OUTCOME-GUIDED ON-POLICY SELF-DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation (OPSD) provides denser token-level supervision and better computational efficiency than Reinforcement Learning with Verifiable Rewards (RLVR). However, this denser supervision may introduce substantial noise and training instability. Existing improvements often rely on high-variance per-token statistics and introduce extra hyperparameters and trade-offs. Based on the advantage formulation in RLVR, we analyze the OPSD objective from the same perspective, incorporating outcome correctness signals. We find that vanilla OPSD imposes insufficient penalties and excessive rewards on incorrect trajectories because it applies a fixed divergence objective regardless of outcome correctness. Furthermore, the reliability of teacher supervision is associated with both trajectory outcome and the cumulative average teacher entropy along the rollout. Based on these observations, we propose Outcome-Guided On-Policy Self-Distillation (OG-OPSD), which dynamically adapts both the divergence objective and distillation position according to binary outcome rewards and the cumulative average teacher entropy. Extensive experiments show that OG-OPSD consistently improves the performance of vanilla OPSD and multiple strong baselines in mathematical reasoning, multimodal reasoning, and out-of-distribution tasks across Qwen3 models at 1.7B, 4B, and 8B scales, as well as Qwen3-VL-2B.

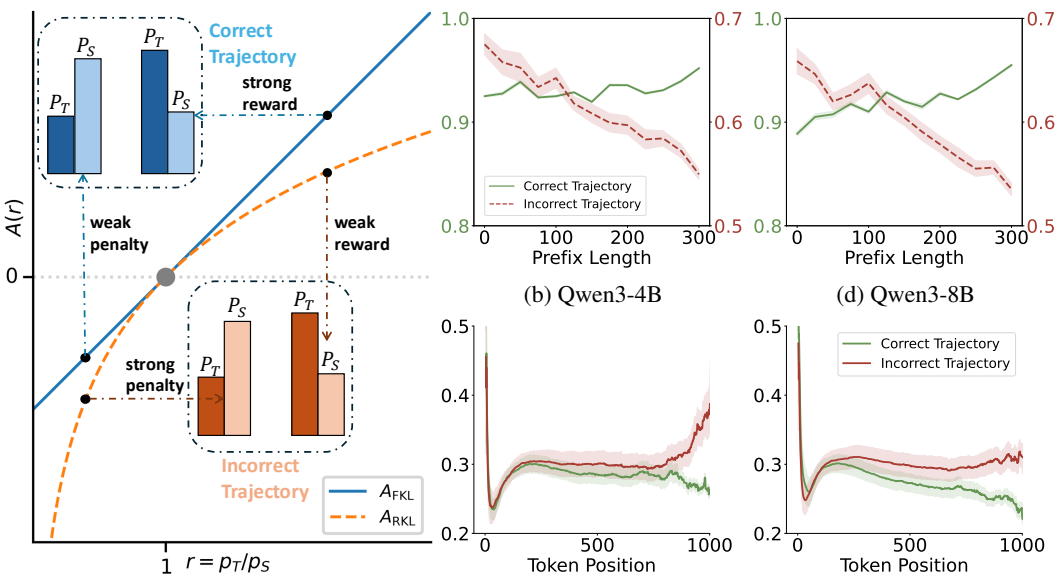
1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) and on-policy distillation (OPD) have emerged as effective approaches to post-training large language models (LLMs). RLVR improves multi-step reasoning by optimizing automatically verifiable outcomes, avoiding costly human preference annotations (Guo et al., 2025; Shao et al., 2024; Jaech et al., 2024). However, its commonly used binary outcome reward provides only a single signal for an entire trajectory, leaving token-level credit assignment unresolved. For example, outcome-supervised GRPO assigns the same group-normalized advantage to every token within a response (Shao et al., 2024). OPD provides denser supervision by matching the teacher's next-token distributions along the student's own rollouts (Gu et al., 2024; Agarwal et al., 2024; Lu & Lab, 2025), but conventional OPD typically relies on a stronger external teacher.

On-policy self-distillation (OPSD) removes the need for a separate teacher by using the same model under different conditioning contexts (Zhao et al., 2026; Shenfeld et al., 2026; Hübotter et al., 2026). The student observes only the question, whereas the teacher additionally receives privileged information, such as a reference solution. This asymmetry enables dense self-supervision, but it does not guarantee that the resulting guidance is useful at every position. Recent studies report optimization instability and degraded reasoning performance in OPSD (Zhu et al., 2026; Kim et al., 2026), highlighting the need to determine how and where teacher supervision should be applied.

The distillation objective directly shapes this supervision. Forward and reverse Kullback-Leibler (KL) divergences weight teacher-student discrepancies differently, inducing distinct token-level learning signals. Vanilla OPSD applies a fixed divergence across all trajectories and positions. Recent work has explored adaptive objectives (Jin et al., 2026; Jung et al., 2025; Jia et al., 2026; Wu et al., 2025), as well as token reweighting and selection (Liu et al., 2026; Wang et al., 2026; Zhang et al., 2026; Ziheng et al., 2026). Token probabilities and entropy characterize local model behavior, whereas outcome correctness provides a direct signal of task success. Although this outcome signal does not identify the contribution of each token, it can inform how dense supervision is used. Our

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.



(a) Outcome-guided divergence selection (b) Qwen3-4B (c) Qwen3-4B (d) Qwen3-8B (e) Qwen3-8B

Figure 1: **(a)** Implicit token-level advantages under forward and reverse KL. P_T and P_S denote the probabilities assigned to a candidate token by the teacher and student, respectively; $A(\cdot)$ denotes the corresponding advantage function. **(b,d)** Teacher completion accuracy (TCA) after resuming generation from student-generated reasoning prefixes of increasing length. Results are averaged over OpenThoughts-Math30K, DeepScaleR, MATH-500, and DeepMath datasets. **(c,e)** Cumulative average teacher entropy along correct and incorrect student trajectories.

central idea is to condition both the choice of divergence and the position of distillation under the guidance of outcome correctness.

We first analyze forward KL (FKL) and reverse KL (RKL) through their implicit token-level advantages (see their conceptual comparisons in Figure 1a). This formulation reveals an asymmetry in their weighting: FKL assigns stronger rewards to tokens underweighted by the student relative to the teacher and bounds penalties, whereas RKL imposes stronger penalties on tokens overestimated by the student. These properties motivate FKL on correct trajectories to learn teacher-preferred continuations while limiting negative feedback on successful behavior, and RKL on incorrect trajectories to reduce student overestimation relative to the teacher.

