E^2-OPSD: TAMING ENTROPY OVERSHOOT IN ON-POLICY SELF-DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation (OPSD) provides dense token-level supervision without a second model: one network acts as teacher with the reference solution and as student with only the problem. We identify a specific failure mode of this recipe. During training, student token entropy rises past the teacher's and remains elevated, a pattern we call ***entropy overshoot***. We trace it to both sides of distillation. The reference-conditioned teacher is confident along its answer-directed reasoning path, but this confidence transfers poorly to student-generated prefixes, making its supervision overly tied to answer-specific cues rather than reusable reasoning patterns; meanwhile, the forward KL used by OPSD continually diffuses the student's predictive distribution without pulling it back. We introduce ***E^2-OPSD*** to address both causes. **Exemplar-guided teaching** replaces the current answer with a retrieved solved neighboring problem, providing transferable reasoning guidance without revealing the destination and better matching student-reachable states. **Entropy-aware distillation** uses the student–teacher entropy gap to determine the direction and strength of each token's correction. ***E^2-OPSD*** improves math reasoning by up to 4.3 points in mean@16 over OPSD, while out-of-domain evaluations show gains over the corresponding base models of up to 4.9 points in mean@16 and 5.5 points in pass@8. Despite these gains, ***E^2-OPSD*** remains simple, requiring no additional forward passes or networks.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) is now the standard recipe for improving LLM reasoning, yet its supervision remains coarse: an entire rollout collapses into a single scalar reward (Shao et al., 2024; Guo et al., 2025). On-policy distillation (OPD) instead offers dense, token-level supervision by having a teacher score the student's own rollouts (Agarwal et al., 2024; Gu et al., 2023), and is now a standard post-training component in Qwen3 (Yang et al., 2025), GLM-5 (GLM-5-Team et al., 2026), and DeepSeek-V4 (DeepSeek-AI et al., 2026). Despite this fine granularity, it still requires a stronger open-weight teacher that shares the student's vocabulary. On-policy self-distillation (OPSD) further removes the need to serve a second model: a copy of the student acts as the teacher, with access to the reference solution, while the student sees only the problem. Teacher and student are evaluated on the same student-generated prefixes and matched with forward KL (Zhao et al., 2026). This makes OPSD an appealing way to turn privileged solution information into dense supervision. In practice, however, its gains are unstable and can deteriorate as training proceeds.

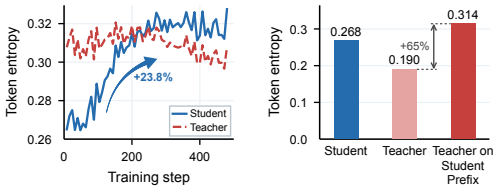
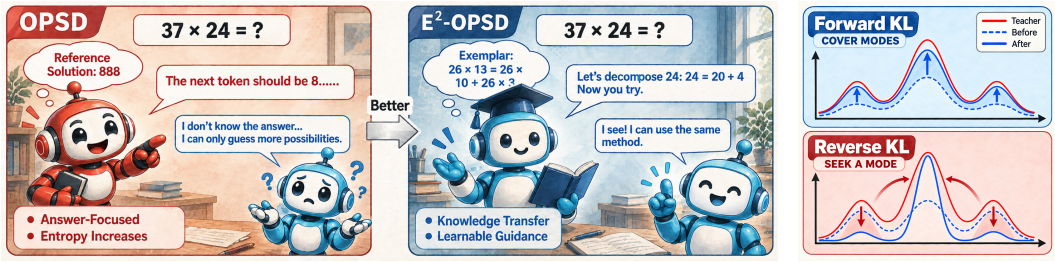


Figure 1: **Left:** Entropy overshoot in Qwen3-1.7B. **Right:** The reference-conditioned teacher is less certain on student prefixes, suggesting a reasoning-path mismatch.

We identify a characteristic failure mode, as shown in Figure 1 (left), ***entropy overshoot***: instead of settling toward the teacher, the student becomes progressively more uncertain, eventually surpassing the teacher's entropy and staying there. Entropy overshoot therefore describes a regime where

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) Exemplar-Guided Teaching (b) Forward/Reverse KL

Figure 2: **(a) Exemplar-Guided Teaching.** Standard OPSD gives the teacher the current reference solution, letting it jump from the student prefix toward the answer. ***E²-OPSD*** instead uses a solved neighbor as a strategy guide, forcing the teacher to follow the student's intermediate step and solve the current problem itself, turning answer leakage into strategy transfer. **(b) Forward and reverse KL play complementary roles.** Forward KL promotes exploration over teacher-supported alternatives, while reverse KL removes unsupported probability mass to concentrate the policy. Entropy-aware distillation switches between them.

useful exploration spills into persistent and excessive uncertainty, gradually eroding the student's decisiveness and amplifying teacher-side uncertainty into noisy student behavior.

We examine this failure from both teacher and student perspectives. First, **OPSD produces supervision mismatched to the student's reasoning path**. The teacher is confident on its own trajectories but becomes substantially more uncertain on student-generated prefixes (Figure 1, right), suggesting that its guidance does not transfer cleanly to the intermediate states the student actually reaches. This mismatch stems from a deeper issue: seeing the current solution lets the teacher narrate toward a known answer rather than reason from scratch, making its supervision answer-specific rather than transferable (Figure 2(a)). Second, **OPSD lacks a mechanism to regulate policy spread**. Forward KL is beneficial when the student is too concentrated as it expands probability toward teacher-supported alternatives. But the same pressure persists even after the policy becomes overly diffuse, which makes exploration a one-way ratchet: uncertainty keeps spreading, with no force pulling the policy back into focus (Figure 2(b)).

These observations motivate ***E²-OPSD***, a simple framework that repairs both sides of the learning process with **E**xemplar-guided teaching and **E**ntropy-aware distillation, as illustrated in Figure 3. More specifically, ***E²-OPSD*** makes three contributions:

- ***What to learn*** — *Exemplar-guided teaching.* We replace the current problem's reference solution in the teacher's context with a retrieved solved neighboring problem. The exemplar provides a transferable problem-solving reference without revealing the current answer. Conditioned on this reference and the student's generated prefix, the teacher's token-level guidance is meant to be more compatible with the student's ongoing reasoning.
- ***How to learn*** — *Entropy-aware distillation.* We use the student–teacher entropy gap as a feedback signal: when the student is more concentrated than the teacher, forward KL encourages exploration; when it becomes more diffuse, reverse KL suppresses unsupported mass and encourages concentration. The gap magnitude downweights uncertain routing decisions and concentrates learning on tokens with larger entropy mismatches. This turns self-distillation from a fixed one-way objective into closed-loop self-balancing control of exploration.
- ***Simple and Generalized.*** ***E²-OPSD*** adds no extra forward passes during distillation and minimum overhead. Across various model families, sizes from 0.6B to 7B, and both thinking and instruction-tuned models, it mitigates persistent entropy overshoot and improves math reasoning by up to 4.3 points over OPSD, with gains transferring to out-of-domain tasks by up to 4.9 mean@16 points over the corresponding base model. These results suggest that effective self-distillation depends not on maximizing imitation, but on **aligning teacher guidance with the student's reasoning path and regulating when the student should explore versus concentrate**.

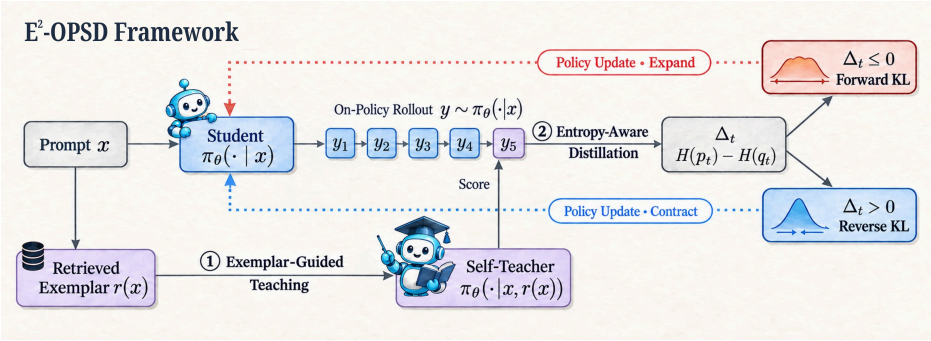


Figure 3: **Overview of *E²-OPSD*.**

2 Preliminaries and Related Work

2.1 KL Divergence

The Kullback–Leibler (KL) divergence is asymmetric, and its two directions induce distinct approximation biases (Minka, 2005; Jin et al., 2026). For a target distribution P and model distribution Q,

$$\mathrm{KL}(P\|Q) = \mathbb{E}_{v\sim P}\left[\log \frac{P(v)}{Q(v)}\right], \qquad \mathrm{KL}(Q\|P) = \mathbb{E}_{v\sim Q}\left[\log \frac{Q(v)}{P(v)}\right]. \tag{1}$$

Forward KL is *mode covering*: if $P(v) > 0$ while $Q(v) \to 0$, its contribution diverges, strongly penalizing target-supported modes that the model fails to cover. Reverse KL has the complementary *mode-seeking* bias: if $Q(v) > 0$ while $P(v) \to 0$, its contribution diverges, strongly penalizing model mass unsupported by the target (Minka, 2005). Thus, forward KL favors coverage, whereas reverse KL favors concentration.

