 0.552. Llama-GSM8K is seed-fragile at the reference learning rate, where two of three replicates collapse, but both conservative-rate replicates survive (0.568/0.552). We therefore treat the tight cells as best-observed parity rather than seed-robust separation.

Table 5: Accuracy of M3-Soft versus the strongest non-M3 baseline.

Architecture	Best non-M3 baseline GSM8K	SVAMP	ARC	M3-Soft GSM8K	SVAMP	ARC	M3-Soft $\geq$ Baseline
Qwen3-1.7B	0.8317	0.940	0.796	**0.8324**	**0.945**	**0.807**	3/3
Qwen2.5-1.5B	0.735	0.860	0.695	**0.752**	**0.905**	**0.736**	3/3
InternLM2.5-1.8B	0.492	0.737	0.604	**0.501**	**0.777**	**0.608**	3/3
Llama-3.2-1B	**0.568**	0.730	0.545	**0.568**	**0.750**	**0.556**	3/3
				Total: 11 strict wins and 1 exact tie			**12/12**

Training reward and magnitude ratio. Table 6 reports per-cell training rewards and the exact $\bar{\kappa}$ measured on the corresponding naive-Hybrid runs. The largest reference-rate separations occur on Llama-GSM8K ($\bar{\kappa} = 7{,}690$; M3-Soft 0.647, $20.9\times$ over Hybrid 0.031) and Llama-SVAMP ($\bar{\kappa} = 12{,}705$; M3-Soft 0.822, $11.4\times$ over GRPO 0.072, while Hybrid reaches 0.000). At the low-κ end, Qwen3-ARC ($\bar{\kappa} = 2{,}465$) favors Hybrid in training reward (0.919 vs. 0.775 for the extended-training M3-Soft candidate). For Llama-GSM8K, this table reports the reference-rate M3-Soft run (0.647), whereas Table 1 uses the conservative-rate SWA chain selected by accuracy (reward 0.591, accuracy 0.568). Because the M3-Soft column includes tuned and, where marked, extended-training candidates, this table is an optimization diagnostic rather than a matched-budget comparison.

Table 6: Per-cell training reward (last-20 mean) and mean gradient-magnitude ratio $\bar{\kappa}$. M3-Soft reports the selected evaluated configuration; * denotes extended training, and $\bar{\kappa}$ is measured on the corresponding naive-Hybrid run.

Architecture	Dataset	GRPO	OPSD	OPSD+GradNorm	M3-Select	M3-Soft	$\bar{\kappa}$
Qwen3-1.7B	GSM8K	0.838	0.825	0.731	0.844	**0.881**	3,404
	SVAMP	0.922	0.894	0.894	0.947	**0.969**	5,080
	ARC	0.741	**0.919**	0.769	0.263	0.775*	2,465
Qwen2.5-1.5B	GSM8K	0.806	0.747	0.653	0.153	**0.828***	2,467
	SVAMP	0.769	0.766	0.681	0.331	**0.778***	2,234
	ARC	0.659	0.650	0.644	0.319	**0.700**	2,184
InternLM2.5-1.8B	GSM8K	0.369	0.428	0.306	0.100	**0.472***	2,745
	SVAMP	0.594	0.563	0.594	0.366	**0.688***	2,690
	ARC	0.616	0.634	0.666	0.416	**0.682***	1,423
Llama-3.2-1B	GSM8K	0.000	0.031	0.006	0.597	**0.647**	7,690
	SVAMP	0.072	0.000	0.038	0.488	**0.822**	12,705
	ARC	0.522	0.469	0.353	0.406	**0.650***	1,562

B.2 Long-Horizon Stability and Runtime Diagnostics

Cross-dataset long-horizon results. Table 7 extends the 500-step evaluation to MATH, SVAMP, and ARC-Challenge using Qwen3-1.7B. Under the reference configurations, prolonged training can still trigger collapse beyond GSM8K: GRPO and OPSD+GradNorm collapse on MATH, while Hybrid collapses on SVAMP. In contrast, both M3 variants remain non-collapsed across all three datasets.

Phase-wise collapse dynamics. Table 8 resolves the Qwen3-1.7B GSM8K runs into 100-step phases. Collapse denotes reward below 0.05 for ten consecutive steps. Unless noted otherwise, the runs use the reference learning rate 5×10^{-5}; all entries are training rewards rather than held-out accuracies.SDPO collapses first at step 50, followed by Hybrid at 221, OPSD+GradNorm at 358, RLSD at 362, and GRPO at 412. In contrast, M3-Select does not trigger the collapse criterion within 500 steps. Phase-wise M3-Soft results are unavailable, but its last-50 reward is reported in Table 2.

Post-collapse conflict diagnostic. After Hybrid collapses, its measured conflict rate falls to zero because the RL gradient vanishes, not because the two signals become compatible. M3-Select instead maintains an active conflict rate near $C_{\mathrm{NTK}} = 0.3$ while preserving reward (Figure 6).

Table 7: Long-horizon reward across datasets on Qwen3-1.7B after 500 steps (last-50 mean). M3-Soft uses the best evaluated gate configuration per dataset; other methods use the reference configuration. Bold marks the column maximum, and ↓ denotes collapse.

Method	MATH	SVAMP	ARC-Challenge
Reference and adaptive baselines			
GRPO	↓0.018	0.900	0.773
Hybrid (OPSD)	0.448	↓0.170	0.765
OPSD+GradNorm	↓0.015	**0.973**	**0.790**
***Ours**: boundary-gated mixing (M3)*			
M3-Select	0.367	0.943	0.282
M3-Soft	**0.523**	0.950	0.667

Table 8: Phase-wise training reward over 500 steps on Qwen3-1.7B GSM8K. "Full" is the all-step mean, "Trend" gives the first collapse step, and † denotes batch size 1.

Method	1–100	101–200	201–300	301–400	401–500	Full	Trend
Pure GRPO	0.840	0.801	0.809	0.802	0.075	0.665	collapse @412
Hybrid (OPSD, $\alpha=0.5$)	0.859	0.762	0.114	0.006	0.001	0.348	collapse @221
OPSD+GradNorm†	0.877	0.823	0.868	0.464	0.000	0.606	collapse @358
RLSD (Yang et al., 2026)	0.855	0.874	0.871	0.259	0.005	0.573	collapse @362
SDPO (Hübotter et al., 2026)	0.306	0.000	0.000	0.000	0.000	0.061	collapse @50
M3-Select ($\alpha_{\max}=0.15$)	0.846	0.818	0.853	0.846	0.810	0.835	stable

B.3 Gate Sensitivity Across Regimes

Controlled sharpness sweep. We first isolate gate sharpness on Qwen3-1.7B GSM8K at 100 steps. With $\alpha_{\max} = 0.5$, the gentle $\beta = 0.5$ gate attains the largest last-20 reward (0.881), compared with 0.856 for $\beta = 1$, 0.850 for GRPO, and 0.844 for hard M3-Select (Table 9). The soft-gate configurations have similar values of the count-based budget proxy α_{eff} (0.439–0.446), consistent with gate shape, rather than a large change in this proxy—driving the observed differences.

Table 9: M3-Soft sharpness sweep on Qwen3-1.7B GSM8K at 100 steps ($\alpha_{\max} = 0.5$). α_{eff} denotes the mean teacher weight. For hard selection, the measured conflict rate is $C_{\text{NTK}} \approx 0.40$, so approximately 60% of tokens retain weight $\alpha_{\max}$, giving $\alpha_{\text{eff}} \approx 0.5 \times 0.60 = 0.30$.

Method	β	Last-10	Last-20	α_{eff}
GRPO	—	0.858	0.850	—
OPSD (uniform)	—	0.833	0.825	0.500
M3-Select (hard)	∞	0.855	0.844	$\sim$0.30
M3-Soft	10	0.875	0.847	0.446
M3-Soft	5	0.813	0.838	0.439
M3-Soft	1	0.906	0.856	0.444
M3-Soft	**0.5**	**0.913**	**0.881**	0.446

Cross-architecture reward sensitivity. Table 10 fixes the horizon at 100 steps and compares $\beta \in \{0.5, 1, 5\}$ across all 12 architecture–dataset cells. A gentle gate ($\beta \in \{0.5, 1\}$) is the best evaluated M3-Soft setting in $11/12$ cells. Both high-κ Llama arithmetic cells favor $\beta = 1$, whereas InternLM-SVAMP is the sole $\beta = 5$ exception. This table diagnoses sensitivity within M3-Soft; it is not the source of the best-observed headline in Table 1.

Sensitivity. Training reward does not by itself select the best gate for accuracy. On Qwen2.5-SVAMP, the very-soft recipe $(\beta, \alpha_{\max}) = (0.1, 0.001)$ reaches 0.884 ± 0.026 across three seeds ($0.855/0.892/0.905$), exceeding RLSD's 0.860 by 2.4pp in the seed mean. On InternLM-SVAMP, the same recipe gives $0.710/0.777/0.722$: the mean (0.736) is at parity with RLSD (0.737), while the best observed seed reaches 0.777. Thus, the Qwen2.5 gain is seed-robust, whereas the InternLM headline is best-observed rather than a mean separation.

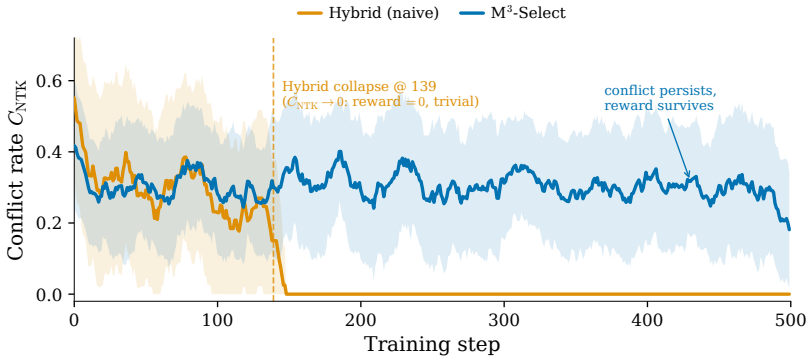


Figure 6: Runtime token-level conflict on Qwen3-0.6B GSM8K. Hybrid's post-collapse drop reflects a vanishing RL gradient; M3-Select remains active near $C_{\mathrm{NTK}} = 0.3$.

Table 10: M3-Soft sharpness sensitivity at 100 steps (last-20 mean reward). $\alpha_{\max} = 0.05$ by default and 0.025 on Qwen3-SVAMP.

Model	Dataset	$\beta = 0.5$	$\beta = 1$	$\beta = 5$	GRPO
Qwen3-1.7B	GSM8K	0.834	**0.856**	0.850	0.838
	SVAMP	**0.991**	0.969	0.981	0.953
	ARC	**0.516**	0.350	0.356	0.741
Llama-3.2-1B	GSM8K	0.569	**0.647**	0.353	↓0.000
	SVAMP	0.456	**0.822**	0.609	0.072
	ARC	**0.559**	0.500	0.388	0.522
Qwen2.5-1.5B	GSM8K	0.697	**0.728**	0.169	0.806
	SVAMP	**0.794**	0.691	0.463	0.769
	ARC	**0.700**	0.475	0.369	0.659
InternLM2.5-1.8B	GSM8K	0.234	**0.269**	0.138	0.369
	SVAMP	0.594	0.366	**0.744**	0.594
	ARC	0.469	**0.575**	0.472	0.616

B.4 Mechanistic Validation

The following diagnostics test the mechanism at progressively coarser levels. We first verify the predicted one-step effect after partitioning tokens by their pre-update cross-signal NTK, then compare response-level and token-level notions of conflict, and finally summarize the aggregate behavior across settings.

Table 11: Response-level semantic conflict (Qwen3-0.6B, 208 trajectories): on-policy = teacher evaluates the student's own trajectory, off-policy = ground-truth contexts; true opposition = wrong trajectories with opposing RL and distillation gradients.

Mode	Conflict %	$\overline{\cos}(\boldsymbol{g}_R, \boldsymbol{g}_D)$	True opposition	κ_{median}	$\bar{\mathcal{L}}_D$
On-policy OPSD	75%	+0.08	25/157 (16%)	388	0.098
Off-policy	75%	−0.01	90/157 (57%)	68	0.810

Response-Level Semantic Conflict. Response-level labels and token-level interactions answer different questions. For each trajectory in a GRPO group, we measure its reward, normalized advantage, trajectory-level gradient cosine $\cos(\boldsymbol{g}_R, \boldsymbol{g}_D)$, and magnitude ratio κ. Across 20 analysis steps with Qwen3-0.6B (batch size 2, group size 8), groups with zero reward variance are excluded because GRPO assigns zero advantage and hence no reward gradient, leaving 208 trajectories. Both modes label 75% of trajectories as semantically conflicting, but they differ sharply in gradient behavior. Among the 157 incorrect trajectories, true opposition occurs in 25 cases (16%) on-policy and 90 cases (57%) off-policy; the corresponding median magnitude ratios are 388 and 68. Within the on-policy sample, the mean cosine is +0.14 on incorrect trajectories and −0.11 on correct tra-

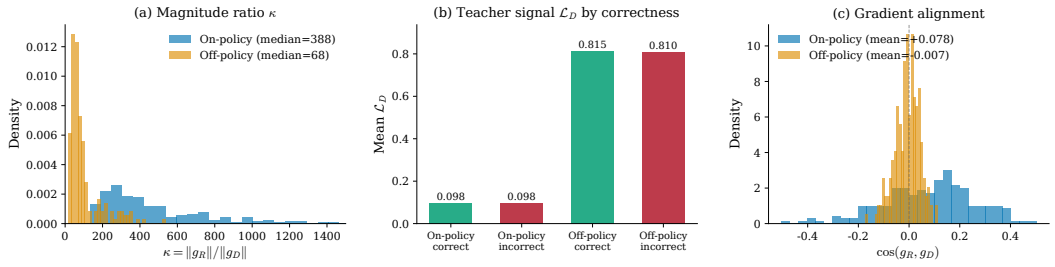


Figure 7: On-policy versus off-policy semantic conflict. On-policy distillation has a higher median κ but fewer incorrect trajectories with genuinely opposed reward and distillation gradients; off-policy supervision reverses this pattern.

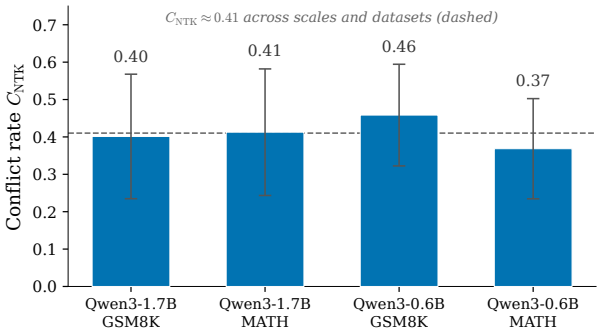


Figure 8: Per-position conflict rate C_{NTK} across model scales and datasets. The fraction of positions with negative cross-signal NTK remains between 0.37 and 0.46 (mean $\approx$ 0.41, dashed line).

jectories, matching the inversion predicted by Proposition 13. Thus, response-level rejection alone does not determine whether the two gradients oppose each other.

Token-Level NTK Conflict. We next measure the same-position cross-signal NTK $K_{DR}(n)$. Positions with $K_{DR}(n) \geq 0$ are locally compatible, whereas positions with $K_{DR}(n) < 0$ receive a teacher component that opposes reward progress. Across the measured model scales and datasets, the count-based conflict rate ranges from 0.37 to 0.46, with a mean of approximately 0.41 (Figure 8). The narrow range shows that the conflict observed in the main-text rollout is not isolated to one model or dataset, without implying that all settings have identical conflict structure.

Figure 9 restores token identities for one correct and one incorrect response to the same prompt. Across the four paired case studies, conflict covers 44.3% of positions on correct rollouts and 54.9% on incorrect rollouts, a difference of 10.6 percentage points. Every rollout nevertheless interleaves compatible and conflicting positions, so these examples support token-level localization rather than a response-level cutoff. The count difference is descriptive for the eight visualized rollouts and is not presented as a population-level estimate.

Aggregate Conflict Diagnostics. Figure 11 summarizes the reference Qwen diagnostics. Across the measured scales and datasets, the mean magnitude ratio remains in the 3,400–4,000 range, while the parameter-level cosine averages -0.009. The M^3-Select trace also shows higher count-based conflict at lower-reward steps. This last relationship is descriptive: the plot does not establish that conflict rate alone causes or predicts method performance.

B.5 Additional Ablations and Illustrative Examples

LoRA Rank Ablation. We vary the LoRA rank from 8 to 128 on Qwen3-1.7B GSM8K while holding the remaining M^3-Soft configuration fixed. The mean reward ranges from 0.824 to 0.844 across this 16$\times$ change in trainable rank, with the best value at $r = 64$ and a total spread of 0.020 (Table 12). Thus, performance in this setting is not sensitive to the tested LoRA rank. Because κ and the token-level conflict rate were not separately recorded for every rank, this ablation supports reward robustness rather than rank invariance of the underlying diagnostics.

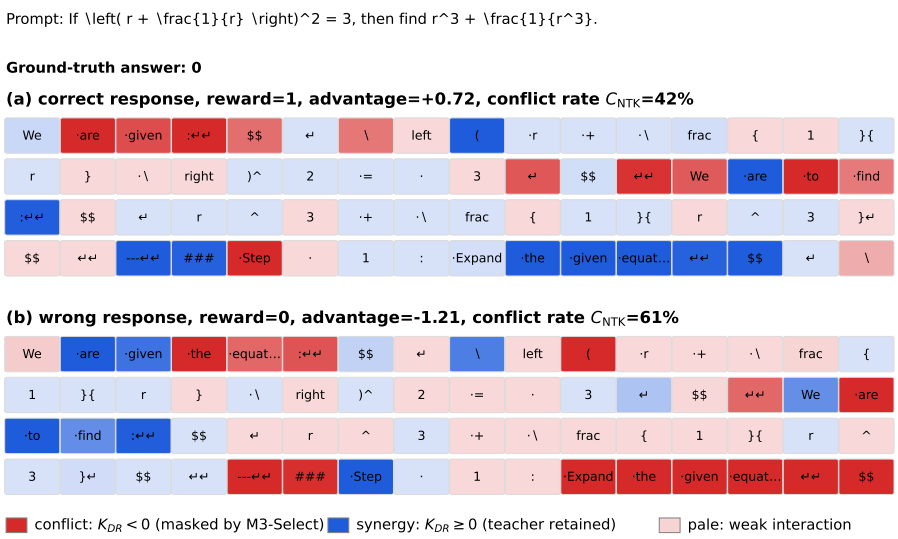


Figure 9: A readable token-level case study for a correct ($A > 0$) and an incorrect ($A < 0$) response to the same prompt. Darker red denotes stronger local conflict ($K_{DR}(n) < 0$), darker blue denotes stronger compatibility, and pale colors denote weak interaction. Both responses interleave the two signal types.

