GR-OPD: GROUP-RELATIVE SEQUENCE CREDIT FOR ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) provides token-level teacher feedback, yet common sampled-token updates assign credit locally and omit the effects of earlier actions on subsequent feedback. We introduce GR-OPD (Group-Relative On-Policy Distillation), which decouples feedback computation from credit assignment by aggregating token feedback into calibrated sequence-level credit. We establish a theoretical basis for this approach by showing that, under ideal on-policy assumptions, broadcasting raw summed feedback to every response token yields an unbiased estimate of the negative gradient of the sequence-level reverse-KL objective. To support effective learning in practice, GR-OPD discounts later feedback that may become less reliable under distribution shift and uses same-prompt rollout groups to calibrate teacher guidance. It preserves relative response preferences while retaining a controlled fraction of the group mean, which we term collective teacher pressure. Across five Qwen3 teacher–student settings and six mathematical reasoning benchmarks, GR-OPD outperforms vanilla OPD in macro Avg@K in every setting, including gains of 10.9 and 11.5 percentage points in two settings with matched thinking patterns. It also maintains stable training under mismatched thinking patterns, where vanilla OPD collapses. Ablations show that retaining an appropriate amount of collective teacher pressure can improve learning, while excessive collective teacher pressure or an unrestricted feedback horizon can destabilize training.

1 INTRODUCTION

On-policy distillation (OPD) has become a common method for post-training large language models (LLMs). Recent frontier LLMs including DeepSeek-V4, MiMo-V2-Flash, and Kimi K3 use OPD to transfer or integrate reasoning capabilities (DeepSeek-AI, 2026; Xiaomi LLM-Core Team, 2026; Kimi Team, 2026). OPD samples responses from the student policy and trains the student to match the teacher's next-token distributions along the generated trajectories. A widely used lightweight variant uses the teacher–student log-probability difference of each student-sampled token as its token-level advantage in a policy-gradient update, avoiding the memory cost of retaining full-vocabulary distributions at every position. Beyond computational efficiency, this sampled-token formulation supports strong post-training gains and provides dense teacher supervision throughout generation, helping to alleviate the credit assignment problem.

Recent work has explored how to make dense teacher supervision more effective in OPD. These efforts include reducing the variance of sampled-token updates (Oh et al., 2026) and addressing the declining reliability of teacher feedback along student-generated trajectories (Liu et al., 2026b). TOPD also addresses limitations of isolated token-level updates by incorporating near-future trajectory guidance (Jiang & Ferraro, 2026). Other approaches refine the transferred signal by isolating changes induced by teacher post-training (Heo et al., 2026) or correcting systematic discrepancies associated with model families and styles (Gu et al., 2026; Wu et al., 2026). These studies suggest that dense teacher feedback is not uniformly beneficial: useful guidance can coexist with teacher-specific preferences that the student cannot readily absorb. Effective distillation therefore calls for considering both how dense teacher feedback is organized into learning signals and how these signals are adapted to the student's current capabilities and reasoning patterns.

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.

Our key insight is that the granularity of teacher feedback computation and the granularity of credit assignment are separate design choices. OPD can compute teacher feedback for every token, but this does not require assigning each action credit only from the feedback computed for that token. We therefore investigate how to aggregate token-level teacher feedback across a response into a trajectory-level signal and use it to assign credit to the actions that produced that response. This allows later teacher feedback to inform earlier decisions while retaining dense token-level scoring. Evidence supports the informativeness of response-level scores: Rethinking OPD finds that averaging teacher feedback over a response distinguishes correct from incorrect answers even when OPD training fails under mismatched thinking patterns (Li et al., 2026a). We observe the same phenomenon in our teacher–student settings with both matched and mismatched thinking patterns.

To understand how aggregation changes credit assignment, we revisit the gradient of the reverse-KL objective underlying OPD. An early action affects both the immediate teacher–student agreement and the later prefixes visited by the student. Building on MiniLLM and Distributional DAgger (Gu et al., 2024; Agrawal et al., 2026), we separate this gradient into local action matching and future state-visitation credit. The common sampled-token update retains only the local term. We show that, under ideal on-policy assumptions, broadcasting the sequence credit obtained by summing teacher feedback to every token yields an unbiased estimate of the negative gradient of OPD's reverse-KL objective. This result establishes the credit-assignment basis for trajectory aggregation and motivates the practical estimator developed below.

Building on this analysis, we propose GR-OPD (Group-Relative On-Policy Distillation), which converts token-level teacher feedback into sequence-level credit that supports stable and effective learning. Our observation that directly broadcasting uncalibrated trajectory-level signals can lead to training collapse motivates two complementary controls. To reduce the influence of later feedback that may become less reliable under distribution shift, GR-OPD applies exponential discounting with parameter γ. To adapt teacher guidance to the student, GR-OPD uses the same-prompt group mean as a reference for separating relative response preferences from a shared score component. This shared component shifts the credit assigned to all responses in the group together; we refer to it as *collective teacher pressure*. GR-OPD preserves the relative preference term and adds a fraction λ of collective teacher pressure before applying a common group normalization. We evaluate GR-OPD across five Qwen3 teacher–student settings and six mathematical reasoning benchmarks. GR-OPD substantially outperforms vanilla OPD in macro Avg@K across all five settings.

Our contributions are threefold:

- **Recasting OPD through sequence-level credit.** We develop a principled sequence-credit formulation of OPD that transforms dense token-level teacher feedback into a response-level signal. By allowing subsequent feedback to inform earlier actions, this formulation captures trajectory-wide effects beyond local token matching. Under ideal on-policy assumptions, broadcasting summed teacher–student log-ratios recovers the full reverse-KL descent gradient in expectation, grounding sequence-level credit in OPD's original objective.
- **Calibrating sequence credit with GR-OPD.** We propose the GR-OPD estimator to calibrate sequence credit by combining feedback-horizon control with collective teacher pressure calibration. Same-prompt groups provide a reference for relative response preferences and collective teacher pressure. The estimator preserves the relative preference term while regulating how much of the shared component enters the student update.
- **Strong long-reasoning performance and stable training.** GR-OPD outperforms vanilla OPD in macro Avg@K across all five teacher–student settings, including gains of 10.9 and 11.5 percentage points in two settings with matched thinking patterns. It also maintains stable training under mismatched thinking patterns, where vanilla OPD collapses. Ablations show that retaining appropriate collective teacher pressure can improve learning, whereas excessive pressure or a full feedback horizon can destabilize training.

2 Related Work

On-policy distillation. MiniLLM and generalized knowledge distillation (GKD) established student-generated trajectories as an effective basis for language-model distillation (Gu et al., 2024; Agarwal et al., 2024). Subsequent work extends this paradigm to multi-teacher and multimodal

learning (Hou et al., 2026), self-distillation with privileged context (Zhao et al., 2026), and asynchronous training (Chen et al., 2026). Alongside these extensions, mechanistic studies investigate gradient variance, teacher–student compatibility, and training pathologies (Oh et al., 2026; Li et al., 2026a; Wang et al., 2026), clarifying when dense teacher supervision translates into effective learning.

Sequence-level credit assignment. DAgger motivates learning on states visited by the student (Ross et al., 2011). MiniLLM and Distributional DAgger further distinguish local action matching from the influence of actions on future state visitation (Gu et al., 2024; Agrawal et al., 2026). To address the limitations of local token-level supervision, TOPD incorporates near-future teacher continuations through optimal-transport alignment to identify trajectory divergence and guide local updates (Jiang & Ferraro, 2026). GR-OPD takes a sequence-level approach to credit assignment, aggregating teacher feedback across the tokens of each student-sampled response so that later feedback can inform earlier actions. Our analysis connects raw-sum broadcasting to the complete sequence-level reverse-KL gradient under ideal on-policy assumptions.

Group-relative optimization and distillation. GRPO and REINFORCE with a leave-one-out baseline use multiple responses to the same prompt to construct relative learning signals (Shao et al., 2024; Ahmadian et al., 2024). Within distillation, GC-OPD separately normalizes trajectory-level OPD scores and verifier rewards within rollout groups, then distributes their disagreement across token-level advantages (Zhang et al., 2026). GR-OPD uses same-prompt groups as a reference for expressing relative response preferences in trajectory-level teacher feedback.

Teacher-signal calibration. G-OPD reformulates reverse-KL distillation as dense KL-regularized reinforcement learning and tunes the balance between the implicit reward and KL regularization (Yang et al., 2026). Other approaches address the content of teacher guidance: CompassOPD uses a teacher-family reference to remove cross-family offsets (Gu et al., 2026); Lightning OPD 2.0 subtracts recurring style-related disagreement (Wu et al., 2026); and On-Policy Delta Distillation isolates the teacher's log-probability changes from its pre-tuning checkpoint (Heo et al., 2026). GR-OPD calibrates the shared component represented by the same-prompt mean, retaining a controlled fraction λ of this collective teacher pressure in the advantages.

