ON-POLICY SELF-DISTILLATION WITHOUT ANY SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy (Self-)Distillation (OPD / OPSD) has shown strong potential for post-training large language models (LLMs). However, existing methods still rely heavily on external supervision, including ground-truth signals, environmental feedback, or guidance from larger models, and therefore fall short of genuine "self"-distillation. In this study, we show that on-policy self-distillation can be achieved using only a model's own generations via internal consistency. We propose unsupervised on-policy self-distillation (U-OPSD). U-OPSD first samples multiple rollouts and constructs a pseudo solution by majority vote under a self-consistency threshold. It then conditions the model's distribution on the pseudo-solution and distills itself on the disagreeing completions, allowing the model to correct itself precisely where it is confidently wrong. Across diverse benchmarks, base models, and training settings, U-OPSD consistently improves over the base models and matches or surpasses supervised methods with ground truth (GT) such as OPSD and GRPO. On five mathematical reasoning benchmarks, i.e., AIME24, AIME25, HMMT25, MATH500, and AMC23, U-OPSD improves over the base model by 8.5% and 10.7% on Qwen3 non-thinking mode at 4B and 8B scales, and outperforms OPSD by 3.2% and 2.3% on average, respectively. In thinking mode, U-OPSD stays on par with OPSD, ahead by 0.9% at 4B and level at 8B and surpassing GRPO by 0.7% and 1.1%, respectively.

1 INTRODUCTION

Post-training has emerged as a key driver of advances in the reasoning capabilities of large language models (LLMs), with progress largely propelled by supervised fine-tuning (SFT) (Ye et al., 2025; Wen et al., 2025; Li et al., 2026), knowledge distillation from stronger teachers (Abdin et al., 2025; Lu and Lab, 2025), and reinforcement learning with verifiable rewards (RLVR) (MiniMax et al., 2025; Guo et al., 2025; Team et al., 2025).

Among the many recipes, *on-policy distillation* bridges SFT and RL by training on the model's own generations, reducing the train-inference mismatch (Gu et al., 2024; Agarwal et al., 2024) and catastrophic forgetting (Shenfeld et al., 2026b) of teacher-forced SFT, while retaining dense token-level supervision instead of sparse scalar rewards. Subsequent work has continued the effort of OPD along its objective, supervision, and systems dimensions. DistiLLM (Ko et al., 2024) introduces skewed KL, while DistiLLM-2 applies asymmetric objectives to teacher- and student-generated responses (Ko et al., 2025). Other work stabilizes long-horizon OPD by restricting supervision to teacher-supported tokens and masking unreliable signals (Fu et al., 2026), or reduces its systems overhead by precomputing teacher scores under a teacher-consistent offline pipeline (Wu et al., 2026). More recent analyses further identify teacher–student mismatch and length exploitation as key failure modes, motivating clipped and log-compressed token-level guidance (Wang et al., 2026). Nevertheless, these methods still heavily rely on a separate strong teacher model for guidance.

On-policy *self*-distillation (OPSD) (Zhao et al., 2026a) pushes this further by removing the need for a stronger teacher model – a single LLM plays both roles, where the teacher (student itself) conditions on the ground-truth solution, whereas the student sees only the problem. Subsequent work identifies information leakage and instability (Yang et al., 2026), redesigns the privileged context (Ye et al., 2026; Penaloza et al., 2026; Sang et al., 2026), and extends it to agentic setups (Liu et al., 2026b; Yang et al., 2026).

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.