Table 12: LoRA-rank ablation for M^3-Soft on Qwen3-1.7B GSM8K over 100 training steps.

LoRA Rank	$r = 8$	$r = 16$	$r = 32$	$r = 64$	$r = 128$
Mean Reward	0.826	0.824	0.828	**0.844**	0.841

Illustrative Failure Patterns. The following cases are stylized examples distilled from qualitative patterns observed in the GSM8K and SVAMP runs; they are not verbatim training trajectories or additional controlled experiments. Their purpose is to show how magnitude imbalance, group-level cancellation, and token-level sign conflict can appear in concrete reasoning traces. In the examples, $+$ and $-$ denote positive and negative group-relative advantages, respectively.

Bad Case 1: length inflation under drowning ($\kappa \gg \kappa^*$)

On Llama-3.2-1B SVAMP ($\bar{\kappa} \approx 1.3 \times 10^4$), the RL gradient dominates the teacher gradient by roughly four orders of magnitude. Because the verifier rewards only the final numeric answer, the group-normalized advantage rewards *any* continuation that eventually reaches the correct digit and penalizes short-but-wrong ones. When distillation is drowned out, the model no longer receives the teacher's "brevity + structure" prior, response length inflates steadily as training proceeds, and reward collapses to 0.072 (Table 6, row Llama/SVAMP/GRPO). A stylized post-drowning rollout of the kind we observe:

> Let x be the answer. We are told that Melanie has $3 \cdot 4$ apples...so $x = 12$. But wait, let us re-verify: $3 \cdot 4 = 12$, so indeed $x = 12$. Actually, checking once more, $x = 12$. ...Answer: **12**.

The re-verification loops are reward-neutral (verifier only checks "**12**") but teacher-negative. Under M3 the magnitude normalization restores the teacher's unit-scale voice—the brevity+structure prior is no longer drowned by the 10^4 norm ratio—and the boundary admits this anti-repetition pressure exactly where it is reward-compatible ($K_{DR}(n) \geq 0$, e.g. in negative-advantage rollouts whose RL residual also pushes against the loops), so length inflation is suppressed and reward reaches 0.822 (same table row, M3-Soft; the hard-gated M3-Select variant recovers 0.488). *This is the token-level dual of the length-drift phenomenon*

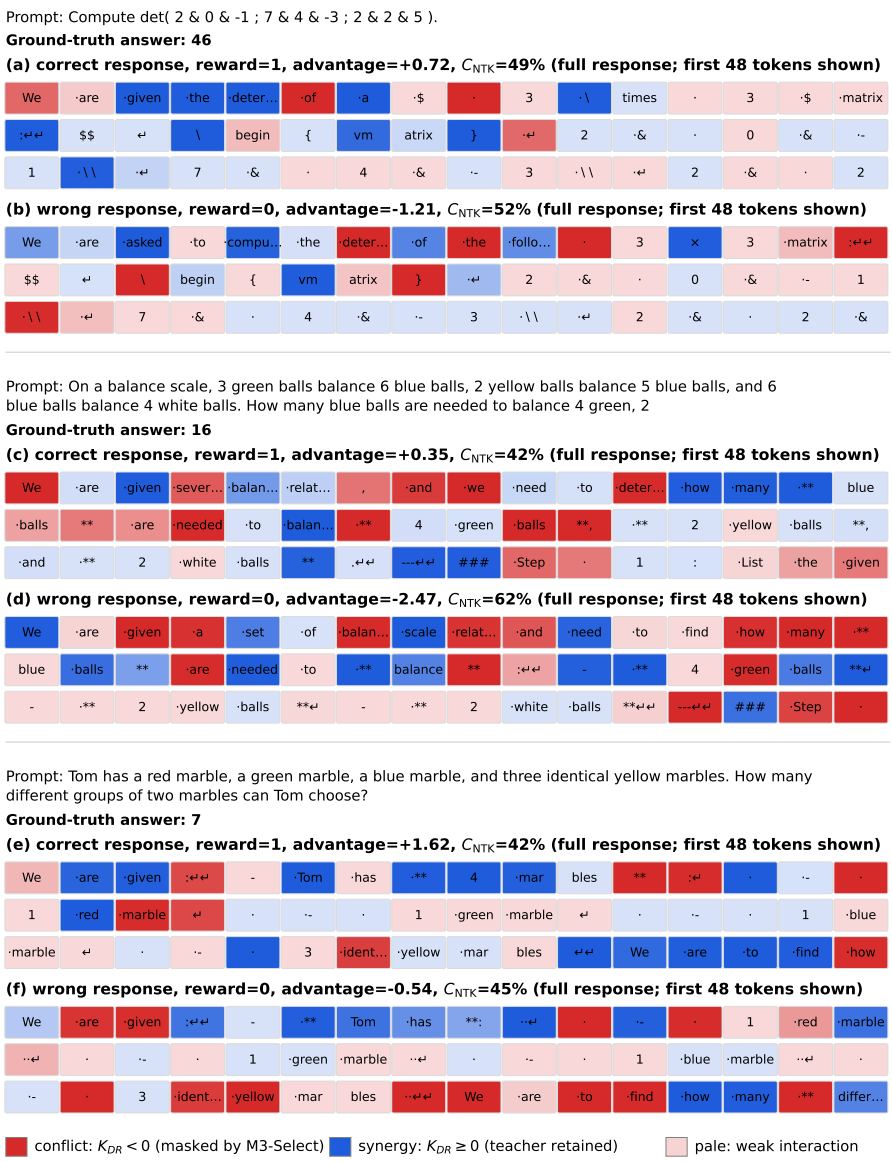


Figure 10: Three additional paired token-level case studies. Headers report C_{NTK} over each full 200-token response; the first 48 tokens are displayed. The incorrect rollout has the higher conflict rate in each pair ($49\% \rightarrow 52\%$, $42\% \rightarrow 62\%$, and $42\% \rightarrow 45\%$), while every rollout contains both compatible and conflicting positions.

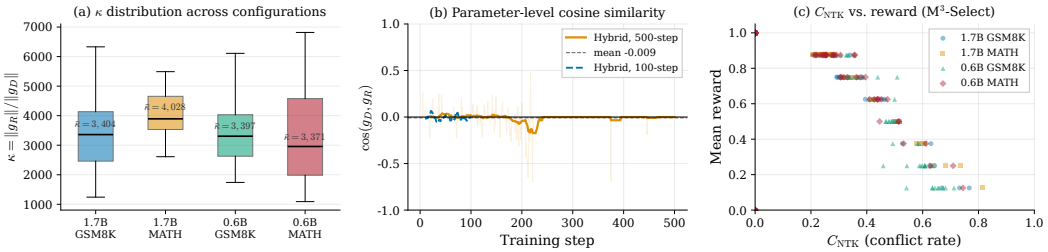


Figure 11: Aggregate conflict diagnostics: *(a)* mean magnitude ratio across settings; *(b)* parameter-level gradient cosine; and *(c)* token-level conflict rate versus reward for M^3-Select.

reported in prior exploration-boundary work: the drowning threshold acts as an implicit "eos-suppression" bias.

Bad Case 2: shared-prefix contamination in the same rollout group

GRPO computes advantages per rollout, so a group of G trajectories that share a long prefix will apply *opposite-sign* updates to that prefix whenever their final answers disagree. Consider a GSM8K group of two rollouts from the same prompt:

`Rollout 1 (+):` $\boxed{\texttt{Let } x \texttt{ be the number of apples.}}$ $x = 5 \cdot 3 = 15$. `Answer:` **15**.
`Rollout 2 (−):` $\boxed{\texttt{Let } x \texttt{ be the number of apples.}}$ $x = 5{+}3 = 8$. `Answer:` **8**.

Denote the shared prefix as s. The naive RL contribution at any s-token y_n is

$$\mathcal{G}_R^n\big|_{y_n \in s} \;=\; -\big[(r_1 - b) + (r_2 - b)\big]\big(\boldsymbol{e}_{y_n} - \boldsymbol{p}_S^n\big) \;=\; 0,$$

since the group-mean baseline $b = \frac{1}{2}(r_1 + r_2)$ centers the two advantages: the RL signal on the shared prefix cancels exactly in this two-rollout group—and nearly so for general G, where unbalanced $+/-$ counts and length-normalization weights leave a small residual. But the distillation signal $\mathcal{G}_D^n = \boldsymbol{p}_S^n - \boldsymbol{p}_T^n$ remains fully alive on the same tokens, and the teacher pushes $\boldsymbol{p}_S$ toward its preferred verbalization—while the small residual RL noise flips sign randomly across steps. The result is an unstable prefix whose gradient variance is dominated by cancellation noise (large γ_n, Definition 3 extension in §F). M3-Select computes $\cos\varphi_n$ between $\mathcal{G}_D^n$ and the *per-rollout* $\mathcal{G}_R^n$, detects $\bar{c}_n = +1$ on the shared prefix of the winning rollout (teacher and RL prefer the same continuation there), and admits the teacher at its full budget $\alpha_{\max}$: after per-token normalization the distillation direction is the only *persistent* signal on the prefix, anchoring it against the sign-flipping residual noise. The RL term is never masked—its group-averaged contribution self-cancels as shown above—so reward supervision effectively acts only on the divergence region ("5·3=15" vs "5+3=8"), where the two rollouts' residuals no longer cancel. *This is the on-policy hybrid analogue of the shared-prefix contamination reported by Ren et al. (2025) ("I-ate-lunch, happy/unhappy" example).*

Bad Case 3: correct intermediate step suppressed by wrong final answer

Even when the group contains only one $(-)$ rollout, an internally-correct intermediate step is penalized because GRPO's outcome reward propagates the trajectory-level sign to every token. Consider a two-step arithmetic rollout:

`Rollout A (−):` $5{\cdot}4 = \boxed{20}$, `then` $20{+}3 = \boxed{24}$. `Answer:` **24**.
`Rollout B (+):` $5{\cdot}4 = \boxed{20}$, `then` $20{+}3 = \boxed{23}$. `Answer:` **23**.

The intermediate token "20" is arithmetically correct in *both* rollouts, and the teacher's distributional prior $\boldsymbol{p}_T$ places > 0.9 mass on it. Under naive hybrid,

$$\mathcal{G}_R^n\big|^{\text{Rollout } A}_{y_n = \text{“20”}} \;=\; -(r_A - b)(\boldsymbol{e}_{20} - \boldsymbol{p}_S^n),$$

Table 13: Stylized failure patterns, observable signatures, and corresponding interventions.

Failure pattern	Observable signature	Intervention
Length inflation *Magnitude drowning*	Large κ; repetitive reasoning not penalized by the final-answer verifier	Magnitude normalization
Shared-prefix cancellation *Group-level cancellation*	Weak aggregate reward signal on tokens shared by opposite-advantage rollouts	Per-token compatibility gate
Correct-intermediate suppression *Token-level sign conflict*	$K_{DR}(n) < 0$ on a locally correct token inside a rejected response	M^3-Select or M^3-Soft
Scaffold-token instability *Boundary uncertainty*	$\lvert K_{DR}(n)\rvert \approx 0$ on structural or low-content tokens	M^3-Soft

with $r_A - b < 0$, so the RL update *lowers* the probability of the correct intermediate. The per-position diagnostic (Proposition 11) yields $\langle \mathcal{G}_D^n, \mathcal{G}_R^n \rangle \propto -(r_A - b)\sigma_n < 0$, so $\bar{c}_n = -1$ and M3-Select withholds teacher supervision from Rollout A's update at this token—the gate acts on the teacher, never on RL. The repair instead comes from the group: Rollout B's equal-and-opposite RL contribution cancels Rollout A's in the aggregate, while on Rollout B the diagnostic reads $\bar{c}_n = +1$ and the teacher is admitted, so the surviving net update pushes $p_S(\text{“20”}) \uparrow$. Correct-intermediate tokens under negative advantage arise whenever a group mixes right and wrong final answers built on shared sub-computations; protecting them through this admitted teacher vote is a driver of the 0.647 vs 0.000 reference-rate gap between M3-Soft and GRPO (Table 6). *This is the dense–sparse counterpart of the “1+1=2” intermediate-step cancellation phenomenon documented in prior exploration-boundary analyses.*

Bad Case 4: directional cancellation on high-entropy scaffold tokens

Certain tokens are structurally common to almost every correct *and* wrong solution: the “=” sign, “Answer:”, “\n”, punctuation, and low-content connectives (“so”, “then”). The teacher assigns $\boldsymbol{p}_T > 0.9$ on these, and the RL signal is noisy and near-zero in expectation (they appear roughly equally in $+$ and $-$ rollouts). Yet the per-step RL contribution is non-zero and has large variance, producing a gradient-cancellation rate

$$\gamma_n \;=\; 1 - \frac{\left\| \frac{1}{G} \sum_g \nabla_{\boldsymbol{\theta}} \ell_g(y_n) \right\|^2}{\frac{1}{G} \sum_g \left\| \nabla_{\boldsymbol{\theta}} \ell_g(y_n) \right\|^2}$$

approaching 1 at these positions, in sharp contrast to content tokens, whose group updates are directionally aligned. Here the soft gate is the right tool: M3-Soft assigns these positions a stable intermediate budget $\alpha_n = \alpha_{\max}\sigma(\beta \cos \varphi_n) \approx \alpha_{\max}/2$ that, at small β, is insensitive to the noisy sign of $\cos \varphi_n$, whereas M3-Select's hard indicator flips with that sign and evicts the teacher on roughly half the scaffold positions at random. Since the group-averaged RL residual is near zero-mean here while the normalized teacher direction is persistent, distillation dominates the *expected* update without any signal being masked—which is why M3-Soft with $\beta = 0.1$ recovers scaffold-token fluency on the low-κ Qwen2.5-SVAMP cell (§B.3).

Table 13 summarizes what each example is intended to illustrate. The first case concerns magnitude imbalance, the next two concern sign structure under outcome-level credit assignment, and the fourth concerns uncertainty near the compatibility boundary.

C Exploration Boundary Framework: Full Statements and Boundary Definition

This appendix states the intra-RL quantities underlying the cross-signal analysis in Section 3.4 and defines the reward–teacher compatibility boundary.

C.1 Exploration Boundary and Gradient Folding in RL

Definition 2 (Gradient Diversity and Exploration Boundary). *For G rollouts, let $\boldsymbol{g}^{(i)} = A_i \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(y_i \mid x)$, with $A_i = (r^{(i)} - \bar{r})/\sigma_r$, and write $\bar{\boldsymbol{g}} = G^{-1} \sum_i \boldsymbol{g}^{(i)}$. For $\bar{\boldsymbol{g}} \neq 0$, define*

$$\Delta(\boldsymbol{\theta}) = \frac{G^{-1} \sum_i \|\boldsymbol{g}^{(i)}\|^2}{\|\bar{\boldsymbol{g}}\|^2} \geq 1, \qquad B_S(\boldsymbol{\theta}) = N_{\max} \Delta(\boldsymbol{\theta}). \tag{16}$$

The inequality is Jensen's inequality; equality holds when all gradients coincide, while equal-norm orthogonal gradients give $\Delta = G$. Large Δ measures cancellation relative to the individual gradient energy. Here $N_{\max}$ is the rollout budget per prompt, and B_S is the exploration-boundary index.

Definition 3 (Gradient Folding and Cancellation Rate). *With binary rewards, positive- and negative-advantage trajectories can oppose one another at shared token positions; we call this* gradient folding. *Its cancellation rate and surviving gradient magnitude satisfy*

$$\gamma = 1 - \Delta^{-1} = 1 - \frac{\|\sum_i \boldsymbol{g}^{(i)}\|^2}{G \sum_i \|\boldsymbol{g}^{(i)}\|^2}, \qquad \|\bar{\boldsymbol{g}}\| = \sqrt{1-\gamma} \left(G^{-1} \sum_i \|\boldsymbol{g}^{(i)}\|^2 \right)^{1/2}. \tag{17}$$

Thus $\gamma \in [0, 1)$ measures the fraction of mean squared gradient magnitude canceled. Complete cancellation gives $\gamma = 1$ and $\Delta = +\infty$, provided the individual gradients are not all zero.

Definition 4 (Token-Level NTK). *Let $\mathbf{J}^n = \nabla_{\boldsymbol{\theta}} \boldsymbol{z}^n \in \mathbb{R}^{d_{\text{par}} \times |\mathcal{V}|}$. The vocabulary-space kernel is $\mathbf{K}(s,t) = (\mathbf{J}^s)^\top \mathbf{J}^t$, and its sampled-token contraction is*

$$K_t(\tau, s, t) = \langle \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_s \mid \tau_{<s}),\, \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_t \mid \tau_{<t}) \rangle\,. \tag{18}$$

For an update from position n alone, $\dot{\boldsymbol{z}}^n = -\mathbf{K}(n,n)\boldsymbol{\delta}^n$; a summed update gives $\dot{\boldsymbol{z}}^n = -\sum_m \mathbf{K}(n,m)\boldsymbol{\delta}^m$. These kernels separate local effects from interactions through shared parameters and provide the geometry for position-level masking.

Definition 5 (Reward–Teacher Compatibility Boundary). *For $K_{DR}(n) = \langle \mathbf{J}^n \boldsymbol{\delta}_D^n, \mathbf{J}^n \boldsymbol{\delta}_R^n \rangle$, define*

$$\partial\mathcal{B} = \mathcal{B}^0 = \{n : K_{DR}(n) = 0\}, \qquad \mathcal{B}^{\pm} = \{n : \pm K_{DR}(n) > 0\}. \tag{19}$$

The compatible region $\mathcal{B}^+$ contributes positively to the local reward projection, the orthogonal region $\mathcal{B}^0$ contributes zero, and the conflicting region $\mathcal{B}^-$ contributes negatively. M3-Select admits the teacher on $\mathcal{B}^+ \cup \mathcal{B}^0$ and masks $\mathcal{B}^-$.

D Extended Theoretical Analysis

This appendix contains extended theoretical results referenced in Sections 3.4 and 3.5.