However, divergence selection alone cannot determine whether the teacher maintains reliability during deployment. We investigate this question by truncating student trajectories at varying lengths and asking the teacher to complete each prefix. Across model scales and mathematical reasoning benchmarks, teacher completion accuracy declines as prefixes from incorrect trajectories grow longer, while remaining comparatively stable on correct trajectories (see empirical results in Figures 1b and 1d). This pattern of prefix failure motivates restricting distillation on incorrect trajectories to shorter prefixes. We further examine the cumulative average teacher entropy, whose curves for correct and incorrect trajectories provide an empirical basis for estimating the truncation cutoff (see empirical results in Figures 1c and 1e). We observe the same trend on Qwen3-1.7B (see empirical results in Figures 4a and 4b).

Building on these findings, we propose *Outcome-Guided On-Policy Self-Distillation* (OG-OPSD), which combines outcome-guided divergence selection with selective prefix truncation. OG-OPSD applies FKL throughout correct trajectories and restricts RKL to prefixes of incorrect trajectories. The cutoff is estimated from the last crossing of the cumulative average teacher entropy curves for the two outcome groups. Trajectory-level correctness thus determines how and where to distill, while the teacher supplies dense token-level supervision within the selected prefix.

Our contributions are summarized as follows:

- We connect the asymmetric token-level weighting of FKL and RKL to trajectory correctness, and identify an outcome-dependent decline in teacher continuation accuracy as incorrect prefixes increase in length.
- We propose a simple but effective method called OG-OPSD, which combines outcome-guided KL divergence selection with an entropy-based prefix cutoff to control how and where self-distillation is applied.
- Experiments show consistent improvements over vanilla OPSD and strong baselines with Qwen3 models at different scales. Evaluations on Qwen3-VL and out-of-distribution datasets further support the effectiveness of the proposed method.

2 Related Work

On-policy distillation. OPD combines student-generated rollouts with dense teacher supervision and has been adopted in recent LLM post-training pipelines (Xu et al., 2026; Zeng et al., 2026; Team et al., 2026; Yang et al., 2025). OPSD removes the need for an external teacher by conditioning the same model on privileged information (Zhao et al., 2026; Shenfeld et al., 2026; Hübotter et al., 2026). Recent works improve the allocation of dense supervision through token weighting, selection, and rollout truncation. Liu et al. (2026) analyze how teacher reliability varies with token position and introduce position-dependent loss weights. Wang et al. (2026) select token positions according to teachability, distinguishing learnable teacher-student disagreement from incompatible predictions. Zhang et al. (2026) use student entropy to determine the direction of supervision, encouraging agreement with the teacher at low-entropy positions and divergence at high-entropy positions to preserve exploration. Other studies address declining teacher reliability along student-generated prefixes. ESR restricts rollout length (Ziheng et al., 2026); Prune-OPD monitors teacher-student compatibility and accumulates drift signals to downweight unreliable supervision and truncate rollouts (Yang et al., 2026); and IW-OPD weights token losses using cumulative prefix discrepancies (Xie et al., 2026). OG-OPSD makes the distillation span conditional on trajectory outcome: it retains full supervision on correct trajectories and uses cumulative average teacher entropy to estimate a prefix cutoff for incorrect trajectories.

Divergence selection. The choice of divergence determines how teacher-student discrepancies translate into learning signals. MiniLLM minimizes sequence-level RKL (Gu et al., 2024), GKD supports alternative divergence objectives (Agarwal et al., 2024), and DistiLLM introduces skewed KL divergences (Ko et al., 2024). Subsequent work adapts the objective to local distributional properties. EOPD augments RKL with FKL at positions with high teacher entropy (Jin et al., 2026). ToDi combines FKL and RKL using token-wise teacher-student log-probability ratios (Jung et al., 2025), while AOPD preserves positive-advantage reinforcement and applies FKL guidance in non-positive-advantage regions (Jia et al., 2026). AKL adjusts the relative weights of FKL and RKL according to discrepancies in the head and tail of the next-token probability distribution (Wu et al., 2025). Guided by an analysis of the implicit token-level advantages of FKL and RKL, OG-OPSD adopts a trajectory-level selection rule: FKL for correct rollouts and RKL for incorrect ones. This choice depends only on the binary verifier outcome and is coupled with selective prefix truncation to account for changes in teacher reliability.

3 Method

OG-OPSD uses the binary outcome of the verifier to select the divergence and determine which token positions receive supervision. It applies forward KL to complete correct trajectories, and reverse KL to prefixes of incorrect trajectories, with the prefix length estimated from teacher entropy.

3.1 Preliminaries

OPSD uses the same model π_θ as both the teacher and the student, with different conditioning contexts (Zhao et al., 2026). Given a question x, the student generates a response $y = (y_1, \ldots, y_N) \sim \pi_\theta(\cdot \mid x)$. The teacher also receives privileged information $y^\star$. At position n, let $s_n = (x, y_{<n})$ denote the question and the student-generated prefix. The student and the teacher define next-token

distributions in the vocabulary $\mathcal{V}$:

$$p_S^n(\cdot) = \pi_\theta(\cdot \mid s_n), \qquad p_T^n(\cdot) = \mathrm{sg}[\pi_\theta(\cdot \mid s_n, y^\star)], \tag{1}$$

where sg denotes the stop-gradient. During optimization, sampled prefixes are treated as fixed, and gradients flow only through the student branch.

Vanilla OPSD averages local distillation losses over the entire response:

$$\mathcal{L}_{\mathrm{OPSD}}(\theta; y) = \frac{1}{N} \sum_{n=1}^{N} d_n, \tag{2}$$

using the same divergence at every position. We consider two forms of KL divergence:

$$d_n^{\mathrm{FKL}} = D_{\mathrm{KL}}(p_T^n \| p_S^n) = \sum_{v \in \mathcal{V}} p_T^n(v) \log \frac{p_T^n(v)}{p_S^n(v)}, \tag{3}$$

$$d_n^{\mathrm{RKL}} = D_{\mathrm{KL}}(p_S^n \| p_T^n) = \sum_{v \in \mathcal{V}} p_S^n(v) \log \frac{p_S^n(v)}{p_T^n(v)}. \tag{4}$$

OPSD provides supervision signals through the distribution differences of the teacher-student model, while RLVR provides supervision through a deterministic outcome reward $R(y)$ that can be automatically verified against ground truth. For a fixed question x, the expected-reward objective is $\mathcal{J}_{\mathrm{RLVR}}(\theta) = \mathbb{E}_{y \sim \pi_\theta(\cdot|x)}[R(y)]$, with REINFORCE gradient

$$\nabla_\theta \mathcal{J}_{\mathrm{RLVR}}(\theta) = \mathbb{E}_{y \sim \pi_\theta(\cdot|x)} \left[R(y) \sum_{n=1}^{N} \nabla_\theta \log p_S^n(y_n) \right]. \tag{5}$$

Here, the same trajectory-level reward weights every sampled token. Distillation provides finer-grained supervision through teacher-student discrepancies.