Feedback reliability and granularity. Teacher supervision can become less reliable as student-generated prefixes grow longer, motivating early-stopped rollouts and position-dependent weighting (Zhou et al., 2026; Liu et al., 2026b; Xie et al., 2026). FiRe-OPD combines trajectory filtering with token-level reweighting (Li et al., 2026b), while position-weighted self-distillation adjusts supervision according to token position (Liu et al., 2026a). These studies motivate accounting for both feedback reliability and optimization granularity when converting dense teacher supervision into learning signals for the student's actions throughout a long response.

3 From Token Credit to Sequence Credit

3.1 Motivation

Sampled-token OPD assigns each action only its local feedback r_t, the teacher–student log-ratio at that token. It therefore cannot use evidence revealed later in the response to revise the credit assigned to an earlier decision. Dense token-level feedback need not be used only for local credit assignment. Sequence aggregation addresses this limitation by pooling teacher feedback over the trajectory and broadcasting the resulting signal to the actions that produced it. Figure 1(a) illustrates the distinction: local feedback favors the incorrect first action B, whereas aggregation reverses the preference once later evidence is included, assigning greater credit to the action that leads to the correct response A.

This distinction is particularly relevant to cross-pattern distillation, where different reasoning patterns can make the teacher's prefix-level preferences poorly aligned with the student's actions. Sequence aggregation does not remove the underlying pattern mismatch; it integrates token-level teacher feedback distributed across the response, allowing the sequence-level preference to counteract isolated, locally misleading signals. Rethinking OPD provides empirical support for this view: sequence-mean teacher feedback distinguishes correct from incorrect student responses with an AUROC of .75 under a cross-pattern setting, comparable to .73 under a setting with matched thinking patterns (Li et al., 2026a). We also observe this phenomenon in our settings. Across all five teacher–student settings, including both with mismatched thinking patterns, Figure 1(b) reports $J_H(H)$, which measures how strongly teacher scores from the first H tokens favor correct responses within a prompt. Higher values indicate stronger separation; Appendix C.5 gives the definition.

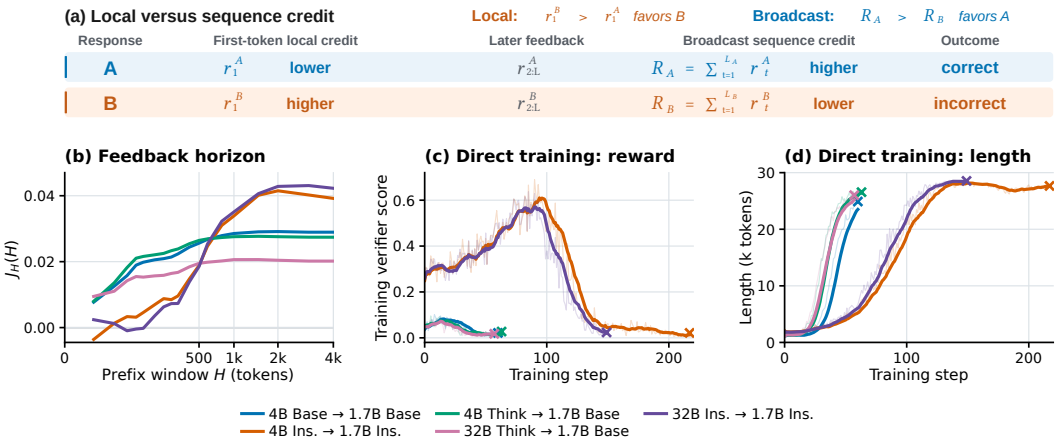

Figure 1: Sequence credit and training stability across five teacher–student pairings. (a) Illustrative local feedback favors incorrect B, whereas trajectory credit favors correct A; outcomes do not enter the objective. (b) $J_H(H)$ measures how strongly prefix scores favor correct responses (higher is better). On fixed responses, it peaks around 1,500–3,000 tokens; the horizontal scale is cube-root. (c,d) Training directly on the discounted trajectory score causes verifier reward to fall as response length approaches the generation cap.

These observations support the informativeness of aggregated teacher feedback. We next develop a theoretical analysis of how trajectory aggregation affects gradient estimation for the reverse-KL objective underlying OPD.

3.2 Unbiased Trajectory Broadcast

For prompt x, let p_θ be the student, q a fixed teacher, and $y = (y_1, \ldots, y_L)$ a student-sampled response with histories $h_t = (x, y_{<t})$. Define token feedback $r_t = \log q(y_t \mid h_t) - \log p_\theta(y_t \mid h_t)$. We revisit the reverse-KL objective underlying OPD:

$$\mathcal{K}(\theta) = D_{\mathrm{KL}}(p_\theta(\cdot \mid x) \,\|\, q(\cdot \mid x)) \,. \tag{1}$$

Let $z_t = \nabla_\theta \log p_\theta(y_t \mid h_t)$ denote the parameter gradient of the sampled token's log-probability; multiplying z_t by its assigned credit gives that token's contribution to the policy update. All expectations below are over responses sampled from $p_\theta(\cdot \mid x)$.

Local feedback captures only part of sequence credit. An action changes both the teacher–student agreement at its own position and the later prefixes visited by the student. Building on MiniLLM and Distributional DAgger (Gu et al., 2024; Agrawal et al., 2026), we write

$$-\nabla_\theta \mathcal{K} = \underbrace{\mathbb{E}\left[\sum_t z_t r_t\right]}_{\text{local action matching}} + \underbrace{\mathbb{E}\left[\sum_t z_t \sum_{u>t} r_u\right]}_{\text{future state-visitation credit}} . \tag{2}$$

The derivation is provided in Appendix A.1. The common sampled-token update retains only the first term. The second attributes later teacher feedback to earlier actions that shaped the subsequent trajectory. We therefore consider broadcasting the full trajectory sum to recover both the local and future components of the reverse-KL gradient.

Proposition 1 (Unbiased trajectory broadcast) *Under on-policy sampling, broadcasting the summed feedback $R = \sum_t r_t$ to every response token yields an unbiased estimate of $-\nabla_\theta \mathcal{K}$.*

Autoregressive factorization gives $\log p_\theta(y \mid x) = \sum_t \log p_\theta(y_t \mid h_t)$, and likewise for q. Thus, $R = \log q(y \mid x) - \log p_\theta(y \mid x)$, so $\mathcal{K} = -\mathbb{E}_{p_\theta}[R]$ by the definition of reverse KL.

Since the teacher is fixed, $\nabla_\theta R = -\sum_t z_t$. Differentiating the expectation accounts for both the sampling distribution and the dependence of R on θ:

$$\begin{aligned} -\nabla_\theta \mathcal{K} = \nabla_\theta \mathbb{E}_{p_\theta}[R] &= \mathbb{E}_{p_\theta}\left[R \sum_t z_t\right] + \mathbb{E}_{p_\theta}[\nabla_\theta R] \\ &= \mathbb{E}_{p_\theta}\left[R \sum_t z_t\right] - \underbrace{\mathbb{E}_{p_\theta}\left[\sum_t z_t\right]}_{=0} = \mathbb{E}_{p_\theta}\left[R \sum_t z_t\right]. \end{aligned} \tag{3}$$

The marked expectation equals the gradient of the total probability over all student responses. This total is one for every θ, so its gradient is zero. Moreover, the past feedback included in R contributes zero in expectation when multiplied by the current token's score gradient: it is fixed by the current history, and the conditional mean of that gradient is zero (Appendix A.2). Thus, broadcasting the same sequence credit R preserves the complete reverse-KL gradient in expectation.

4 Constructing Reliable Group-Relative Sequence Credit

The previous section shows that broadcasting summed teacher feedback preserves the complete reverse-KL gradient under ideal on-policy assumptions. In practice, however, directly using aggregated feedback for policy updates can still lead to unstable training. This motivates GR-OPD, an estimator that calibrates trajectory feedback to support stable and effective learning.

4.1 From Sequence Scores to Stable Learning Signals

Teacher feedback can become less reliable along the trajectory. Sequence credit incorporates later teacher feedback, which may become less reliable under distribution shift as student-generated prefixes grow longer. Prior work reports deteriorating supervision on later student prefixes and improved stability from truncating rollouts (Zhou et al., 2026); the first 30% of tokens can match full OPD while the last 30% alone learns little (Xie et al., 2026); and teacher fidelity decays deeper in the sequence (Liu et al., 2026b).

We next examine how this reliability concern affects aggregated teacher feedback. We rescore the same responses using progressively longer prefixes and measure the resulting scores' within-prompt association with answer correctness. Figure 1(b) shows that this association initially strengthens, then begins to decline beyond pairing-specific peaks between roughly 1,500 and 3,000 tokens. These observations motivate retaining downstream evidence while limiting the influence of later feedback. GR-OPD therefore aggregates feedback using discounted weights, retaining feedback from the full response while progressively reducing the influence of later positions.

Informative trajectory scores do not guarantee stable learning. Reducing the influence of later feedback does not by itself make a trajectory score suitable for direct policy optimization. When we train directly on the discounted trajectory score, all five teacher–student settings exhibit the same failure pattern: the training verifier score rises briefly and then declines, while mean response length approaches the generation cap (Figure 1(c,d)). These results motivate further calibration of trajectory scores before using them as learning signals.