Corollary 1 (One-Step Logits Degradation). *For advantage $A \neq 0$, write $\boldsymbol{g}_R^n = \mathbf{J}^n \boldsymbol{\delta}_R^n$ and $\boldsymbol{g}_D^n = \mathbf{J}^n \boldsymbol{\delta}_D^n$. The first-order contribution of position n's hybrid update to its sampled-token log-probability is*

$$\begin{aligned} \Delta_n \log p_S(\hat{y}_n \mid \hat{y}_{<n}) &\approx \left\langle -\frac{\boldsymbol{g}_R^n}{A}, -\eta[(1-\alpha)\boldsymbol{g}_R^n + \alpha \boldsymbol{g}_D^n] \right\rangle \\ &= \frac{\eta}{A}\left[(1-\alpha)K_{RR}(n) + \alpha K_{DR}(n)\right], \end{aligned} \tag{20}$$

*where $K_{RR}(n) = \|\boldsymbol{g}_R^n\|^2$ and $\nabla_{\boldsymbol{\theta}} \log p_S(\hat{y}_n \mid \hat{y}_{<n}) = -\boldsymbol{g}_R^n / A$. At a conflicting position, distillation lowers the sampled-token log-probability when $A > 0$ and raises it when $A < 0$. For $A > 0$, the total diagonal contribution becomes negative exactly when $\alpha|K_{DR}(n)| > (1-\alpha)K_{RR}(n)$; Section **??** measures this effect.*

Exact κ-growth identity. For nonzero gradients, direct logarithmic differentiation gives

$$\frac{d \log \kappa}{dt} = \frac{d \log \|\boldsymbol{g}_R\|}{dt} - \frac{d \log \|\boldsymbol{g}_D\|}{dt}. \tag{21}$$

Thus κ grows whenever the reward norm decays more slowly than the teacher norm. If the difference of these logarithmic rates is the positive constant λ, then $\kappa(t) = \kappa(0)e^{\lambda t}$.

Corollary 2 (Aggregate Synergy Condition). *Let $\Phi_\tau = \cos(\boldsymbol{g}_D(\tau), \boldsymbol{g}_R(\tau))$ be the trajectory-level alignment, with fixed conditional means $\mu_0 = \mathbb{E}[\Phi_\tau \mid r = 0] > 0$ and $\mu_1 = \mathbb{E}[\Phi_\tau \mid r = 1] < 0$. At accuracy p,*

$$\mathbb{E}[\Phi_\tau] = (1-p)\mu_0 + p\mu_1 > 0 \quad \Longleftrightarrow \quad p < p^* := \frac{\mu_0}{\mu_0 - \mu_1}. \tag{22}$$

Corollary 3 (Diminishing Synergy Under Increasing Accuracy)**.** *Under the fixed-conditional-mean model of Corollary 2, $d\mathbb{E}[\Phi_\tau]/dp = \mu_1 - \mu_0 < 0$, so expected trajectory alignment crosses zero at p^* and approaches μ_1 as $p \to 1$. The threshold is determined by the two conditional alignments.*

Remark 2 (Why NTK, not just cosine?)**.** *The aggregate cosine summarizes the same parameter-space inner product represented by the NTK: $\langle \boldsymbol{g}_D, \boldsymbol{g}_R \rangle = \sum_{n,m} (\boldsymbol{\delta}_D^n)^\top \mathbf{K}(n,m) \boldsymbol{\delta}_R^m$ for summed gradients. The kernel decomposition exposes which positions and cross-position interactions produce that scalar, enabling local gating that an aggregate cosine alone cannot specify.*

The measured threshold separates cells in which hard masking helps from cells in which a softer gate preserves more teacher signal. The local score identifies the teacher's immediate reward projection; gradient magnitude, estimation noise, and cross-position interactions determine how this local decision translates into training progress. The experiments in Sections 5.3 and 5.3 compare these regimes.

D.1 Conflict as Information Destruction

At conflicting positions of a positive-advantage trajectory, the teacher contribution lowers the sampled-token log-probability (Corollary 1). Repeated contributions of this sign can erode rewarded behavior. The 500-step experiments (Section 5.3) show collapse under uniform mixing in high-κ cells, while M3-Select removes these negative local teacher projections before the update. This links the local mechanism to the observed training trajectories.

E Extended Method Theory

This appendix presents the full theoretical analysis of M3 and M3-Select. All formal statements and proof sketches summarized in Section 4 are collected here.

E.1 M3-EG: Fast–Slow Extragradient Update

The extragradient variant of Section 4 keeps the boundary gate α_n^* of Eq. (12) but replaces the synchronous mixture with a two-timescale schedule. From the anchor $\boldsymbol{\theta}$, the inner *fast* step applies only the boundary-gated teacher to reach a look-ahead point

$$\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta} - \eta_{\text{in}} \tfrac{1}{N} \textstyle\sum_{n=1}^{N} \mathbf{J}^n \alpha_n^* \hat{\boldsymbol{\delta}}_D^n. \tag{23}$$

The on-policy reward group is then scored *at* $\tilde{\boldsymbol{\theta}}$, but its gradient is applied as an outer correction anchored at the original $\boldsymbol{\theta}$,

$$\boldsymbol{\theta}^+ = \boldsymbol{\theta} - \eta_{\text{out}} \tfrac{1}{N} \textstyle\sum_{n=1}^{N} \mathbf{J}^n \hat{\boldsymbol{\delta}}_R^n \big|_{\tilde{\boldsymbol{\theta}}}, \tag{24}$$

using a first-order approximation that does not backpropagate through the inner step: in practice one caches the slow-gradient direction $\frac{1}{N}\sum_n \mathbf{J}^n \hat{\boldsymbol{\delta}}_R^n|_{\tilde{\boldsymbol{\theta}}}$, restores $\boldsymbol{\theta}$, and then applies Eq. (24). The step sizes $(\eta_{\text{in}}, \eta_{\text{out}})$ play the roles of the inner (look-ahead) and outer (correction) rates of a standard extragradient scheme. Algorithm 1 states the full procedure.

Algorithm 1 M3-EG: Fast–Slow Extragradient Update

Input: gate ceiling $\alpha_{\max}$, inner (look-ahead) rate η_{in}, outer (correction) rate η_{out}, smoothing ϵ

1: **For** step $t = 1, \ldots, T$ **do**

2: Compute per-position residuals $\boldsymbol{\delta}_D^n, \boldsymbol{\delta}_R^n$, Jacobians $\mathbf{J}^n$, and scores $K_{DR}(n) \leftarrow \langle \mathbf{J}^n \boldsymbol{\delta}_D^n, \mathbf{J}^n \boldsymbol{\delta}_R^n \rangle$.

3: Gate $\alpha_n^* \leftarrow \alpha_{\max} \mathbf{1}\{K_{DR}(n) \geq 0\}$ (hard) or $\alpha_{\max}\sigma(\beta \cos\varphi_n)$ (soft); normalize $\hat{\boldsymbol{\delta}}_D^n, \hat{\boldsymbol{\delta}}_R^n$.

4: **Fast / look-ahead (teacher only):** $\tilde{\boldsymbol{\theta}} \leftarrow \boldsymbol{\theta}_t - \eta_{\text{in}} \frac{1}{N} \sum_n \mathbf{J}^n \alpha_n^* \hat{\boldsymbol{\delta}}_D^n$. ▷ hard gate masks Ω_-

5: **Re-score at $\tilde{\boldsymbol{\theta}}$ (reward only):** $\boldsymbol{h}_R(\tilde{\boldsymbol{\theta}}) \leftarrow \frac{1}{N} \sum_n \mathbf{J}^n \hat{\boldsymbol{\delta}}_R^n|_{\tilde{\boldsymbol{\theta}}}$ (no backprop through $\tilde{\boldsymbol{\theta}}$).

6: **Slow / correction (anchored at $\boldsymbol{\theta}_t$):** $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \eta_{\text{out}} \boldsymbol{h}_R(\tilde{\boldsymbol{\theta}})$.

End for

Conflict isolation. Write $\boldsymbol{h}_R(\boldsymbol{\theta}) = N^{-1}\sum_n \mathbf{J}^n\hat{\boldsymbol{\delta}}_R^n$ for the normalized reward field and $\boldsymbol{m}_D = N^{-1}\sum_n \mathbf{J}^n\alpha_n^*\hat{\boldsymbol{\delta}}_D^n$ for the gated teacher field. The executed move is $-\eta_{\text{out}}\boldsymbol{h}_R(\boldsymbol{\theta} - \eta_{\text{in}}\boldsymbol{m}_D)$, so teacher information enters through the look-ahead location. Positive rescaling of an unsmoothed teacher residual leaves $\boldsymbol{m}_D$ unchanged.

Proposition 1 (Local Reward Descent of the Extragradient Field)**.** *Let $\boldsymbol{g}_R = \nabla\mathcal{L}_R(\boldsymbol{\theta})$, assume $a_R = \langle \boldsymbol{h}_R(\boldsymbol{\theta}), \boldsymbol{g}_R\rangle > 0$, and let $\boldsymbol{h}_R$ be L_h-Lipschitz along the inner step. Then*

$$\begin{aligned}\langle \boldsymbol{\theta} - \boldsymbol{\theta}^+, \boldsymbol{g}_R\rangle &= \eta_{\text{out}}\langle \boldsymbol{h}_R(\boldsymbol{\theta} - \eta_{\text{in}}\boldsymbol{m}_D), \boldsymbol{g}_R\rangle \\ &\geq \eta_{\text{out}}\left(a_R - \eta_{\text{in}}L_h\|\boldsymbol{m}_D\|\|\boldsymbol{g}_R\|\right) > 0\end{aligned} \tag{25}$$

whenever $\eta_{\text{in}}L_h\|\boldsymbol{m}_D\|\|\boldsymbol{g}_R\| < a_R$. For sufficiently small outer step η_{out}, this is a reward-descent step. When $\boldsymbol{h}_R = \boldsymbol{g}_R$, its first-order expansion is $\eta_{\text{out}}[\|\boldsymbol{g}_R\|^2 - \eta_{\text{in}}\langle \mathbf{H}_R\boldsymbol{m}_D, \boldsymbol{g}_R\rangle] + O(\eta_{\text{out}}\eta_{\text{in}}^2)$, with $\mathbf{H}_R = \nabla^2\mathcal{L}_R(\boldsymbol{\theta})$.

The inner normalization controls the teacher field's scale, while the hard gate removes its negative local reward projections. The outer descent condition above quantifies the additional effect of moving the reward evaluation point (empirical comparison: Table 1).

Algorithm 2 M3-Select: Boundary-Guided Token-Level Mixing

Input: $\alpha_{\max}$, step size η, smoothing ϵ, EMA update scale $\bar{s}$

1: **For** step $t = 1, \ldots, T$ **do**
2: Compute per-position residuals $\boldsymbol{\delta}_D^n, \boldsymbol{\delta}_R^n$ and Jacobians $\mathbf{J}^n$.
3: Compute compatibility scores $K_{DR}(n) \leftarrow \langle \mathbf{J}^n\boldsymbol{\delta}_D^n, \mathbf{J}^n\boldsymbol{\delta}_R^n\rangle$.
4: Set $\alpha_n \leftarrow \alpha_{\max}\mathbf{1}\{K_{DR}(n) \geq 0\}$ and record C_{NTK}.
5: Normalize residuals $\hat{\boldsymbol{\delta}}_D^n \leftarrow \boldsymbol{\delta}_D^n/(\|\boldsymbol{\delta}_D^n\| + \epsilon)$ and $\hat{\boldsymbol{\delta}}_R^n \leftarrow \boldsymbol{\delta}_R^n/(\|\boldsymbol{\delta}_R^n\| + \epsilon)$.
6: Update $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \eta\bar{s}\frac{1}{N}\sum_n \mathbf{J}^n[\alpha_n\hat{\boldsymbol{\delta}}_D^n + (1-\alpha_n)\hat{\boldsymbol{\delta}}_R^n]$.
End for

Quadratic sub-optimality of a fixed mixing coefficient. Let $\boldsymbol{d} = \boldsymbol{g}_D - \boldsymbol{g}_R$ and let $\alpha^* \in [0,1]$ minimize $\|\boldsymbol{g}_H(\alpha)\|^2$. Expanding around the constrained minimizer gives

$$\begin{aligned}\|\boldsymbol{g}_H(\alpha)\|^2 - \|\boldsymbol{g}_H(\alpha^*)\|^2 &= \|\boldsymbol{d}\|^2(\alpha - \alpha^*)^2 + 2(\alpha - \alpha^*)\langle \boldsymbol{g}_H(\alpha^*), \boldsymbol{d}\rangle \\ &\geq \|\boldsymbol{d}\|^2(\alpha - \alpha^*)^2.\end{aligned} \tag{26}$$

The last term on the first line is nonnegative by constrained optimality and vanishes at an interior minimizer. The curvature is $\|\boldsymbol{d}\|^2 = {r_D}^2 + {r_R}^2 - 2\Phi_{DR}r_D r_R$. For a random vector $\boldsymbol{g}$, we use the total variance

$$\text{Var}(\boldsymbol{g}) := \mathbb{E}\|\boldsymbol{g} - \mathbb{E}\boldsymbol{g}\|^2 = \text{tr}\,\text{Cov}(\boldsymbol{g}),$$

with Var_t denoting conditioning on the history available before the current gradient sample.

Theorem 1 (Boundary-Gated Reward Convergence)**.** *Assume $\mathcal{L}_R$ is L-smooth and bounded below. In the conditional model of Appendix L, each position has equal-norm reward and deterministic teacher means $\boldsymbol{u}_n, \boldsymbol{d}_n$, position subspaces are orthogonal, gates are fixed before fresh reward noise is sampled, and $\nabla\mathcal{L}_R = N^{-1}\sum_n \boldsymbol{u}_n$. Define $c_n = \langle \boldsymbol{u}_n, \boldsymbol{d}_n\rangle/\|\boldsymbol{u}_n\|^2$ and $w_n = \|\boldsymbol{u}_n\|^2/\sum_m\|\boldsymbol{u}_m\|^2$, taking $c_n = 0$ at zero-norm positions. The population-sign gate is $\alpha_n = \alpha_{\max}\mathbf{1}\{c_n \geq 0\}$, and $\rho_{R,t} = \sum_n w_n[(1-\alpha_n) + \alpha_n c_n]$. For $\alpha_{\max} < 1$, choose deterministic bounds $0 < \underline{\rho}_R \leq \rho_{R,t}$ and $\text{Var}_t(\hat{\boldsymbol{g}}_H^{\text{sel}}) \leq v_\alpha\sigma_R^2$ valid at every iteration and history; $\underline{\rho}_R = 1 - \alpha_{\max}$ is admissible. If $\eta \leq \underline{\rho}_R/L$, then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla\mathcal{L}_R(\boldsymbol{\theta}_t)\|^2 \leq \frac{1}{\underline{\rho}_R}\left[\frac{2(\mathcal{L}_R(\boldsymbol{\theta}_0) - \mathcal{L}_R^*)}{\eta T} + L\eta v_\alpha\sigma_R^2\right]. \tag{27}$$

Proposition 2 (Conditional Variance of Boundary-Gated Mixing)**.** *In the preceding conditional model, let $\nu_n = \mathbb{E}_t\|\boldsymbol{\xi}_n\|^2$ be the fresh reward-noise variance at position n. Then*

$$\text{Var}_t(\boldsymbol{g}_H^{\text{sel}}) = \frac{1}{N^2}\sum_n(1-\alpha_n)^2\nu_n = v_t\sigma_{R,t}^2, \tag{28}$$

where $\sigma_{R,t}^2 = N^{-2}\sum_n \nu_n$ *and* $v_t = \sum_n (1-\alpha_n)^2\nu_n / \sum_n \nu_n \le 1$ *when noise is nonzero; take* $v_t = N^{-1}\sum_n (1-\alpha_n)^2$ *when all* $\nu_n = 0$. *Choose a deterministic* $v_\alpha \ge v_t$ *valid uniformly over iterations and histories;* $v_\alpha = 1$ *is always admissible when* $\sigma_{R,t}^2 \le \sigma_R^2$. *For equal position variances and admitted fraction* $\rho_{\text{adm}} = |\Omega_+ \cup \Omega_0|/N$, $v_t = 1 - (2\alpha_{\max} - \alpha_{\max}^2)\rho_{\text{adm}} = (1-\bar\alpha)^2 + \alpha_{\max}^2\rho_{\text{adm}}(1-\rho_{\text{adm}})$, *where* $\bar\alpha = \alpha_{\max}\rho_{\text{adm}}$.

The local reward-projection guarantee underlying both statements is

$$\Pi_R(n) \ge (1-\alpha_{\max})\|\mathbf{J}^n\hat{\boldsymbol{\delta}}_R^n\|^2 \ \text{ for } n \in \Omega_+ \cup \Omega_0, \qquad \Pi_R(n) = \|\mathbf{J}^n\hat{\boldsymbol{\delta}}_R^n\|^2 \ \text{ for } n \in \Omega_-. \tag{29}$$

Theorem 2 (M3-Norm Convergence Guarantee). *Assume* $\mathcal{L}_R$ *is* L*-smooth and bounded below, with an unbiased reward estimator of conditional variance at most* σ_R^2. *Conditional on each iterate, let the teacher direction be deterministic, matched to the population reward-gradient norm, and have alignment at least* ϕ_* *(Appendix J). For fixed* $\alpha \in (0,1)$, *put* $c = 1-\alpha+\alpha\phi_* > 0$. *If* $\eta \le c/L$, *then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla\mathcal{L}_R(\boldsymbol{\theta}_t)\|^2 \le \frac{1}{c}\left[\frac{2(\mathcal{L}_R^0 - \mathcal{L}_R^*)}{\eta T} + L\eta(1-\alpha)^2\sigma_R^2\right]. \tag{30}$$

Proposition 3 (Variance Reduction via Distillation). *Under the conditional deterministic-teacher model of Theorem 2,* $\mathrm{Var}_t(\hat{\boldsymbol{g}}_H) = (1-\alpha)^2\,\mathrm{Var}_t(\hat{\boldsymbol{g}}_R) \le (1-\alpha)^2\sigma_R^2$.