3.2 Outcome-Guided Divergence Selection

To compare how FKL and RKL allocate supervision, we express their local gradients in the score-function form of Eq. 5. For a fixed prefix, define $r_n(v) = p_T^n(v)/p_S^n(v)$. With gradients stopped through the teacher, the descent directions are

$$-\nabla_\theta d_n^{\mathrm{FKL}} = \sum_{v \in \mathcal{V}} p_S^n(v) \big(r_n(v) - 1\big) \nabla_\theta \log p_S^n(v), \tag{6}$$

$$-\nabla_\theta d_n^{\mathrm{RKL}} = \sum_{v \in \mathcal{V}} p_S^n(v) \log r_n(v) \nabla_\theta \log p_S^n(v). \tag{7}$$

The unit baseline in Eq. 6 leaves the gradient unchanged because the score function has zero expectation under the student distribution. Detailed derivations appear in Appendix A. We refer to

$$A_n^{\mathrm{FKL}}(v) = r_n(v) - 1, \qquad A_n^{\mathrm{RKL}}(v) = \log r_n(v) \tag{8}$$

as the implicit token-level advantages.

For any candidate token v, the implicit advantages $A_n^{\mathrm{FKL}}(v)$ and $A_n^{\mathrm{RKL}}(v)$ have the same sign: both are positive when the teacher assigns greater probability to the token than the student, negative when it assigns less, and zero when the probabilities coincide. In this representation, the two divergences never disagree on the direction of the update for a given token and differ in the magnitude of the coefficients assigned to each teacher-student discrepancy.

The FKL advantage $r_n(v) - 1$ is bounded below by -1, limiting the magnitude of negative coefficients for tokens that the student overweights relative to the teacher ($r_n(v) < 1$). For underweighted tokens ($r_n(v) > 1$), its positive advantage grows linearly without bound. In contrast, the RKL advantage $\log r_n(v)$ is unbounded in both directions: it becomes arbitrarily negative as $r_n(v)$ approaches zero, while its positive values grow only logarithmically.

Since $\log r \leq r - 1$ for all $r > 0$, with equality only at $r = 1$, FKL assigns larger positive advantages when $r_n(v) > 1$, whereas RKL assigns more negative advantages when $r_n(v) < 1$. Thus, in this

advantage representation, divergence selection controls the balance between positive and negative token-level weighting: FKL emphasizes positive feedback for student-underweighted tokens, while RKL emphasizes negative feedback for student-overweighted tokens.

We next analyze the requirements for the properties of the advantage function under different values of the verifier's judgment $R(y) \in \{0, 1\}$ for the output y sampled from the student. This motivates the selection of the divergence function according to the verifier's judgment $R(y)$, rather than applying a fixed divergence function for all trajectories as in vanilla OPSD.

Correct trajectories ($R(y) = 1$). For a successful trajectory, we aim to reinforce teacher-preferred tokens and absorb the extra probability mass it assigns to valid alternative continuations, without aggressively disrupting an already successful trajectory due to teacher conservatism. This requires a divergence with a gentle, bounded penalty when $r_n(v) < 1$ and a strong reward when $r_n(v) > 1$. FKL supports this goal through its larger positive advantages and bounded negative advantages. We therefore apply FKL to correct trajectories.

Incorrect trajectories ($R(y) = 0$). For an unsuccessful trajectory, we place greater emphasis on suppressing tokens where the student is overconfident relative to the teacher ($r_n(v) < 1$), the likely culprits behind the error. The FKL's penalty has a lower bound, limiting the correction speed. However, the unbounded penalty in RKL can achieve a stronger suppression effect. Conversely, for teacher-preferred tokens in a failed rollout, the teacher's rationalization of $y^\star$ may be locally noisy. Therefore, we tend to adopt a cautious logarithmic damping update rather than an aggressive reinforcement, which is in line with RKL. We thus apply RKL to incorrect trajectories.

This yields an outcome-guided divergence selection rule. When $R(y) = 1$, $d_n = d_n^{\mathrm{FKL}}$, and when $R(y) = 0$, $d_n = d_n^{\mathrm{RKL}}$. It lets the model allocate token-level credit in a manner consistent with the coarse and trajectory-level outcome correctness signal, while retaining the dense and token-level granularity of self-distillation. The analysis of the advantage function $A_n(v)$ above is based on a necessary precondition: the guidance signal $p_T^n(v)$ must be reliable, which we examine next.

3.3 Selective Prefix Distillation

Diagnose teacher reliability. The teacher is conditioned on student-generated prefixes, which may contain errors or deviate from its preferred reasoning paths. As these prefixes lengthen, the teacher may become less able to produce a correct continuation, a phenomenon known as *prefix failure*. We examine whether this degradation differs between correct and incorrect student trajectories.

We sample responses from Qwen3 models (Yang et al., 2025) on OpenThoughts-Math30K (Guha et al., 2026), DeepScaleR (Tan et al., 2026), MATH-500 (Hendrycks et al., 2021), and DeepMath (He et al., 2026). We divide the responses into correct and incorrect pools, $\mathcal{D}_c$ and $\mathcal{D}_i$, according to their verifier outcomes. We then concatenate the teacher prompt with student-generated prefixes truncated at varying lengths, and let the teacher resume generation from different token positions across both correct and incorrect student trajectories to complete the remaining reasoning. Teacher completion accuracy (TCA) measures the accuracy rate of these completion results.

Figures 1b and 1d show a clear difference between the two groups. For incorrect student trajectories, the TCA degrades markedly with truncation length. In contrast, for correct trajectories, it remains stable or even improves regardless of truncation position. These observations motivate distillation over the full response when the student succeeds and restrict distillation to a prefix when it fails.