The privileged context, however, is also its own bottleneck. Existing OPD and OPSD methods rely on external supervision, such as ground-truth solutions, environmental feedback, or guidance from larger models. Thus, the model is self-distilled only in the sense that the teacher and student share parameters; the information that makes the teacher more capable still comes from outside the model. This dependence limits scalability to unlabeled problems and restricts applicability in domains where supervision is costly, unreliable, or unavailable. We therefore ask: *Does the teacher in on-policy self-distillation need a ground-truth solution? Can a model construct its own privileged context and perform genuine self-distillation?*

We answer with U-OPSD, an unsupervised on-policy self-distillation method that requires no external supervision. Our key observation is that although an individual rollout may be unreliable, agreement among multiple independently sampled rollouts provides an endogenous confidence signal. This makes it possible to derive both the teacher reference and the student trajectories entirely from the model's own on-policy samples. Formally, for each unlabeled problem $x \sim \mathcal{U}$, we sample G independent rollouts $y^{(1)}, \ldots, y^{(G)} \overset{\text{i.i.d.}}{\sim} \bar{\pi}(\cdot \mid x)$, where $\mathcal{U}$ is the unlabeled problem distribution, G is the number of rollouts per problem, and $\bar{\pi} \triangleq \pi_{\text{sg}[\theta]}$ denotes a stop-gradient copy of the current policy π_θ. We extract the final answers from each rollout as $a^{(g)} = \text{Ans}(y^{(g)})$, where $\text{Ans}(\cdot)$ denotes the answer-extraction and canonicalization function, and obtain the vote-based *pseudo-answer* $\tilde{a}(x) = \arg\max_a \sum_{g=1}^{G} \mathbb{1}\big[a^{(g)} = a\big]$, where $\mathbb{1}[\cdot]$ is the indicator function.

The rollouts are then partitioned into agreeing and disagreeing multisets $\mathcal{Y}_x^+ = \{y^{(g)} : a^{(g)} = \tilde{a}(x)\}$, $\mathcal{Y}_x^- = \{y^{(g)} : a^{(g)} \neq \tilde{a}(x)\}$. When the winning vote fraction reaches a self-consistency threshold τ, the longest agreeing rollout $y^+ \in \arg\max_{y \in \mathcal{Y}_x^+} |y|$ serves as the *pseudo-solution*, while the disagreeing rollouts in $\mathcal{Y}_x^-$ serve as the student trajectories. By distilling the teacher distribution conditioned on the pseudo-solution into the student along prefixes of the model's disagreeing completions, U-OPSD achieves continual improvement through on-policy self-distillation without any external supervision.

$$\mathcal{L}_{\text{U-OPSD}}(\theta) = \mathbb{E}_{x\sim\mathcal{U}}\, \mathbb{E}_{\{y^{(g)}\}_{g=1}^{G}\sim\bar{\pi}(\cdot|x)} \Bigg[\mathbb{1}\big[\mathcal{Y}_x^- \neq \emptyset\big] \frac{1}{|\mathcal{Y}_x^-|} \sum_{y^- \in \mathcal{Y}_x^-} \frac{1}{|y^-|} \sum_{n=1}^{|y^-|} D_\beta\big(\bar{\pi}\big(\cdot \mid x,\, y^+,\, y^-_{<n}\big) \,\big\|\, \pi_\theta\big(\cdot \mid x,\, y^-_{<n}\big)\big) \Bigg]. \tag{1}$$

Here, $|y^-|$ denotes the token length of a disagreeing rollout, and $y^-_{<n}$ denotes its prefix before token position n. The function $D_\beta(P \,\|\, Q)$ denotes a divergence between the teacher and student next-token distributions, such as the forward KL divergence $D_{\text{KL}}(P \,\|\, Q)$.

We evaluate U-OPSD on five mathematical reasoning benchmarks across six Qwen3 configurations: 4B and 8B models in both non-thinking and thinking modes, together with Qwen3-4B-Instruct-2507 and Qwen3-30B-A3B-Instruct-2507. Across all configurations, U-OPSD improves over the base model, by 8.5–10.7% in non-thinking mode, 1.9–2.2% in thinking mode, and 1.7–1.8% on the Instruct models. U-OPSD also improves over prior self-rewarding RL methods by a large margin, i.e., on average 7.0–11.3% in non-thinking mode and 0.8–1.4% in thinking mode, showing that consensus is substantially more effective as conditioning context for token-level distillation than as a scalar reward for policy optimization. Without any external supervision, U-OPSD is still on par with or better than supervised SFT, GRPO, and OPSD with GT labels, with gains of up to 10.9%, 8.9%, and 3.2%, respectively, while tying OPSD only in the 8B thinking setting.