Proposition 4 (Optimal Per-Position Mixing). *For unit parameter-space directions with cosine* c_n, *let* $\Pi_R(n;\alpha) = 1-\alpha+\alpha c_n$. *Maximizing the calibrated objective* $\Pi_R(n;\alpha)+\alpha = 1+\alpha c_n$ *over* $[0,\alpha_{\max}]$ *gives the hard gate* $\alpha_n^* = \alpha_{\max}\mathbf{1}\{c_n \ge 0\}$, *with ties assigned to the teacher. Under additive logistic score noise of scale* $1/\beta$, *its expected allocation is* $\alpha_{\max}\sigma(\beta c_n)$, *the M3-Soft gate.*

Proposition 5 (GradNorm Degeneracy under $\kappa \gg 1$). *At equal target training rates, impose* $w_D + w_R = 1$ *and norm balance* $w_D\|\boldsymbol{g}_D\| = w_R\|\boldsymbol{g}_R\|$. *Then* $w_D^* = \kappa/(1+\kappa)$ *and* $w_R^* = 1/(1+\kappa)$, *giving* $w_R^* \approx 0.03\%$ *at* $\kappa = 3{,}400$. *While* $w_D/w_R = \Theta(1)$, *the teacher's share of weighted gradient magnitude is* $w_D/(w_D + w_R\kappa) = \Theta(\kappa^{-1})$.

A natural loss-level alternative to fixed mixing is to gate α on the positive part of the alignment functional, ${\Phi_{DR}}^+ = \max(\Phi_{DR}, 0)$, yielding the adaptive schedule

$$\alpha_{\text{adaptive}} = \alpha_{\max}\cdot {\Phi_{DR}}^+ / \max({\Phi_{DR}}^+, \epsilon). \tag{31}$$

When the gate uses an exponentially smoothed alignment estimate, its response to a sign change has the following delay.

Proposition 6 (Phase Delay in Adaptive Mixing). *Let* $\widehat{\Phi}_{t+1} = \rho\widehat{\Phi}_t + (1-\rho)\Phi_t$, *with* $0 < \rho < 1$. *If alignment changes from* $\Phi_+ > 0$ *to* $\Phi_- < 0$ *at* t_1 *and* $\widehat{\Phi}_{t_1} = \Phi_+$, *then*

$$\widehat{\Phi}_{t_1+k} = \Phi_- + (\Phi_+ - \Phi_-)\rho^k, \qquad k_* = \left\lceil \frac{\log[(\Phi_+ - \Phi_-)/|\Phi_-|]}{\log(1/\rho)} \right\rceil. \tag{32}$$

The estimate remains positive for $k < k_*$, *so the teacher gate stays active during that interval.*

Proposition 7 (Empirical Invariance of κ under LoRA Rank). *Within the measured LoRA-rank range* $r \in \{8, 16, 32, 64\}$, $\bar\kappa = 3{,}476 \pm 301$ *(CV* $= 8.7\%$*). On Qwen3-GSM8K, the corresponding scale comparison is* $\kappa \approx 3{,}397$ *at* 0.6B *and* $\approx 3{,}400$ *at* 1.7B. *These are within-task observations; at rank* 128 *the measured ratio rises to* 6,296, *and ratios vary substantially across tasks (Section B.1).*

Remark 3 (κ Invariance Across LoRA Ranks). *A shared rank factor cancels from* κ *when the reward and teacher gradients have the same rank-dependent norm scaling. Together with weak token correlations, this supplies the approximation developed in Appendix O. A common LoRA subspace alone does not enforce equal scaling for two different directions.*

Proposition 8 (Distillation Mode Determines κ). *Holding* $\boldsymbol{g}_R$ *fixed gives* $\kappa_{\text{on}}/\kappa_{\text{off}} = \|\boldsymbol{g}_D^{\text{off}}\|/\|\boldsymbol{g}_D^{\text{on}}\|$. *For mode* $s \in \{\text{on}, \text{off}\}$, *let* $E_s = \mathcal{L}_D^s - \min\mathcal{L}_D^s$ *and assume positive local curvature bounds* $\lambda_s^-\mathbf{I} \preceq \nabla^2\mathcal{L}_D^s \preceq \lambda_s^+\mathbf{I}$. *Then*

$$\frac{\kappa_{\text{on}}}{\kappa_{\text{off}}} \ge \sqrt{\frac{\lambda_{\text{off}}^- E_{\text{off}}}{\lambda_{\text{on}}^+ E_{\text{on}}}}. \tag{33}$$

A large off-policy excess KL relative to the on-policy excess therefore raises this lower bound, with the curvature ratio accounting for the different prefix distributions.

Proposition 9 (Conflict-Free Guarantee for M3-Select). *With $\widetilde{\boldsymbol{g}}_D^n = \mathbf{1}\{K_{DR}(n) \geq 0\}\boldsymbol{g}_D^n$, masking gives $\langle \widetilde{\boldsymbol{g}}_D^n, \boldsymbol{g}_R^n \rangle = \max\{K_{DR}(n), 0\} \geq 0$ at every position.*

Proposition 10 (Distillation as Implicit Regularization). *In the matched-context isotropic quadratic model $\mathcal{L}_D(\boldsymbol{\theta}_T + \Delta\boldsymbol{\theta}) = \frac{1}{2} f_D \|\Delta\boldsymbol{\theta}\|^2$, with $f_D > 0$, constant $\boldsymbol{g}_R$, and $\Delta\boldsymbol{\theta}(0) = 0$, hybrid gradient flow satisfies*

$$\Delta\boldsymbol{\theta}_H(T) = -\frac{(1-\alpha)\boldsymbol{g}_R}{\alpha f_D}(1 - e^{-\eta\alpha f_D T}), \qquad \|\Delta\boldsymbol{\theta}_H(T)\| \leq \frac{(1-\alpha)\|\boldsymbol{g}_R\|}{\alpha f_D}. \tag{34}$$

If degradation is proportional to parameter deviation with a common coefficient, then $\delta_H \leq \delta_R/(1+\alpha\rho_{\text{reg}})$, where $\rho_{\text{reg}} = \eta f_D T/2$. A finite rank-dependent reversal follows in the model when the pure-RL degradation grows continuously without bound, the drowning penalty is bounded, ρ_{reg} is bounded away from zero, and the initial reward gap is positive; Appendix AD gives this conditional argument and the observed reversal at rank 128.

F Gradient Folding, Five Bridges, and Unified Framework: Full Statements

This section collects the formal statements of the propositions, theorems, and corollaries whose proofs appear in subsequent appendix sections and whose summaries appear in Section 3.5 and the method discussion.

Proposition 11 (Per-Position Conflict Decomposition). *The per-position residual inner product has the exact decomposition*

$$(\boldsymbol{\delta}_D^n)^\top \boldsymbol{\delta}_R^n = -A(\sigma_n + \eta_n), \tag{35}$$

where $\sigma_n = p_S^n(\hat{y}_n) - p_T^n(\hat{y}_n)$ measures student-teacher disagreement on the sampled token, and $\eta_n = (\boldsymbol{p}_T^n)^\top \boldsymbol{p}_S^n - \|\boldsymbol{p}_S^n\|^2$ captures cross-token probability redistribution. Under $\mathbf{K}(n,n) = \lambda_n \mathbf{I}$ with $\lambda_n > 0$, conflict ($K_{DR}(n) < 0$) on positive-advantage trajectories ($A > 0$) occurs exactly when $\sigma_n + \eta_n > 0$.

Proposition 12 (Asymmetric Harm from Magnitude Drowning). *For $\boldsymbol{g}_H = (1-\alpha)\boldsymbol{g}_R + \alpha\boldsymbol{g}_D$ with $0 < \alpha < 1$ and $\Phi_{DR} < 0$, dividing the harmful cross-terms by their respective self-progress terms gives*

$$H_{R\to D} = \frac{(1-\alpha)|\langle \boldsymbol{g}_D, \boldsymbol{g}_R \rangle|}{\alpha\|\boldsymbol{g}_D\|^2} = \frac{(1-\alpha)\kappa|\Phi_{DR}|}{\alpha}, \quad H_{D\to R} = \frac{\alpha|\Phi_{DR}|}{(1-\alpha)\kappa}. \tag{36}$$

Hence $H_{R\to D}/H_{D\to R} = [(1-\alpha)/\alpha]^2\kappa^2$. At $\alpha = 1/2$ and $\kappa = 3{,}400$, the ratio is about 1.16×10^7 (Eq. 6).

Proposition 13 (Cosine Inversion in On-Policy OPSD). *For binary rewards with baseline $p \in (0,1)$, write $\boldsymbol{g}_R = -(r-p)\boldsymbol{s}$, where $\boldsymbol{s} = \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\hat{y})$, and let $q = \langle \boldsymbol{s}, \boldsymbol{g}_D \rangle / (\|\boldsymbol{s}\|\|\boldsymbol{g}_D\|)$. If $\mathbb{E}[q \mid r = 0] + \mathbb{E}[q \mid r = 1] > 0$, then*

$$\mathbb{E}[\cos(\boldsymbol{g}_R, \boldsymbol{g}_D) \mid r = 0] - \mathbb{E}[\cos(\boldsymbol{g}_R, \boldsymbol{g}_D) \mid r = 1] = \mathbb{E}[q \mid r = 0] + \mathbb{E}[q \mid r = 1] > 0. \tag{37}$$

Thus the conditional score–teacher alignment determines whether the teacher aligns more strongly with reward updates on incorrect trajectories.

Proposition 14 (Folding–Drowning Coupling). *Let γ be the cancellation rate and κ_{raw} the ratio formed from the root-mean-square trajectory gradient. Then $\kappa_{\text{eff}} = \sqrt{1-\gamma}\,\kappa_{\text{raw}}$. Under the concentrated two-class model with equal class-mean score norms and overlap cosine c_{overlap},*

$$\gamma = 1 - 2p(1-p)(1 - c_{\text{overlap}}). \tag{38}$$

The overlap contribution $2p(1-p)c_{\text{overlap}}$ peaks at balanced accuracy when $c_{\text{overlap}} > 0$. Reduced effective magnitude and positive cross-signal alignment jointly favor hybridization whenever both conditions hold.

Proposition 15 (Distillation Boundary Bound). *Let $\Delta_{DR} = (\|\boldsymbol{g}_D\|^2 + \|\boldsymbol{g}_R\|^2)/\|\boldsymbol{g}_D + \boldsymbol{g}_R\|^2$ and $\Delta_\tau = (1-\gamma)^{-1}$. If $\kappa_{\text{raw}} \gg \sqrt{\Delta_\tau}$, then*

$$\Delta_{DR} \leq 1 + \frac{2|\Phi_{DR}|\sqrt{\Delta_\tau}}{\kappa_{\text{raw}}} + O\left(\frac{\Delta_\tau}{\kappa_{\text{raw}}^2}\right). \tag{39}$$

The index $B_D = \alpha_{\max}\Delta_{DR}$ is the cross-signal analogue of B_S. Its reference value $\Delta_{DR} = 1$ separates negative from positive gradient cross-terms.

Proposition 16 (NTK-Guided Token Masking (Bridge 4)). *For $\boldsymbol{u}_n = \mathbf{J}^n \hat{\boldsymbol{\delta}}_D^n$, $\boldsymbol{v}_n = \mathbf{J}^n \hat{\boldsymbol{\delta}}_R^n$, and $S^- = \{n : \langle \boldsymbol{u}_n, \boldsymbol{v}_n \rangle < 0\}$, the average local reward-projection gain over uniform M3-Norm is*

$$\Delta\Pi_R = \frac{\alpha_{\max}}{N} \sum_{n \in S^-} \left(\|\boldsymbol{v}_n\|^2 - \langle \boldsymbol{u}_n, \boldsymbol{v}_n \rangle \right). \tag{40}$$

It is positive when $S^- \neq \varnothing$ and $\alpha_{\max} > 0$. For unit parameter-space directions this reduces to $\Delta\Pi_R = \alpha_{\max} C_{\mathrm{NTK}}(1 + |\bar{c}^-|)$, where $|\bar{c}^-|$ is the mean absolute cosine on S^-.

Proposition 17 (RL Projection Under M3-Norm). *For unit directions, bilinearity gives $\Pi_R(\alpha) = \langle \alpha \hat{\boldsymbol{g}}_D + (1-\alpha)\hat{\boldsymbol{g}}_R, \hat{\boldsymbol{g}}_R \rangle = 1 - \alpha + \alpha \Phi_{DR}$. With the same reward-norm reference scale, norm-balanced GradNorm has projection $(1 + \Phi_{DR})/(1 + \kappa)$. Hence their projection ratio is $\Theta(\kappa)$ when both cosine factors remain positive and bounded away from zero.*

Theorem 3 (Unified NTK Learning Efficiency). *Define the composite efficiency index by*

$$\eta^{\mathrm{eff}} = \underbrace{(1-\gamma)}_{\eta_{\mathrm{explore}} = 1/\Delta_\tau} \underbrace{[1 - \alpha + \alpha \Phi_{DR}]}_{\eta_{\mathrm{hybrid}}}. \tag{41}$$

Both factors have NTK decompositions: cross-trajectory interactions determine γ, and cross-signal interactions determine Φ_{DR}. Under a normalized hybrid step, first-order reward progress relative to the root-mean-square reward-gradient scale is instead proportional to $\sqrt{1-\gamma}\,\eta_{\mathrm{hybrid}}$. For fixed α, the composite index decreases with accuracy only when $-\gamma'(p)\eta_{\mathrm{hybrid}} + (1-\gamma)\alpha\Phi'_{DR}(p) \leq 0$.

Corollary 4 (C_{NTK}-Based α Scheduling). *The hard gate allocates the mean coefficient $\alpha_{\mathrm{eff}} = \alpha_{\max}(1 - C_{\mathrm{NTK}})$ and admits the fraction $1 - \widetilde{C}_{\mathrm{NTK}}$ of absolute cross-signal NTK mass. Under the fixed-cohort population model of Proposition 21, $\widetilde{C}_{\mathrm{NTK}}$ increases with accuracy, so this admitted mass fraction decreases automatically. The count-based coefficient budget follows the unweighted conflict rate.*

Full proofs of the above statements appear in the dedicated appendix sections below, together with the PCGrad degeneracy analysis in Appendix M. The five bridges connecting exploration boundary theory to hybrid dynamics are: (1) shared NTK geometry (Theorem 3); (2) cancellation-modulated drowning (Proposition 14); (3) the distillation boundary (Proposition 15); (4) token masking (Proposition 16); and (5) conflict-rate–driven α scheduling (Corollary 4).

G Extended Discussion

This appendix collects the optimality and benefit-condition results referenced in the Conclusion, together with the practitioner's decision tree that operationalises them.

Corollary 5 (Local Projection and Teacher Allocation under Low Accuracy). *Assume the unit-direction setting of Proposition 16 and the fixed conditional alignments of Corollary 2. For $p < p^*$, trajectory alignment is positive in expectation, and M3-Select weakly improves the average local reward projection over M3-Norm. At the same ceiling $\alpha_{\max} \in (0, 1)$, its mean teacher allocation relative to naive mixing's norm-based teacher share is*

$$\frac{\alpha_{\max}(1 - C_{\mathrm{NTK}})}{\alpha_{\max}/[\alpha_{\max} + (1-\alpha_{\max})\kappa]} = (1 - C_{\mathrm{NTK}})[\alpha_{\max} + (1-\alpha_{\max})\kappa]. \tag{42}$$

Thus normalization restores teacher allocation at large κ, and the gate retains only positions with nonnegative local compatibility.

Proof. The two projection claims follow from $\mathbb{E}[\Phi_\tau] = (1-p)\mu_0 + p\mu_1 > 0$ and $\Pi_R^{\mathrm{Select}} - \Pi_R^{\mathrm{Norm}} = \alpha_{\max} C_{\mathrm{NTK}}(1 + |\bar{c}^-|) \geq 0$. The allocation identity follows by dividing the admitted mean coefficient by naive mixing's norm-based teacher share. □

Corollary 6 (Hybrid Benefit Condition). *At the same step size $\eta \leq c/L$, the M3-Norm bound of Theorem 2, with $c = 1 - \alpha + \alpha\phi_* > 0$, is strictly smaller than its pure-GRPO instance exactly when*

$$L\eta\sigma_R^2(1 - \alpha + \phi_*) > \frac{2(1-\phi_*)(\mathcal{L}_R^0 - \mathcal{L}_R^*)}{\eta T}. \tag{43}$$

For fixed η, positive variance, and $\phi_* > -(1-\alpha)$, this condition holds for sufficiently large T.