Estimate the prefix cutoff. We want to seek a boundary separating two regimes of teacher reliability on incorrect trajectories: one where the teacher signal remains valid, and one beyond which guidance quality begins to deteriorate. Following common practice, we use entropy as an indicator of the internal state of the model during token generation. Unlike most work that analyzes token-level entropy (Liu et al., 2026; Jin et al., 2026), we argue that single-token entropy is a high-variance and fluctuating signal whose value at any position cannot accurately represent the overall state of the model at that point. We use cumulative average teacher entropy to summarize uncertainty along a prefix and reduce sensitivity to fluctuations at individual positions:

$$\bar{H}_n(y) = \frac{1}{n} \sum_{k=1}^{n} H(p_T^k), \qquad H(p) = -\sum_{v \in \mathcal{V}} p(v) \log p(v). \tag{9}$$

Algorithm 1 Outcome-Guided On-Policy Self-Distillation

Require: Training data $\mathcal{S}$; model π_θ; verifier R; cutoff N_t from Eq. 11
for each training iteration **do**
  Sample a minibatch $\mathcal{B} \subset \mathcal{S}$.
  for all $(x, y^\star) \in \mathcal{B}$ **do**
    Generate a complete response $y \sim \pi_\theta(\cdot \mid x)$.
    Set $N \leftarrow |y|$ and obtain $b \leftarrow R(y)$.
    if $b = 1$ **then**
      $T \leftarrow N$; $\kappa \leftarrow \mathrm{FKL}$.
    else
      $T \leftarrow \min\{N, N_t\}$; $\kappa \leftarrow \mathrm{RKL}$.
    end if
    Initialize $\ell(x, y^\star) \leftarrow 0$.
    if $T > 0$ **then**
      Compute p_S^n and p_T^n for $n = 1, \ldots, T$.
      $\ell(x, y^\star) \leftarrow T^{-1} \sum_{n=1}^{T} d_n^\kappa$.
    end if
  end for
  $\mathcal{L} \leftarrow |\mathcal{B}|^{-1} \sum_{(x,y^\star)\in\mathcal{B}} \ell(x, y^\star)$.
  Update θ using $\nabla_\theta \mathcal{L}$.
end for

For variable-length responses, let $\mathcal{D}_a(n) = \{y \in \mathcal{D}_a : |y| \geq n\}$ contain the trajectories that reach position n, where $a \in \{c, i\}$. The outcome-conditioned averages are

$$\bar{H}_n^a = \frac{1}{|\mathcal{D}_a(n)|} \sum_{y \in \mathcal{D}_a(n)} \bar{H}_n(y), \qquad a \in \{c, i\}, \tag{10}$$

evaluated at positions where both pools are nonempty.

As shown in Figures 1c and 1e, the two curves almost overlap in the early stage. This indicates that even incorrect reasoning trajectories remain in a regime where the teacher can still provide relatively reliable supervision in early stages, yielding supervision signals of comparable quality to those on correct trajectories. Beyond their last intersection, the two curves gradually diverge, with the incorrect trajectories showing higher cumulative average entropy than the correct trajectories. This divergence is consistent with the hypothesis that accumulating errors cause the trajectory prefix to deviate from the conditional distribution expected by the teacher model, increasing its uncertainty. At this time, the teacher's supervision signal is no longer consistent with the reliable signal on correct trajectories. We then consider that the reliability of the teacher signal begins to deteriorate at this point. We use the two curves' last crossing as an empirical cutoff. Let $\Delta_n = \bar{H}_n^i - \bar{H}_n^c$, and let L be the largest position represented in both pools. We define

$$N_t = \begin{cases} \max \mathcal{C}, & \mathcal{C} \neq \varnothing, \\ 0, & \mathcal{C} = \varnothing. \end{cases} \quad \text{where} \quad \mathcal{C} = \{n \in \{2, \ldots, L\} : \Delta_{n-1}\Delta_n \leq 0\} \tag{11}$$

The resulting cutoff N_t is shared across incorrect trajectories.

Overall objective. Combining outcome-guided divergence selection with prefix restriction gives OG-OPSD. For a response of length N, the number of supervised positions $T(y)$ equals N when $R(y) = 1$ and equals $\min\{N, N_t\}$ when $R(y) = 0$. The per-response loss is

$$\mathcal{L}_{\text{OG-OPSD}}(\theta; y) = \frac{1}{\max\{1, T(y)\}} \sum_{n=1}^{T(y)} \Big[R(y) d_n^{\mathrm{FKL}} + \big(1 - R(y)\big) d_n^{\mathrm{RKL}} \Big]. \tag{12}$$

Thus, correct responses receive FKL supervision at every position, while incorrect responses receive RKL supervision only within the retained prefix. The truncation applies only when computing the distillation loss.

We average the per-response losses over each minibatch and update the model through the student branch. Algorithm 1 summarizes the training procedure.

Table 1: Main results on AIME 2024/2025 and HMMT 2025 across Qwen3-1.7B, Qwen3-4B and Qwen3-8B. Scores are Avg@12 and Average is the unweighted mean across benchmarks.

	Qwen3-1.7B				Qwen3-4B				Qwen3-8B			
Method	AIME24	AIME25	HMMT25	Avg.	AIME24	AIME25	HMMT25	Avg.	AIME24	AIME25	HMMT25	Avg.
Base	51.5	36.7	23.1	37.1	74.9	66.4	42.2	61.1	75.8	65.6	43.9	61.7
SFT	48.4	36.3	22.7	35.8	70.2	62.3	43.4	58.6	72.3	64.2	42.9	59.8
GRPO	51.1	38.3	23.7	37.7	75.6	68.1	44.4	62.7	76.4	68.9	46.7	64.0
OPSD	57.2	41.1	28.8	42.3	75.6	67.9	44.5	62.6	77.8	67.5	45.8	63.7
EOPD	52.2	39.1	26.9	39.4	75.7	66.5	43.9	62.0	77.5	70.0	46.6	64.7
AVSD	57.2	38.3	24.4	39.9	75.5	68.0	44.4	62.6	75.2	69.1	46.1	63.4
REOPOLD	-	-	-	-	74.4	63.8	42.8	60.3	-	-	-	-
Purified OPSD	-	-	-	-	76.3	67.2	44.7	62.7	-	-	-	-
DOPD	-	-	-	-	74.4	66.1	41.6	60.7	-	-	-	-
ESR	51.6	39.4	23.8	38.2	73.3	68.0	43.0	61.4	76.3	69.4	45.5	63.7
PW-OPSD	57.5	41.3	31.1	43.3	76.7	68.3	44.5	63.1	78.3	70.8	47.2	65.4
OG-OPSD	**58.8**	**41.9**	**32.5**	**44.4**	**77.5**	**70.2**	**46.9**	**64.9**	**78.9**	**71.4**	**48.9**	**66.4**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We conduct experiments with the Qwen3-1.7B, Qwen3-4B, and Qwen3-8B Instruct models (Yang et al., 2025). All three models are trained on OpenThoughts-Math30K (Guha et al., 2026), which contains 29,434 mathematical reasoning examples, each comprising a question, a reference solution, and a final answer. At each distillation iteration, the student model receives only the question and generates its own response, while the teacher model additionally receives the reference solution containing reasoning steps and the final answer. To evaluate the method on multimodal tasks, we also train Qwen3-VL-2B-Instruct (Bai et al., 2025) on FinMME (Luo et al., 2025). Dataset preparation details are provided in Appendix B.