2 Related work

Self-rewarding Reinforcement Learning. Although RLVR has proven effective at strengthening LLM reasoning (Shao et al., 2024; Guo et al., 2025), it hinges on curated ground-truth labels, whose cost and scarcity quickly become the limiting factor (Yue et al., 2025). A growing body of work replaces external verification with reward signals the model derives from its own behavior on unlabeled data. Among these intrinsic signals, confidence-based rewards are especially appealing: a model's certainty in its own output—captured directly as self-certainty (Zhao et al., 2026b; Li et al., 2025) or inversely through predictive entropy (Prabhudesai et al., 2025; Zhang et al., 2025)—yields

a dense, label-free reward that requires no extra sampling infrastructure and correlates well with answer correctness. Agreement across sampled solutions offers a complementary signal, as exploited by majority voting (Wang et al., 2023) and test-time training on self-consistency (Zuo et al., 2025). These methods build on earlier self-rewarding language models (Yuan et al., 2024) and self-play supervision (Chen et al., 2024), and the paradigm has since broadened to unsupervised self-training (Xu et al., 2025; Fang et al., 2025), self-correction guided by the model's own judgments (Xiong et al., 2025), and zero-data self-evolution in which models construct their own curricula (Zhao et al., 2025; Huang et al., 2025; Liu et al., 2026a).

On-policy (self-) distillation. On-policy distillation trains on student-generated trajectories, allowing a teacher to provide dense token-level supervision at states visited by the student and reducing the train–inference mismatch of off-policy distillation (Gu et al., 2024; Agarwal et al., 2024). On-policy self-distillation removes the separate teacher by letting the same model act under asymmetric contexts, with the teacher conditioned on verified solutions (Zhao et al., 2026a), demonstrations or in-context examples (Shenfeld et al., 2026a), privileged observations (Penaloza et al., 2026), or environmental feedback (Hübotter et al., 2026). Subsequent work explores alternative conditioning signals (Sang et al., 2026), derives process supervision from externally verified successful and unsuccessful trajectories (Tan and Hong, 2026), or modulates token-level supervision according to teacher reliability (Ke et al., 2026; Liu et al., 2026c). U-OPSD instead constructs the teacher reference from rollout agreement and identifies correction targets through rollout disagreement, without annotations, demonstrations, environmental feedback, gold answers, or externally verified outcomes.

Self-training, self-distillation and consistency. Self-training improves models using supervision derived from their own generations. Existing methods select or refine self-generated data using answer verification and reward signals (Zelikman et al., 2022; Yuan et al., 2023; Gulcehre et al., 2023), as well as confidence, consistency, and iterative self-improvement (Huang et al., 2023; Liang et al., 2026). More broadly, self-distillation transfers knowledge across different instances or views of the same model, including previous model generations and intermediate optimization snapshots (Furlanello et al., 2018; Yang et al., 2018). Other approaches distill across architectural branches or augmented views (Zhang et al., 2019; Li et al., 2023; Caron et al., 2021), while recent work extends self-distillation to foundation models using diverse reasoning traces or multiple self-teachers (Wu et al., 2025; Jin et al., 2026). Self-consistency originally aggregates multiple reasoning paths through majority voting at inference time (Wang et al., 2023), while subsequent methods turn agreement into sequence-level preferences (Prasad et al., 2025), pseudo-labels or scalar reinforcement-learning rewards (Zuo et al., 2025; Shafayat et al., 2025), or model-generated feedback and preferences (Yuan et al., 2024). U-OPSD connects self-consistency with on-policy self-distillation by using consensus to construct an instance-specific self-teacher and transferring dense next-token distributional supervision along prefixes of consensus-disagreeing rollouts.