This need for calibration is consistent with recent OPD methods that refine teacher guidance by isolating changes induced by teacher post-training (Heo et al., 2026) or correcting systematic discrepancies associated with model families and styles (Gu et al., 2026; Wu et al., 2026). The teacher–student log-ratio can reflect differences in capability, style, and reasoning pattern, which need not all provide equally useful guidance to the student.

We use responses to the same prompt as a reference for calibrating this guidance. Subtracting the group mean from each trajectory score expresses the relative preference among responses, while the group mean represents a shared score component. We refer to this shared component as *collective teacher pressure*. It may contain useful guidance as well as differences incompatible with the student's current behavior, so its magnitude alone does not determine its teaching value. GR-OPD preserves relative response preferences and adds back a controlled amount of collective teacher pressure to construct the advantages.

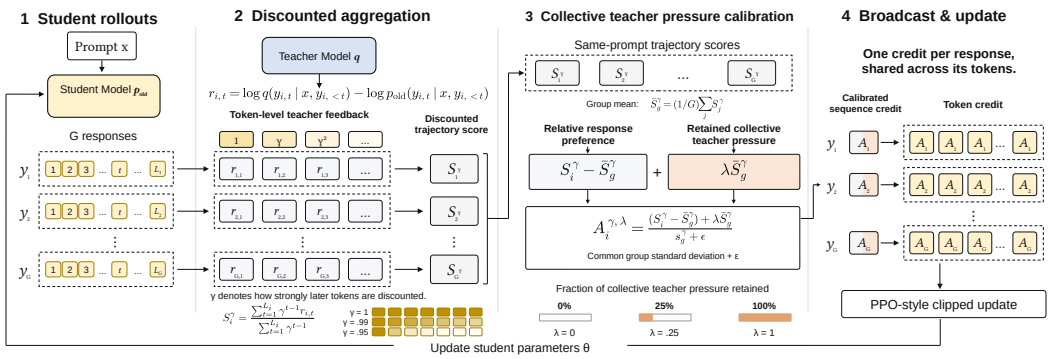

Figure 2: **GR-OPD overview.** (1) The student samples responses to the same prompt and receives token feedback from teacher–student log-probability differences. (2) Discounting with γ aggregates feedback into a trajectory score while down-weighting later positions. (3) Calibration preserves relative response preferences and retains a fraction λ of the group mean, termed *collective teacher pressure*. The two components are added and normalized by a common group standard deviation. (4) Each response's advantage is broadcast to its tokens for a PPO-style clipped update.

4.2 The GR-OPD Estimator

As shown in Figure 2, GR-OPD comprises four components: student rollouts and teacher feedback, discounted trajectory scoring, group-based calibration, and trajectory broadcast for the policy update.

Student rollouts and teacher feedback. For prompt g, the rollout policy p_{old} samples G responses $y_1, \ldots, y_G$. Response i has length L_i and histories $h_{i,t} = (g, y_{i,<t})$. We compute token feedback as $r_{i,t} = \log q(y_{i,t} \mid h_{i,t}) - \log p_{\text{old}}(y_{i,t} \mid h_{i,t})$ on the same history.

Discounted trajectory scoring. We aggregate token feedback into one score per response:

$$S_i^\gamma = \frac{\sum_{t=1}^{L_i} \gamma^{t-1} r_{i,t}}{\sum_{t=1}^{L_i} \gamma^{t-1}}. \tag{4}$$

Dividing by the accumulated discount weight prevents the score from growing merely with response length. Setting $\gamma = 1$ gives the full-response mean, whereas $0 < \gamma < 1$ reduces the influence of later feedback (Appendix A.3). We empirically choose $\gamma = 0.99$ as the default in our experiments.

Group-based calibration. Let $\bar{S}_g^\gamma$ and s_g^γ denote the mean and population standard deviation of the G trajectory scores for prompt g. For a retention fraction $0 \le \lambda \le 1$, the response-level advantage is

$$A_i^{\gamma,\lambda} = \frac{\underbrace{S_i^\gamma - \bar{S}_g^\gamma}_{\text{relative response preference}} + \underbrace{\lambda \bar{S}_g^\gamma}_{\text{retained collective teacher pressure}}}{s_g^\gamma + \epsilon}, \tag{5}$$

Both terms derive from teacher–student feedback: the first expresses relative response preference, while the second shifts all advantages together and can be positive or negative. The common denominator follows GRPO's within-prompt scaling (Shao et al., 2024), with $\epsilon > 0$ for numerical stability. At $\lambda = 0$, GR-OPD gives a GRPO-style relative update that favors responses with above-average teacher–student trajectory scores. At $\lambda = 1$, the full trajectory score is used with the same scaling, retaining all collective teacher pressure.

Trajectory broadcast and policy update. We broadcast $A_i^{\gamma,\lambda}$ to every token in y_i, allowing trajectory feedback to inform each action. Feedback and advantages are held fixed during the update. With token probability ratio $\rho_{i,t}(\theta) = p_\theta(y_{i,t} \mid h_{i,t}) / p_{\text{old}}(y_{i,t} \mid h_{i,t})$, the PPO-style clipped surrogate (Schulman et al., 2017) is

$$\ell_{i,t}(\theta) = \min\Big(\rho_{i,t}(\theta) A_i^{\gamma,\lambda}, \text{clip}(\rho_{i,t}(\theta),\, 1-\delta,\, 1+\delta)\, A_i^{\gamma,\lambda}\Big), \tag{6}$$

where δ is the clipping threshold. We maximize the surrogate averaged within each response and then across responses. Thus, γ controls the feedback horizon, while λ controls the fraction of collective teacher pressure retained.

5 EXPERIMENTS

5.1 SETUP

Teacher–student pairings. Three pairings match the teacher and student families: Qwen3-4B-Base to Qwen3-1.7B-Base, and Qwen3-4B- or 32B-Instruct (nothink) to Qwen3-1.7B-Instruct (nothink). Two further pairings combine a think-mode Instruct teacher with the Qwen3-1.7B-Base student; we regard these as cross-pattern pairings because the teacher uses Instruct thinking mode while the Base student follows its default reasoning protocol. For compactness, B_s and I_s denote Base and Instruct models with s billion parameters, and a superscript records the think or nothink mode of each Instruct model; both Instruct students are used in nothink mode.

Datasets and metrics. All runs are trained on the DAPO-Math prompt set (Yu et al., 2025). We evaluate performance on six mathematical reasoning benchmarks: AIME 2024, AIME 2025, AMC 2023, MATH-500 (Hendrycks et al., 2021), Minerva Math (Lewkowycz et al., 2022), and OlympiadBench (He et al., 2024), which comprise 1,042 problems in total. We report the macro average of per-problem sample accuracy (Avg@K) across the six suites. We tag a run "Collapse" when its response length grows toward the training limit while the task-verifier reward falls to a near-zero level. The tag marks a failed training trajectory.

Implementation details. Training uses 221 steps and a maximum response length of 28,672 tokens, and evaluation uses the same response length. The main-table GR-OPD runs use $\gamma = .99$; they retain no collective teacher pressure ($\lambda = 0$) for the base-to-base and cross-pattern pairings and retain 25% ($\lambda = .25$) for the matched Instruct pairings. The vanilla sampled-token OPD reference uses the same prompts, teacher, rollout batch, step count, and clipped-objective optimization, and differs only in the credit rule: its per-token advantage is the raw token-level teacher feedback $r_{i,t}$ of Section 3, with neither trajectory aggregation nor group normalization. Appendix C clarifies pairing modes and reports model-level reference evaluations and feedback-horizon diagnostics.

5.2 MAIN RESULTS

Table 1 reports the per-suite and aggregate results for the five evaluated pairings, comparing vanilla sampled-token OPD against GR-OPD.

Broad gains over vanilla OPD. Across all five evaluated pairings, GR-OPD raises the macro average and achieves the higher displayed score on every suite. On the compatible pairings, macro Avg@K increases from 23.333 to 27.800 for Base-to-Base, from 34.118 to 45.066 for Qwen3-4B-Instruct, and from 30.320 to 41.774 for Qwen3-32B-Instruct.

Stable learning under pattern mismatch. In both think-Instruct-to-Base pairings, vanilla OPD collapses, ending at 8.45 and 10.07 macro Avg@K. GR-OPD remains trainable and reaches 29.571 and 24.463 under the same two pairings.

Larger teachers do not guarantee larger gains. Holding the student family, teacher mode, λ, and training setup fixed, scaling the teacher from 4B to 32B lowers macro Avg@K from 29.571 to 24.463 for think-Instruct-to-Base and from 45.066 to 41.774 for nothink-Instruct-to-Instruct. This non-monotonic pattern also appears in vanilla OPD (Li et al., 2026a), indicating that teacher size is an unreliable proxy for teaching value.

Table 1: Per-suite Avg@K (%) across five teacher–student pairings. Bold marks the higher score within each pairing; Macro is the unweighted mean over the six suites before rounding. "Collapse" marks a failed training trajectory.