Baseline. We compare OG-OPSD with ten baselines for mathematical reasoning. *SFT* (Ouyang et al., 2022) trains on reference solutions, while *GRPO* (Shao et al., 2024) optimizes verifiable rewards based on final-answer correctness. The remaining eight baselines use distillation. *OPSD* (Zhao et al., 2026) applies a fixed FKL objective uniformly along student-generated trajectories. *EOPD* (Jin et al., 2026) selects the divergence at each token position using teacher entropy. *AVSD* (Nguyen et al., 2026) combines supervision from multiple privileged teacher views. *REOPOLD* (Ko et al., 2026) uses the teacher-student log-likelihood ratio as a token-level reward, together with reward clipping and entropy-based sampling. *Purified OPSD* (Shen et al., 2026) uses pointwise mutual information to construct distillation targets that reduce reference-induced interference. *DOPD* (Yu et al., 2026) dynamically routes teacher-student supervision according to token-level privilege advantage gaps. *ESR* (Ziheng et al., 2026) applies a uniform truncation rule to all trajectories. *PW-OPSD* (Liu et al., 2026) weights token-level distillation losses according to position. For multimodal evaluation, we compare against *SFT* and *SDFT* (Shenfeld et al., 2026), a self-distillation method developed for continual learning. We also report the corresponding base model in all comparisons.

Benchmarks. We evaluate mathematical reasoning on AIME 2024, AIME 2025, and HMMT 2025, each containing 30 problems, using the official answer keys to determine correctness. Following the OPSD evaluation protocol (Zhao et al., 2026), we set the maximum generation length to 38,912 tokens. We generate 12 responses per problem and report the average accuracy over these samples and all problems (Avg@12). We also report the unweighted mean across the three benchmarks.

For multimodal evaluation, we use the FinMME test set and three out-of-distribution (OOD) benchmarks: MathVerse (Zhang et al., 2024), MathVision (Wang et al., 2024), and MMMU_{Pro} (Yue et al., 2025). To assess cross-domain retention after mathematical post-training, we additionally evaluate the Qwen3 models on LiveCodeBench v5 (LCB v5) (Jain et al., 2025).

4.2 MAIN RESULTS

Table 1 reports the main results on Qwen3-4B. In the experiment, we use the cumulative average teacher entropy to obtain the cutoff $N_t = 100$. OG-OPSD achieves the best performance in all three benchmarks, reaching 77.5, 70.2 and 46.9 on AIME24, AIME25 and HMMT25. OG-OPSD improves over vanilla OPSD by 1.9, 2.3, and 2.4 points on AIME24, AIME25, and HMMT25, respectively, and outperforms the strongest baseline by 1.8 points on average across the three benchmarks. OG-OPSD directly conditions the choice between FKL and RKL on the verifier's correctness

signal $R(y)$, which provides a coarser but more reliable anchor than purely distributional heuristics, and no additional hyperparameters or trade-offs are required.

We further investigate how performance gains scale with model size across all three benchmarks. For each model size, our method consistently achieves the maximum performance gain over the evaluated baselines. Moreover, as the model parameter size increases, the advantage gap over vanilla OPSD continues to widen. In the HMMT25 dataset, the gap over vanilla OPSD (Zhao et al., 2026) increases from 2.4 on Qwen3-4B to 3.1 on Qwen3-8B, and on AIME25 it grows even more markedly from 2.3 to 3.9. Averaged over the three benchmarks, the gain over vanilla OPSD increases steadily with model size, from 2.1 on Qwen3-1.7B to 2.3 on Qwen3-4B and 2.7 on Qwen3-8B. This indicates that the gains from outcome-guided divergence switching and selective truncation are not a phenomenon specific to small models. A plausible explanation is that larger models possess stronger contextual reasoning capacity and can provide higher-quality teacher supervision signals. This reflects the rationality and effectiveness of our dynamic selection of the KL divergence based on the outcome to adapt to the requirements for token reward or penalty magnitudes on different sampling trajectories. Due to limitations in computing resources, we are unable to compare all the baselines on the 1.7B and 8B models. However, existing results have already demonstrated to some extent that the observed gain of our method is not limited to the Qwen3-4B. We also notice that the gain of GRPO is relatively smooth, and even shows a slight negative growth at 1.7B scale on AIME24. This suggests that optimizing solely with sparse final-outcome rewards is challenging for smaller models. By leveraging this binary outcome signal to guide dense distillation supervision, our method remains effective even for small student models.

4.3 Ablation Study

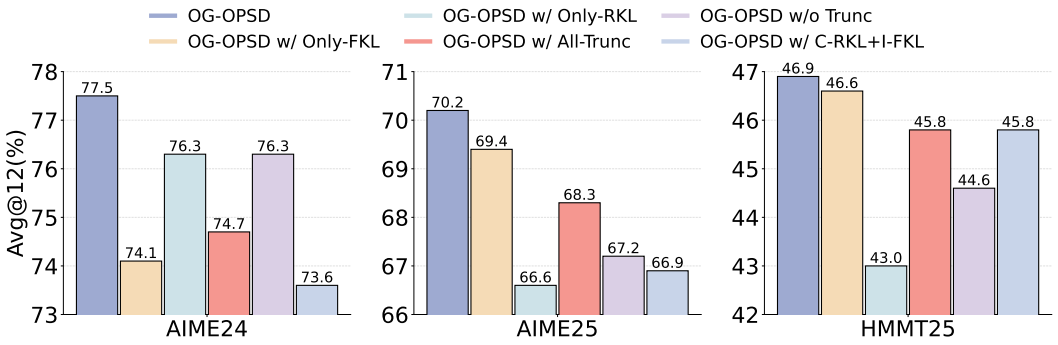


Figure 2: Ablation results of divergence selection strategy and distillation position on AIME24, AIME25 and HMMT25 benchmarks.

We conduct ablation studies on both the divergence selection strategy and the distillation position of OG-OPSD. We define five variants: *OG-OPSD w/ Only-FKL* and *OG-OPSD w/ Only-RKL* replace the outcome-guided divergence switching with a fixed FKL and a fixed RKL respectively for all trajectories, while still truncating the incorrect trajectories at position N_t. *OG-OPSD w/ All-Trunc* retains our dynamic divergence selection but applies truncation at N_t across all trajectories regardless of correctness. *OG-OPSD w/o Trunc* only uses dynamic divergence selection without any truncation. *OG-OPSD w/ C-RKL+I-FKL* reverses our divergence assignment, using RKL for correct trajectories and FKL for incorrect ones.

The selection of the divergence. As shown in Figure 2, under identical truncation treatment, OG-OPSD consistently outperforms all alternative divergence-selection variants across the three benchmarks. Both fixing a single divergence (w/ Only-FKL, w/ Only-RKL) and reversing the divergence assignment (w/ C-RKL+I-FKL) lead to clear performance drops. This provides evidence that the gains of OG-OPSD do not simply come from having some adaptive mechanism, but specifically from correctly pairing FKL with correct trajectories and RKL with incorrect ones, consistent with our analysis in Section 3.2.