2.2 On-Policy Distillation

On-policy distillation (OPD) trains a student π_θ to match a teacher π_T on the student's own generations (Agarwal et al., 2024; Gu et al., 2023; Lu & Lab, 2025). Given a problem $x \sim \mathcal{D}$, the student samples $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_{|\hat{y}|})$, and both models score the same student-generated prefix $s_t = (x, \hat{y}_{<t})$:

$$\mathcal{L}(\theta) = \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{\hat{y}\sim\pi_\theta(\cdot|x)}\left[\frac{1}{|\hat{y}|}\sum_{t=1}^{|\hat{y}|} D\big(\pi_T(\cdot \mid s_t) \,\big\|\, \pi_\theta(\cdot \mid s_t)\big)\right], \tag{2}$$

where D denotes a per-token divergence between the teacher and student distributions, instantiated as forward KL $\mathrm{KL}(\pi_T\|\pi_\theta)$ or reverse KL $\mathrm{KL}(\pi_\theta\|\pi_T)$, and gradients flow only through the student. Reverse KL is a common choice (Gu et al., 2023; Lu & Lab, 2025), while prior work also explores mixed or skewed divergences as alternative distillation objectives (Agarwal et al., 2024; Ko et al., 2024; Shing et al., 2025; Wu et al., 2025; Jin et al., 2026). OPSD computes D over the full vocabulary and applies per-token pointwise clipping, capping each vocabulary entry's contribution at τ so that a few stylistic tokens do not dominate the signal (Zhao et al., 2026). We keep this pointwise clipping and, following recent sparse-support distillation practice, approximate D on the teacher's top-K_{v} tokens together with a tail bucket carrying the remaining probability mass (Li et al., 2026b).

2.3 Related Work

On-policy self-distillation. On-policy self-distillation (OPSD) instantiates Eq. 2 with a copy of the student as teacher, $\pi_T = \pi_{\theta_0}$, whose context includes the reference solution y^* unseen by the student and is matched via forward KL (Zhao et al., 2026). This privileged teacher creates an irreducible objective gap (Yang et al., 2026), overweights stylistic and answer-implied tokens over genuine decisions (Pan et al., 2026; Harne et al., 2026), and suppresses uncertainty important to

thinking models (Kaur et al., 2026; Kim et al., 2026). Existing remedies largely patch these effects downstream. *Verifier-directed* methods outsource update control to outcome rewards (Yang et al., 2026; Ke et al., 2026; He et al., 2026), effectively reintroducing RLVR. *Signal-purification* methods cancel privileged bias with an additional teacher pass (Pan et al., 2026; Shen et al., 2026), paying extra compute after the bias is already introduced. *Context-curation* methods expose less of the solution or replace it with a successful sibling rollout (Han et al., 2026; Hübotter et al., 2026), but still rely on solution-specific supervision and auxiliary control. These approaches mitigate symptoms without removing their source: teacher guidance tied to the current problem's solution. ***E²-OPSD*** removes this mismatch at the source, restores the concentrating correction absent from forward KL, and retains a single frozen-teacher pass with no reward.

Entropy-aware post-training. Entropy has become a common control signal in post-training, but serves different roles across RLVR and distillation. In RLVR, it is primarily used to regulate exploration, from preventing entropy collapse to concentrating updates on high-entropy or decision-critical tokens (Cui et al., 2025; Yu et al., 2025; Wang et al., 2025; Hao et al., 2025). In distillation, entropy has been used to guide supervision selection and adaptation: EOPD activates additional forward-KL supervision at high-teacher-entropy tokens, while subsequent work further adapts supervision across uncertainty regimes (Jin et al., 2026; Zhang et al., 2026; Lee et al., 2026). ***E²-OPSD*** instead targets the relative entropy dynamics of the student and teacher, using their signed token-level entropy gap to correct entropy overshoot during on-policy self-distillation.

Training-time contextual guidance. A broad line of work improves post-training by providing auxiliary context that is available during training but absent at inference. Context distillation internalizes prompts, exemplars, or scratchpads into model weights (Askell et al., 2021; Snell et al., 2022; Shin et al., 2025; Kalwar et al., 2026). More recent methods diversify the form of this guidance: OPCD distills experiential knowledge from historical solution traces, SDFT learns on-policy from demonstrations, SKALD conditions a self-teacher on abstract skill cards, and PMD accumulates reusable procedural memories across training episodes (Ye et al., 2026; Shenfeld et al., 2026; Jiang et al., 2026; Liu et al., 2026). Other approaches inject more instance-specific scaffolds directly into optimization, such as partial solutions in QuestA, off-policy reasoning traces in LUFFY, in-context exemplars in XRPO, or retrieved analogous demonstrations in RA-RFT (Li et al., 2026a; Yan et al., 2025; Bamba et al., 2025; Xiao et al., 2026). OPSD conditions its self-teacher on the current problem's reference solution; ***E²-OPSD*** instead uses a retrieved solved neighbor, transferring reusable strategy across instances without revealing the current answer.

3 METHOD

E²-OPSD repairs both sides of the entropy-overshoot failure identified in Section 1: *exemplar-guided teaching* changes what the teacher is conditioned on so its guidance is grounded in transferable reasoning rather than answer-specific cues, while *entropy-aware distillation* changes how the student matches it, using the per-token entropy gap to set both the direction and strength of the correction. Both run inside the standard OPSD loop with one frozen teacher and no additional forward passes or extra network.

Setup. The student π_θ and frozen teacher π_ϕ, initialized from the same model ($\phi = \theta_0$), score the same student-generated prefix $\hat{y}_{<t}$ but under different contexts:

$$p_t(v) = \pi_\theta(v \mid x, \hat{y}_{<t}), \qquad q_t(v; r) = \pi_\phi(v \mid x, r(x), \hat{y}_{<t}), \tag{3}$$

where $r(x)$ denotes teacher-only privileged context. Standard OPSD sets $r(x) = y^*$, giving the teacher the reference solution to the current problem.

3.1 EXEMPLAR-GUIDED TEACHING

Conditioning on y^* lets the teacher narrate toward a known answer, producing guidance that depends on information absent from the student's context. Exemplar-Guided Teaching (EGT) instead replaces the current solution with a solved neighboring problem, so the teacher must still solve x itself while using the exemplar only as transferable strategy (Figure 2(a)).

We embed all training problems and compute pairwise cosine similarity offline, keeping each problem's K nearest neighbors (self-excluded) as $\mathcal{N}_K(x)$. For the g-th of G rollouts per problem, we draw one neighbor uniformly,

$$r_g(x) = (x'_g,\, y'^*_g), \qquad x'_g \sim \mathrm{Uniform}(\mathcal{N}_K(x))\,, \quad x'_g \neq x, \tag{4}$$

seeded deterministically by problem and rollout index, so different rollouts of the same problem see different exemplars at no extra retrieval cost. Thus the teacher is conditioned on $(x, x'_g, y'^*_g, \hat{y}_{<t})$, while the student still sees only $(x, \hat{y}_{<t})$. EGT therefore removes the current answer from teacher conditioning and replaces answer-specific privilege with cross-problem strategy guidance. Prompts and offline preprocessing cost are in Appendix A.4.

3.2 Entropy-Aware Distillation

EGT improves *what* the teacher provides; Entropy-Aware Distillation (EAD) controls *how* the student follows it. Building on the complementary mode-covering and mode-seeking biases of forward and reverse KL (Section 2.1), EAD has two components: *Bidirectional Entropy Routing* (BER) selects the KL direction from the student–teacher entropy gap, while *Adaptive Gap Damping* (AGD) uses its magnitude to scale the update.

Bidirectional entropy routing. We use the entropy gap $\Delta_t = H(p_t) - H(q_t)$ to determine which correction the student needs. When $\Delta_t \leq 0$, the student is relatively more concentrated than the teacher, so BER applies forward KL to broaden coverage. When $\Delta_t > 0$, the student is relatively more diffuse, so BER switches to reverse KL to refocus the policy (Figure 2(b)). Using $\widetilde{\mathcal{V}}_t$, the teacher's top-$K_{\mathrm{v}}$ tokens plus a tail bucket, and the per-element clip τ inherited from OPSD (Section 2.2),

$$D_t^{\mathrm{BER}} = \begin{cases} \sum_{v\in\widetilde{\mathcal{V}}_t} \min\left(q_t(v)\log\frac{q_t(v)}{p_t(v)},\, \tau\right), & \Delta_t \leq 0, \\ \sum_{v\in\widetilde{\mathcal{V}}_t} \min\left(p_t(v)\log\frac{p_t(v)}{q_t(v)},\, \tau\right), & \Delta_t > 0. \end{cases} \tag{5}$$

BER thus replaces OPSD's fixed one-way pressure with a state-dependent correction that can either broaden or refocus the student policy.

Adaptive gap damping. The sign of Δ_t selects the correction direction, while its magnitude indicates the need for correction: near-zero gaps make routing sensitive to small perturbations, whereas larger gaps indicate more pronounced student–teacher entropy mismatches that call for stronger correction. We therefore use the gap magnitude as a confidence weight to downweight uncertain routing decisions and concentrate learning on tokens with larger entropy mismatches. Over the valid response tokens $\mathcal{T}$ in the current mini-batch, we define the typical gap scale $\sigma^2 = \frac{1}{|\mathcal{T}|}\sum_{(i,t)\in\mathcal{T}} \Delta_{i,t}^2$ and weight each routed update by

$$c_t = \frac{\Delta_t^2}{\Delta_t^2 + \max(\sigma^2, \epsilon^2)} \in [0, 1), \tag{6}$$

where ϵ is the FP32 machine epsilon, with $c_t = 0$ when $\Delta_t = 0$. Gaps small relative to the batch scale are therefore suppressed, while those much larger than the scale retain most of their update. Because σ^2 is recomputed every mini-batch, the damping adapts to the training dynamics without a tunable threshold.