Proof. Let $D = 2(\mathcal{L}_R^0 - \mathcal{L}_R^*)/(\eta T)$ and $V = L\eta\sigma_R^2$. Comparing the two bounds and using $\alpha > 0$ gives the single equivalence chain

$$\begin{aligned} \frac{D + (1-\alpha)^2 V}{c} < D + V &\iff V[c - (1-\alpha)^2] > D(1-c) \\ &\iff V(1 - \alpha + \phi_*) > D(1 - \phi_*), \end{aligned} \tag{44}$$

which is the stated condition. □

Remark 4 (Practitioner's Decision Tree). *Estimate $\bar{\kappa}$ from a 10-step probe run on the target cell, then apply the following gate-selection rule (thresholds are the ones used throughout this work; $\kappa^* = 5{,}000$ is the catastrophic threshold of Remark 1).*

(1) ***Catastrophic regime,*** *$\bar{\kappa} > \kappa^* = 5{,}000$: use **M3-Select** (hard gate). Empirically verified on Llama-GSM8K ($\bar{\kappa} = 7{,}690$) and Llama-SVAMP ($\bar{\kappa} \approx 1.3 \times 10^4$), where every uniform-mixing baseline collapses to reward < 0.05 while M3-Select survives (§5.3, §5.3); at these κ the soft-gate rescue is seed-fragile (two of three Llama-GSM8K replicates collapse during training, §B.1), making the hard gate the robust long-horizon choice.*

(2) ***Stable regime,*** *$1{,}000 < \bar{\kappa} \leq \kappa^*$: use **M3-Soft** with the empirically effective sharpness $\beta \approx 1$; Proposition **??** describes its local variance sensitivity. Empirically verified on Qwen3-GSM8K ($\bar{\kappa} = 3{,}404$), Qwen3-SVAMP ($\bar{\kappa} \approx 5{,}080$, sitting essentially at the threshold), and every InternLM/Qwen2.5 cell in this range. Hard masking over-prunes here (Remark 1); Soft-gate preserves the beneficial synergy tokens the estimator misclassifies.*

(3) ***Low-κ regime,*** *$\bar{\kappa} \leq 1{,}000$: naive uniform mixing (Hybrid, $\alpha \approx 0.5$) is already sufficient; specialized gating offers diminishing marginal returns. This regime is not instantiated in our sweep—the nearest cells are the ARC cells across architectures ($\bar{\kappa} \in [1.4,\ 2.5] \times 10^3$), where gentle M3-Soft ($\beta \in \{0.5, 1\}$, small $\alpha_{\max}$) still helps but the gap over Hybrid is within the seed-level standard deviation (§B.1).*

(4) ***On-policy vs. off-policy OPSD choice.*** *Use the measured conditional alignments to locate the synergy threshold (Corollary 2); the observed on-policy crossover is near $p^* \approx 0.56$. Compare on- and off-policy validation curves when choosing the sampling mode.*

(5) ***Conflict-rate override.*** *If the observed conflict rate $C_{\mathrm{NTK}} > 50\%$ persists into training, switch from M3-Soft to M3-Select regardless of $\bar{\kappa}$: the mass of Ω_- tokens is large enough that the Soft gate's residual bias on Ω_- dominates its variance reduction on Ω_+.*

*Two calibration remarks. (i) The Hybrid→Soft boundary at $\bar{\kappa} \approx 1{,}000$ is set by an empirical gate-calibration heuristic, and is more forgiving than the Soft→Select boundary at κ^*: below it the Soft gate is still safe, only unnecessary. (ii) In cells where dynamic-gating M3-Soft is at parity with static-mixing Hybrid at the peak-training regime (InternLM-GSM8K is the canonical example), apply the SWA post-step to reduce late-stage checkpoint fluctuations (Proposition **??**, §B.3) before falling back to Hybrid.*

H Proof of Proposition 18: Spectral Conflict Bound

Proposition 18 (Spectral Conflict Bound). *Let $\mathbf{K} = [\mathbf{K}(n,m)]_{n,m=1}^N \in \mathbb{R}^{N|\mathcal{V}| \times N|\mathcal{V}|}$ be the global token-level NTK, and let $\boldsymbol{u}_D$ and $\boldsymbol{u}_R$ stack the residuals $\boldsymbol{\delta}_D^n$ and $\boldsymbol{\delta}_R^n$, respectively. Then*

$$\begin{aligned} \langle \boldsymbol{g}_D, \boldsymbol{g}_R \rangle &\geq \frac{\lambda_{\min}(\mathbf{K})\|\boldsymbol{u}_D + \boldsymbol{u}_R\|^2 - \lambda_{\max}(\mathbf{K})\|\boldsymbol{u}_D - \boldsymbol{u}_R\|^2}{4N^2}, \\ \langle \boldsymbol{g}_D, \boldsymbol{g}_R \rangle &\leq \frac{\lambda_{\max}(\mathbf{K})\|\boldsymbol{u}_D + \boldsymbol{u}_R\|^2 - \lambda_{\min}(\mathbf{K})\|\boldsymbol{u}_D - \boldsymbol{u}_R\|^2}{4N^2}. \end{aligned} \tag{45}$$

Thus local residual alignment alone does not determine the aggregate interaction.

Proof. Stacking the Jacobians gives $\mathbf{J} = [\mathbf{J}^1, \dots, \mathbf{J}^N]$ and $\mathbf{K} = \mathbf{J}^\top \mathbf{J} \succeq 0$. Polarization and the Rayleigh bounds yield the single chain

$$\begin{aligned}\langle \boldsymbol{g}_D, \boldsymbol{g}_R \rangle &= N^{-2} \boldsymbol{u}_D^\top \mathbf{K} \boldsymbol{u}_R \\ &= \frac{(\boldsymbol{u}_D + \boldsymbol{u}_R)^\top \mathbf{K} (\boldsymbol{u}_D + \boldsymbol{u}_R) - (\boldsymbol{u}_D - \boldsymbol{u}_R)^\top \mathbf{K} (\boldsymbol{u}_D - \boldsymbol{u}_R)}{4N^2} \\ &\geq \frac{\lambda_{\min}(\mathbf{K}) \|\boldsymbol{u}_D + \boldsymbol{u}_R\|^2 - \lambda_{\max}(\mathbf{K}) \|\boldsymbol{u}_D - \boldsymbol{u}_R\|^2}{4N^2}.\end{aligned}$$

Interchanging the upper and lower Rayleigh bounds proves the second inequality. □

Geometric reading. Conflict occurs exactly when $(\boldsymbol{u}_D - \boldsymbol{u}_R)^\top \mathbf{K} (\boldsymbol{u}_D - \boldsymbol{u}_R) > (\boldsymbol{u}_D + \boldsymbol{u}_R)^\top \mathbf{K} (\boldsymbol{u}_D + \boldsymbol{u}_R)$: the difference field has more kernel-weighted energy than the sum field. This depends on the residuals' projections onto the kernel eigenspaces. If the trainable parameter dimension is below $N|\mathcal{V}|$, then $\operatorname{rank}(\mathbf{K}) \leq \dim(\boldsymbol{\theta})$ implies $\lambda_{\min}(\mathbf{K}) = 0$, reducing the lower bound to $-\lambda_{\max}(\mathbf{K}) \|\boldsymbol{u}_D - \boldsymbol{u}_R\|^2 / (4N^2)$.

I Proof of Proposition 11: Per-Position Conflict Decomposition

Proof. Substitution of the two residuals directly gives

$$\begin{aligned}(\boldsymbol{\delta}_D^n)^\top \boldsymbol{\delta}_R^n &= -A (\boldsymbol{p}_S^n - \boldsymbol{p}_T^n)^\top (\boldsymbol{e}_{\hat{y}_n} - \boldsymbol{p}_S^n) \\ &= -A \left[p_S^n(\hat{y}_n) - p_T^n(\hat{y}_n) + (\boldsymbol{p}_T^n)^\top \boldsymbol{p}_S^n - \|\boldsymbol{p}_S^n\|^2 \right] \\ &= -A (\sigma_n + \eta_n).\end{aligned}$$

The residual identity is exact. Under $\mathbf{K}(n,n) = \lambda_n \mathbf{I}$ with $\lambda_n > 0$, $K_{DR}(n)$ has the same sign, so a positive-advantage trajectory conflicts precisely when $\sigma_n + \eta_n > 0$. □

Interpretation. σ_n isolates disagreement on the sampled token, whereas $\eta_n = \sum_v [p_T^n(v) - p_S^n(v)] p_S^n(v)$ aggregates vocabulary-wide redistribution weighted by student confidence. The measured ratio $|\eta_n|/|\sigma_n| \approx 8.6$ (Section 5) identifies redistribution as the larger contribution in our pilot experiments.

J Proof of Theorem 2: M3-Norm Convergence Guarantee

Proof. Let $\mathbb{E}_t$ condition on the history before the fresh reward-gradient sample, and write $\boldsymbol{g}_R^t = \nabla \mathcal{L}_R(\boldsymbol{\theta}^t)$. The population-norm-matched update is $\hat{\boldsymbol{g}}_H^t = (1-\alpha)(\boldsymbol{g}_R^t + \boldsymbol{\xi}^t) + \alpha \tilde{\boldsymbol{g}}_D^t$, where $\tilde{\boldsymbol{g}}_D^t$ is conditionally deterministic, $\|\tilde{\boldsymbol{g}}_D^t\| = \|\boldsymbol{g}_R^t\|$, $\mathbb{E}_t \boldsymbol{\xi}^t = 0$, and $\mathbb{E}_t \|\boldsymbol{\xi}^t\|^2 \leq \sigma_R^2$. At a stationary point set $\tilde{\boldsymbol{g}}_D^t = 0$. For the uniform alignment lower bound ϕ_*, put $c = 1 - \alpha + \alpha\phi_* > 0$ and $\bar{\boldsymbol{g}}_H^t = \mathbb{E}_t \hat{\boldsymbol{g}}_H^t$. Then

$$\langle \boldsymbol{g}_R^t, \bar{\boldsymbol{g}}_H^t \rangle = (1-\alpha) \|\boldsymbol{g}_R^t\|^2 + \alpha \langle \boldsymbol{g}_R^t, \tilde{\boldsymbol{g}}_D^t \rangle \geq c \|\boldsymbol{g}_R^t\|^2, \tag{46}$$

$$\mathbb{E}_t \|\hat{\boldsymbol{g}}_H^t\|^2 = \|\bar{\boldsymbol{g}}_H^t\|^2 + (1-\alpha)^2 \mathbb{E}_t \|\boldsymbol{\xi}^t\|^2 \leq \|\boldsymbol{g}_R^t\|^2 + (1-\alpha)^2 \sigma_R^2. \tag{47}$$

The last inequality uses $\|\bar{\boldsymbol{g}}_H^t\| \leq (1-\alpha)\|\boldsymbol{g}_R^t\| + \alpha \|\tilde{\boldsymbol{g}}_D^t\| = \|\boldsymbol{g}_R^t\|$. Smoothness, $\eta \leq c/L$, and telescoping now give

$$\begin{aligned}\mathbb{E}_t \mathcal{L}_R(\boldsymbol{\theta}^{t+1}) &\leq \mathcal{L}_R(\boldsymbol{\theta}^t) - \eta \langle \boldsymbol{g}_R^t, \bar{\boldsymbol{g}}_H^t \rangle + \tfrac{L\eta^2}{2} \mathbb{E}_t \|\hat{\boldsymbol{g}}_H^t\|^2 \\ &\leq \mathcal{L}_R(\boldsymbol{\theta}^t) - \tfrac{\eta c}{2} \|\boldsymbol{g}_R^t\|^2 + \tfrac{L\eta^2}{2} (1-\alpha)^2 \sigma_R^2, \\ \frac{\eta c}{2} \sum_{t=0}^{T-1} \mathbb{E} \|\boldsymbol{g}_R^t\|^2 &\leq \mathcal{L}_R^0 - \mathbb{E} \mathcal{L}_R(\boldsymbol{\theta}^T) + \tfrac{TL\eta^2}{2} (1-\alpha)^2 \sigma_R^2 \\ &\leq \mathcal{L}_R^0 - \mathcal{L}_R^* + \tfrac{TL\eta^2}{2} (1-\alpha)^2 \sigma_R^2.\end{aligned}$$

Dividing by $\eta c T/2$ proves the rate. This argument applies to the stated population-normalized update; the EMA scale in Algorithm 2 estimates its scale. □

Remark 5 (Additional properties of the M3 update)**.** *At the minimum-norm coefficient α^*, the convex-hull optimality condition gives $\langle \boldsymbol{g}_H(\alpha^*), \boldsymbol{g}_i \rangle \geq \|\boldsymbol{g}_H(\alpha^*)\|^2$ for $i \in \{D, R\}$, establishing*

first-order Pareto descent (Sener & Koltun, 2018). By comparison, naive mixing with $\alpha = \frac{1}{2}$ increases distillation loss to first order exactly when $\Phi_{DR} < -1/\kappa$, since $\langle \boldsymbol{g}_H, \boldsymbol{g}_D \rangle = \frac{1}{2} {r_D}^2 (1 + \kappa \Phi_{DR})$. The normalized reward projection instead follows Eq. 46. For an interior minimum-norm coefficient, completing the square gives Eq. 26; at $\alpha = \frac{1}{2}$ its gap is $({r_R}^2 - {r_D}^2)^2/(4\|\boldsymbol{g}_D - \boldsymbol{g}_R\|^2)$. A clipped boundary minimizer also contributes the corresponding one-sided linear term.

K Proof of Proposition 3: Variance Reduction via Distillation

Proof. In the conditional model above, $\hat{\boldsymbol{g}}_H^t - \mathbb{E}_t \hat{\boldsymbol{g}}_H^t = (1-\alpha)\boldsymbol{\xi}^t$, hence

$$\mathrm{Var}_t(\hat{\boldsymbol{g}}_H^t) = (1-\alpha)^2 \mathbb{E}_t \|\boldsymbol{\xi}^t\|^2 \le (1-\alpha)^2 \sigma_R^2.$$

Its contribution to the convergence bound is $L\eta(1-\alpha)^2\sigma_R^2/c$. Relative to the pure-RL bound at the same admissible step size, the noise-floor factor satisfies

$$\frac{(1-\alpha)^2}{c} < 1 \quad \Longleftrightarrow \quad c - (1-\alpha)^2 = \alpha(1 - \alpha + \phi_*) > 0 \quad \Longleftrightarrow \quad \phi_* > -(1-\alpha).$$

For $\phi_* \ge 0$ this factor is at most $1-\alpha$. The variance contraction follows from the reward weight; normalization additionally controls the mean projection through c. □

L Joint Proof of Theorem 1 and Proposition 2

We use the conditional, orthogonal-position model of the two statements. At each iteration, let $\boldsymbol{u}_n$ be the mean reward contribution and $\boldsymbol{d}_n$ its conditionally deterministic teacher counterpart, with $\|\boldsymbol{d}_n\| = \|\boldsymbol{u}_n\|$. The reward noise $\boldsymbol{\xi}_n$ has $\mathbb{E}_t \boldsymbol{\xi}_n = 0$ and $\nu_n = \mathbb{E}_t \|\boldsymbol{\xi}_n\|^2$. Contributions from distinct positions lie in mutually orthogonal parameter subspaces, as in the exact block-diagonal NTK model. The gates are fixed before this fresh noise is sampled. Suppressing t locally, define

$$\alpha_n = \alpha_{\max} \mathbf{1}[\langle \boldsymbol{u}_n, \boldsymbol{d}_n \rangle \ge 0], \qquad \rho_{\mathrm{adm}} = \frac{|\Omega_+ \cup \Omega_0|}{N}, \qquad \bar{\alpha} = \alpha_{\max} \rho_{\mathrm{adm}} = \alpha_{\max}(1 - C_{\mathrm{NTK}}). \tag{48}$$

Here $\boldsymbol{u} = N^{-1} \sum_n \boldsymbol{u}_n = \nabla \mathcal{L}_R(\boldsymbol{\theta}^t)$ and $\hat{\boldsymbol{g}}_H = N^{-1} \sum_n [(1-\alpha_n)(\boldsymbol{u}_n + \boldsymbol{\xi}_n) + \alpha_n \boldsymbol{d}_n]$. For a zero-norm position set $\boldsymbol{d}_n = 0$ and $c_n = 0$; it has zero reward-energy weight. At $\boldsymbol{u} = 0$, all mean contributions vanish and the projection bound holds directly.

Proof. The same decomposition yields both projection and variance, so it suffices to establish their constants once. With $w_n = \|\boldsymbol{u}_n\|^2 / \sum_m \|\boldsymbol{u}_m\|^2$, $c_n = \langle \boldsymbol{u}_n, \boldsymbol{d}_n \rangle / \|\boldsymbol{u}_n\|^2$, and $\bar{\boldsymbol{g}}_H = \mathbb{E}_t \hat{\boldsymbol{g}}_H$, orthogonality gives

$$\langle \boldsymbol{u}, \bar{\boldsymbol{g}}_H \rangle = \frac{1}{N^2} \sum_n [(1-\alpha_n) + \alpha_n c_n] \|\boldsymbol{u}_n\|^2 = \rho_{R,t} \|\boldsymbol{u}\|^2, \quad \rho_{R,t} := \sum_n w_n [(1-\alpha_n) + \alpha_n c_n], \tag{49}$$

$$\rho_{R,t} \ge 1 - \alpha_{\max} \sum_{n \in \Omega_+ \cup \Omega_0} w_n \ge 1 - \alpha_{\max} > 0, \tag{50}$$

$$\mathrm{Var}_t(\hat{\boldsymbol{g}}_H) = \mathbb{E}_t \Big\| \frac{1}{N} \sum_n (1-\alpha_n) \boldsymbol{\xi}_n \Big\|^2 = \frac{1}{N^2} \sum_n (1-\alpha_n)^2 \nu_n = v_t \sigma_{R,t}^2, \tag{51}$$

where $\sigma_{R,t}^2 = N^{-2} \sum_n \nu_n$ and $v_t = \sum_n (1-\alpha_n)^2 \nu_n / \sum_n \nu_n \le 1$; use $v_t = N^{-1} \sum_n (1-\alpha_n)^2$ when all $\nu_n = 0$. Cross-position noise terms vanish by the subspace orthogonality. These identities show why the aggregate projection uses reward-energy weights and the variance uses noise-energy weights.

For equal position variances, the variance factor has the closed form

$$v_t = (1-\alpha_{\max})^2 \rho_{\mathrm{adm}} + 1 - \rho_{\mathrm{adm}} = 1 - (2\alpha_{\max} - \alpha_{\max}^2)\rho_{\mathrm{adm}} = (1-\bar{\alpha})^2 + \alpha_{\max}^2 \rho_{\mathrm{adm}}(1 - \rho_{\mathrm{adm}}). \tag{52}$$

This proves Proposition 2, including the correction due to heterogeneity of the gate.

For convergence, take uniform bounds $\underline{\rho}_R \le \rho_{R,t}$, $v_t \le v_\alpha$, and $\sigma_{R,t}^2 \le \sigma_R^2$. The local triangle inequality gives $\|(1-\alpha_n)\boldsymbol{u}_n + \alpha_n \boldsymbol{d}_n\| \le \|\boldsymbol{u}_n\|$, so orthogonality and Eq. 51 imply

$$\mathbb{E}_t \|\hat{\boldsymbol{g}}_H^t\|^2 = \|\bar{\boldsymbol{g}}_H^t\|^2 + \mathrm{Var}_t(\hat{\boldsymbol{g}}_H^t) \le \|\nabla \mathcal{L}_R(\boldsymbol{\theta}^t)\|^2 + v_\alpha \sigma_R^2. \tag{53}$$

Substituting Eqs. 49 and 53 into the smoothness-and-telescoping chain in Appendix J, with c replaced by $\underline{\rho}_R$ and $(1-\alpha)^2$ by v_α, gives for $\eta \le \underline{\rho}_R/L$,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla\mathcal{L}_R(\boldsymbol{\theta}^t)\|^2 \le \frac{1}{\underline{\rho}_R}\left[\frac{2(\mathcal{L}_R^0-\mathcal{L}_R^*)}{\eta T}+L\eta v_\alpha\sigma_R^2\right]. \tag{54}$$

The choice $\underline{\rho}_R = 1-\alpha_{\max}$ is always valid in this model. □

Comparison with uniform mixing. For the same matched local directions, masking increases the energy-weighted reward projection by $\alpha_{\max}\sum_{n\in\Omega_-} w_n(1-c_n) \ge 0$. It also leaves more reward noise than uniform mixing: $v_t \ge (1-\alpha_{\max})^2$. The convergence comparison therefore depends on the ratio of variance to projection. Against pure RL, the variance contracts strictly whenever an admitted position carries nonzero reward noise. These are reward-descent guarantees; at a rejected position the pure reward direction can still oppose the teacher direction.