The selection of the distillation position. Replacing outcome-guided truncation with all truncation (w/ All-Trunc) also yields consistent performance drops across all benchmarks,

Table 2: Evaluation on FinMME and OOD datasets after training on FinMME.

	Base	SFT	SDFT	OG-OPSD
FinMME	35.0	32.2(↓ 2.8)	37.2(↑ 2.2)	38.2(↑ 3.2)
MathVerse	31.2	24.5(↓ 6.7)	30.3(↓ 0.9)	32.4(↑ 1.2)
MathVision	14.8	14.5(↓ 0.3)	12.8(↓ 2.0)	17.4(↑ 2.6)
MMMU_{Pro}	33.7	32.5(↓ 1.2)	33.5(↓ 0.2)	34.4(↑ 0.7)

Table 3: Evaluation on LCB v5 after training on OpenThoughts-Math30K.

Method	Qwen3-1.7B	Qwen3-4B	Qwen3-8B
Base	32.8	60.3	53.5
OPSD	18.9(↓ 13.9)	49.2(↓ 11.1)	45.5(↓ 8.0)
PW-OPSD	17.7(↓ 15.1)	46.2(↓ 14.1)	47.8(↓ 5.7)
OG-OPSD	24.6(↓ 8.2)	54.1(↓ 6.2)	52.5(↓ 1.0)

showing that applying cutoff to all trajectories regardless of their correctness, such as ESR (Ziheng et al., 2026), discards the valid supervision signal from the correct trajectories. The decline in the performance of *OG-OPSD w/o Trunc* indicates the necessity of considering the reliability of the teacher's guidance signal. Before training, using Eq. 11, we calculate the truncation parameter N_t as 100 on the training data. We further examine the sensitivity of N_t in Figure 3. We use different values of N_t to train Qwen3-4B. After calculating Pass@1 on OpenThoughts, AIME24, AIME25, and HMMT25, we normalize the results and also take the average. The performance of the model peaks at $N_t = 100$ and remains relatively stable within the adjacent range. This indicates that N_t, derived from the cumulative average teacher entropy, is robust to moderate perturbation and does not require precise tuning to be effective.

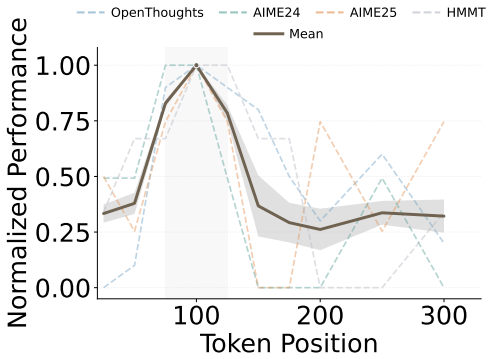


Figure 3: Train with different N_t and evaluate on multiple datasets.

4.4 Multimodal and OOD results

Table 2 reports the results after training Qwen3-VL-2B-Instruct on FinMME. OG-OPSD achieves the highest scores on both FinMME and the three OOD benchmarks among the evaluated methods. On FinMME, it scores 38.2, improving over the base model by 3.2 points. On MathVerse, MathVision, and MMMU_{Pro}, it exceeds SDFT by 2.1, 4.6, and 0.9 points, respectively. It also improves over the base model on all three OOD benchmarks, demonstrating its applicability to multimodal tasks and its benefits beyond the training domain.

Table 3 evaluates coding performance after mathematical post-training on OpenThoughts-Math30K. All three distillation methods reduce LiveCodeBench v5 scores relative to the base models, but OG-OPSD consistently incurs the smallest decline. Its score reductions are 8.2, 6.2, and 1.0 points at the 1.7B, 4B, and 8B scales, compared with 13.9, 11.1, and 8.0 points for vanilla OPSD. These results indicate that OG-OPSD better preserves coding performance while improving mathematical reasoning. This might be because we truncate the incorrect trajectories by using cumulative average entropy, which may reduce interference with the model's existing knowledge caused by the low-quality guidance from the teacher in the later stages of the incorrect trajectories.

5 Conclusion

We present OG-OPSD, which uses binary outcome rewards to guide divergence selection and determine which token positions receive supervision. Our analysis shows that the implicit token-level advantages of FKL and RKL have matching signs but different magnitudes, motivating FKL for correct trajectories and RKL for incorrect ones. We further identify an outcome-dependent pattern of prefix failure: teacher continuation accuracy declines with prefix length on incorrect trajectories but remains stable or improves on correct ones. Accordingly, OG-OPSD distills correct trajectories in full and restricts distillation on incorrect trajectories to prefixes whose length is estimated from cumulative average teacher entropy. Experiments on Qwen3 and Qwen3-VL show consistent improvements over the evaluated baselines, with ablations supporting the contributions of both divergence selection and selective truncation. These findings highlight how coarse but reliable outcome signals can guide the use of dense self-distillation supervision. Future work could extend this approach to graded rewards and per-trajectory adaptive truncation.

AI USE STATEMENT

Large Language Models (LLMs) were used to aid in the polishing of the manuscript to improve the readability and ensure clarity in various sections of the paper. The LLM was not involved in the ideation, research methodology, or experimental design. All the edited content based on LLM was reviewed and confirmed by the authors. The authors take full responsibility for the content of the manuscript.

REPRODUCIBILITY STATEMENT

We provide detailed information to facilitate reproduction of our results. We have provided detailed descriptions of the construction details of some datasets, as well as the specific parameters for training and evaluation in the Appendix B. Anonymous source code for reproducing the experiments is provided in the supplementary material.

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*, volume 2024, pp. 21246–21263, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. Openthoughts: Data recipes for reasoning models. In *International Conference on Learning Representations*, volume 2026, pp. 108059–108130, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026.

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. *NeurIPS*, 2021.

Jonas Hübotter, Frederike Lübeck, Lejs Deen Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. In *Forty-third International Conference on Machine Learning*, 2026.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Naman Jain, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, volume 2025, pp. 58791–58831, 2025.

Nan Jia, Haojin Yang, Xing Ma, Jiesong Lian, Shuailiang Zhang, Weipeng Zhang, Ke Zeng, Xunliang Cai, and Zequn Sun. Asymmetric on-policy distillation: Bridging exploitation and imitation at the token level. *arXiv preprint arXiv:2605.06387*, 2026.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. In *Forty-third International Conference on Machine Learning*, 2026.

Seongryong Jung, Suwan Yoon, DongGeon Kim, and Hwanhee Lee. Todi: Token-wise distillation via fine-grained divergence control. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8089–8102, 2025.