Training objective. Combining exemplar-guided teaching, bidirectional entropy routing, and adaptive gap damping, ***E²-OPSD*** minimizes

$$\mathcal{L}_{\mathcal{T}}(\theta) = \frac{1}{|\mathcal{T}|}\sum_{(i,t)\in\mathcal{T}} \mathrm{sg}(c_{i,t})\, D_{i,t}^{\mathrm{BER}}, \tag{7}$$

where $\mathcal{T}$ contains the valid response tokens in the current loss computation, indexed by response i and position t. Sampled responses and exemplars are held fixed during optimization. The teacher conditions on the sampled solved neighbor (x', y'^*) through q_t, while the student context remains unchanged. Both the routing decision and c_t are treated as stop-gradient quantities, so optimization acts only through the student probabilities in D_t^{BER}. Compared with OPSD, ***E²-OPSD*** changes only

the teacher context and per-token divergence, requiring the same rollouts and frozen-teacher scoring with no additional forward pass, verifier, or reward during distillation. See Appendix A.4 for more overhead analysis. At inference, all privileged context is removed.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Models & Datasets. We evaluate four configurations across the Qwen3 (Yang et al., 2025) and Olmo3 (Olmo et al., 2026) families: Qwen3-0.6B and Qwen3-1.7B with thinking enabled, Qwen3-4B-Instruct-2507, and Olmo-3-7B-Instruct. The reasoning mode is consistent across the student, teacher, and evaluation. Following Pan et al. (2026), we use DeepMath-103K (He et al., 2025) for training, with a difficulty band assigned to each configuration (Appendix A.1). We evaluate in-domain mathematical reasoning on AMC23 (math-ai, 2023), AIME'24 (Zhang & Math-AI, 2024), AIME'25 (Zhang & Math-AI, 2025), and HMMT'25 Feb (Harvard–MIT Mathematics Tournament, 2025). Out-of-domain evaluation covers logical reasoning on a fixed, randomly sampled 100-puzzle subset of ZebraLogic (Lin et al., 2025)for computational tractability, and code generation on HumanEval (Chen et al., 2021).

Baselines. We compare against (1) **Base**, the untrained checkpoints; (2) **OPSD** (Zhao et al., 2026), which distills a frozen teacher conditioned on the current problem's reference solution using forward KL; (3) **SDPO** (Hübotter et al., 2026), which uses an EMA teacher conditioned on a successful sibling rollout and a generalized JSD loss; (4) **RLSD** (Yang et al., 2026), which uses verifiable rewards to determine update directions and self-distillation signals to modulate token-level update magnitudes; and (5) **OPSD + EOPD**, our self-distillation adaptation of EOPD (Jin et al., 2026), which combines reverse KL with forward KL at high-teacher-entropy positions. We retain this entropy-gated divergence but replace EOPD's external stronger teacher with OPSD's reference-conditioned self-teacher for a controlled comparison.

Implementation Details. We use veRL (Sheng et al., 2025) with vLLM (Kwon et al., 2023) on 8×H200 GPUs and a learning rate of 10^{-6}. We faithfully implement each baseline following the algorithmic design in its original paper, preserving its method-specific teacher construction, update rules, and training objective. All trained methods and ablations share the same training settings and a budget of 360 optimizer steps. Detailed baseline configurations, including the explicit adaptation of EOPD to the OPSD setting, are provided in Appendix A.4 for reproducibility. The baseline results in Table 1 are obtained from these implementations under the shared protocol. The teacher is frozen at its initial weights except in SDPO and RLSD. We use $K = 15$ for a diverse exemplar selection, and supply retrieved exemplars only to the teacher. We evaluate each model with three evaluation seeds to reduce noise. We provide the model with only the original questions during evaluation inference. See Appendix A for detailed training and evaluation settings.

4.2 MAIN RESULTS

E²-OPSD* delivers robust and transferable gains.** As shown in Tables 1 and 2, ***E²-OPSD achieves the highest average math mean@16 across all four model configurations, spanning two model families and both thinking and instruct modes. In particular, ***E²-OPSD*** improves over Base by 4.4 and 2.1 percentage points on Qwen3-1.7B Thinking and Qwen3-4B-Instruct-2507, respectively, while leading on all four math benchmarks for both models. Its consistent effectiveness across distinct reasoning modes suggests that ***E²-OPSD*** addresses a fundamental failure mode of OPSD, rather than relying on a model- or reasoning-format-specific recipe. More importantly, the gains extend beyond math reasoning: ***E²-OPSD*** improves over Base in both logical reasoning and code generation across all four configurations, with logical reasoning gains of 4.9 and 3.1 points on Qwen3-1.7B Thinking and Qwen3-4B-Instruct-2507, respectively, alongside consistent gains in code generation as well. Together, these gains support our central premise that effective self-distillation should transfer reasoning strategies and regulate distillation dynamics, rather than merely imitate solutions.

***E²-OPSD* mitigates persistent entropy overshoot.** Figures 4(a) and 4(b) revisit the failure identified in the introduction. Under OPSD, student entropy crosses the teacher early and continues to

Table 1: Main results across model scales and types on math reasoning, reported as mean@16 over three seeds. Green subscripts show our gains over Base, computed from the displayed values. Best values are bold and second-best distinct values underlined. Full evaluation results are in Appendix B.

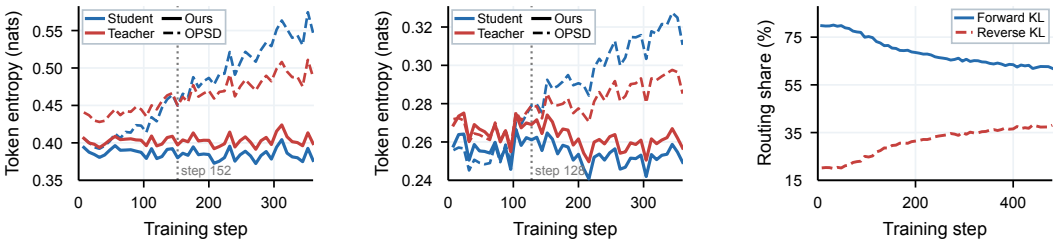
		Math Reasoning (mean@16)				
Model	**Method**	AMC23	AIME24	AIME25	HMMT25	Avg.
Qwen3-0.6B (Thinking)	Base	45.8	9.8	14.2	8.3	19.5
	OPSD	45.9	<u>10.9</u>	14.2	**9.2**	20.0
	SDPO	45.5	9.4	14.6	7.8	19.3
	RLSD	**46.8**	**11.0**	15.1	7.8	<u>20.2</u>
	OPSD + EOPD	46.4	9.8	<u>15.9</u>	8.5	20.1
	***E²-OPSD* (Ours)**	<u>46.6</u>$_{+0.8}$	**11.0**$_{+1.2}$	**16.2**$_{+2.0}$	<u>8.9</u>$_{+0.6}$	**20.7**$_{+1.2}$
Qwen3-1.7B (Thinking)	Base	75.5	46.0	34.6	21.9	44.5
	OPSD	74.9	46.0	35.0	<u>22.7</u>	44.6
	SDPO	<u>76.1</u>	48.0	<u>35.3</u>	21.9	<u>45.3</u>
	RLSD	75.9	<u>48.3</u>	34.0	21.8	45.0
	OPSD + EOPD	74.3	47.4	34.2	21.5	44.3
	***E²-OPSD* (Ours)**	**79.0**$_{+3.5}$	**52.3**$_{+6.3}$	**39.2**$_{+4.6}$	**25.0**$_{+3.1}$	**48.9**$_{+4.4}$
Qwen3-4B-Instruct-2507	Base	85.4	61.7	46.5	29.0	<u>55.7</u>
	OPSD	84.8	<u>62.4</u>	45.6	<u>29.5</u>	55.6
	SDPO	83.9	61.3	45.8	27.5	54.6
	RLSD	85.1	62.2	<u>46.9</u>	28.4	<u>55.7</u>
	OPSD + EOPD	<u>85.5</u>	62.0	46.5	28.4	55.6
	***E²-OPSD* (Ours)**	**87.1**$_{+1.7}$	**62.8**$_{+1.1}$	**51.0**$_{+4.5}$	**30.2**$_{+1.2}$	**57.8**$_{+2.1}$
Olmo-3-7B-Instruct	Base	<u>80.5</u>	52.8	37.8	23.3	48.6
	OPSD	<u>80.5</u>	52.8	38.3	22.4	48.5
	SDPO	80.3	51.3	<u>39.9</u>	**24.5**	49.0
	RLSD	80.4	<u>53.6</u>	38.8	23.8	<u>49.2</u>
	OPSD + EOPD	79.6	53.5	39.2	<u>24.0</u>	49.1
	***E²-OPSD* (Ours)**	**81.2**$_{+0.7}$	**56.3**$_{+3.5}$	**40.0**$_{+2.2}$	23.9$_{+0.6}$	**50.3**$_{+1.7}$

Table 2: **Out-of-domain evaluation on logical reasoning and code generation. *E²-OPSD*** consistently improves both mean@16 and pass@8 across ZebraLogic and HumanEval, showing that its gains extend beyond in-domain mathematical reasoning.

		Logic: ZebraLogic		**Coding: HumanEval**	
Model	**Method**	mean@16	pass@8	mean@16	pass@8
Qwen3-0.6B (Thinking)	Base	28.8	56.5	52.6	78.4
	E²-OPSD	**29.0**$_{+0.2}$	56.5$_{+0.0}$	**53.1**$_{+0.5}$	**79.0**$_{+0.6}$
Qwen3-1.7B (Thinking)	Base	62.5	78.3	83.8	93.2
	E²-OPSD	**67.4**$_{+4.9}$	**83.8**$_{+5.5}$	**84.5**$_{+0.7}$	**93.8**$_{+0.6}$
Qwen3-4B Instruct-2507	Base	79.1	92.4	84.0	89.9
	E²-OPSD	**82.2**$_{+3.1}$	**92.8**$_{+0.4}$	**84.8**$_{+0.8}$	**90.7**$_{+0.8}$
OLMo-3-7B Instruct	Base	32.7	56.6	78.2	92.1
	E²-OPSD	**33.9**$_{+1.2}$	**57.9**$_{+1.3}$	**78.7**$_{+0.5}$	**92.5**$_{+0.4}$

drift upward, creating a persistent entropy gap. In contrast, ***E²-OPSD*** keeps student entropy tightly regulated around the teacher throughout training, preventing runaway diffusion while preserving sufficient exploration. This curbs entropy overshoot and stabilizes distillation.