Method	AIME24	AIME25	AMC23	MATH500	Minerva	Olympiad	Macro
Qwen3-4B-Base → Qwen3-1.7B-Base							
Vanilla OPD	3.750	3.229	29.219	56.225	20.956	26.618	23.333
GR-OPD ($\lambda = 0$)	**5.521**	**3.958**	**36.094**	**64.125**	**26.517**	**30.588**	**27.800**
Qwen3-4B-Instruct (nothink) → Qwen3-1.7B-Instruct (nothink)							
Vanilla OPD	17.604	12.708	44.453	68.775	20.358	40.809	34.118
GR-OPD ($\lambda = .25$)	**25.000**	**26.563**	**58.516**	**77.550**	**25.781**	**56.985**	**45.066**
Qwen3-4B-Instruct (think) → Qwen3-1.7B-Base *(pattern mismatch)*							
Vanilla OPD (Collapse)	0.938	1.354	7.813	22.750	8.502	9.338	8.449
GR-OPD ($\lambda = 0$)	**8.229**	**5.833**	**37.891**	**66.375**	**25.276**	**33.824**	**29.571**
Qwen3-32B-Instruct (think) → Qwen3-1.7B-Base *(pattern mismatch)*							
Vanilla OPD (Collapse)	1.875	3.333	13.828	23.050	6.250	12.059	10.066
GR-OPD ($\lambda = 0$)	**4.896**	3.333	**32.266**	**57.350**	**21.140**	**27.794**	**24.463**
Qwen3-32B-Instruct (nothink) → Qwen3-1.7B-Instruct (nothink)							
Vanilla OPD	12.292	11.667	40.313	63.575	18.704	35.368	30.320
GR-OPD ($\lambda = .25$)	**22.188**	**25.000**	**57.031**	**72.850**	**20.267**	**53.309**	**41.774**

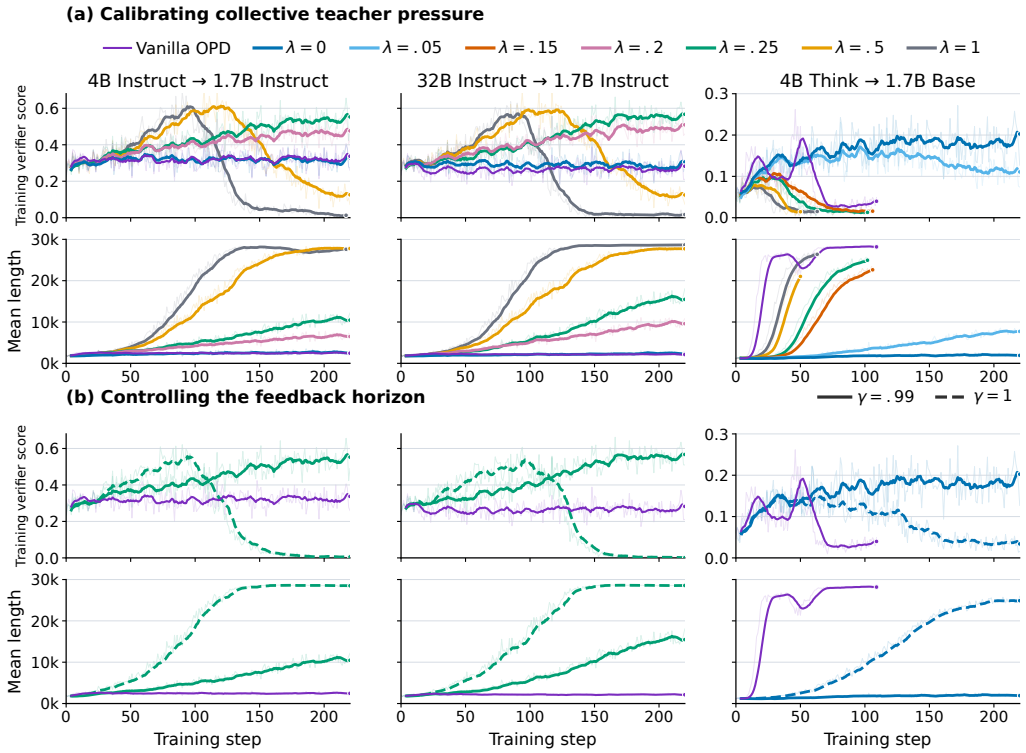

Figure 3: Training dynamics across two matched Instruct pairings and one cross-pattern pairing. Purple curves denote vanilla OPD, and GR-OPD colors denote λ. (a) Calibrating collective teacher pressure at $\gamma = .99$: λ is the retained fraction of the shared component. (b) Controlling the feedback horizon: solid and dashed GR-OPD curves denote $\gamma = .99$ and $\gamma = 1$. Faint lines show per-step values, bold lines show ten-step moving averages, and endpoint dots mark the last recorded step.

5.3 Calibrating collective teacher pressure

An appropriate level of collective teacher pressure sustains improvement. In both matched Instruct pairings, removing all collective teacher pressure ($\lambda = 0$) keeps response length stable but yields limited gains in training-verifier reward. At $\lambda = .25$ or .2, retaining 25% or 20% of collective teacher pressure alongside the relative preference term supports sustained improvements in verifier reward as responses grow longer (Figure 3(a)). Compared with $\lambda = .2$, $\lambda = .25$ generally produces longer responses and a higher final training-verifier reward. The normalized collective teacher pressure estimates in Table 2 are roughly twice as large for the completed cross-pattern runs as for the matched runs; at $\lambda = 0$, no collective teacher pressure is retained in their advantages, and training remains stable in both verifier reward and response length.

Excessive collective teacher pressure can cause training collapse. With full pressure retention ($\lambda = 1$), the shared component is fully retained alongside the relative preference term, and both pattern-matched pairings show rapid length growth followed by reward collapse. In the $\mathrm{I}_4^{\mathrm{nothink}} \rightarrow \mathrm{I}_{1.7}^{\mathrm{nothink}}$ pairing, reducing λ to .5 makes the deterioration less abrupt, but responses still grow excessively long and reward ultimately falls. Even at its training-reward peak, the $\lambda = .5$ checkpoint performs worse in evaluation than the final $\lambda = .25$ model. An early rise in training reward therefore does not establish that the student can sustain the additional collective teacher pressure. One possible explanation is that the student extends its reasoning trajectories with more teacher-preferred content. At an appropriate level of collective teacher pressure, this expansion may produce useful additional reasoning; beyond that range, greater length no longer brings corresponding gains in correctness.

Pattern mismatch requires stronger suppression of collective teacher pressure. For the mismatched $\mathrm{I}_4^{\mathrm{think}} \rightarrow \mathrm{B}_{1.7}$ pairing, even $\lambda = .15$ leads to reward decline and rapid length growth. Reducing λ further to .05 improves training stability, but the final training-verifier reward remains lower than when collective teacher pressure is fully removed ($\lambda = 0$). One possible explanation is that, in such pairings, the shared component carries incompatible reasoning-pattern differences.

The result indicates that the amount of collective teacher pressure a student can absorb depends on teacher–student compatibility rather than on a universal retention fraction.

Table 2: Collective teacher pressure before and after calibration ($\gamma = .99$). Cross-pattern runs show larger normalized pressure than matched runs; λ controls the amount retained alongside relative response preferences. Both columns report batch-level magnitude proxies (Appendix C.7).

Teacher → student	λ	Normalized pressure magnitude proxy	Retained pressure magnitude proxy
$\mathrm{I}_{4}^{\mathrm{nothink}} \to \mathrm{I}_{1.7}^{\mathrm{nothink}}$	.5	3.21	1.60
$\mathrm{I}_{4}^{\mathrm{nothink}} \to \mathrm{I}_{1.7}^{\mathrm{nothink}}$	.25	3.16	0.79
$\mathrm{I}_{4}^{\mathrm{nothink}} \to \mathrm{I}_{1.7}^{\mathrm{nothink}}$	.2	3.17	0.63
$\mathrm{I}_{32}^{\mathrm{nothink}} \to \mathrm{I}_{1.7}^{\mathrm{nothink}}$	.25	3.61	0.90
$\mathrm{I}_{32}^{\mathrm{nothink}} \to \mathrm{I}_{1.7}^{\mathrm{nothink}}$	.2	3.56	0.71
$\mathrm{I}_{4}^{\mathrm{think}} \to \mathrm{B}_{1.7}$	0	7.23	0
$\mathrm{I}_{32}^{\mathrm{think}} \to \mathrm{B}_{1.7}$	0	6.08	0

5.4 Controlling the feedback horizon

A full feedback horizon can destabilize training. Figure 3(b) compares $\gamma = .99$ and $\gamma = 1$ while holding λ fixed within each pairing. In the two pattern-matched pairings at $\lambda = .25$, the full-horizon runs initially gain verifier reward, but reward later declines as response length continues to grow. With $\gamma = .99$, reward continues to improve and length grows more gradually. Our appendix analysis (Figure 4) shows a corresponding change in the $\mathrm{I}_{32}^{\mathrm{nothink}} \to \mathrm{I}_{1.7}^{\mathrm{nothink}}$ pairing: under $\gamma = 1$, trajectory scores become more associated with response length and less aligned with verifier reward.