Jeonghye Kim, Xufang Luo, Minbeom Kim, Sangmook Lee, Dohyung Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. Why does self-distillation (sometimes) degrade the reasoning capability of LLMs? In *Third Conference on Language Modeling*, 2026.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. DistiLLM: Towards streamlined distillation for large language models. In *Forty-first International Conference on Machine Learning*, 2024.

Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. Scaling reasoning efficiently via relaxed on-policy distillation. *arXiv preprint arXiv:2603.11137*, 2026.

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*, 2026.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

Junyu Luo, Zhizhuo Kou, Liming Yang, Xiao Luo, Jinsheng Huang, Zhiping Xiao, Jingshu Peng, Chengzhong Liu, Jiaming Ji, Xuanzhe Liu, et al. Finmme: Benchmark dataset for financial multimodal reasoning evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29465–29489, 2025.

Duy Nguyen, Hanqi Xiao, Archiki Prasad, Zaid Khan, Anirban Das, Austin Zhang, Sambit Sahu, Hyunji Lee, Elias Stengel-Eskin, and Mohit Bansal. Avsd: Adaptive-view self-distillation by balancing consensus and teacher-specific privileged signals. *arXiv preprint arXiv:2605.20643*, 2026.

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

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

Zhanming Shen, Jintao Tong, Shaotian Yan, Chen Shen, Hao Chen, Wentao Ye, Xiaomeng Hu, Rui Miao, Haobo Wang, Junbo Zhao, et al. Purified opsd: On-policy self-distillation without losing how to think. *arXiv preprint arXiv:2607.02234*, 2026.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. In *Forty-third International Conference on Machine Learning*, 2026.

Sijun Tan, Michael Luo, Justin Wong, Colin Cai, Xiaoxiang Shi, William Yuan Tang, Manan Roongta, Tianjun Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Effective RL scaling of reasoning models via iterative context lengthening, 2026.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024.

Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. Not all disagreement is learnable: Token teachability in on-policy distillation. *arXiv preprint arXiv:2605.26844*, 2026.

Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. Rethinking kullback-leibler divergence in knowledge distillation for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5737–5755, 2025.

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.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhicheng Yang, Zhijiang Guo, Yifan Song, Minrui Xu, Yongxin Wang, Yiwei Wang, Xiaodan Liang, and Jing Tang. Prune-opd: Efficient and reliable on-policy distillation for long-horizon reasoning. *arXiv preprint arXiv:2605.07804*, 2026.

Xinlei Yu, Gen Li, Qingyi Si, Guibin Zhang, Yuqi Xu, Congcong Wang, Shuai Dong, Kaiwen Tuo, Xiangyu Zeng, Kaituo Feng, et al. Dopd: Dual on-policy distillation. *arXiv preprint arXiv:2606.30626*, 2026.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Hongbin Zhang, Chaozheng Wang, Kehai Chen, Youcheng Pan, Yang Xiang, Jinpeng Wang, and Min Zhang. Tailoring teaching to aptitude: Direction-adaptive self-distillation for llm reasoning. *arXiv preprint arXiv:2605.22263*, 2026.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024.

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.

Siqi Zhu, Xuyan Ye, Hongyu Lu, Weiye Shi, and Ge Liu. The many faces of on-policy distillation: Pitfalls, mechanisms, and fixes. *arXiv preprint arXiv:2605.11182*, 2026.

Zhou Ziheng, 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.

A GRADIENT DERIVATION

In this section we derive Eq. 6 and Eq. 7. For analytical clarity, we first derive the gradients of the unclipped KL objectives. In implementation, we additionally clip individual KL components as described in Appendix B.

Notation and assumptions. Fix a position n in a student-generated rollout, so that the prefix s_n is a given constant. For brevity, write

$$p_i \triangleq p_S^n(v_i), \qquad q_i \triangleq p_T^n(v_i), \qquad r_i \triangleq r_n(v_i) = \frac{q_i}{p_i}. \tag{13}$$

Both are categorical distributions over $\mathcal{V}$, $p_i, q_i > 0$ and $\sum_{i=1}^{|\mathcal{V}|} p_i = \sum_{i=1}^{|\mathcal{V}|} q_i = 1$. The teacher distribution is treated as a fixed target:

$$\nabla_\theta q_i = 0, \qquad \nabla_\theta p_i = p_i \nabla_\theta \log p_i. \tag{14}$$

The second equality is the log-derivative trick.

Since $\sum_i^{|\mathcal{V}|} p_i = 1$ holds for all θ,

$$\sum_i^{|\mathcal{V}|} p_i \nabla_\theta \log p_i = \sum_i^{|\mathcal{V}|} \nabla_\theta p_i = \nabla_\theta \sum_i^{|\mathcal{V}|} p_i = \nabla_\theta 1 = 0. \tag{15}$$

Consequently, for any constant c that does not depend on i,

$$\sum_i^{|\mathcal{V}|} p_i \, c \, \nabla_\theta \log p_i = 0. \tag{16}$$

A.1 FORWARD KL

According to the definition of Forward KL divergence,

$$d_n^{\mathrm{FKL}} = D_{\mathrm{KL}}(p_T^n \| p_S^n) = \sum_i^{|\mathcal{V}|} q_i \left(\log q_i - \log p_i\right). \tag{17}$$

The q_i is a constant with respect to θ, so the term $q_i \log q_i$ vanishes under differentiation.

$$\nabla_\theta d_n^{\mathrm{FKL}} = \nabla_\theta \sum_i^{|\mathcal{V}|} q_i \left(\log q_i - \log p_i\right) \tag{18}$$

$$= \sum_i^{|\mathcal{V}|} \left(\nabla_\theta \left(q_i \log q_i\right) - \nabla_\theta \left(q_i \log p_i\right)\right) \tag{19}$$

$$= -\sum_i^{|\mathcal{V}|} q_i \nabla_\theta \log p_i. \tag{20}$$

To write this in the form $\sum_i p_i A(v_i) \nabla_\theta \log p_i$, we use $q_i = p_i r_i$:

$$-\nabla_\theta d_n^{\mathrm{FKL}} = \sum_i^{|\mathcal{V}|} p_i \, r_i \, \nabla_\theta \log p_i. \tag{21}$$

Subtracting the zero term in Eq. 16 with $c = 1$:

$$-\nabla_\theta d_n^{\mathrm{FKL}} = \sum_i^{|\mathcal{V}|} p_i \, r_i \, \nabla_\theta \log p_i - \sum_i^{|\mathcal{V}|} p_i \, \nabla_\theta \log p_i \tag{22}$$

$$= \sum_i^{|\mathcal{V}|} p_i \left(r_i - 1\right) \nabla_\theta \log p_i, \tag{23}$$

which is exactly Eq. 6, with advantage $A_n^{\mathrm{FKL}}(v) = r_n(v) - 1$. □

A.2 REVERSE KL

According to the definition of Reverse KL divergence,

$$d_n^{\mathrm{RKL}} = D_{\mathrm{KL}}(p_S^n \| p_T^n) = \sum_i^{|\mathcal{V}|} p_i \left(\log p_i - \log q_i\right). \tag{24}$$

Now p_i appears both as the weight and inside the logarithm, so we apply the product rule:

$$\nabla_\theta d_n^{\mathrm{RKL}} = \sum_i^{|\mathcal{V}|} \nabla_\theta p_i \left(\log p_i - \log q_i\right) + \sum_i^{|\mathcal{V}|} p_i \nabla_\theta \left(\log p_i - \log q_i\right). \tag{25}$$

For the second sum, $\nabla_\theta \log q_i = 0$ and $\sum_i^{|\mathcal{V}|} p_i \nabla_\theta \log p_i = 0$ by Eq. 15, so it's equal to zero.