***E²-OPSD* exhibits an emergent exploration-to-concentration transition.** As shown in Figure 4(c), forward KL dominates early, promoting exploration toward teacher-supported alternatives, while reverse KL becomes increasingly prevalent later, suppressing excess probability mass and concentrating the student distribution. This shift emerges without an explicit schedule and natu-

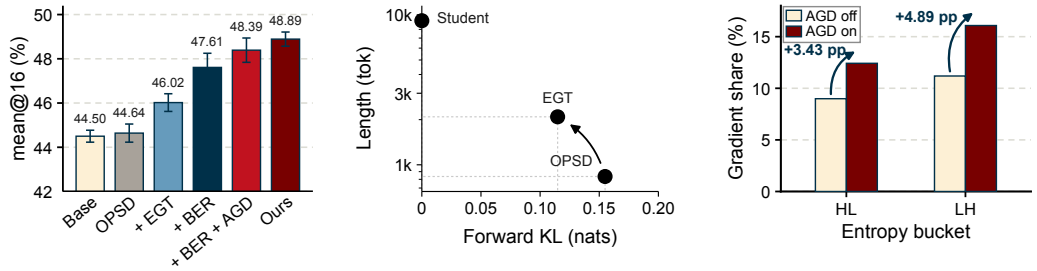


(a) Entropy dynamics on Olmo-3-7B-Instruct.

(b) Entropy dynamics on Qwen3-4B-Instruct-2507.

(c) Emergent KL-routing dynamics.

Figure 4: **Training dynamics.** (a,b) Student and teacher token entropy under OPSD (dashed) and ***E²-OPSD*** (solid); the dotted line marks the first student–teacher entropy crossing under OPSD. (c) On Qwen3-1.7B Thinking, routing share denotes the fraction of response tokens assigned to forward or reverse KL by BER. The routing naturally shifts from forward-dominated exploration toward increased reverse-KL concentration as training proceeds.



(a) Component contributions.

(b) Teacher-context probe.

(c) AGD gradient probe.

Figure 5: **Component ablation and mechanism probes.** (a) Performance contribution of each design component; "+" denotes the component added on top of OPSD. (b) EGT improves teacher–student compatibility. The x-axis denotes teacher-to-student forward KL evaluated on student prefixes; the y-axis shows response length from each setting's own rollouts (log scale). (c) AGD sharpens entropy-aware routing. HL denotes high-student/low-teacher entropy and LH the reverse.

rally tracks the evolving student state. As both routing curves flatten, ***E²-OPSD*** further exhibits a self-equilibrating diffusion–concentration dynamic.

4.3 Ablation Study

Performance breakdown by design components. Figure 5(a) isolates the contribution of each design choice on Qwen3-1.7B Thinking. Starting from OPSD at 44.6 macro mean@16, exemplar-guided teaching raises performance to 46.0, while bidirectional entropy routing reaches 47.6. Adding adaptive gap damping further improves the score to 48.4, and the full combination achieves 48.9. The steady gains suggest that the components provide complementary benefits, improving both the supervision the student receives and how that supervision is applied. For detailed results, please refer to Table 11.

State-dependent KL routing outperforms fixed directions. With the reference-answer teacher fixed and adaptive gap damping disabled, our entropy-aware routing clearly outperforms all other methods that apply fixed KL directions (see Figure 6). To be more precise, forward KL keeps expanding the student and eventually drives entropy overshoot, making the student overly diffuse. Reverse KL suppresses excess mass but also restricts exploration when the student has not yet covered enough plausible alternatives. Their fixed mixture softens both extremes, yet applies expansion and contraction simultaneously regardless of the student's current state, causing the two corrections to partially counteract each other rather than selecting the one that is needed. In contrast, ***E²-OPSD*** uses the student–teacher entropy gap to select the needed correction, expanding when the student is

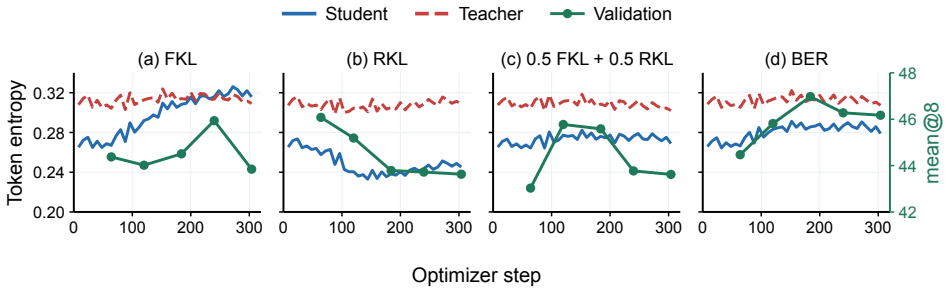


Figure 6: **State-dependent KL routing avoids fixed directional bias.** KL-direction dynamics on Qwen3-1.7B Thinking with the teacher context fixed and AGD disabled. The x-axis shows optimizer steps; the left y-axis reports student and teacher token entropy, while the right y-axis of panel (d) reports validation mean@8. Panels compare (a) forward KL, (b) reverse KL, (c) an equal-weight forward/reverse KL mixture, and (d) BER. The trajectories reveal distinct distributional dynamics under fixed versus state-dependent KL routing.

too concentrated and contracting when it is too diffuse, yielding more stable entropy dynamics and stronger validation performance.

Exemplar guidance makes the teacher locally learnable. Figure 5(b) reveals why replacing the current reference solution matters. When conditioned on the reference answer, the teacher produces much shorter and more confident trajectories than the student, with an entropy of 0.190 nats versus the student's 0.268 nats. EGT instead keeps the teacher closer to the student's reasoning regime, at 2,090 tokens and 0.261 nats. Moreover, EGT reduces teacher–student forward KL by 25.9% (0.1549 to 0.1148 nats) on the same student prefixes. Together, these observations suggest that EGT brings the teacher's native uncertainty closer to the student's reasoning regime while providing more closely aligned token-level guidance on student-visited states.

AGD sharpens entropy-aware routing. Figure 5(c) isolates how AGD redistributes token-level learning signals. Here, high entropy means $H \geq \ln 2$, corresponding to at least two effective next-token choices. Entropy-discordant positions—HL and LH, where student and teacher fall into opposite high/low entropy regimes—account for only 7.67% of all response tokens, yet contribute 20.18% of the relative gradient mass without AGD. Enabling AGD further raises this share to 28.50%, with both HL and LH increasing consistently. This shows that AGD does not uniformly rescale the distillation objective; instead, it makes BER more selective by shifting learning toward positions with clearer student–teacher disagreement in policy spread. To rule out that small entropy gaps merely hide distributional mismatch, we measure probability-mass overlap: strongly damped positions have 0.985 mean overlap (Appendix C), supporting the entropy gap as a simple yet principled proxy for local distributional mismatch. In this sense, BER determines *which direction* to correct, while AGD determines *which routing decisions deserve stronger updates*.

5 Conclusion

We identify entropy overshoot as a key failure mode of on-policy self-distillation: a teacher that has read the answer supplies guidance the student cannot fully learn, and OPSD's forward-KL objective keeps pushing the student outward, turning useful exploration into persistent distributional mismatch. We introduce ***E²-OPSD***, which controls this process by constructing more locally learnable teacher targets and adapting both the direction and the strength of distillation to the student's training state. Across both thinking and instruct models, ***E²-OPSD*** consistently improves model's reasoning capability, with gains transferring to out-of-domain tasks. Our analyses show that effective self-distillation requires matching teacher guidance to student-reachable states and uncertainty structure, rather than simply increasing or suppressing entropy. These results suggest that controlling where exploration occurs is as important as encouraging exploration itself.

AI USE STATEMENT

In this work, we used generative AI tools to polish the writing. We have not used generative AI tools for establishing theoretical models, conceptual frameworks or generate synthetic data sets. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We will release our implementation and experiment scripts to facilitate reproduction of our results. All datasets, model checkpoints, and evaluation benchmarks used in this work are publicly available. Detailed training and evaluation settings, prompts, and method-specific configurations are provided in Appendix. Our method is lightweight and straightforward to implement, introducing no additional forward passes or networks within the standard on-policy self-distillation loop.