Including more late feedback may create a length-based score shortcut. Length alone cannot raise a trajectory score that averages token feedback. The risk is that $\gamma = 1$ gives late feedback the same positional weight as early feedback, even as student-generated prefixes move into states where the teacher may be less reliable. On these prefixes, repetitive or circuitous continuations may receive teacher–student log-ratios above the preceding trajectory average. Adding such continuations raises the trajectory score without advancing the solution. Broadcasting the resulting score across the trajectory can then reinforce the behavior that produced those continuations.

Discounting weakens this feedback loop. With $\gamma = .99$, later log-ratios have less influence on the trajectory score, while earlier downstream evidence remains available for credit assignment. The pattern-mismatched $\mathrm{I}_{4}^{\mathrm{think}} \to \mathrm{B}_{1.7}$ pairing makes the distinction clear: both runs use $\lambda = 0$, so collective teacher pressure is removed, yet $\gamma = 1$ still leads to reward decline and length growth toward the training limit. The $\gamma = .99$ run continues to improve with relatively stable length. Horizon control therefore remains important when the update removes all collective teacher pressure and uses only within-prompt relative advantages: those advantages still depend on which parts of the trajectory supply their evidence.

6 Conclusion

We presented GR-OPD, a sequence-credit estimator for on-policy distillation that decouples token-level feedback computation from credit assignment. We showed that, under ideal on-policy assumptions, broadcasting summed teacher feedback recovers the negative gradient of the sequence-level reverse-KL objective in expectation. GR-OPD discounts later feedback and calibrates sequence credit by preserving relative response preferences while controlling the retained collective teacher pressure. Experiments across five Qwen3 teacher–student settings demonstrate consistent improvements over vanilla OPD on mathematical reasoning benchmarks and stable training under mismatched thinking patterns.

Our evaluation is limited to mathematical reasoning and Qwen3 models, leaving generalization to broader domains and model families to be established. Our qualitative analysis suggests that the appropriate level of retained collective teacher pressure depends on teacher–student compatibility. Building on these observations, future work could develop principled criteria for selecting λ across teacher–student pairings. Moreover, broadcasting a single sequence-level credit does not explicitly differentiate credit among tokens within a response. Future work could combine calibrated sequence-level credit with token-level credit assignment to preserve trajectory-wide information while enabling more localized updates. GR-OPD's same-prompt group structure also provides a natural basis for integration with GRPO: shared rollout groups could support both teacher-derived credit and verifier-based rewards, enabling joint optimization of knowledge transfer and task performance.

Reproducibility Statement

Appendix C clarifies pairing modes and records model-level reference evaluations and feedback-horizon diagnostics. The accompanying repository stores the provenance for every table and figure. Theoretical derivations and practical-estimator boundaries appear in Appendices A and B.

AI Use Statement

Large language models were used to help find and check relevant literature, discuss experimental design and interpret results, organize experimental records, and draft and edit parts of the manuscript. They also assisted with checking mathematical exposition and preparing figures and code. The authors reviewed the AI-assisted work, checked citations and reported results against their sources, and take responsibility for the final manuscript.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. In *International Conference on Learning Representations*, pp. 21246–21263, 2024.

Rishabh Agrawal, Jacob Fein-Ashley, and Paria Rashidinejad. Reinforcement Learning from Rich Feedback with Distributional DAgger. arXiv preprint arXiv:2606.05152, 2026. URL https://arxiv.org/abs/2606.05152.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to Basics: Revisiting REINFORCE-Style Optimization for Learning from Human Feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, 2024. doi: 10.18653/v1/2024.acl-long.662.

Xianwei Chen, Shimin Zhang, and Jibin Wu. f-OPD: Stabilizing Long-Horizon On-Policy Distillation with Freshness-Aware Control. arXiv preprint arXiv:2605.17862, 2026. URL https://arxiv.org/abs/2605.17862.

DeepSeek-AI. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence. arXiv preprint arXiv:2606.19348, 2026. URL https://arxiv.org/abs/2606.19348.

Naibin Gu, Qingyi Si, Chenxu Yang, Chuanyu Qin, Junhao Zhou, Peng Fu, Zheng Lin, and Weiping Wang. CompassOPD: Cross-Family On-Policy Distillation via Within-Family Likelihood Shifts. arXiv preprint arXiv:2609.10154, 2026. URL https://arxiv.org/abs/2609.10154.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge Distillation of Large Language Models. In *International Conference on Learning Representations*, pp. 32694–32717, 2024.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024. doi: 10.18653/v1/2024.acl-long.211.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Byeongho Heo, Jaehui Hwang, Sangdoo Yun, and Dongyoon Han. On-Policy Delta Distillation. arXiv preprint arXiv:2607.15161, 2026. URL https://arxiv.org/abs/2607.15161.

Wenjin Hou, Shangpin Peng, Weinong Wang, Zheng Ruan, Yue Zhang, Zhenglin Zhou, Mingqi Gao, Yifei Chen, Kaiqi Wang, Hongming Yang, Chengquan Zhang, Zhuotao Tian, Han Hu, Yi Yang, Fei Wu, and Hehe Fan. Uni-OPD: Unifying On-Policy Distillation with a Dual-Perspective Recipe. arXiv preprint arXiv:2605.03677, 2026. URL https://arxiv.org/abs/2605.03677.

Yuxuan Jiang and Francis Ferraro. Bridging Reasoning Trajectories in On-Policy Distillation via Near-Future Guidance. arXiv preprint arXiv:2606.00305, 2026. URL https://arxiv.org/abs/2606.00305.

Kimi Team. Kimi K3: Open Frontier Intelligence. arXiv preprint arXiv:2607.24653, 2026. URL https://arxiv.org/abs/2607.24653.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving Quantitative Reasoning Problems with Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3843–3857, 2022. doi: 10.52202/068431-0278.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking On-Policy Distillation of Large Language Models: Phenomenology, Mechanism, and Recipe. arXiv preprint arXiv:2604.13016, 2026a. URL https://arxiv.org/abs/2604.13016.

Yuying Li, Leqi Zheng, Yongzi Yu, Wenrui Zhou, Xuchang Zhong, Xing Hu, Jing Jin, Hangjie Yuan, and Tao Feng. Filter, Then Reweight: Rethinking Optimization Granularity in On-Policy Distillation. arXiv preprint arXiv:2606.02684, 2026b. URL https://arxiv.org/abs/2606.02684.

Xiaogeng Liu, Xinyan Wang, Yingzi Ma, Yechao Zhang, and Chaowei Xiao. When Are Teacher Tokens Reliable? Position-Weighted On-Policy Self-Distillation for Reasoning. arXiv preprint arXiv:2605.21606, 2026a. URL https://arxiv.org/abs/2605.21606.

Yanjiang Liu, Jie Lou, Xinyan Guan, Yuqiu Ji, Hongyu Lin, Ben He, Xianpei Han, Le Sun, Xing Yu, and Yaojie Lu. Your Teacher Can't Help You Here: Combating Supervision Fidelity Decay in On-Policy Distillation. arXiv preprint arXiv:2605.30833, 2026b. URL https://arxiv.org/abs/2605.30833.

Minjae Oh, Sangjun Song, Gyubin Choi, Yunho Choi, and Yohan Jo. KL for a KL: On-Policy Distillation with Control Variate Baseline. arXiv preprint arXiv:2605.07865, 2026. URL https://arxiv.org/abs/2605.07865.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635. PMLR, 2011.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. arXiv preprint arXiv:1707.06347, 2017. URL https://arxiv.org/abs/1707.06347.

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. URL https://arxiv.org/abs/2402.03300.

Rui Wang, Hongru Wang, Yi Chen, Boyang Xue, Tianqing Fang, Wenhao Yu, and Kam-Fai Wong. Demystifying On-Policy Distillation: Roles, Pathologies, and Regulations. arXiv preprint arXiv:2607.13399, 2026. URL https://arxiv.org/abs/2607.13399.

Yecheng Wu, Song Han, and Han Cai. Lightning OPD 2.0: Mitigating Style Bias in Cross-Teacher On-Policy Distillation for Large Reasoning Models. arXiv preprint arXiv:2607.28449, 2026. URL https://arxiv.org/abs/2607.28449.

Xiaomi LLM-Core Team. MiMo-V2-Flash Technical Report. arXiv preprint arXiv:2601.02780, 2026. URL https://arxiv.org/abs/2601.02780.

Yan Xie, Sijie Zhu, Tiansheng Wen, Bo Chen, and Yifei Wang. On the Position Bias of On-Policy Distillation. arXiv preprint arXiv:2606.22600, 2026. URL https://arxiv.org/abs/2606.22600.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond Teacher: Generalized On-Policy Distillation with Reward Extrapolation. arXiv preprint arXiv:2602.12125, 2026. URL https://arxiv.org/abs/2602.12125.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru 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, Yonghui Wu, and Mingxuan Wang. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 113222–113244, 2025. doi: 10.52202/085713-3775.

Zhu Zhang, Jixun Wang, Xiaoang Xu, Xiaorong Wang, Zihan Zhou, Zhiyuan Wang, Shuo Wang, Chaojun Xiao, and Yuezhi Zhou. Beyond Teacher Likelihood: Group-Calibrated On-Policy Distillation for Long-Context Reasoning. arXiv preprint arXiv:2608.19181, 2026. URL https://arxiv.org/abs/2608.19181.