3 METHOD

3.1 PRELIMINARIES

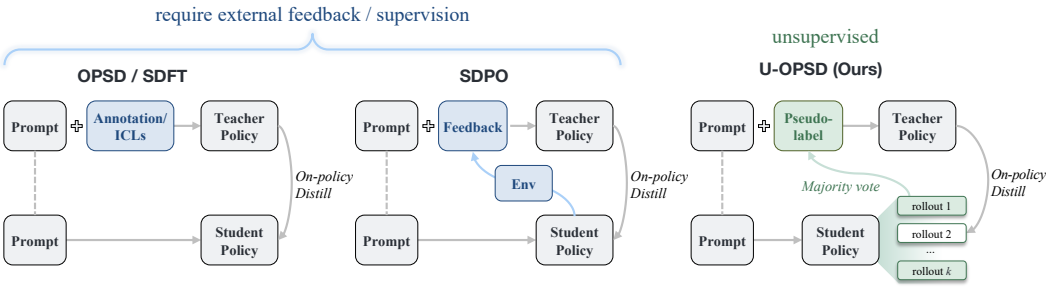

Figure 1: Comparison between OPSD / SDFT with ground-truth solution or ICLs (left), SDPO with rich feedback from the environment (middle) and our U-OPSD without any supervision.

Notation. Let π_θ be an LLM defining next-token distributions $\pi_\theta(\cdot \mid c)$ over a vocabulary $\mathcal{V}$ for a context c, and let $\bar{\pi} \triangleq \pi_{\mathrm{sg}[\theta]}$ be the same network with gradients detached, so that sampling from or

scoring under $\bar{\pi}$ carries no gradient. For a sequence y, $y_{<t}$ denotes its prefix, and $\mathrm{Ans}(y) \in \mathcal{A} \cup \{\varnothing\}$ the final answer parsed from it, where $\mathcal{A}$ is the space of admissible answers and $\varnothing$ marks a generation from which no answer can be parsed. Supervised post-training draws from a labeled corpus $\mathcal{S} = \{(x, y^\star)\}$ of problems with reference solutions and gold answers $a^\star = \mathrm{Ans}(y^\star)$; U-OPSD uses only unlabeled problems $x \sim \mathcal{U}$. Throughout, D_β denotes the generalized Jensen–Shannon divergence between distributions on $\mathcal{V}$, whose $\beta \to 0$ limit is the forward KL.

Supervised on-policy baselines. GRPO (Shao et al., 2024) uses gold-answer-verified sequence-level rewards, whereas OPD (Gu et al., 2024; Agarwal et al., 2024; Lu and Lab, 2025) provides dense token-level supervision from a typically stronger teacher. OPSD (Zhao et al., 2026a) removes the separate teacher but still conditions a detached copy of the same model on a gold solution $y^\star$, while the student sees only x. Appendix A gives the full GRPO, OPD, and OPSD objectives. Our method retains OPSD's dense on-policy objective but replaces $y^\star$ with a reference constructed solely from the model's own rollouts.

3.2 U-OPSD: On-Policy Self-Distillation without Any Supervision

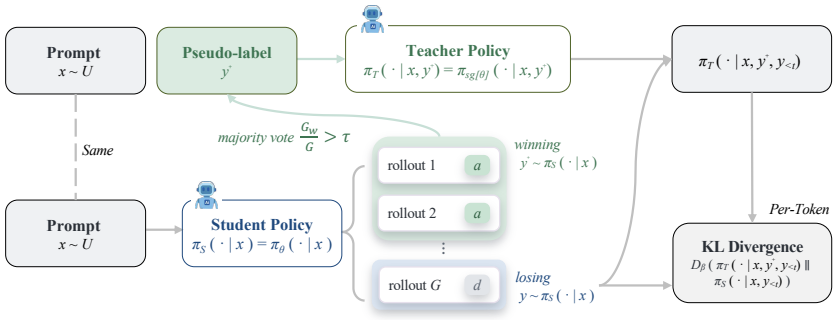

Figure 2: Overview of Unsupervised On-policy Self-Distillation (U-OPSD), which replaces ground-truth supervision in On-Policy Self-distillation with pseudo-labels generated from the model's own majority-vote consensus.