M PCGrad Degeneracy in the Asymmetric Regime

Proposition 19 (PCGrad Magnitude Preservation)**.** *Let $\boldsymbol{g}_D, \boldsymbol{g}_R$ be nonzero gradients with cosine $-1 < \Phi < 0$. Their symmetric PCGrad projections preserve the magnitude ratio:*

$$\frac{\|\boldsymbol{g}_D^\perp\|}{\|\boldsymbol{g}_R^\perp\|} = \frac{\|\boldsymbol{g}_D\|}{\|\boldsymbol{g}_R\|} = \frac{r_D}{r_R} = \frac{1}{\kappa}. \tag{55}$$

Proof. For $(i,j) \in \{(D,R),(R,D)\}$, orthogonal projection gives

$$\|\boldsymbol{g}_i^\perp\|^2 = \left\|\boldsymbol{g}_i - \frac{\langle \boldsymbol{g}_i,\boldsymbol{g}_j\rangle}{\|\boldsymbol{g}_j\|^2}\boldsymbol{g}_j\right\|^2 = \|\boldsymbol{g}_i\|^2 - \frac{\langle \boldsymbol{g}_i,\boldsymbol{g}_j\rangle^2}{\|\boldsymbol{g}_j\|^2} = r_i{}^2(1-\Phi^2). \tag{56}$$

Canceling the common positive factor proves the ratio. Consequently, for $0<\alpha<1$, the norm-based teacher share in $\alpha\boldsymbol{g}_D^\perp + (1-\alpha)\boldsymbol{g}_R^\perp$ is $w_D = \alpha/[\alpha+(1-\alpha)\kappa]$, approximately 0.03% at $\alpha = 0.5$ and $\kappa = 3{,}400$. At exact antiparallelity both projections vanish. □

N Proof of Proposition 6: Phase Delay in Adaptive Mixing

Proof. Consider an alignment estimate with exponential smoothing $\hat{\Phi}_{t+1} = \rho\hat{\Phi}_t + (1-\rho)\Phi_t$, $0<\rho<1$. Suppose the input changes from $\Phi_+ > 0$ to $\Phi_- < 0$ at t_1, with $\hat{\Phi}_{t_1} = \Phi_+$. Solving the recurrence and its zero-crossing condition in one chain gives

$$\hat{\Phi}_{t_1+k} = \rho^k\Phi_+ + (1-\rho)\Phi_-\sum_{j=0}^{k-1}\rho^j = \Phi_- + (\Phi_+-\Phi_-)\rho^k, \tag{57}$$

$$\hat{\Phi}_{t_1+k} \le 0 \iff \rho^k \le \frac{|\Phi_-|}{\Phi_+-\Phi_-} \iff k \ge \frac{\log[(\Phi_+-\Phi_-)/|\Phi_-|]}{\log(1/\rho)}. \tag{58}$$

Thus the first nonpositive estimate occurs after the ceiling of the last expression. A rule that retains the high teacher weight while $\hat{\Phi} > 0$ continues doing so during this delay, although the current alignment is negative. For a boxcar average of width W, replacing k old observations gives $\hat{\Phi} = \Phi_+ + k(\Phi_- - \Phi_+)/W$, so the corresponding delay is $\lceil W\Phi_+/(\Phi_+-\Phi_-)\rceil$; equal transition magnitudes give approximately $W/2$. The result applies to loss-based schedules when their smoothed control statistic follows this assumed alignment transition. □

O Heuristic Derivation for Proposition 7: Rank-Invariance of κ

Heuristic derivation. Write the distillation gradient as $\boldsymbol{g}_D = N^{-1}\sum_n \boldsymbol{d}_n$. To isolate the role of the LoRA subspace, suppose both signals share a rank-dependent second-moment factor $s_r > 0$, with $\mathbb{E}\|\boldsymbol{g}_R\|^2 = a_R s_r$, $\mathbb{E}\|\boldsymbol{d}_n\|^2 = a_D s_r$, and $\mathbb{E}\langle \boldsymbol{d}_n,\boldsymbol{d}_m\rangle = 0$ for $n \ne m$. Here $a_R, a_D > 0$ are rank-independent signal constants. Then

$$\mathbb{E}\|\boldsymbol{g}_D\|^2 = \frac{1}{N^2}\sum_{n,m}\mathbb{E}\langle \boldsymbol{d}_n,\boldsymbol{d}_m\rangle = \frac{a_D s_r}{N}, \qquad \sqrt{\frac{\mathbb{E}\|\boldsymbol{g}_R\|^2}{\mathbb{E}\|\boldsymbol{g}_D\|^2}} = \sqrt{\frac{a_R}{a_D}N}. \tag{59}$$

The common factor cancels; concentration of the squared norms transfers this root-mean-square ratio to typical observed ratios. For example, an isotropic LoRA model may give $s_r \propto Lr/h$ for L adapted layers of width h. The cancellation depends on shared scaling and weak cross-token correlations, which are the modeling assumptions behind this heuristic. The measured rank sweep, $r \in \{8, 16, 32, 64\}$ and $\bar{\kappa} = 3{,}476 \pm 301$ (CV 8.7%), supplies the empirical evidence in Remark 3; the signal constants determine its absolute scale. □

P Proof of Proposition 8: Distillation Mode Determines κ

Proof. For mode $s \in \{\text{on}, \text{off}\}$, let $E_s = \mathcal{L}_D^s(\boldsymbol{\theta}) - \mathcal{L}_D^s(\boldsymbol{\theta}_s^*)$ be the excess distillation loss above its local minimum. In an exact local quadratic model, $E_s = \frac{1}{2}\Delta\boldsymbol{\theta}_s^\top \mathbf{H}_s \Delta\boldsymbol{\theta}_s$ and $\boldsymbol{g}_D^s = \mathbf{H}_s\Delta\boldsymbol{\theta}_s$, where $\Delta\boldsymbol{\theta}_s = \boldsymbol{\theta} - \boldsymbol{\theta}_s^*$ and $\mathbf{H}_s$ is positive definite on the active parameter subspace. Writing λ_s^-, λ_s^+ for its extremal eigenvalues yields

$$2\lambda_s^- E_s = \lambda_s^- \Delta\boldsymbol{\theta}_s^\top \mathbf{H}_s \Delta\boldsymbol{\theta}_s \le \Delta\boldsymbol{\theta}_s^\top \mathbf{H}_s^2 \Delta\boldsymbol{\theta}_s = \|\boldsymbol{g}_D^s\|^2 \le \lambda_s^+ \Delta\boldsymbol{\theta}_s^\top \mathbf{H}_s \Delta\boldsymbol{\theta}_s = 2\lambda_s^+ E_s, \tag{60}$$

$$\frac{\kappa_{\text{on}}}{\kappa_{\text{off}}} = \frac{\|\boldsymbol{g}_D^{\text{off}}\|}{\|\boldsymbol{g}_D^{\text{on}}\|} \ge \sqrt{\frac{\lambda_{\text{off}}^- E_{\text{off}}}{\lambda_{\text{on}}^+ E_{\text{on}}}}, \tag{61}$$

where the second line uses the same nonzero reward gradient in both modes. The same inequality follows from local strong-convexity and smoothness bounds with the corresponding constants. If both losses have zero minimum and share a curvature matrix $\mathbf{F}$, this specializes to $\sqrt{D_{\text{KL}}^{\text{off}}/D_{\text{KL}}^{\text{on}}}/\sqrt{\chi(\mathbf{F})}$.

A nearly deterministic teacher on ground-truth tokens gives $D_{\text{KL}}^{\text{off}}(n) \approx -\log p_S(y_n^*)$. On-policy evaluation changes the prefix distribution, and its KL gap must be measured. Table 11 reports mean teacher losses 0.098 on-policy and 0.810 off-policy, with median magnitude ratios 388 and 68, respectively. These observations establish the cross-mode gap in the evaluated setting; the bound explains how excess loss and curvature jointly control gradient magnitude. □

Q Proof of Proposition 9: Conflict-Free Guarantee

Proof. By Definition 1, $\boldsymbol{g}_D^n = \mathbf{J}^n\boldsymbol{\delta}_D^n$, $\boldsymbol{g}_R^n = \mathbf{J}^n\boldsymbol{\delta}_R^n$, and $\langle \boldsymbol{g}_D^n, \boldsymbol{g}_R^n\rangle = K_{DR}(n)$. With the gate treated as a fixed coefficient during the update,

$$\big\langle \mathbf{1}[K_{DR}(n) \ge 0]\boldsymbol{g}_D^n, \boldsymbol{g}_R^n \big\rangle = \mathbf{1}[K_{DR}(n) \ge 0]K_{DR}(n) = \max\{K_{DR}(n), 0\} \ge 0. \tag{62}$$

Positive norm rescaling preserves this sign, proving the token-level guarantee. □

For comparison, if the aggregate normalized gradients have cosine $\Phi_{DR} \ge 0$ and $0 < \alpha < 1$, their mixture satisfies $\langle \boldsymbol{g}_H, \hat{\boldsymbol{g}}_R\rangle = 1 - \alpha + \alpha\Phi_{DR} > 0$ and $\langle \boldsymbol{g}_H, \hat{\boldsymbol{g}}_D\rangle = \alpha + (1-\alpha)\Phi_{DR} > 0$. The latter projection changes sign at $\Phi_{DR} = -\alpha/(1-\alpha)$, compared with $-\alpha/[(1-\alpha)\kappa]$ for unnormalized mixing. Cross-position interactions enter the aggregate condition through Proposition 20.

R Proof of Proposition 5: GradNorm Degeneracy

Proof. At a norm-balanced GradNorm equilibrium with equal target training rates and $w_D + w_R = 1$, the weighted norms coincide. Consequently,

$$w_D r_D = w_R r_R \implies (w_D^*, w_R^*) = \frac{(\kappa, 1)}{1+\kappa}, \tag{63}$$

$$\boldsymbol{g}_H^{\text{GN}} = \frac{r_R}{1+\kappa}(\hat{\boldsymbol{g}}_D + \hat{\boldsymbol{g}}_R), \qquad \|\boldsymbol{g}_H^{\text{GN}}\| \le \frac{2r_R}{1+\kappa}, \tag{64}$$

$$\langle \boldsymbol{g}_R, \boldsymbol{g}_H^{\text{GN}}\rangle = {r_R}^2 \frac{1+\Phi_{DR}}{1+\kappa}. \tag{65}$$

For an L-smooth reward loss, $\mathcal{L}_R(\boldsymbol{\theta}) - \mathcal{L}_R(\boldsymbol{\theta} - \eta\boldsymbol{g}_H) = \eta\langle \boldsymbol{g}_R, \boldsymbol{g}_H\rangle + O(L\eta^2\|\boldsymbol{g}_H\|^2)$. The first-order reward decrease is therefore $(1+\Phi_{DR})/(1+\kappa)$ of the pure-RL decrease at the same learning rate. Before equilibrium, comparable positive weights give teacher share $w_D/(w_D + w_R\kappa) = \Theta(1/\kappa)$.

To compare with M3-Norm at a common reward scale, use $r_R[\alpha\hat{\boldsymbol{g}}_D + (1-\alpha)\hat{\boldsymbol{g}}_R]$. Its reward projection relative to GradNorm is

$$\frac{\Pi_R^{\mathrm{M3}}}{\Pi_R^{\mathrm{GN}}} = \frac{(1+\kappa)[1-\alpha+\alpha\Phi_{DR}]}{1+\Phi_{DR}}, \tag{66}$$

which is $\Theta(\kappa)$ when both cosine-dependent factors stay positive and bounded away from zero. At $\alpha = 0.15$, $\kappa = 4{,}381$, and near-zero cosine, this first-order ratio is about 3,725. It describes the specified update scaling; learning-rate rescaling or a different weight optimizer changes the comparison. □

S Proof of Proposition 14: Folding–Drowning Coupling

Proof. Let $M_R^2 = G^{-1}\sum_i \|A_i \nabla_{\boldsymbol{\theta}} \log \pi(y_i \mid x)\|^2 > 0$. The definitions of cancellation and raw magnitude immediately give

$$\kappa_{\mathrm{eff}} = \frac{\|\boldsymbol{g}_R\|}{\|\boldsymbol{g}_D\|} = \frac{\sqrt{(1-\gamma)M_R^2}}{\|\boldsymbol{g}_D\|} = \sqrt{1-\gamma}\,\kappa_{\mathrm{raw}}. \tag{67}$$

For the binary-outcome model, let $p \in (0,1)$ be the correct fraction, $\bar{\boldsymbol{g}}_\pm$ the class-mean score gradients, and $\sigma_r = \sqrt{p(1-p)}$. The standardized advantages imply

$$\boldsymbol{g}_R = p\frac{1-p}{\sigma_r}\bar{\boldsymbol{g}}_+ - (1-p)\frac{p}{\sigma_r}\bar{\boldsymbol{g}}_- = \sqrt{p(1-p)}(\bar{\boldsymbol{g}}_+ - \bar{\boldsymbol{g}}_-), \tag{68}$$

$$\|\boldsymbol{g}_R\|^2 = p(1-p)\left(\|\bar{\boldsymbol{g}}_+\|^2 + \|\bar{\boldsymbol{g}}_-\|^2 - 2\langle \bar{\boldsymbol{g}}_+, \bar{\boldsymbol{g}}_-\rangle\right). \tag{69}$$

If score gradients concentrate at class means with common norm s, then $M_R^2 = s^2$ and

$$\gamma = 1 - 2p(1-p)(1-c_{\mathrm{overlap}}) = \underbrace{1-2p(1-p)}_{\text{averaging contribution}} + \underbrace{2p(1-p)c_{\mathrm{overlap}}}_{\text{cross-class overlap contribution}}. \tag{70}$$

For nonnegative overlap this yields $\gamma \geq 2p(1-p)c_{\mathrm{overlap}}$; the overlap contribution peaks at $p = 1/2$ and decreases for $p > 1/2$. It is this contribution, rather than the full cancellation rate, that vanishes as $p \to 1$ in the model.

Combining $\gamma > 0$ with the expected-alignment condition of Corollary 2 gives reduced effective magnitude imbalance and positive expected alignment whenever that corollary's low-accuracy condition holds. As accuracy increases beyond $1/2$, the overlap contribution decreases; Corollary 3 separately describes the decline in expected alignment. These are the two quantities tracked by the accuracy-adaptive interpretation of Eq. 31. □

T Proof of Proposition 20: Aggregate Conflict as a Cross-Signal NTK Sum

Proposition 20 (Aggregate Conflict as Cross-Signal NTK Sum)**.** *For nonzero aggregate gradients, their cosine is the normalized sum of diagonal cross-signal NTK terms and cross-position interactions. Dropping the latter gives the diagonal NTK approximation.*

Proof. Define $\mathcal{E}_{\mathrm{cross}} = N^{-2}\sum_{n\neq m}(\boldsymbol{\delta}_D^n)^\top \mathbf{K}(n,m)\boldsymbol{\delta}_R^m$. Using $\mathbf{K}(n,m) = (\mathbf{J}^n)^\top \mathbf{J}^m$ and Definition 1, the decomposition follows directly:

$$\Phi_{DR} = \frac{1}{N^2\|\boldsymbol{g}_D\|\|\boldsymbol{g}_R\|}\sum_{n,m}(\boldsymbol{\delta}_D^n)^\top \mathbf{K}(n,m)\boldsymbol{\delta}_R^m \tag{71}$$

$$= \frac{N^{-2}\sum_n \langle \mathbf{J}^n\boldsymbol{\delta}_D^n, \mathbf{J}^n\boldsymbol{\delta}_R^n\rangle + \mathcal{E}_{\mathrm{cross}}}{\|\boldsymbol{g}_D\|\|\boldsymbol{g}_R\|} = \frac{N^{-2}\sum_n K_{DR}(n) + \mathcal{E}_{\mathrm{cross}}}{\|\boldsymbol{g}_D\|\|\boldsymbol{g}_R\|}. \tag{72}$$

The diagonal approximation is accurate to the extent that the normalized cross-position remainder is small. □

U Proof of Proposition 15: Distillation Boundary Bound

Proof. Let $c = \Phi_{DR}$ and $q = \sqrt{\Delta_\tau}/\kappa_{\text{raw}} = 1/\kappa_{\text{eff}}$, using $1 - \gamma = 1/\Delta_\tau$ and Proposition 14. For $q \to 0$, the exact diversity identity and its expansion are

$$\Delta_{DR} = \frac{1 + \kappa_{\text{eff}}^2}{1 + \kappa_{\text{eff}}^2 + 2c\kappa_{\text{eff}}} = \frac{1 + q^2}{1 + 2cq + q^2} = 1 - 2cq + O(q^2) \tag{73}$$

$$\leq 1 + \frac{2|c|\sqrt{\Delta_\tau}}{\kappa_{\text{raw}}} + O\left(\frac{\Delta_\tau}{\kappa_{\text{raw}}^2}\right). \tag{74}$$

The expansion is uniform for $c \in [-1, 1]$ with q sufficiently small; its regime is $\kappa_{\text{raw}} \gg \sqrt{\Delta_\tau}$. Hence $B_D = \alpha_{\max}\Delta_{DR} = \alpha_{\max}[1 + O(q)]$.