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. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Jpxfof0EaS.

Ziheng Zhou, Jiaqi Li, Huacong Tang, Ying Nian Wu, and Demetri Terzopoulos. Less is More: Early Stopping Rollout for On-Policy Distillation. arXiv preprint arXiv:2605.27028, 2026. URL https://arxiv.org/abs/2605.27028.

A PROOFS

A.1 SEQUENCE-GRADIENT DECOMPOSITION

MiniLLM gives a detailed policy-gradient derivation of return-to-go for sequence reverse KL and then separates the single-step and long-term terms (Gu et al., 2024, Sec. 2.2 and Apps. A.2–A.3). Distributional DAgger independently formulates the corresponding local/future-credit decomposition for on-policy self-distillation (Agrawal et al., 2026, Sec. 3.2). We give the conditional-KL derivation of Eq. (2) in our notation.

Let $d_\theta(h) = D_{\mathrm{KL}}(p_\theta(\cdot \mid h) \| q(\cdot \mid h))$, and write $D_{p,t}(h)$ for the probability of visiting history h before position t. Then

$$\mathcal{K}(\theta) = \sum_t \sum_h D_{p,t}(h) d_\theta(h). \tag{7}$$

Differentiating the conditional mismatch contributes the local term $\mathbb{E} \sum_t \nabla_\theta d_\theta(h_t)$. For the state distribution,

$$\nabla D_{p,t}(h) = D_{p,t}(h) \sum_{u<t} \nabla \log p_\theta(y_u \mid h_u). \tag{8}$$

Substitution and exchanging the order of u and t yields

$$\sum_t \sum_h \nabla D_{p,t}(h) d_\theta(h) = \mathbb{E} \sum_u z_u \sum_{t>u} d_\theta(h_t), \tag{9}$$

which is the future state-visitation term.

At a fixed history,

$$\nabla d_\theta(h) = \nabla \sum_a p_\theta(a \mid h)\big(\log p_\theta(a \mid h) - \log q(a \mid h)\big) \tag{10}$$

$$= \mathbb{E}_{a\sim p_\theta}\left[\nabla \log p_\theta(a \mid h)\big(\log p_\theta(a \mid h) - \log q(a \mid h)\big)\right], \tag{11}$$

because the extra score term has expectation zero. Hence $-\nabla d_\theta(h) = \mathbb{E}[z_t r_t \mid h_t = h]$. For $u > t$, z_t is measurable at h_u, giving

$$\mathbb{E}[z_t r_u] = \mathbb{E}[z_t \mathbb{E}[r_u \mid h_u]] = -\mathbb{E}[z_t d_\theta(h_u)]. \tag{12}$$

These identities prove Eq. (2). They also show that the return-to-go estimator $\widehat{G}_{\mathrm{RTG}} = \sum_t z_t R_t$, with $R_t = \sum_{u=t}^{L} r_u$, satisfies $\mathbb{E}[\widehat{G}_{\mathrm{RTG}}] = -\nabla_\theta \mathcal{K}$.

A.2 Alternative return-to-go proof of Proposition 1

Let $Z = \sum_t z_t$ and $P_t = \sum_{u<t} r_u$. Expanding the broadcast estimator gives

$$RZ = \sum_t z_t(P_t + R_t) = \widehat{G}_{\mathrm{RTG}} + \sum_t z_t P_t. \tag{13}$$

The past-feedback sum P_t is determined by h_t, and $\mathbb{E}[z_t \mid h_t] = 0$. Hence $\mathbb{E}[z_t P_t] = 0$ and $\mathbb{E}[RZ] = \mathbb{E}[\widehat{G}_{\mathrm{RTG}}] = -\nabla_\theta \mathcal{K}$, as in Eq. (3). The direct sequence-likelihood proof appears in Section 3.

A.3 Tail influence under exponential discounting

We quantify the influence of later feedback on the trajectory score in Eq. (4). Fix a response length L, a prefix length $1 \le H \le L$, and $0 < \gamma < 1$. Write $W_L = \sum_{t=1}^{L} \gamma^{t-1}$ and $a_{t,L} = \gamma^{t-1}/W_L$ for the normalized position weights. The total weight after H is

$$\alpha_{\gamma,H,L} = \sum_{t=H+1}^{L} a_{t,L} = \frac{\gamma^H - \gamma^L}{1-\gamma^L} \le \min\left\{\gamma^H, \frac{L-H}{L}\right\}. \tag{14}$$

Derivation of the tail weight. Summing the geometric series gives the equality. The first bound follows from

$$\alpha_{\gamma,H,L} = \gamma^H \frac{1-\gamma^{L-H}}{1-\gamma^L} \le \gamma^H.$$

For the second bound, the weights γ^{t-1} are nonincreasing, so their average over the last $L-H$ positions is no larger than their average over all L positions. Dividing the tail sum by W_L therefore gives $\alpha_{\gamma,H,L} \le (L-H)/L$. The case $H = L$ has zero tail weight.

Sensitivity to changes in later feedback. Suppose two feedback sequences have the same length and agree through position H, with $|r'_t - r_t| \le \eta$ for $t > H$. Their trajectory scores satisfy

$$|S_L^{\gamma\prime} - S_L^{\gamma}| = \left|\sum_{t=H+1}^{L} a_{t,L}(r'_t - r_t)\right| \le \eta \alpha_{\gamma,H,L}. \tag{15}$$

This follows directly from the triangle inequality and nonnegative weights. It requires no statistical noise model for the feedback.

The normalization also changes when a response is extended. If $|r_t| \le B$, compare the score S_H^γ of a fixed prefix with the score S_L^γ after appending a tail. For $H < L$, let $S_{H+1:L}^\gamma$ be the normalized weighted mean over that tail. Then

$$S_L^\gamma = (1-\alpha_{\gamma,H,L})S_H^\gamma + \alpha_{\gamma,H,L} S_{H+1:L}^\gamma, \tag{16}$$

$$|S_L^\gamma - S_H^\gamma| \le 2B\alpha_{\gamma,H,L} \le 2B\gamma^H. \tag{17}$$

Thus the influence bound includes the change in the score denominator. At $\gamma = 1$, the tail weight is $(L-H)/L$. These bounds quantify score sensitivity, not whether later feedback is useful. They do not establish an optimal discount or a guarantee of stable policy updates.

A.4 WEIGHTED OBJECTIVES

To relate position weighting to the teacher-matching objective, consider fixed nonnegative weights $w_1, \ldots, w_T$ that do not depend on the sampled response or its realized length. As in Section 3, the teacher is fixed and the student is sampled on-policy. We assume a common sequence space, teacher support wherever the student has positive probability, and interchangeability of differentiation and expectation. Terminated responses are padded to a fixed maximum length T by a shared absorbing state with zero subsequent feedback. With $d_\theta(h_t) = D_{\mathrm{KL}}(p_\theta(\cdot \mid h_t) \| q(\cdot \mid h_t))$, define

$$R_w = \sum_{t=1}^{T} w_t r_t, \qquad \mathcal{K}_w = -\mathbb{E}_p[R_w] = \sum_{t=1}^{T} w_t \mathbb{E}_p[d_\theta(h_t)]. \tag{18}$$

The last equality follows from $\mathbb{E}[r_t \mid h_t] = -d_\theta(h_t)$.

Decomposition into prefix objectives. Let $K_n = D_{\mathrm{KL}}(p_{1:n} \| q_{1:n})$, where $p_{1:n}$ and $q_{1:n}$ are the distributions of the first n positions, and set $K_0 = 0$. The KL chain rule gives

$$K_n = \sum_{t=1}^{n} \mathbb{E}_p[d_\theta(h_t)], \qquad \mathbb{E}_p[d_\theta(h_t)] = K_t - K_{t-1}.$$

Substitution and rearrangement of the finite sum, with $w_{T+1} = 0$, yield

$$\begin{aligned} \mathcal{K}_w &= \sum_{t=1}^{T} w_t (K_t - K_{t-1}) \\ &= \sum_{n=1}^{T} (w_n - w_{n+1}) K_n. \end{aligned} \tag{19}$$

For nonincreasing w, all coefficients are nonnegative. A fixed prefix window $w_t = \mathbf{1}\{t \le H\}/H$ gives $\mathcal{K}_w = K_H/H$. Exponential weights $w_t = \gamma^{t-1}/W_T$ give coefficients $(1-\gamma)\gamma^{n-1}/W_T$ for $n < T$ and γ^{T-1}/W_T for $n = T$. Thus these weighting schemes correspond to nonnegative combinations of prefix KL objectives.

Broadcast gradient for fixed weights. Write $Z = \sum_{t=1}^{T} z_t$. Since the weights are fixed and $\nabla_\theta R_w = -\sum_t w_t z_t$,

$$-\nabla_\theta \mathcal{K}_w = \mathbb{E}_p[R_w Z] + \mathbb{E}_p[\nabla_\theta R_w] = \mathbb{E}_p[R_w Z], \tag{20}$$

where the second term vanishes because $\mathbb{E}_p[z_t] = 0$ at every position. This extends the raw broadcast identity to a fixed weighted objective. The practical score instead divides by the weight mass of each realized response length, and its update also uses group statistics and clipping. Those operations are outside this fixed-weight result, as discussed in Appendix B.