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.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Udbhav Bamba, Ming-Hao Fang, Yi-Fan Yu, Haizhong Zheng, and Fan Lai. Xrpo: Pushing the limits of grpo with targeted exploration and exploitation. *ArXiv*, abs/2510.06672, 2025. URL https://api.semanticscholar.org/CorpusID:281891555.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mo Bavarian, Clemens S. Winter, Phil Tillet, Felipe Petroski Such, David W. Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew M. Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *ArXiv*, abs/2107.03374, 2021. URL https://api.semanticscholar.org/CorpusID:235755472.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chengyu Hou, Chenhao Xu, Chenze Shao, Chong Ruan, Conner Sun, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Donghao Li, Dongjie Ji, Erhang Li, Fang Wei, Fangyun Lin, Fangzhou Yuan, Feiyu Xia, Fucong Dai, Guangbo Hao, Guanting Chen, Guoai Cao, Guolai Meng, Guowei Li, Han Yu, Han Zhang, Hanwei Xu, Hao Li, Haofen Liang, Haoling Zhang, Haoming Luo, Haoran Wei, Haotian Yuan, Haowei Zhang, Haowen Luo, Haoyu Chen, Haozhe Ji, Hengqing Zhang, Honghui Ding, Hongxuan Tang, Huanqi Cao, Huazuo Gao, Hui Qu, Hui Zeng, J Yang, JQ Zhu, Jia Luo, Jia Song, Jia Yu, Jialiang Huang, Jialu Cai, Jian Liang, Jiangting Zhou, Jiasheng Ye, Jiashi Li, Jiaxin Xu, Jiewen Hu, Jieyu Yang, Jin Chen, Jin Yan, Jingchang Chen, Jingli Zhou, Jingting Xiang, Jingyang Yuan, Jingyuan Cheng, Jingzi Zhou, Jinhua Zhu, Jiping Yu, Joseph Sun, Jun Ran, Junguang Jiang, Junjie Qiu, Junlong Li, Junmin Zheng, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Kexing Zhou, Kezhao Huang, Kuai Yu, Lean Wang,

Lecong Zhang, Lei Wang, Leyi Xia, Li Zhang, Liang Zhao, Lihua Guo, Lingxiao Luo, Linwang Ma, Linyan Zhu, Litong Wang, Liyu Cai, Liyue Zhang, Longhao Chen, MS Di, MY Xu, Max Mei, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Mingxu Zhou, Minmin Han, Ning Wang, Panpan Huang, Panpan Wang, Peixin Cong, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qingyang Li, Qinyu Chen, Qiushi Du, Qiwei Jiang, Rui Tian, Ruifan Xu, Ruijie Lu, Ruiling Xu, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runqian Chen, Runqiu Yin, Runxin Xu, Ruomeng Shen, Ruoyu Zhang, Ruyi Chen, SH Liu, Shanghao Lu, Shangmian Sun, Shangyan Zhou, Shanhuang Chen, Shaofei Cai, Shaoheng Nie, Shaoqing Wu, Shaoyuan Chen, Shengding Hu, Shengyu Liu, Shiqiang Hu, Shirong Ma, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, Shuying Yu, Songyang Zhou, Tao Ni, Tao Yun, Tian Jin, Tian Pei, Tian Ye, Tianle Lin, Tianran Ji, Tianyi Cui, Tianyuan Yue, Tingting Yu, Tun Wang, W Zhang, WL Xiao, Wangding Zeng, Wei An, Weilin Zhao, Wen Liu, Wenfeng Liang, Wenjie Pang, Wenjing Luo, Wenjing Yao, Wenjun Gao, Wenkai Yang, Wenlve Huang, Wenqing Hou, Wentao Zhang, Wenting Ma, Xi Gao, Xiang He, Xiangwen Wang, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaokang Zhang, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingchen Liu, Xingkai Yu, Xingyou Li, Xinyu Yang, Xinyu Zhang, Xu Chen, Xuanyu Wang, Xuecheng Su, Xueyin Chen, Xuheng Lin, Xuwei Fu, YC Yan, YQ Wang, YW Ma, Yanfeng Luo, Yang Zhang, Yanhong Xu, Yanru Ma, Yanwen Huang, Yao Li, Yao Li, Yao Xu, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Qian, Yi Shao, Yi Yu, Yichao Zhang, Yifan Ding, Yifan Shi, Yijia Wu, Yiliang Xiong, Yiling Ma, Ying He, Ying Tang, Ying Zhou, Yingjia Luo, Yinmin Zhong, Yishi Piao, Yisong Wang, Yixiang Zhang, Yixiao Chen, Yixuan Tan, Yixuan Wei, Yiyang Ma, Yiyuan Liu, Yonglun Yang, Yongqiang Guo, Yongtong Wu, Yu Wu, YuKun Li, Yuan Cheng, Yuan Ou, Yuanfan Xu, Yuanhao Li, Yuduan Wang, Yuehan Yang, Yuer Xu, Yuhan Wu, Yuhao Meng, Yuheng Zou, Yukun Zha, Yunfan Xiong, Yupeng Chen, Yuping Lin, Yuqian Cao, Yuqian Wang, Yushun Zhang, Yuting Yan, Yutong Lin, Yuxian Gu, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuxuan Zhou, Yuyang Zhou, Yuzhen Huang, ZF Wu, Zehao Wang, Zehua Zhao, Zehui Ren, Zekai Zhang, Zhangli Sha, Zhe Fu, Zhe Ju, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zheren Gao, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhixian Huang, Zhixuan Chen, Zhiyu Wu, Zhizhou Ren, Zhongyu Wu, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihua Qu, Zihui Gu, Zijia Zhu, Zilin Li, Zipeng Zhang, Ziwei Xie, Ziyi Gao, Ziyi Wan, Zizheng Pan, and Zongqing Yao. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, Jiajie Zhang, Jian Jiao, Jiaqi Guo, Jingsen Wang, Jingzhao Du, Jinzhu Wu, Kedong Wang, Lei Li, Lin Fan, Lucen Zhong, Mingdao Liu, Mingming Zhao, Pengfan Du, Qian Dong, Rui Lu, Shuang-Li, Shulin Cao, Song Liu, Ting Jiang, Xiaodong Chen, Xiaohan Zhang, Xuancheng Huang, Xuezhen Dong, Yabo Xu, Yao Wei, Yifan An, Yilin Niu, Yitong Zhu, Yuanhao Wen, Yukuo Cen, Yushi Bai, Zhongpei Qiao, Zihan Wang, Zikang Wang, Zilin Zhu, Ziqiang Liu, Zixuan Li, Bojie Wang, Bosi Wen, Can Huang, Changpeng Cai, Chao Yu, Chen Li, Chengwei Hu, Chenhui Zhang, Dan Zhang, Daoyan Lin, Dayong Yang, Di Wang, Ding Ai, Erle Zhu, Fangzhou Yi, Feiyu Chen, Guohong Wen, Hailong Sun, Haisha Zhao, Haiyi Hu, Hanchen Zhang, Hanrui Liu, Hanyu Zhang, Hao Peng, Hao Tai, Haobo Zhang, He Liu, Hongwei Wang, Hongxi Yan, Hongyu Ge, Huan Liu, Huanpeng Chu, Jia'ni Zhao, Jiachen Wang, Jiajing Zhao, Jiamin Ren, Jiapeng Wang, Jiaxin Zhang, Jiayi Gui, Jiayue Zhao, Jijie Li, Jing An, Jing Li, Jingwei Yuan, Jinhua Du, Jinxin Liu, Junkai Zhi, Junwen Duan, Kaiyue Zhou, Kangjian Wei, Ke Wang, Keyun Luo, Laiqiang Zhang, Leigang Sha, Liang Xu, Lindong Wu, Lintao Ding, Lu Chen, Minghao Li, Nianyi Lin, Pan Ta, Qiang Zou, Rongjun Song, Ruiqi Yang, Shangqing Tu, Shangtong Yang, Shaoxiang Wu, Shengyan Zhang, Shijie Li, Shuang Li, Shuyi Fan, Wei Qin, Wei Tian, Weining Zhang, Wenbo Yu, Wenjie Liang, Xiang Kuang, Xiangmeng Cheng, Xiangyang Li, Xiaoquan Yan, Xiaowei Hu, Xiaoying Ling, Xing Fan, Xingye Xia, Xinyuan Zhang, Xinze Zhang, Xirui Pan, Xu Zou, Xunkai Zhang, Yadi Liu, Yandong Wu, Yanfu Li, Yidong Wang, Yifan Zhu, Yijun Tan, Yilin Zhou, Yiming Pan, Ying Zhang, Yinpei Su, Yipeng Geng, Yong Yan, Yonglin Tan, Yuean Bi, Yuhan Shen, Yuhao Yang, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yurong Wu, Yutao Zhang, Yuxi Duan, Yuxuan Zhang, Zezhen Liu, Zhengtao Jiang, Zhenhe Yan, Zheyu Zhang, Zhixiang Wei, Zhuo Chen, Zhuoer Feng, Zijun Yao, Ziwei Chai, Ziyuan Wang, Zuzhou Zhang, Bin Xu, Minlie Huang, Hongning Wang,

Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: On-policy distillation of large language models. *arXiv preprint arXiv:2306.08543*, 2023.

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.

Zihao Han, Tiangang Zhang, Huaibin Wang, and Yilun Sun. Adaptive teacher exposure for self-distillation in llm reasoning. *arXiv preprint arXiv:2605.11458*, 2026.

Zhezheng Hao, Hong Wang, Haoyang Liu, Jian Luo, Jiarui Yu, Hande Dong, Qiang Lin, Can Wang, and Jiawei Chen. Rethinking entropy interventions in rlvr: An entropy change perspective. *ArXiv*, abs/2510.10150, 2025. URL https://api.semanticscholar.org/CorpusID:282057600.

Sarthak Harne, Chinmay Karkar, Yash Pandya, Ahmed Awadallah, and Akshay Nambi. Privileged, but biased: How pi-conditioned teachers break self-distillation. *arXiv preprint arXiv:2608.04794*, 2026.

Harvard–MIT Mathematics Tournament. HMMT February 2025: Problems and solutions, 2025. URL https://www.hmmt.org/www/archive/282.

Yinghui He, Simran Kaur, Adithya Bhaskar, Yongjin Yang, Jiarui Liu, Narutatsu Ri, Liam Fowl, Abhishek Panigrahi, Danqi Chen, and Sanjeev Arora. Self-distillation zero: Self-revision turns binary rewards into dense supervision. *arXiv preprint arXiv:2604.12002*, 2026.

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. *ArXiv*, abs/2504.11456, 2025. URL https://api.semanticscholar.org/CorpusID:277787455.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Yubo Jiang, Fengying Xie, Zhiguo Jiang, and Haopeng Zhang. Distill skills into weights, not prompts: Abstract skills as privileged signals for on-policy self-distillation. *arXiv preprint arXiv:2608.09826*, 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. *arXiv preprint arXiv:2603.07079*, 2026.