For nonzero gradients with nonzero sum, the exact denominator also gives $\Delta_{DR} < 1, = 1$, or > 1 according as $c > 0, = 0$, or < 0. Thus unity marks the sign change of the cross-signal contribution to $\|\boldsymbol{g}_D + \boldsymbol{g}_R\|^2$. Relative to pure RL, the separate condition for a smaller squared update norm is $\|\boldsymbol{g}_D\|^2 + 2\langle \boldsymbol{g}_D, \boldsymbol{g}_R\rangle < 0$. □

V Proof of Proposition 16: NTK-Guided Token Masking

Proof. Write $\boldsymbol{u}_n = \mathbf{J}^n \hat{\boldsymbol{\delta}}_D^n$, $\boldsymbol{v}_n = \mathbf{J}^n \hat{\boldsymbol{\delta}}_R^n$, and $S^- = \{n : K_{DR}(n) < 0\}$. These are parameter-space directions obtained by positive residual rescaling, so $\langle \boldsymbol{u}_n, \boldsymbol{v}_n\rangle$ has the sign of $K_{DR}(n)$. The gate replaces $\alpha_{\max}\boldsymbol{u}_n + (1 - \alpha_{\max})\boldsymbol{v}_n$ by $\boldsymbol{v}_n$ on S^- and leaves the other positions unchanged. Consequently, the mean local reward-projection gain is

$$\Delta\Pi_R = \frac{1}{N}\sum_n \langle \boldsymbol{g}_H^{\text{sel},n} - \boldsymbol{g}_H^{\text{M3},n}, \boldsymbol{v}_n\rangle \tag{75}$$

$$= \frac{\alpha_{\max}}{N}\sum_{n\in S^-}\langle \boldsymbol{v}_n - \boldsymbol{u}_n, \boldsymbol{v}_n\rangle = \frac{\alpha_{\max}}{N}\sum_{n\in S^-}\left(\|\boldsymbol{v}_n\|^2 - \langle \boldsymbol{u}_n, \boldsymbol{v}_n\rangle\right) > 0 \tag{76}$$

whenever $\alpha_{\max} > 0$ and S^- is nonempty. In the unit-direction model, $\|\boldsymbol{u}_n\| = \|\boldsymbol{v}_n\| = 1$, this specializes to

$$\Delta\Pi_R = \frac{\alpha_{\max}}{N}\sum_{n\in S^-}(1 + |\cos\varphi_n|) = \alpha_{\max}C_{\text{NTK}}(1 + |\bar{c}^-|), \tag{77}$$

where $C_{\text{NTK}} = |S^-|/N$ and $|\bar{c}^-|$ is the mean absolute parameter-space cosine on S^-. The gain is zero when S^- is empty. Under the diagonal NTK model, cross-position inner products vanish, so the aggregate projection onto $N^{-1}\sum_n \boldsymbol{v}_n$ has the same sign, with gain $\Delta\Pi_R/N$. □

W M3-Select Projection and Convergence Guarantees

Theorem 4 (M3-Select Projection Improvement over M3-Norm)**.** *Under the decoupled-position model of Proposition 16, with $0 < \alpha_{\max} < 1$, M3-Select satisfies:*

- *(a)* ***Reward projection:*** *its aggregate update has at least the reward projection of uniform M3-Norm, strictly larger whenever $C_{\text{NTK}} > 0$.*
- *(b)* ***Teacher allocation:*** *the sum of its teacher coefficients is the fraction $1 - C_{\text{NTK}}$ of the uniform allocation, entirely on positions with $K_{DR}(n) \geq 0$.*
- *(c)* ***Reward convergence:*** *under the additional smoothness, matched-norm, and noise assumptions of Theorem 1, it satisfies the reward-stationarity bound in Eq. 54.*

Proof. Part (a) is Proposition 16, using the positive rescaling from the mean local projection to the aggregate reward projection. For part (b), summing the gate gives $\sum_n \alpha_n = \alpha_{\max}(N - |S^-|) = N\alpha_{\max}(1 - C_{\text{NTK}})$, and every retained coefficient has nonnegative local compatibility. Part (c) is the descent-and-telescoping argument of Theorem 1, applied to the same gate and reward objective. □

X NTK Conflict Rate Dynamics

Proposition 21 (Accuracy Dependence of the Population Weighted Conflict Rate). *Let $\bar{M}^{\pm} > 0$ be the expected per-trajectory absolute cross-signal NTK masses on correct and incorrect trajectories, and let $\widetilde{C}_{\pm}$ be the corresponding ratios of expected negative mass to expected absolute mass. If these four quantities are fixed as accuracy p varies, the pooled population conflict rate is*

$$\widetilde{C}_{\mathrm{NTK}}(p) = \frac{p\bar{M}^{+}\widetilde{C}_{+} + (1-p)\bar{M}^{-}\widetilde{C}_{-}}{p\bar{M}^{+} + (1-p)\bar{M}^{-}}. \tag{78}$$

It increases strictly with p if $\widetilde{C}_{+} > \widetilde{C}_{-}$, interpolating between $\widetilde{C}_{-}$ at $p = 0$ and $\widetilde{C}_{+}$ at $p = 1$.

Proof. Conditioning the expected negative and absolute masses on trajectory correctness gives the displayed ratio. Differentiating and canceling the common terms yields

$$\frac{d\widetilde{C}_{\mathrm{NTK}}}{dp} = \frac{\bar{M}^{+}\bar{M}^{-}(\widetilde{C}_{+} - \widetilde{C}_{-})}{[p\bar{M}^{+} + (1-p)\bar{M}^{-}]^2} > 0. \tag{79}$$

The endpoint values follow by substitution. Exact sign masking retains the fraction $1 - \widetilde{C}_{\mathrm{NTK}}$ of absolute NTK mass, which therefore decreases under the same assumptions. This mass fraction differs from the position fraction $1 - C_{\mathrm{NTK}}$ used in α_{eff}: a change in mass allocation need not change the number of admitted positions. □

Y Proof of Proposition 13: Cosine Inversion in On-Policy OPSD

Proof. For a sampled trajectory, write $\boldsymbol{s} = \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\hat{y} \mid x)$, $\boldsymbol{g}_R = -A\boldsymbol{s}$, and $\boldsymbol{g}_D = N^{-1}\sum_n \nabla_{\boldsymbol{\theta}} \mathrm{KL}(\boldsymbol{p}_T^n \| \boldsymbol{p}_S^n)$, where $A = (r-b)/\sigma_r$, $0 < b < 1$, and $\sigma_r > 0$. Assume $\boldsymbol{s}$ and $\boldsymbol{g}_D$ are nonzero, and define the normalized score–distillation alignment $q = \langle \boldsymbol{s}, \boldsymbol{g}_D \rangle / (\|\boldsymbol{s}\| \|\boldsymbol{g}_D\|)$. Then

$$\cos(\boldsymbol{g}_R, \boldsymbol{g}_D) = -\operatorname{sign}(A) q, \tag{80}$$

$$\mathbb{E}[\cos(\boldsymbol{g}_R, \boldsymbol{g}_D) \mid r = 0] - \mathbb{E}[\cos(\boldsymbol{g}_R, \boldsymbol{g}_D) \mid r = 1] = \mathbb{E}[q \mid r = 0] + \mathbb{E}[q \mid r = 1] > 0, \tag{81}$$

where the last inequality is the proposition's conditional alignment assumption. In particular, positive conditional means of q give positive cosine on incorrect trajectories and negative cosine on correct ones.

The inner product underlying this condition includes all token pairings:

$$\langle \boldsymbol{s}, \boldsymbol{g}_D \rangle = \frac{1}{N} \sum_{m,n} \langle \nabla_{\boldsymbol{\theta}} \log p_S^m(\hat{y}_m),\ \nabla_{\boldsymbol{\theta}} \mathrm{KL}(\boldsymbol{p}_T^n \| \boldsymbol{p}_S^n) \rangle \,. \tag{82}$$

Thus the assumption concerns gradient-weighted alignment, including cross-position interactions. For an on-policy teacher, positive q means its descent direction reduces the sampled trajectory's log probability; changing the sign of the reward advantage reverses whether that change agrees with RL. Table 11 reports the corresponding conditional cosines $+0.14$ and -0.11. □

Z Proof of Theorem 3: Unified NTK Learning Efficiency

Proof. Let $S_R^2 = G^{-1}\sum_i \|\boldsymbol{g}_R^{(i)}\|^2$ and $\boldsymbol{g}_R = G^{-1}\sum_i \boldsymbol{g}_R^{(i)}$. The definitions of diversity, folding, and the normalized hybrid direction give the complete factorization

$$\eta_{\mathrm{explore}} = \frac{\|\boldsymbol{g}_R\|^2}{S_R^2} = \Delta_\tau^{-1} = 1 - \gamma, \tag{83}$$

$$\eta_{\mathrm{hybrid}} = \langle (1-\alpha)\hat{\boldsymbol{g}}_R + \alpha \hat{\boldsymbol{g}}_D, \hat{\boldsymbol{g}}_R \rangle = (1-\alpha) + \alpha \Phi_{DR}, \tag{84}$$

$$\eta^{\mathrm{eff}} := \eta_{\mathrm{explore}} \eta_{\mathrm{hybrid}} = (1-\gamma)\big[(1-\alpha) + \alpha \Phi_{DR}\big]. \tag{85}$$

Both factors depend on NTK inner products. Writing $\boldsymbol{s}_i = \nabla \log \pi(\hat{y}^{(i)} \mid x)$ and $\boldsymbol{g}_R^{(i)} = -A_i \boldsymbol{s}_i$ gives

$$G^2 \|\boldsymbol{g}_R\|^2 = \sum_{i,j} A_i A_j \langle \boldsymbol{s}_i, \boldsymbol{s}_j \rangle = \sum_{i,j} A_i A_j \sum_{n,m} K_t(\tau_i, n; \tau_j, m), \tag{86}$$

while Proposition 20 expands Φ_{DR} into within-position and cross-position signal interactions. These two expansions establish the shared geometry.

The composite index uses the squared cancellation factor. For a unit-scale hybrid update, the actual first-order RL decrease relative to the RMS single-trajectory scale is instead $\sqrt{1-\gamma}\,\eta_{\text{hybrid}}$. At fixed α, differentiation of the composite index yields

$$\frac{d\eta^{\text{eff}}}{dp} = -\gamma'(p)\big[(1-\alpha) + \alpha\Phi_{DR}(p)\big] + (1-\gamma(p))\alpha\Phi_{DR}{}'(p). \tag{87}$$

Hence decreasing alignment lowers the hybrid factor; monotonicity of the product follows when the displayed derivative is nonpositive. Under fixed conditional trajectory cosines $\mu_0 > \mu_1$, their mixture has derivative $\mu_1 - \mu_0$, with zero crossing $p^* = \mu_0/(\mu_0 - \mu_1)$ when $\mu_0 > 0 > \mu_1$. This crossing characterizes the alignment factor, while the cancellation factor contributes separately through $\gamma'(p)$. □

AA Proof of Proposition 4: Optimal Per-Position Mixing

Proof. Write $c_n = \cos\varphi_n$. Valuing admitted distillation by $\rho\alpha_n$ in units of RL projection gives the separable allocation objective

$$\mathcal{J} = \frac{1}{N}\sum_n \big[(1-\alpha_n) + \alpha_n c_n + \rho\alpha_n\big] = 1 + \frac{1}{N}\sum_n \alpha_n(c_n - \tau), \qquad \tau = 1-\rho, \tag{88}$$

subject to $0 \le \alpha_n \le \alpha_{\max}$. If the optional budget $N^{-1}\sum_n \alpha_n \le \bar{\alpha}$ is imposed, its multiplier $\lambda \ge 0$ shifts each coefficient to $c_n - \tau - \lambda$. Maximizing these linear terms gives

$$\alpha_n^* = \begin{cases} \alpha_{\max}, & c_n > \tau + \lambda, \\ 0, & c_n < \tau + \lambda, \\ \text{any feasible value in } [0, \alpha_{\max}], & c_n = \tau + \lambda. \end{cases} \tag{89}$$

For $\rho = 1$ and a slack budget, choose the upper endpoint at ties to obtain $\alpha_n^* = \alpha_{\max}\mathbf{1}[c_n \ge 0]$. Under the isotropic within-position kernel assumption $\mathbf{K}(n,n) = \lambda_n\mathbf{I}$ with $\lambda_n > 0$, $K_{DR}(n) = \lambda_n\langle \boldsymbol{\delta}_D^n, \boldsymbol{\delta}_R^n\rangle$ has the same sign, yielding the M3-Select gate.

For a noisy score $\hat{c}_n = c_n + \varepsilon_n$ with logistic error of scale β^{-1}, the expected hard allocation is

$$\mathbb{E}[\alpha_{\max}\mathbf{1}[\hat{c}_n \ge 0]] = \alpha_{\max}\Pr(\varepsilon_n \ge -c_n) = \alpha_{\max}\sigma(\beta c_n). \tag{90}$$

This is M3-Soft. As $\beta \to \infty$, it approaches the hard gate for $c_n \ne 0$ and assigns half the budget at the indifferent point $c_n = 0$. □

AB Convergence Rate Comparison: M3-Norm vs GradNorm

Theorem 5 (Descent-Bound Comparison at a Common Gradient Scale)**.** *Let $\mathcal{L}_R$ be L-smooth and bounded below, with $\boldsymbol{g}_R = \nabla\mathcal{L}_R$. Compare the norm-restored M3 direction $\boldsymbol{h}_{\text{M3}} = \|\boldsymbol{g}_R\|[(1-\alpha)\hat{\boldsymbol{g}}_R + \alpha\hat{\boldsymbol{g}}_D]$ with the norm-balanced GradNorm direction $\boldsymbol{h}_{\text{GN}} = \|\boldsymbol{g}_R\|(\hat{\boldsymbol{g}}_R + \hat{\boldsymbol{g}}_D)/(1+\kappa)$. Assume $0 \le \alpha \le \alpha_0 < 1/2$, $\Phi_{DR} \ge -1+\delta$ for fixed $\delta > 0$, and a fixed magnitude ratio $\kappa > 0$. For method j, let the stochastic update be $\boldsymbol{h}_j + \boldsymbol{\xi}_j$ with conditional mean $\mathbb{E}[\boldsymbol{\xi}_j] = 0$ and $\mathbb{E}\|\boldsymbol{\xi}_j\|^2 \le s_j^2$. Define*

$$(c_{\text{M3}}, b_{\text{M3}}) = (1-2\alpha_0, 1), \qquad (c_{\text{GN}}, b_{\text{GN}}) = \left(\frac{\delta}{1+\kappa}, \frac{2}{1+\kappa}\right). \tag{91}$$

For a common step size $\eta \le \min_j c_j/(Lb_j^2)$ and $\Delta_R = \mathcal{L}_R(\boldsymbol{\theta}^0) - \inf\mathcal{L}_R$,

$$\frac{1}{T}\sum_{t<T}\mathbb{E}\|\nabla\mathcal{L}_R(\boldsymbol{\theta}^t)\|^2 \le \frac{2\Delta_R}{\eta c_j T} + \frac{L\eta s_j^2}{c_j}. \tag{92}$$

When $L\eta s_j^2/c_j \le \epsilon/2$, sufficient iteration budgets scale as $O(\Delta_R/(\eta\epsilon))$ for M3 and $O((1+\kappa)\Delta_R/(\eta\epsilon))$ for GradNorm.

Proof. Bilinearity and the triangle inequality give $\langle \boldsymbol{g}_R, \boldsymbol{h}_j \rangle \geq c_j \|\boldsymbol{g}_R\|^2$ and $\|\boldsymbol{h}_j\| \leq b_j \|\boldsymbol{g}_R\|$. Applying smoothness conditionally on the current iterate and using the step-size restriction yields the single descent chain

$$\mathbb{E}_t[\mathcal{L}_R(\boldsymbol{\theta}^{t+1})] \leq \mathcal{L}_R(\boldsymbol{\theta}^t) - \eta \langle \boldsymbol{g}_R, \boldsymbol{h}_j \rangle + \frac{L\eta^2}{2}(\|\boldsymbol{h}_j\|^2 + s_j^2) \tag{93}$$

$$\leq \mathcal{L}_R(\boldsymbol{\theta}^t) - \eta \left(c_j - \frac{L\eta b_j^2}{2} \right) \|\boldsymbol{g}_R\|^2 + \frac{L\eta^2 s_j^2}{2} \tag{94}$$

$$\leq \mathcal{L}_R(\boldsymbol{\theta}^t) - \frac{\eta c_j}{2} \|\boldsymbol{g}_R\|^2 + \frac{L\eta^2 s_j^2}{2}. \tag{95}$$

Telescoping and dividing by $\eta c_j T/2$ proves the bound; taking $T \geq 4\Delta_R/(\eta c_j \epsilon)$ gives the stated sufficient budgets. Their optimization terms differ by a factor proportional to $1 + \kappa$ under the specified common scale and step size. □

AC Concentration of the NTK Conflict Rate

Proposition 22 (Conflict-Rate Variance under Mixing). *Let* $\iota_n = \mathbf{1}[K_{DR}(n) < 0]$ *and* $C_{\mathrm{NTK}} = N^{-1} \sum_n \iota_n$. *Suppose* $|\operatorname{Corr}(\iota_n, \iota_{n+k})| \leq C_\rho e^{-\beta_{\mathrm{mix}} k}$ *whenever both variances are nonzero. Define* $D_\rho = 1 + 2C_\rho/(e^{\beta_{\mathrm{mix}}} - 1)$. *Then*

$$\operatorname{Var}(C_{\mathrm{NTK}}) \leq \frac{D_\rho}{4N}, \qquad \Pr[|C_{\mathrm{NTK}} - \mathbb{E}C_{\mathrm{NTK}}| > t] \leq \min\left\{1, \frac{D_\rho}{4Nt^2}\right\}. \tag{96}$$

For the mean over B independent length-N rollouts, with probability at least $1 - \delta$, the deviation is at most $\sqrt{D_\rho/(4BN\delta)}$.

Proof. Since $\operatorname{Var}(\iota_n) \leq 1/4$, the covariance expansion reduces to a geometric series:

$$\operatorname{Var}\left(\frac{1}{N} \sum_n \iota_n \right) = \frac{1}{N^2} \left[\sum_n \operatorname{Var}(\iota_n) + 2 \sum_{k=1}^{N-1} \sum_{n=1}^{N-k} \operatorname{Cov}(\iota_n, \iota_{n+k}) \right] \tag{97}$$

$$\leq \frac{1}{4N} \left[1 + 2 \sum_{k=1}^{N-1} \left(1 - \frac{k}{N} \right) C_\rho e^{-\beta_{\mathrm{mix}} k} \right] \leq \frac{D_\rho}{4N}. \tag{98}$$

Chebyshev's inequality gives the tail bound; independence divides the variance by B. Using $D_\rho \leq 1 + 2C_\rho/\beta_{\mathrm{mix}}$, the illustrative setting $N = 256$, $C_\rho = 1$, $\beta_{\mathrm{mix}} = 0.5$, and $\delta = 0.05$ gives deviations at most 0.313 for one rollout and 0.079 for $B = 16$. □

AD Proof of Proposition 10: Distillation as Implicit Regularization

Proof. Consider matched teacher and student distributions evaluated on the same contexts, with $\pi_{\boldsymbol{\theta}_T} = \pi_T$. The local KL expansion is $\mathcal{L}_D(\boldsymbol{\theta}_T + \Delta\boldsymbol{\theta}) = \frac{1}{2} \Delta\boldsymbol{\theta}^\top \mathbf{F}_D \Delta\boldsymbol{\theta} + o(\|\Delta\boldsymbol{\theta}\|^2)$, where $\mathbf{F}_D \succeq 0$ is the Fisher matrix. In the proposition's isotropic quadratic model, $\mathbf{F}_D = f_D \mathbf{I}$ with $f_D > 0$, so $\nabla\mathcal{L}_D = f_D \Delta\boldsymbol{\theta}$ and $\langle \nabla\mathcal{L}_D, \Delta\boldsymbol{\theta} \rangle = f_D \|\Delta\boldsymbol{\theta}\|^2$: distillation provides a restoring direction.