A.5 ALGEBRAIC INTERPRETATION OF COLLECTIVE TEACHER PRESSURE

The scalar estimator in Eq. (5) has an equivalent group-level interpretation. Let $\mathbf{S}_g^\gamma = (S_1^\gamma, \ldots, S_G^\gamma)^\top$ and let $\mathbf{1}$ be the all-ones vector. Define

$$\mathbf{C}_g^\gamma = \mathbf{S}_g^\gamma - \bar{S}_g^\gamma \mathbf{1}, \qquad \mathbf{P}_g^\gamma = \bar{S}_g^\gamma \mathbf{1}, \qquad \mathbf{S}_g^\gamma = \mathbf{C}_g^\gamma + \mathbf{P}_g^\gamma.$$

The relative response preference component $\mathbf{C}_g^\gamma$ records deviations from the group mean, while $\mathbf{P}_g^\gamma$ records collective teacher pressure. Both are constructed from the same teacher–student feedback. They are orthogonal under the Euclidean inner product because

$$\langle \mathbf{C}_g^\gamma, \mathbf{P}_g^\gamma \rangle = \bar{S}_g^\gamma \sum_{i=1}^{G} (S_i^\gamma - \bar{S}_g^\gamma) = 0.$$

For $\epsilon > 0$, the advantage vector is therefore

$$\mathbf{A}_g^{\gamma,\lambda} = \frac{\mathbf{S}_g^\gamma - (1-\lambda)\bar{S}_g^\gamma \mathbf{1}}{s_g^\gamma + \epsilon} = \frac{\mathbf{C}_g^\gamma + \lambda \mathbf{P}_g^\gamma}{s_g^\gamma + \epsilon}.$$

Its group mean and pairwise differences satisfy

$$\frac{1}{G}\sum_{i=1}^{G} A_i^{\gamma,\lambda} = \frac{\lambda \bar{S}_g^{\gamma}}{s_g^{\gamma}+\epsilon}, \qquad A_i^{\gamma,\lambda} - A_j^{\gamma,\lambda} = \frac{S_i^{\gamma} - S_j^{\gamma}}{s_g^{\gamma}+\epsilon}.$$

Thus, for a fixed group of scores and its standard deviation, λ changes only the shared offset and preserves all pairwise advantage differences and the response ordering. At $\lambda = 0$, the shared offset vanishes and only the relative response preference component remains; this component still depends on teacher feedback. The shared offset is signed and may raise or lower the advantages together. These algebraic properties do not imply invariant PPO gradients or subsequent training trajectories. The decomposition interprets the scalar estimator without changing its computation or identifying separate semantic sources of guidance.

B Practical Estimator Boundaries

The expectation results above apply to raw returns and a single score-function update. The practical estimator divides the weighted trajectory score by its realized weight mass, scales advantages by a within-prompt group standard deviation, and uses clipping and repeated PPO updates. These data-dependent operations change the gradient estimator; the raw-return unbiasedness result does not transfer to the practical loss. The proofs motivate sequence-level credit assignment without asserting that the implemented update is the exact reverse-KL gradient.

C Experimental Details

The five teacher–student pairings appear in Table 1. For Instruct models, think or nothink mode is part of the pairing and uses the corresponding chat template. Base models use their default reasoning protocol, which is distinct from the Instruct think/nothink mode switch. Both Instruct students use nothink mode; only the Instruct-teacher-to-Base-student pairings use a thinking Instruct teacher. Teacher and student log probabilities are evaluated on the same conditional history.

Parameter convention. Throughout the paper, λ is the fraction of collective teacher pressure retained. Archived run labels and training configurations use the complementary group-mean subtraction coefficient $c = 1 - \lambda$. All reported parameter values and figure legends use the retention convention; the underlying runs and measured results are unchanged.

C.1 Training data

We use the English configuration of open-r1/DAPO-Math-17k-Processed, pinned to revision 31dd3095, with 14,116 questions in its canonical training split. Each example retains its problem statement and reference answer and is converted into a system instruction followed by the user question. The instruction asks the model to reason step by step and place its final answer in a boxed expression. We apply no additional difficulty or correctness filtering. After chat-template rendering and tokenization, the training loader retains prompts with more than five and at most 4,096 tokens.

The fixed-response diagnostic draws questions from the same pool. Its candidate questions are deduplicated by normalized prompt text, and exact normalized matches to evaluation questions are excluded. A deterministic hash ordering with seed 20260910 selects 1,024 questions, whose alternating entries form two equal subsets. Figure 1(b) uses the 512-question confirmation subset.

C.2 Training configuration

Table 3 summarizes the GR-OPD training configuration. Each rollout batch contains 64 prompts with eight responses per prompt, yielding 512 student responses. Training uses the ROLL framework with Megatron for model optimization and teacher scoring, and vLLM for student generation.

The 512 responses are divided into eight global optimization minibatches of 64 responses each. The student uses eight data-parallel replicas, a configured per-device batch size of one, and eight

Table 3: GR-OPD training configuration. Sampling settings refer to student rollouts during training.

Setting	Value
Prompts per rollout batch	64
Responses per prompt, G	8
Configured per-device training batch size	1
Gradient-accumulation steps	8
Student data-parallel replicas	8
Responses per global optimization minibatch	64
Optimizer updates per rollout step	8
Training rollout steps	221
Maximum prompt length	4,096 tokens
Maximum response length	28,672 tokens
Optimizer	Adam (Megatron implementation)
Learning rate	5×10^{-7}
Weight decay	0.1
Learning-rate schedule	Linear warm-up followed by linear decay to zero
Warm-up	20 optimizer steps
PPO epochs per rollout batch	1
PPO probability-ratio clipping, δ	0.2
Sampling temperature / top-p	1.0 / 1.0
Top-k sampling	Disabled
Random seed	42
Model precision	bfloat16
Reference hardware configuration	8 NVIDIA H20 GPUs

gradient-accumulation steps. With one PPO epoch per rollout batch, each rollout step therefore produces eight optimizer updates.

For GR-OPD, token-level teacher feedback is clipped to $[-5, 5]$ before trajectory aggregation. After group calibration, the resulting teacher advantage is also clipped to $[-5, 5]$ before the policy update. These two value-clipping operations are separate from the PPO probability-ratio clipping in Eq. (6).

Terminal-token alignment. Qwen3 Base and Instruct models use different end-of-sequence (EOS) conventions. For the Base-student/Instruct-teacher pairings, we map the student's EOS target (token ID 151643) to the teacher's EOS target (151645) when computing teacher log probabilities. This aligns the probability of termination under the two models. The mapping applies only to active response positions in the teacher-scoring copy; the sampled student response is preserved. Student generation also accepts token ID 151645 as an additional stopping token.

Model and tokenizer versions. We use models from the original Qwen3 release (April 2025). Table 4 maps the model names used in this paper to their official checkpoints. Each link pins a public revision containing the model weights and tokenizer files; we use each model's associated tokenizer. The models denoted Instruct correspond to the repositories without the `-Base` suffix.

Table 4: Official checkpoints for the models and tokenizers. Repository links point to fixed public revisions on Hugging Face.

Model used in this paper	Official checkpoint	Public revision
Qwen3-1.7B-Base	`Qwen/Qwen3-1.7B-Base`	`ea980cb`
Qwen3-4B-Base	`Qwen/Qwen3-4B-Base`	`906bfd4`
Qwen3-1.7B-Instruct	`Qwen/Qwen3-1.7B`	`70d244c`
Qwen3-4B-Instruct	`Qwen/Qwen3-4B`	`1cfa9a7`
Qwen3-32B-Instruct	`Qwen/Qwen3-32B`	`9216db5`

Software versions. Our implementation is based on a modified ROLL 0.3.0 checkout with `mcore_adapter` 0.10.0.dev0. The recorded training environment uses PyTorch 2.8.0, Transformers 4.57.1, Megatron-Core 0.16.1, and vLLM `0.11.1rc2.dev0+gc3a722fcb.d20251021`.

C.3 EVALUATION PROTOCOL

Table 5 lists the six evaluation suites and the number of sampled responses per question. For the main comparison, generation uses temperature 0.6, top-$p = 0.9$, no top-k restriction, and a maximum response length of 28,672 tokens. Avg@K is the fraction of correct responses among the K samples for each question, averaged over questions within a suite. Macro Avg@K is the unweighted mean of the six suite scores.

Table 5: Evaluation question counts and samples per question. The six suites contain 1,042 questions in total.

Benchmark	Questions	K
AIME 2024	30	32
AIME 2025	30	32
AMC 2023	40	32
MATH-500	500	8
Minerva Math	272	8
OlympiadBench (math subset)	170	8

Correctness is determined by the ROLL mathematical-answer verifier using Math-Verify. After validating the thinking-tag structure, the verifier examines the text after the closing thinking tag, falling back to the final paragraph when no closing tag is present. It first extracts the last boxed expression; if no complete boxed answer is available, it applies the mathematical-expression parser to that answer text. The extracted result is checked for mathematical equivalence to the reference answer. Failed extraction, parsing errors, and verification timeouts receive a score of zero. Responses that reach the generation limit remain in the evaluation and are judged from the generated text using the same procedure.