Durgesh Kalwar, Vardhan Palod, and Subbarao Kambhampati. Masked distillation: Internalizing the chain-of-thought in language models. *arXiv preprint arXiv:2607.22629*, 2026.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking on-policy self-distillation for thinking models. *arXiv preprint arXiv:2607.05184*, 2026.

Junlong Ke, Zichen Wen, Weijia Li, Conghui He, and Linfeng Zhang. Respecting self-uncertainty in on-policy self-distillation for efficient llm reasoning. *arXiv preprint arXiv:2605.13255*, 2026.

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? *arXiv preprint arXiv:2603.24472*, 2026.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention, 2023. URL https://arxiv.org/abs/2309.06180.

Chia-Hsuan Lee, Zelei Cheng, Yu Wang, Renkun Ni, Sambit Sahu, Shi-Xiong Zhang, and William Campbell. Sead: Competence-aware on-policy distillation via entropy-guided supervision. *arXiv preprint arXiv:2606.28562*, 2026.

Jiazheng Li, Hongzhou Lin, Hong Lu, Kaiyue Wen, Zaiwen Yang, Jiaxuan Gao, Yi Wu, and Jingzhao Zhang. Questa: Expanding reasoning capacity in llms via question augmentation, 2026a. URL https://arxiv.org/abs/2507.13266.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan ang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe, 2026b. URL https://arxiv.org/abs/2604.13016.

Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. Zebralogic: On the scaling limits of llms for logical reasoning. *ArXiv*, abs/2502.01100, 2025. URL https://api.semanticscholar.org/CorpusID:276094621.

Ye Liu, Srijan Bansal, Bo Pang, Yang Li, Zeyu Leo Liu, Yifei Ming, Zixuan Ke, Shafiq Joty, and Semih Yavuz. Procedural memory distillation: Online reflection for self-improving language models. *arXiv preprint arXiv:2607.01480*, 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.

math-ai. AMC 2023, 2023. URL https://huggingface.co/datasets/math-ai/amc23. Dataset compiled from the 2023 American Mathematics Competitions.

Thomas P. Minka. Divergence measures and message passing. 2005. URL https://api.semanticscholar.org/CorpusID:7585417.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2026. URL https://arxiv.org/abs/2512.13961.

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. Rlcsd: Reinforcement learning with contrastive on-policy self-distillation. *arXiv preprint arXiv:2606.11709*, 2026.

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. *arXiv preprint arXiv:2601.19897*, 2026.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, pp. 1279–1297. ACM, March 2025. doi: 10.1145/3689031.3696075. URL http://dx.doi.org/10.1145/3689031.3696075.

Haebin Shin, Lei Ji, Yeyun Gong, Sungdong Kim, Eunbi Choi, and Minjoon Seo. Generative prompt internalization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7338–7363, 2025.

Makoto Shing, Kou Misaki, Han Bao, Sho Yokoi, and Takuya Akiba. Taid: Temporally adaptive interpolated distillation for efficient knowledge transfer in language models, 2025. URL https://arxiv.org/abs/2501.16937.

Charlie Snell, Dan Klein, and Ruiqi Zhong. Learning by distilling context. *arXiv preprint arXiv:2209.15189*, 2022.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning, 2025. URL https://arxiv.org/abs/2506.01939.

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.

Zilin Xiao, Qi Ma, Chun cheng Jason Chen, Xintao Chen, Avinash Atreya, Hanjie Chen, and Vicente Ordonez. Learning to reason by analogy via retrieval-augmented reinforcement fine-tuning, 2026. URL https://arxiv.org/abs/2606.13680.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance, 2025. URL https://arxiv.org/abs/2504.14945.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

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

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026.

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

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.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *ArXiv*, abs/2506.05176, 2025. URL https://api.semanticscholar.org/CorpusID:279243736.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2024, 2024.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2025, 2025.

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

A Experimental Details

This section expands the experimental protocol in Section 4.1. Complete results and standard deviations are given in Appendix B.

A.1 Data and Model Configurations

Training prompts come from DeepMath-103K (He et al., 2025), filtered to the difficulty bands in Table 3. Within each configuration, the same reasoning mode is used for the student, teacher, and evaluation. The Instruct configurations use Qwen3-4B-Instruct-2507 and Olmo-3-7B-Instruct.

Table 3: Backbone configurations and DeepMath training difficulty bands.

Backbone	Reasoning mode	Difficulty levels
Qwen3-0.6B	Thinking	1–3
Qwen3-1.7B	Thinking	5–7
Qwen3-4B-Instruct-2507	Instruct	6–8
Olmo-3-7B-Instruct	Instruct	7–10

For in-domain mathematical reasoning, we evaluate on AMC23, AIME'24, AIME'25, and HMMT'25 Feb. For out-of-domain evaluation, we use ZebraLogic for logical reasoning and HumanEval for code generation. We construct a fixed subset of 100 ZebraLogic puzzles by sampling four puzzles per grid size, using the same subset for all configurations.

A.2 Shared Training Parameters and Budgets

All methods and ablations use veRL with vLLM rollouts on a single 8×H200 node. They share the training settings in Table 4 and a budget of 360 optimizer steps.

Table 4: Shared training parameters.

Parameter	Value
Optimizer	AdamW
Learning rate	1×10^{-6}
Warmup	100 steps
Weight decay	0.01
Gradient-norm clipping	0.1
Prompts per rollout round	256
Responses per prompt	$G = 8$
Mini-batch size	32 prompt groups
Optimizer updates per round	8
Rollout temperature	1.0
Rollout top-p / top-k	0.95 / 20
Maximum rollout response length	16,384 tokens

A.3 Evaluation Protocol

We evaluate models with the settings in Table 5. We report mean@16 averaged over three evaluation seeds. Evaluation uses the student alone with the original question and the backbone's corresponding reasoning mode.

A.4 Method-Specific Implementation and Exemplars

Table 6 summarizes the settings that differ across methods. OPSD + EOPD adapts EOPD (Jin et al., 2026) by using OPSD's reference-conditioned self-teacher. Teacher forward passes are gradient-

Table 5: Evaluation parameters. Batch sizes count prompts.

Group	Parameter	Value
Evaluation	Responses per problem	16
	Evaluation seeds	17, 42, 1234
	Temperature	0.6
	Top-p / top-k	0.95 / 20
	Maximum prompt length	2,048 tokens
Maximum response length	Math reasoning	38,912 tokens
	Logic (ZebraLogic)	38,912 tokens
	Coding (HumanEval)	16,384 tokens
Evaluation batch size	Math reasoning	4 prompts
	Logic (ZebraLogic)	16 prompts
	Coding (HumanEval)	8 prompts

free; EMA and snapshot updates follow the schedules below. Entropy is computed over the full vocabulary.

Table 6: Method-specific teacher and distillation settings.

Method	Setting	Value	Description
OPSD	Teacher	Frozen initial weights	Current reference solution as context.
	Objective	Forward KL	Teacher-to-student distillation.
	Support / clipping	Top-100 + tail; 0.05	Teacher-selected support; per-element cap.
SDPO	Teacher / EMA decay	EMA; 0.95	Updated after each optimizer step.
	Teacher context	Successful sibling	Excludes the rollout being scored.
	JSD mixing α	0.5	Equal student–teacher mixture.
	Support	Top-100 + tail	Student-selected support.
	Importance-ratio cap	2.0	Upper cap on the distillation ratio.
RLSD	Teacher / refresh	Snapshot; every 10 steps	Copies student weights every 10 optimizer steps.
	PPO / evidence clips	$\epsilon = 0.2$; $\epsilon_w = 0.2$	Policy-ratio and evidence-weight clipping.
	Modulation λ	$0.5 \to 0$	Linear decay over 50 training steps.
OPSD + EOPD	Teacher	Frozen initial weights	Current reference solution as context.
	Entropy gate	$H_T \geq 0.8$ nats	Adds forward KL to sampled reverse KL.
	Forward-KL coefficient	1.0	Weight of the entropy-gated term.
	Forward-KL support	Top-16; no tail	Renormalized teacher, full-vocabulary student; no per-element cap.

Continued on next page

Table 6 (continued)

Method	Setting	Value	Description
E²-OPSD	Teacher	Frozen initial weights	One retrieved solved neighbor as context.
	Objective	BER + AGD	Entropy gap sets direction and weight.
	Support / clipping	Top-100 + tail; 0.05	Teacher-selected support; per-element cap.
	Exemplar candidates	15 neighbors	Self-excluded; one draw per rollout.

Exemplar construction and teacher prompt. Using Qwen3-Embedding-8B (Zhang et al., 2025), we build a self-excluded 15-neighbor index within each difficulty band. Each rollout draws one neighbor and its dataset reference solution, with sampling seeded by the training seed, step, problem, and rollout slot. The box reproduces the fixed user message with variable fields in braces. After applying the backbone's native chat template and reasoning mode, the teacher scores the student's continuation conditioned on {student_prefix}. Exemplars are used only during training. For Qwen3-1.7B, one-time embedding and neighbor-index construction took 219 seconds on eight H200 GPUs, accounting for 0.41% of the GPU time spent on rollout generation and actor updates over 360 optimizer steps.

EGT teacher prompt

```
Problem: {current_problem}

Here is a reference solution to a DIFFERENT but related
problem. It is not the problem you must solve, and its final
answer is NOT the answer to your problem -- use it only as a
guide to the strategies and techniques that transfer:

=== Reference Solution Begin ===
Related problem: {related_problem}

Solution to the related problem:
{related_solution}
=== Reference Solution End ===

After reading the reference solution above, make sure you
understand the reasoning behind each step, then solve the
original problem with whatever transfers from it.

Please reason step by step, and put your final answer within
\boxed{}.
```

Assistant continuation: {student_prefix}
The teacher scores the next token after this prefix; this line is not part of the user message.