For the first sum, we apply the log-derivative trick:

$$\sum_i^{|\mathcal{V}|} \nabla_\theta p_i \left(\log p_i - \log q_i\right) = \sum_i^{|\mathcal{V}|} p_i \left(\log p_i - \log q_i\right) \nabla_\theta \log p_i \tag{26}$$

$$= \sum_i^{|\mathcal{V}|} p_i \left(-\log r_i\right) \nabla_\theta \log p_i. \tag{27}$$

Combining the two,

$$-\nabla_\theta d_n^{\mathrm{RKL}} = \sum_i^{|\mathcal{V}|} p_i \log r_i \, \nabla_\theta \log p_i, \tag{28}$$

which is exactly Eq. 7, with advantage $A_n^{\mathrm{RKL}}(v) = \log r_n(v)$. □

B IMPLEMENTATION DETAILS

Details of FinMME Dataset Processing. FinMME (Luo et al., 2025) is a comprehensive benchmark dataset designed for evaluating multimodal large language models in the financial domain. It comprises approximately 11,000 high-quality financial samples covering 18 subfields of finance and six major asset classes. In terms of question types, the dataset includes single-choice questions, multiple-choice questions, and numerical answer questions. In this work, we first filter out the samples with excessively high image resolution. Then, we adjust the prompts for different question types and extracted subsets. The sampling process ensures that options A, B, C, and D appear equally frequently in all answers. Finally, a training set of 1200 samples is constructed, including 600 multiple-choice questions and 600 numerical questions. A test set of 600 samples is also obtained, consisting of 300 multiple-choice questions and 300 numerical questions. We use the Qwen3-VL-32B-Instruct to generate reference solutions for this dataset as privileged information for the teacher model.

Training. Our training setup follows vanilla OPSD (Zhao et al., 2026). All attention and multilayer perceptron projections are adapted using LoRA with rank 64 and scale 128. The privileged-teacher forward pass is evaluated with the LoRA adapter disabled, so gradient updates are applied only to the student's LoRA parameters. We optimize with a learning rate of 5×10^{-6}, gradient-norm clipping at 0.1, bfloat16 mixed precision, and gradient checkpointing. The student samples one on-policy rollout per problem at temperature 1.1, top-p 0.95, top-k 20, generating up to $1,024$ new tokens and the training context length is capped at $20,000$ tokens. During rollout generation, the student receives only the problem statement, rendered with the Qwen3 chat template and `enable_thinking=False`. The teacher receives the problem together with the reference solution (including the final answer). The specific prompts are presented in Table 4. We use a per-device batch size of 4 with 2 gradient accumulation steps, and the main experiments utilize 4 NVIDIA RTX Pro 6000 GPUs. The data collator right-pads prompts within a batch. Every directed loss term is computed over the full vocabulary at each non-padding completion token. Both teacher and student logits are temperature-scaled by 1.1 prior to computing the KL divergence, and, consistent with

View	User-message template
Student	**Problem:** {problem} Please reason step by step, and put your final answer within \boxed{}.
Teacher	**Problem:** {problem} Here is a reference solution to this problem: === Reference Solution Begin === {reference completion} === Reference Solution End === After reading the reference solution above, make sure you truly understand the reasoning behind each step — do not copy or paraphrase it. Now, using your own words and independent reasoning, derive the same final answer to the problem above. Think step by step, explore different approaches, and don't be afraid to backtrack or reconsider if something doesn't work out: Please reason step by step, and put your final answer within \boxed{}.

Table 4: User-message templates for the *Student* view without privileged information and the reference-conditioned *Teacher* view.

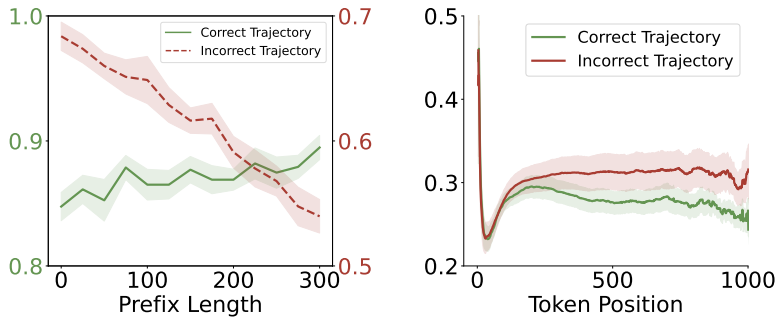
vanilla OPSD (Zhao et al., 2026), individual KL components are clipped with a cap of $c = 0.05$. All baselines use their published default hyperparameters.

Evaluation. We use a maximum generation length of $38,912$ tokens: `max_new_tokens` $=$ 38912, $M = 12$ samples per problem, temperature 1.0, top-p 0.95, top-k disabled, and `enable_thinking=True`, matching the vanilla OPSD evaluation setup. For a single problem with $M = 12$ generated solutions $\{y^{(1)}, \ldots, y^{(M)}\}$, the predicted answer is extracted as the content of the last $\texttt{\textbackslash boxed}\{\ldots\}$ in $y^{(i)}$. Let $c_i \in \{0, 1\}$ indicate whether $y^{(i)}$ is graded correct against the gold answer. We report the Avg@12 on AIME24, AIME25, and HMMT25.

$$\text{Avg@12} = \frac{1}{12} \sum_{i=1}^{12} c_i. \tag{29}$$

C More Results On Qwen3-1.7B

We observe the same trend on Qwen3-1.7B as the 4B and 8B scale models in the main text. The results of TCA are averaged over OpenThoughts-Math30K, DeepScaleR, MATH-500, and DeepMath datasets consistent with Figures 1b and 1d.



(a) Teacher completion accuracy (TCA) (b) Cumulative average teacher entropy

Figure 4: TCA and cumulative average teacher entropy on Qwen3-1.7B.