With constant $\boldsymbol{g}_R$, initial condition $\Delta\boldsymbol{\theta}(0) = 0$, and $0 < \alpha < 1$, solve the resulting linear flow in one step:

$$\dot{\Delta\boldsymbol{\theta}}_H = -\eta[(1 - \alpha)\boldsymbol{g}_R + \alpha f_D \Delta\boldsymbol{\theta}_H], \tag{99}$$

$$\Delta\boldsymbol{\theta}_H(T) = -\frac{(1 - \alpha)\boldsymbol{g}_R}{\alpha f_D}(1 - e^{-\eta\alpha f_D T}), \qquad \|\Delta\boldsymbol{\theta}_H(T)\| \leq \frac{(1 - \alpha)\|\boldsymbol{g}_R\|}{\alpha f_D}. \tag{100}$$

Pure RL has $\Delta\boldsymbol{\theta}_R(T) = -\eta T \boldsymbol{g}_R$. If degradation is proportional to deviation with the same coefficient for both flows, setting $x = \eta\alpha f_D T$ and $\rho_{\mathrm{reg}} = \eta f_D T/2$ gives

$$\frac{\delta_H}{\delta_R} = (1 - \alpha)\frac{1 - e^{-x}}{x} \leq \frac{1 - \alpha}{1 + x/2} \leq \frac{1}{1 + \alpha\rho_{\mathrm{reg}}}. \tag{101}$$

The scalar inequality follows from $(1+x/2)(1-e^{-x}) \le x$ for $x \ge 0$; the difference has derivative $[1-(1+x)e^{-x}]/2 \ge 0$ and vanishes at zero.

For the rank-dependent model $G(r) = \epsilon_{\text{drown}} - [\delta_R(r) - \delta_H(r)]$, a finite crossing follows if $\epsilon_{\text{drown}} > 0$ is bounded, $\delta_R(r)$ is continuous and grows without bound, and $\rho_{\text{reg}}(r) \ge \rho_0 > 0$. Indeed,

$$G(r) \le \epsilon_{\text{drown}} - \frac{\alpha\rho_0}{1+\alpha\rho_0}\delta_R(r) \longrightarrow -\infty. \tag{102}$$

Together with a positive initial gap, continuity gives a crossing of this model. The reported positive gap at $r = 64$ and negative gap at $r = 128$ locate the observed reversal between the tested ranks. □

AE Proof of Proposition: LSGV Variance Amplification

Proof. Write $\boldsymbol{u} = \boldsymbol{g}_R$, $\boldsymbol{v} = \boldsymbol{g}_D$, $\boldsymbol{\mu}_u = \mathbb{E}\boldsymbol{u}$, $\boldsymbol{\mu}_v = \mathbb{E}\boldsymbol{v}$, and let the joint estimator covariance be $\boldsymbol{\Omega}/B$, including its cross-signal blocks. Assume $\lambda_- \mathbf{I} \preceq \boldsymbol{\Omega} \preceq \lambda_+ \mathbf{I}$ for fixed positive constants, $0 < m \le \|\boldsymbol{\mu}_v\| \le M$, and $|c_0| \le 1 - \delta_c$, where $c_0 = \langle \boldsymbol{\mu}_u, \boldsymbol{\mu}_v \rangle / (\|\boldsymbol{\mu}_u\| \|\boldsymbol{\mu}_v\|)$. Set $\bar{\kappa} = \|\boldsymbol{\mu}_u\|/\|\boldsymbol{\mu}_v\| \le 1$. In the small-noise regime, with delta-method remainders negligible relative to the leading variance, the cosine gradient, evaluated at $(\boldsymbol{\mu}_u, \boldsymbol{\mu}_v)$ with $\hat{\boldsymbol{\mu}}_j = \boldsymbol{\mu}_j/\|\boldsymbol{\mu}_j\|$, and its squared norm are

$$\nabla_{\boldsymbol{u}} f = \frac{\hat{\boldsymbol{\mu}}_v - c_0 \hat{\boldsymbol{\mu}}_u}{\|\boldsymbol{\mu}_u\|}, \qquad \nabla_{\boldsymbol{v}} f = \frac{\hat{\boldsymbol{\mu}}_u - c_0 \hat{\boldsymbol{\mu}}_v}{\|\boldsymbol{\mu}_v\|}, \tag{103}$$

$$\|\nabla f\|^2 = (1 - c_0^2)\left(\frac{1}{\|\boldsymbol{\mu}_u\|^2} + \frac{1}{\|\boldsymbol{\mu}_v\|^2}\right), \qquad f(\boldsymbol{u}, \boldsymbol{v}) = \frac{\langle \boldsymbol{u}, \boldsymbol{v} \rangle}{\|\boldsymbol{u}\| \|\boldsymbol{v}\|}. \tag{104}$$

Consequently, retaining the joint covariance throughout,

$$\mathrm{Var}(c) = \frac{1}{B} \nabla f^\top \boldsymbol{\Omega} \nabla f + o\left(\frac{\|\nabla f\|^2}{B}\right) = \Theta\left(\frac{1-c_0^2}{B}\left[\frac{1}{\|\boldsymbol{\mu}_u\|^2} + \frac{1}{\|\boldsymbol{\mu}_v\|^2}\right]\right) = \Theta\left(\frac{1}{B\bar{\kappa}^2}\right). \tag{105}$$

This rate describes amplification while relative noise is small. The global bound $\mathrm{Var}(c) \le 1$ continues to hold when that approximation ceases to apply.

For $\alpha = \alpha_{\max}\sigma(\beta c)$, a second delta expansion and the Lipschitz constant $\alpha_{\max}\beta/4$ give, respectively,

$$\mathrm{Var}(\alpha) = \alpha_{\max}^2 \beta^2 \sigma'(\beta c_0)^2 \,\mathrm{Var}(c) + o(\mathrm{Var}(c)), \tag{106}$$

$$\mathrm{Var}(\alpha) \le \min\left\{\frac{\alpha_{\max}^2}{4}, \frac{\alpha_{\max}^2 \beta^2}{16}\,\mathrm{Var}(c)\right\}. \tag{107}$$

Static Hybrid has $\mathrm{Var}(\alpha) = 0$. To identify the corresponding contribution to update variance, let $a_0 = \alpha_{\max}\sigma(\beta c_0)$, $\boldsymbol{d} = \boldsymbol{\mu}_v - \boldsymbol{\mu}_u$, and $\delta\boldsymbol{h}_0 = (1-a_0)(\boldsymbol{u} - \boldsymbol{\mu}_u) + a_0(\boldsymbol{v} - \boldsymbol{\mu}_v)$. Linearizing $\boldsymbol{h} = (1-\alpha)\boldsymbol{u} + \alpha\boldsymbol{v}$ yields

$$\mathrm{Cov}(\boldsymbol{h}) \simeq \mathrm{Cov}(\delta\boldsymbol{h}_0) + \mathrm{Var}(\alpha)\boldsymbol{d}\boldsymbol{d}^\top + \mathrm{Cov}(\delta\boldsymbol{h}_0, \alpha)\boldsymbol{d}^\top + \boldsymbol{d}\,\mathrm{Cov}(\alpha, \delta\boldsymbol{h}_0). \tag{108}$$

A constant gate removes the gate fluctuation terms. When the gate error is uncorrelated with $\delta\boldsymbol{h}_0$, the additional covariance is the positive semidefinite term $\mathrm{Var}(\alpha)\boldsymbol{d}\boldsymbol{d}^\top$; for a gate computed from the same gradients, the displayed cross-covariances determine its net effect. □

AF Proof of Proposition: Cumulative Conflict Budget

Proof. At step s, let $q_s = N_s^{-1} \sum_n \alpha_{n,s} \mathbf{1}[K_{DR}^s(n) < 0]$ measure retained conflicting teacher weight, and define the scalar budget $E(t) = \sum_{s \le t} \kappa_s q_s$. For naive mixing $q_s = \alpha {C_{\text{NTK}}}^s$; exact sign masking gives $q_s = 0$. If the time-averaged product stabilizes, the entire accumulation law is

$$E(t) = t\left(\frac{1}{t}\sum_{s \le t} \kappa_s q_s\right) = t\bar{d} + o(t), \qquad T_\Gamma \approx \frac{\Gamma}{\bar{d}}, \tag{109}$$

where T_Γ is the first crossing of a fixed budget threshold Γ in this model. For approximately constant κ and conflict rate, $\bar{d}_{\text{naive}} = \alpha\bar{\kappa}\bar{C}_{\text{NTK}}$. If imperfect masking retains weight $q_s \approx \alpha_{\text{eff}}\delta_{\text{mask}}$, then under the same $\bar{\kappa}$,

$$\frac{E_{\text{M3}}(t)}{E_{\text{naive}}(t)} \approx \frac{\alpha_{\text{eff}}\delta_{\text{mask}}}{\alpha\bar{C}_{\text{NTK}}}. \tag{110}$$

The illustrative values $\alpha_{\text{eff}} = 0.09$, $\delta_{\text{mask}} = 0.05$, $\alpha = 0.5$, and $\bar{C}_{\text{NTK}} = 0.40$ give 0.0225, or about 44 times slower budget accumulation.

This calculation concerns retained conflict. A parameter deviation formed from conflict updates obeys $\|\sum_s \eta_s \boldsymbol{d}_s\| \le \sum_s \eta_s \|\boldsymbol{d}_s\|$, so connecting the scalar budget to reward collapse requires a model of update directions and a collapse threshold. The reported ordering—Hybrid at step 221, GradNorm at 358, GRPO at 412, and no M3 collapse through step 500—is an experimental observation. In particular, pure GRPO has $q_s = 0$ and its collapse is governed by dynamics outside this conflict budget. □

AG Supplementary Details for the Main Analysis and Method

AG.1 Distillation Divergence and Local Residual Model

For $D_\lambda(\boldsymbol{p}\|\boldsymbol{q}) = \lambda \mathrm{KL}(\boldsymbol{p}\|\boldsymbol{m}) + (1-\lambda)\mathrm{KL}(\boldsymbol{q}\|\boldsymbol{m})$, with $\boldsymbol{m} = \lambda\boldsymbol{p} + (1-\lambda)\boldsymbol{q}$, strictly positive distributions satisfy

$$\frac{D_\lambda(\boldsymbol{p}\|\boldsymbol{q})}{\lambda} \to \mathrm{KL}(\boldsymbol{p}\|\boldsymbol{q}) \quad (\lambda \to 0), \qquad \frac{D_\lambda(\boldsymbol{p}\|\boldsymbol{q})}{1-\lambda} \to \mathrm{KL}(\boldsymbol{q}\|\boldsymbol{p}) \quad (\lambda \to 1).$$

The experiments use the symmetric point $\lambda = 1/2$. For an unclipped token near student–teacher agreement,

$$\nabla_{\boldsymbol{z}^n} D_\lambda(\boldsymbol{p}_T^n \| \boldsymbol{p}_S^n) = \lambda(1-\lambda)(\boldsymbol{p}_S^n - \boldsymbol{p}_T^n) + O(\|\boldsymbol{p}_S^n - \boldsymbol{p}_T^n\|^2).$$

The main analysis absorbs the leading constant into the teacher update scale and uses the local forward-KL residual $\boldsymbol{\delta}_D^n = \boldsymbol{p}_S^n - \boldsymbol{p}_T^n$. Tokens above the clipping threshold have zero distillation gradient. This is a local approximation, not an identity for arbitrary distributions.

For completeness, the GRPO group statistics in Section 3.1 are $\bar{r} = G^{-1}\sum_i r^{(i)}$ and $\sigma_r^2 = G^{-1}\sum_i (r^{(i)} - \bar{r})^2$. Their values and the sampled responses are held fixed during differentiation.

AG.2 Gradient Geometry and Magnitude Asymmetry

The NTK sums in Eqs. 7–8 form the gradient Gram matrix

$$\begin{pmatrix} \|\boldsymbol{g}_D\|^2 & \langle \boldsymbol{g}_D, \boldsymbol{g}_R \rangle \\ \langle \boldsymbol{g}_R, \boldsymbol{g}_D \rangle & \|\boldsymbol{g}_R\|^2 \end{pmatrix}.$$

Its diagonal entries measure gradient strength; its off-diagonal entries measure interaction. The residual quadratic sums are kernel energies, equal to squared gradient norms up to the common factor N^{-2}. Together, the entries determine the first-order loss changes for given α and η. The normalized score Φ_{DR} retains direction but removes magnitude.

For $0 < \alpha < 1$ and $\Phi_{DR} < 0$, normalize each harmful cross-effect in Eq. 6 by the corresponding objective's own descent term. The ratios are

$$q_D = \frac{1-\alpha}{\alpha}\kappa|\Phi_{DR}|, \qquad q_R = \frac{\alpha}{1-\alpha}\frac{|\Phi_{DR}|}{\kappa}, \qquad \frac{q_D}{q_R} = \left(\frac{1-\alpha}{\alpha}\right)^2 \kappa^2.$$

Thus the relative asymmetry scales as $\Theta(\kappa^2)$ for fixed interior mixing weights. A shrinking teacher residual can increase κ when the reward gradient remains substantial, but the reward residual can also vanish when $A = 0$. Both residual directions and the kernel affect their parameter-gradient norms. Proposition 7 reports the observed LoRA-rank and model-scale comparisons, and Appendix F develops the extended connections.

AG.3 Minimum-Norm Global Mixing

For $\boldsymbol{g}_H(\alpha) = \alpha\boldsymbol{g}_D + (1-\alpha)\boldsymbol{g}_R$, let $r_D = \|\boldsymbol{g}_D\|$, $r_R = \|\boldsymbol{g}_R\|$, and $\Phi_{DR} = \langle \boldsymbol{g}_D, \boldsymbol{g}_R \rangle / (r_D r_R)$, assuming nonzero gradients. The minimum-norm mixture is the point closest to zero on the segment joining the two gradients. Minimizing $\|\boldsymbol{g}_H(\alpha)\|^2$ over $[0,1]$ gives, for $\boldsymbol{g}_D \neq \boldsymbol{g}_R$,

$$\alpha^* = \mathrm{clip}_{[0,1]}\left(\frac{r_R^2 - \Phi_{DR} r_D r_R}{r_D^2 + r_R^2 - 2\Phi_{DR} r_D r_R}\right), \tag{111}$$

The clipping operation restricts the unconstrained minimizer to $[0, 1]$. If the gradients coincide, every coefficient produces the same update. For $\kappa = r_R/r_D \gg 1$, $1 - \alpha^* = O(\kappa^{-1})$. This direction-dependent criterion need not equal the norm-balancing choice $\alpha = \kappa/(1+\kappa)$ in Section 4.1. Neither global coefficient can remove teacher contributions only at conflicting positions. The quadratic gap around α^* is given in Eq. 26.

AG.4 Normalization and Continuous-Gate Limits

The full-gradient normalization principle is

$$\boldsymbol{g}_H^{\mathrm{MA}}(\alpha) = \bar{s}[\alpha\hat{\boldsymbol{g}}_D + (1-\alpha)\hat{\boldsymbol{g}}_R], \qquad \hat{\boldsymbol{g}}_D = \boldsymbol{g}_D/\|\boldsymbol{g}_D\|, \quad \hat{\boldsymbol{g}}_R = \boldsymbol{g}_R/\|\boldsymbol{g}_R\|,$$

for nonzero gradients, with $\bar{s}$ an EMA-smoothed update scale. Algorithm 2 implements normalization on residuals before applying the Jacobians. This preserves compatibility signs but need not equalize parameter-gradient norms. The population-norm-matched convergence model is therefore distinct from the residual-normalized implementation.

For M3-Soft, $\beta \to \infty$ recovers hard selection wherever $K_{DR}(n) \neq 0$. At an exact zero score the weight remains $\alpha_{\max}/2$, whereas M3-Select admits it at weight $\alpha_{\max}$. As $\beta \to 0$, all positions receive $\alpha_{\max}/2$. The mean teacher weight is $\alpha_{\text{eff}} = N^{-1}\sum_n \alpha_n$. Larger β makes selection sharper but increases sensitivity near zero alignment; Appendix B.3 evaluates this trade-off.

AG.5 Conditional Variance and Global Rebalancing

In Theorem 1, $\underline{\rho}_R$ lower-bounds the mean reward projection weighted by each position's squared mean reward-gradient norm; $\underline{\rho}_R = 1 - \alpha_{\max}$ is admissible. The noise factor v_α uniformly bounds the noise-energy-weighted mean of $(1-\alpha_n)^2$. The effective step size includes the update scale, estimated in the algorithm by an EMA.

For equal position noise variances, admitted fraction ρ_{adm}, and $\bar{\alpha} = \alpha_{\max}\rho_{\mathrm{adm}}$, the exact factor is

$$v_t = (1-\rho_{\mathrm{adm}}) + \rho_{\mathrm{adm}}(1-\alpha_{\max})^2 = (1-\bar{\alpha})^2 + \alpha_{\max}^2\rho_{\mathrm{adm}}(1-\rho_{\mathrm{adm}}).$$

Rejected positions retain the full reward-noise contribution, while admitted positions reduce it. These statements condition on the history before fresh reward noise and treat the teacher and gates as fixed. Extra variability in estimated gates or teacher directions is not included.

Remark 6 (Limitations of Global Gradient Rebalancing)**.** *Global rebalancing does not select teacher contributions by position. For symmetric two-gradient projections with $-1 < \Phi_{DR} < 0$, PCGrad (Yu et al., 2020) removes conflicting components but preserves the norm ratio (Appendix M). At $\kappa = 3400$ and equal mixing weights, the teacher accounts for approximately 0.03% of the sum of weighted gradient norms. Under the equal-target-rate, unit-sum weighting model of Proposition 5, GradNorm (Chen et al., 2018) balances norms with reward coefficient $1/(1+\kappa)$, reducing its step scale by order $1/\kappa$. M3 combines residual normalization with token-level control of the teacher signal.*