B Full Evaluation Results

B.1 Evaluation Metrics

For N problems, let $c_i(m)$ be the number of correct answers among m independently sampled responses to problem i. We compute

$$\text{mean@}k = \frac{1}{N}\sum_{i=1}^{N}\frac{c_i(k)}{k}, \qquad \widehat{\text{pass@}k} = \frac{1}{N}\sum_{i=1}^{N}\left[1 - \frac{\binom{n-c_i(n)}{k}}{\binom{n}{k}}\right].$$

Here $k \le n = 16$, with $\binom{a}{k} = 0$ for $a < k$. Estimated pass@1 equals mean@16, so its duplicate table is omitted. Entries are mean $\pm$ population standard deviation (%) over three seeds; bold/underlined means indicate best/second-best distinct results.

B.2 In-domain Mathematical Reasoning

Table 7: Mathematical reasoning mean@16 (%). Avg. equally weights the four benchmarks.

		Math Reasoning (mean@16)				
Model	**Method**	AMC23	AIME24	AIME25	HMMT25	Avg.
Qwen3-0.6B (Thinking)	Base	45.81±0.98	9.79±0.29	14.17±1.12	8.26±0.43	19.51±0.19
	OPSD	45.88±0.86	10.90±0.86	14.17±0.59	**9.17**±0.51	20.03±0.49
	SDPO	45.48±0.27	9.38±0.34	14.58±1.39	7.78±1.70	19.30±0.62
	RLSD	**46.84**±1.95	<u>10.97</u>±0.49	15.14±0.64	7.78±0.26	<u>20.18</u>±0.53
	OPSD + EOPD	46.36±0.62	9.79±0.95	<u>15.90</u>±0.35	8.54±1.03	20.15±0.24
	***E²-OPSD* (Ours)**	<u>46.56</u>±1.37	**11.04**±0.29	**16.18**±0.35	<u>8.89</u>±1.39	**20.67**±0.42
Qwen3-1.7B (Thinking)	Base	75.50±0.28	46.04±1.78	34.58±0.29	21.88±0.78	44.50±0.27
	OPSD	74.87±0.46	45.97±1.58	35.00±1.36	<u>22.71</u>±0.61	44.64±0.41
	SDPO	<u>76.05</u>±0.98	47.99±2.04	<u>35.28</u>±1.72	21.88±1.28	<u>45.30</u>±0.29
	RLSD	75.85±0.20	<u>48.26</u>±0.10	33.96±1.39	21.81±0.55	44.97±0.52
	OPSD + EOPD	74.27±0.15	47.43±1.28	34.24±1.21	21.46±0.78	44.35±0.09
	***E²-OPSD* (Ours)**	**79.02**±0.41	**52.29**±0.45	**39.24**±0.26	**25.00**±0.90	**48.89**±0.32
Qwen3-4B-Instruct-2507	Base	85.39±0.48	61.74±0.49	46.46±2.04	29.03±0.71	55.65±0.79
	OPSD	84.81±0.09	<u>62.36</u>±1.39	45.63±1.47	<u>29.51</u>±0.84	55.58±0.52
	SDPO	83.86±0.69	61.32±1.32	45.83±1.45	27.50±1.06	54.63±0.66
	RLSD	85.12±0.84	62.22±0.94	<u>46.94</u>±0.84	28.40±0.26	<u>55.67</u>±0.30
	OPSD + EOPD	<u>85.47</u>±0.49	62.01±2.21	46.46±0.29	28.40±0.26	55.59±0.71
	***E²-OPSD* (Ours)**	**87.10**±0.63	**62.85**±1.09	**50.97**±1.13	**30.21**±0.61	**57.78**±0.34
Olmo-3-7B-Instruct	Base	<u>80.52</u>±0.29	52.85±1.21	37.78±0.20	23.26±0.35	48.60±0.39
	OPSD	<u>80.52</u>±0.40	52.78±1.32	38.26±1.79	22.36±0.49	48.48±0.69
	SDPO	80.27±0.12	51.32±0.79	<u>39.93</u>±1.74	**24.51**±1.43	49.01±0.14
	RLSD	80.37±0.51	<u>53.61</u>±1.09	38.82±0.10	23.82±0.26	<u>49.16</u>±0.31
	OPSD + EOPD	79.62±0.47	53.47±1.19	39.17±2.17	<u>23.96</u>±1.47	49.05±0.37
	***E²-OPSD* (Ours)**	**81.22**±1.05	**56.25**±1.35	**40.00**±2.13	23.89±0.35	**50.34**±0.75

Table 8: Mathematical reasoning unbiased pass@2 (%).

Model	Method	**Math Reasoning** (pass@2) AMC23	AIME24	AIME25	HMMT25	Avg.
Qwen3-0.6B (Thinking)	Base	56.41±0.97	14.05±1.10	19.74±1.63	11.48±0.17	25.42±0.35
	OPSD	56.63±1.46	16.39±1.23	19.95±1.05	**12.79**±0.80	26.44±0.76
	SDPO	56.11±0.10	14.44±0.80	20.19±1.88	10.79±2.17	25.38±0.70
	RLSD	**57.57**±2.43	16.79±0.73	21.13±0.36	10.72±0.40	26.55±0.39
	OPSD + EOPD	56.52±0.56	14.85±2.14	21.94±0.24	11.03±1.75	26.08±0.31
	***E²-OPSD* (Ours)**	56.81±0.77	**17.17**±0.82	**22.28**±0.54	12.20±1.39	**27.11**±0.37
Qwen3-1.7B (Thinking)	Base	82.44±0.36	59.14±1.90	41.86±0.19	25.96±1.24	52.35±0.24
	OPSD	82.34±0.48	58.31±1.94	43.56±1.89	27.72±1.12	52.99±0.79
	SDPO	82.85±0.72	59.87±2.21	42.85±2.48	26.75±2.21	53.08±0.70
	RLSD	83.14±0.39	61.35±0.21	41.59±2.17	26.10±1.01	53.05±0.66
	OPSD + EOPD	81.14±0.85	60.64±1.71	41.79±1.00	25.77±1.21	52.33±0.14
	***E²-OPSD* (Ours)**	**86.10**±0.28	**62.86**±0.63	**49.32**±0.98	**29.95**±1.08	**57.06**±0.52
Qwen3-4B-Instruct-2507	Base	89.44±0.20	71.11±0.17	56.45±2.38	33.00±0.93	62.50±0.82
	OPSD	89.14±0.41	71.94±1.83	56.09±1.29	**34.26**±1.42	62.86±0.68
	SDPO	88.79±0.19	71.28±0.33	55.66±2.04	31.97±1.21	61.92±0.71
	RLSD	89.54±0.83	**72.16**±0.19	56.64±0.95	32.80±0.21	62.78±0.17
	OPSD + EOPD	89.60±0.40	71.25±2.79	55.30±1.00	32.44±0.48	62.14±0.64
	***E²-OPSD* (Ours)**	**90.92**±0.39	70.65±0.76	**60.89**±1.48	34.25±0.65	**64.18**±0.33
Olmo-3-7B-Instruct	Base	87.40±0.30	63.84±0.65	47.07±0.73	28.08±0.92	56.60±0.50
	OPSD	87.37±0.41	63.90±1.77	46.72±1.64	26.38±0.31	56.09±0.93
	SDPO	87.38±0.13	62.15±1.27	**49.31**±1.54	**29.88**±0.94	57.18±0.65
	RLSD	87.53±0.45	63.67±1.76	48.24±0.81	29.07±0.34	57.13±0.40
	OPSD + EOPD	87.00±0.28	63.10±0.77	48.56±2.67	28.94±1.52	56.90±0.54
	***E²-OPSD* (Ours)**	**87.92**±0.81	**67.39**±1.35	48.86±1.97	28.86±0.86	**58.26**±0.59

Table 9: Mathematical reasoning unbiased pass@4 (%).

Model	Method	**Math Reasoning** (pass@4) AMC23	AIME24	AIME25	HMMT25	Avg.
Qwen3-0.6B (Thinking)	Base	66.30±0.67	19.70±2.15	24.67±1.98	15.02±0.31	31.42±0.71
	OPSD	66.28±1.94	23.39±2.33	25.01±1.61	**16.84**±1.19	32.88±1.12
	SDPO	65.01±0.57	21.34±1.46	25.35±2.85	14.25±2.30	31.49±0.91
	RLSD	**66.82**±2.26	23.79±1.18	27.00±0.90	14.42±0.77	33.01±0.01
	OPSD + EOPD	65.42±0.67	21.59±3.90	27.59±0.92	13.86±2.47	32.11±0.75
	***E²-OPSD* (Ours)**	65.85±0.28	**25.02**±1.33	**27.65**±0.67	16.09±1.30	**33.65**±0.16
Qwen3-1.7B (Thinking)	Base	87.14±0.43	68.65±2.28	48.73±0.64	31.31±2.08	58.96±0.75
	OPSD	87.65±0.31	67.16±2.89	51.35±2.24	33.75±1.32	59.97±1.17
	SDPO	87.72±0.87	69.17±1.62	49.76±2.75	32.14±2.90	59.69±0.87
	RLSD	88.02±0.72	**70.87**±0.65	49.51±1.84	31.29±1.54	59.92±0.48
	OPSD + EOPD	85.68±1.81	**70.87**±2.23	49.64±1.00	31.13±2.25	59.33±0.40
	***E²-OPSD* (Ours)**	**90.91**±0.73	70.54±1.37	**57.71**±1.98	**34.83**±1.26	**63.50**±0.99
Qwen3-4B-Instruct-2507	Base	91.67±0.12	77.81±0.24	64.22±2.16	36.89±1.13	67.65±0.64
	OPSD	91.36±0.77	78.85±1.39	64.51±1.05	**38.20**±2.27	68.23±0.72
	SDPO	91.75±0.48	78.45±1.12	62.88±2.11	36.11±1.10	67.30±0.71
	RLSD	91.99±0.94	**79.73**±0.65	64.67±1.17	36.79±0.69	68.30±0.27
	OPSD + EOPD	91.93±0.33	77.64±3.05	61.66±1.89	36.41±1.15	66.91±0.34
	***E²-OPSD* (Ours)**	**93.00**±0.52	76.29±0.32	**67.94**±1.38	38.04±0.97	**68.82**±0.36
Olmo-3-7B-Instruct	Base	90.69±0.33	73.01±1.02	55.42±1.90	32.72±1.30	62.96±0.74
	OPSD	90.76±0.72	73.33±1.87	54.37±1.36	29.51±0.54	61.99±1.07
	SDPO	90.77±0.25	72.20±1.28	**57.58**±1.42	**34.62**±1.90	63.79±0.86
	RLSD	91.04±0.54	72.45±1.66	56.75±0.87	33.39±1.12	63.41±0.63
	OPSD + EOPD	91.02±0.13	71.88±0.18	56.71±2.54	33.55±0.92	63.29±0.63
	***E²-OPSD* (Ours)**	**91.25**±0.66	**76.34**±1.44	56.06±1.78	33.49±1.73	**64.28**±0.74