C.4 TEACHER AND STUDENT BASE CAPABILITIES

Table 6 provides pre-OPD model-level reference scores. Its rows have different generation caps and are not used as matched method comparisons. Mean response length is computed by re-tokenizing the saved response text with the corresponding tokenizer, without adding special tokens; prompt tokens are excluded.

Table 6: Pre-OPD reference capability. Macro Avg@K and mean response length are averaged equally over the six suites. Mean response length is measured in tokens.

Model	Mode	Role	Macro Avg@K	Mean tokens
Qwen3-1.7B-Base	default	Student	23.099	4,166.4
Qwen3-4B-Base	default	Teacher	31.244	3,245.5
Qwen3-1.7B-Instruct	nothink	Student	32.995	1,662.1
Qwen3-4B-Instruct	nothink	Teacher	42.970	1,937.5
Qwen3-4B-Instruct	think	Teacher	74.574	10,138.4
Qwen3-32B-Instruct	think	Teacher	78.954	9,221.9
Qwen3-32B-Instruct	nothink	Teacher	44.163	2,152.6

C.5 FIXED-RESPONSE HORIZON DIAGNOSTIC

Figure 1(b) measures how teacher scores from different prefix lengths distinguish correct from incorrect responses on a fixed rollout bank. For each of 512 prompts, we score 32 responses over their first $\min(H, L_i)$ tokens, averaging teacher–student log-ratios clipped to $[-5, 5]$. Responses are organized into groups of eight. Let S^H_{gi} be the score of response i in group g, let $C_{gi} \in \{0, 1\}$ indicate correctness of its complete answer, and let $\mathcal{G}_x$ be the complete groups for prompt x. The plotted statistic is

$$J_H(H) = \frac{1}{512} \sum_{x=1}^{512} \frac{1}{|\mathcal{G}_x|} \sum_{g \in \mathcal{G}_x} \frac{\mathrm{Cov}_{i \in g}(S^H_{gi}, C_{gi})}{\mathrm{Std}_{i \in g}(S^H_{gi})}. \tag{21}$$

Covariance and standard deviation use their population definitions. A group contributes zero if its score standard deviation is at most 10^{-6} or if all its correctness labels agree; such groups remain in the average. We first average groups within each prompt and then average prompts equally. Positive

values mean that correct responses receive higher standardized teacher scores on average. This fixed-response test compares scoring horizons on the same responses; it does not compare policies learned by training with different horizons.

C.6 Training diagnostics for feedback horizon

Figure 4 compares $\gamma = 1$ and $\gamma = .99$ on $\mathrm{I}_{32}^{\mathrm{nothink}} \to \mathrm{I}_{1.7}^{\mathrm{nothink}}$. Both runs use full group centering ($\lambda = 0$). At each rollout step, let L_i be the number of active response tokens, v_i the task-verifier reward, and S_i^γ the trajectory score before group calibration. The score aggregates token-level teacher feedback clipped to $[-5, 5]$ using Eq. (4). Panels (A) and (B) report the batch means of L_i and v_i. Panels (C) and (D) compute

$$r_{SL} = \mathrm{Corr}_{i \in \mathcal{B}}(S_i^\gamma, L_i), \qquad \rho_{Sv} = \mathrm{Corr}_{i \in \mathcal{B}}\big(\mathrm{rank}(S_i^\gamma), \mathrm{rank}(v_i)\big), \tag{22}$$

where Corr is Pearson correlation and $\mathcal{B}$ contains the responses in the rollout batch. Spearman correlation uses average ranks for ties; the implementation returns zero when the correlation denominator is at or below the numerical tolerance. These correlations pool responses across prompts, without within-prompt centering, and differ from the grouped correctness statistic $J_H(H)$. For visualization, we use centered moving averages over seven rollout steps for (A), (C), and (D), and eleven for (B), using available observations at the endpoints.

Under $\gamma = 1$, response length grows and the trajectory score becomes increasingly associated with length, while its rank association with verifier reward turns negative. With $\gamma = .99$, length grows more slowly, score–length correlation remains weaker, and score–verifier correlation remains positive. This contrast suggests that a trajectory score can increasingly favor length without continuing to favor correct answers. The correlations describe the evolving batches and do not establish that length causes the change or isolate the contribution of particular continuations.

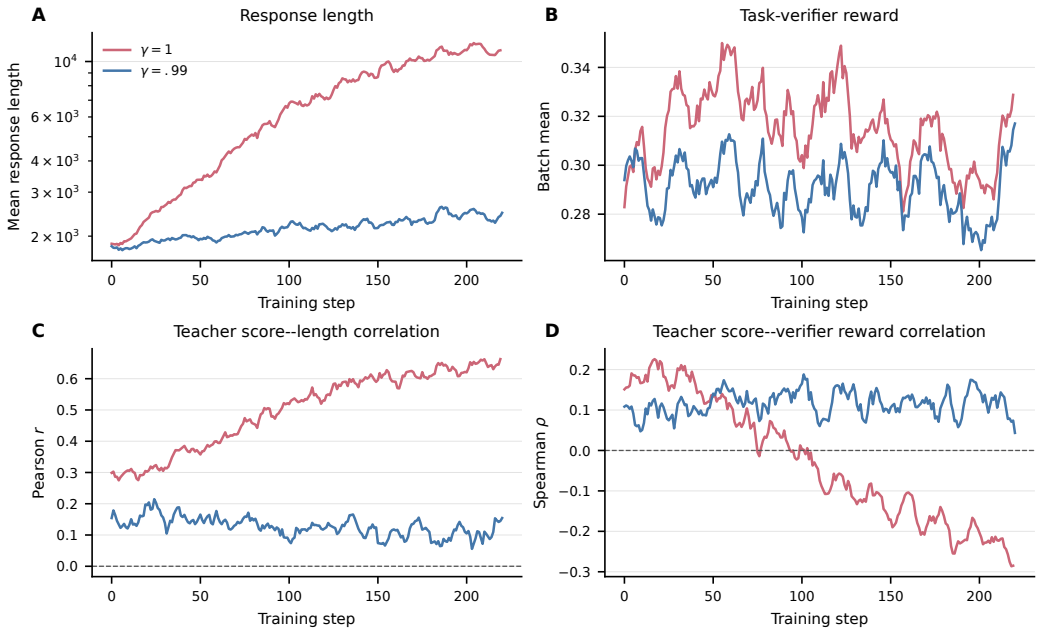

Figure 4: Training dynamics for $\gamma = 1$ and $\gamma = .99$ on $\mathrm{I}_{32}^{\mathrm{nothink}} \to \mathrm{I}_{1.7}^{\mathrm{nothink}}$. (A) Mean response length in tokens (logarithmic axis); (B) mean task-verifier reward; (C) Pearson correlation between trajectory score and response length; and (D) Spearman correlation between trajectory score and verifier reward. Correlations pool responses within each rollout batch. Curves use centered moving averages as specified above.

C.7 PRESSURE MAGNITUDE DIAGNOSTICS

Table 2 reports batch-level magnitude proxies at $\gamma = .99$ for completed 221-step runs. For each training batch, we divide the absolute batch mean trajectory score by the batch mean within-prompt group standard deviation plus 10^{-8}. The normalized proxy is the median of this ratio over training steps; the retained proxy multiplies it by λ. This ratio of batch summaries is not an average of per-prompt normalized pressure magnitudes and does not measure the signed shared component for an individual prompt.

Pressure magnitude varies across pairings. The normalized proxy ranges from 3.16 to 3.61 in the two matched Instruct pairings, compared with 6.08 and 7.23 in the two cross-pattern pairings. Thus, a larger shared score offset relative to within-prompt variation accompanies pattern mismatch in these runs. Within each matched pairing, the normalized proxy varies little across the reported retention settings: 3.16–3.21 for $\mathrm{I}_4^{\mathrm{nothink}} \to \mathrm{I}_{1.7}^{\mathrm{nothink}}$ and 3.56–3.61 for $\mathrm{I}_{32}^{\mathrm{nothink}} \to \mathrm{I}_{1.7}^{\mathrm{nothink}}$. Most of the difference in the retained proxy therefore comes from the retention fraction. The same fraction can nevertheless retain different magnitudes across pairings: at $\lambda = .25$, the proxies are 0.79 and 0.90, respectively.

Magnitude must be interpreted with training behavior. For $\mathrm{I}_4^{\mathrm{nothink}} \to \mathrm{I}_{1.7}^{\mathrm{nothink}}$, $\lambda = .5$ retains a proxy magnitude of 1.60 and eventually leads to reward decline, whereas $\lambda = .25$ and $.2$ retain 0.79 and 0.63 and support sustained improvement (Figure 3). The completed cross-pattern runs retain no shared component at $\lambda = 0$, while their relative response preference term still uses teacher feedback. Together, these observations support calibrating collective teacher pressure to the pairing. They do not establish a universal magnitude threshold or show that a large shared component is intrinsically harmful: the proxy summarizes the magnitude of collective teacher pressure, while its teaching value also depends on compatibility with the student's current behavior.