B.3 Out-of-domain Evaluation

Table 10: Out-of-domain evaluation on logical reasoning and code generation. ZebraLogic uses whole-puzzle correctness; best means are bold within each model and dataset.

Model	Method	mean@16	pass@2	pass@4	pass@8
Logic: ZebraLogic					
Qwen3-0.6B (Thinking)	Base	28.83±0.30	38.93±0.35	48.54±0.23	56.47±0.63
	***E²-OPSD* (Ours)**	**28.96**±0.21	**39.27**±0.32	**48.86**±0.53	**56.54**±0.71
Qwen3-1.7B (Thinking)	Base	62.54±1.05	69.14±0.99	74.19±1.21	78.26±1.32
	***E²-OPSD* (Ours)**	**67.35**±0.43	**73.57**±0.62	**78.75**±0.84	**83.77**±0.98
Qwen3-4B-Instruct-2507	Base	79.15±0.16	85.65±0.23	89.61±0.41	92.35±0.64
	***E²-OPSD* (Ours)**	**82.23**±0.21	**87.57**±0.10	**90.65**±0.35	**92.83**±0.82
Olmo-3-7B-Instruct	Base	32.69±0.39	42.38±0.37	50.61±0.25	56.56±0.05
	***E²-OPSD* (Ours)**	**33.90**±0.43	**43.96**±0.53	**52.16**±0.71	**57.94**±0.95
Coding: HumanEval					
Qwen3-0.6B (Thinking)	Base	52.55±0.31	63.96±0.45	72.60±0.34	78.42±0.48
	***E²-OPSD* (Ours)**	**53.09**±0.70	**64.26**±0.76	**72.84**±0.71	**79.01**±0.72
Qwen3-1.7B (Thinking)	Base	83.75±0.21	88.91±0.22	91.55±0.17	93.17±0.48
	***E²-OPSD* (Ours)**	**84.50**±0.13	**89.12**±0.19	**91.79**±0.34	**93.77**±0.34
Qwen3-4B-Instruct-2507	Base	83.96±0.16	86.41±0.24	88.37±0.36	89.90±0.51
	***E²-OPSD* (Ours)**	**84.76**±0.22	**87.13**±0.03	**88.92**±0.12	**90.67**±0.20
Olmo-3-7B-Instruct	Base	78.18±0.51	85.33±0.47	89.68±0.44	92.05±0.25
	***E²-OPSD* (Ours)**	**78.66**±0.19	**85.78**±0.34	**90.14**±0.52	**92.52**±0.74

B.4 Ablation Results

Table 11: Per-benchmark mean@16 with standard deviations for the component ablation on Qwen3-1.7B Thinking (Figure 5(a)). Avg. equally weights the four benchmarks. Best means are bold and second-best distinct means underlined. AGD denotes adaptive gap damping.

Method	AMC23	AIME24	AIME25	HMMT25	Avg.
Base	75.50 ±0.28	46.04 ±1.78	34.58 ±0.29	21.88 ±0.78	44.50 ±0.27
OPSD	74.87 ±0.46	45.97 ±1.58	35.00 ±1.36	22.71 ±0.61	44.64 ±0.41
+ EGT	76.46 ±0.99	48.26 ±1.21	36.81 ±1.49	22.57 ±0.20	46.02 ±0.40
+ BER	77.26 ±0.38	52.01 ±1.13	37.15 ±0.94	24.03 ±0.35	47.61 ±0.64
+ BER + AGD	<u>77.66</u> ±0.68	<u>52.15</u> ±1.58	**39.44** ±0.69	<u>24.31</u> ±1.83	<u>48.39</u> ±0.55
***E²-OPSD* (Ours)**	**79.02** ±0.41	**52.29** ±0.45	<u>39.24</u> ±0.26	**25.00** ±0.90	**48.89** ±0.32

C Mechanism Probe Settings and Results

C.1 Rollout and Probe Settings

Figures 5(b,c) use the same diagnostic batch of student responses. We sample 256 distinct problems from the DeepMath level 5–7 training pool with seed 42 and generate eight responses per problem. All three conditions use the same frozen Qwen3-1.7B base weights in Thinking mode, with no parameter updates. The student receives only the original problem; the reference-conditioned teacher additionally receives its reference solution and answer; the EGT teacher instead receives one retrieved solved neighbor. We retain the same problems and exemplar assignments when generating the teachers' own responses.

Table 12: Diagnostic rollout settings. Indices i and g start at zero. The two teacher conditions share sampling seeds. Teacher generation budgets are also limited by the remaining context window.

Parameter	Value
Problems / responses per problem	256 / 8
Sampling temperature	1.0
Top-p / top-k	0.95 / 20
Maximum response length	16,384 tokens
Maximum total context length	40,960 tokens
Student prompt length limit	2,048 tokens
Student sampling seed	$42 + i$ for problem index i
Teacher sampling seed	$42 + 8i + g$ for response slot g
Entropy and KL scoring temperature	1.0

Teacher-context probe. For Figure 5(b), we score each teacher both on its own generated response and on the corresponding student response. Entropy and teacher-to-student forward KL are computed over the full vocabulary, without top-k truncation or clipping. We average over response tokens within each trajectory, then equally over 2,047 matched response slots; one slot is excluded because the EGT prompt and student response exceed the context window. Response lengths use all 2,048 generations per condition and include generated end-of-sequence tokens. Relative entropy increase is the difference between the two mean entropies divided by the own-prefix mean, rather than an average of per-response ratios.

AGD gradient probe. For Figure 5(c), we reuse all 2,048 student responses with the reference-conditioned teacher. At each response position, we compute the ℓ_2 norm of the routed loss gradient with respect to the full student logit vector, using teacher top-100 support plus a tail bucket and a per-element cap of 0.05. We compare the same fixed distributions with AGD off and on, stopping gradients through the routing decision and damping weight. The gap scale is computed over valid tokens in fixed batches of 32 responses, ordered by problem and response slot. Gradient shares are normalized within each response, averaged over responses within a problem, and then over problems. High entropy means $H \geq \ln 2$; the first and second letters in HH, HL, LH, and LL refer to student and teacher entropy, respectively. This is an offline logit-gradient probe, not a replay of parameter updates during training.

C.2 Detailed Probe Results

Table 13 reports the teacher-context comparison. EGT reduces forward KL on the same student prefixes by 25.9%. Own-prefix entropy and response length describe each condition's separate trajectories; the forward-KL comparison uses shared student prefixes.

Table 13: Teacher-context probe underlying Figure 5(b). Lengths use 2,048 responses per condition; entropy and KL use 2,047 matched slots. Student zeros denote identical-prefix or self comparisons.

Metric	Student	Reference teacher	EGT teacher
Mean response length (tokens)	9,139.7	835.6	2,089.5
Own-prefix entropy (nats)	0.2679	0.1900	0.2605
Forward KL to student (nats)	0	0.1549	0.1148

Table 14 reports all four entropy buckets for Figure 5(c). HL and LH together contain 1,436,048 of 18,718,121 response tokens (7.67%). Their gradient shares increase by 3.43 and 4.89 percentage points, respectively.

Table 14: AGD gradient probe on 2,048 fixed student responses. Token shares pool all response tokens; gradient shares use the response- and problem-level normalization described in Appendix C.1. Changes are computed from the displayed gradient shares.

Bucket	Token share (%)	AGD off (%)	AGD on (%)	Change (pp)
HH	14.40	45.57	43.26	-2.31
HL	2.33	8.99	12.42	$+3.43$
LH	5.34	11.19	16.08	$+4.89$
LL	77.93	34.25	28.24	-6.01

Distributional agreement at strongly damped positions. Using the same reference-teacher gradient probe, we compute full-vocabulary probability-mass overlap $O_t = \sum_v \min\{p_t(v), q_t(v)\}$ on student prefixes. Table 15 reports token-pooled statistics for strongly damped positions ($c_t \le 0.1$).

Table 15: Overlap at strongly damped positions. Percentages are computed within each row.

Subset	Positions	Mean overlap	$O_t < 0.5$ (%)
All strongly damped positions	12,708,962	0.9848	0.380
with $H(p_t) \ge \ln 2$	766,682	0.8343	3.662

Positions with both $c_t \le 0.1$ and $O_t < 0.5$ account for 0.464% of $\sum_t g_t$ pooled over all response tokens, where $g_t = \|\nabla_{z_t} D_t^{\mathrm{BER}}\|_2$ is the logit-gradient norm before damping.