Despite on-policy training, these prior objectives remain externally supervised (Appendix A). RLVR relies on gold answers for verification, OPD on a stronger teacher, and OPSD and related work on privileged teacher contexts such as GT solutions (Zhao et al., 2026a), demonstrations (Shenfeld et al., 2026a), or environmental feedback (Hübotter et al., 2026) (see the left of Figure 1). U-OPSD removes this dependence by enabling the model to construct its own privileged context through internal consistency, i.e., a majority vote over its own rollouts that identifies a pseudo-solution in place of $y^\star$, while the resulting conditional teacher distribution is distilled into the student on the disagreeing rollouts (Figure 1, right).

(1) Sample. For each prompt, draw G rollouts $y^{(1)}, \ldots, y^{(G)} \sim \bar{\pi}(\cdot \mid x)$ at the training temperature and parse $a^{(g)} = \mathrm{Ans}(y^{(g)}) \in \mathcal{A} \cup \{\varnothing\}$ (i.e., extract the final answer $a^{(g)}$ from `\boxed{...}` with normalization and canonicalization). $\varnothing$ marks rollouts without a parsable final answer (usually truncations), which we call *invalid*.

(2) Vote. Among the valid answers, we define the pseudo-answer $\tilde{a}(x)$ as the majority vote, breaking ties uniformly at random (Wang et al., 2023). This vote partitions the rollouts into agreeing and disagreeing sets $\mathcal{Y}_x^+ = \{y^{(g)} : a^{(g)} = \tilde{a}(x)\}$, $\mathcal{Y}_x^- = \{y^{(g)} : a^{(g)} \neq \tilde{a}(x)\}$. Invalid rollouts belong to neither set, as an incomplete generation provides evidence of neither a correct nor an incorrect belief. We quantify vote confidence using the self-consistency score $c(x) = \frac{1}{G}\sum_{g=1}^{G} \mathbb{1}\left[a^{(g)} = \tilde{a}(x)\right]$. The score is normalized by the total number of sampled rollouts G, rather than only by the number of valid rollouts where truncated generations reduce confidence. If $c(x) < \tau$, the prompt is treated as unlabeled and contributes no gradient in the training step. We set $\tau = 1/2$ for an absolute majority.

(3) Distill. Given a valid majority vote, we select an agreeing rollout $y^+ \in \mathcal{Y}_x^+$ as the teacher reference in place of $y^\star$, and a subset $\mathcal{B}_x^- \subseteq \mathcal{Y}_x^-$ of disagreeing rollouts as the distillation targets. Using the OPSD construction in Appendix A, but replacing the gold reference $y^\star$ with (x, y^+), yields

$$\mathcal{L}_{\text{U-OPSD}}(\theta) \;=\; \frac{1}{|\mathcal{B}_x^-|} \sum_{y^- \in \mathcal{B}_x^-} \sum_{t=1}^{|y^-|} D_\beta\Big(\bar{\pi}\big(\cdot \mid x, y^+, y^-_{<t}\big) \;\Big\|\; \pi_\theta\big(\cdot \mid x, y^-_{<t}\big)\Big). \tag{2}$$

The above skips two classes of samples: those for which the model cannot form a sufficiently consistent vote ($c(x) < \tau$), making the pseudo-answer unreliable, and those for which all valid rollouts already agree ($\mathcal{Y}_x^- = \emptyset$), leaving no disagreement to correct. Training therefore focuses on the model's competence frontier: prompts for which it can consistently identify a plausible solution but still assigns substantial probability to conflicting trajectories. This curriculum emerges directly from the model's own voting statistics, without external difficulty labels or scheduling.

Unlike OPSD, U-OPSD requires no labeled pair $(x, y^\star)$; its privileged context is constructed entirely from the model's own on-policy rollouts. Unlike conventional majority-vote self-training, it does not simply imitate the selected response under teacher-forced prefixes. Instead, it transfers the solution-conditioned next-token distribution along the model's own disagreeing trajectories, providing dense corrective supervision precisely at the prefixes that lead toward answers inconsistent with the model's consensus. In this way, U-OPSD enables iterative on-policy self-evolution and self-improvement.

Algorithm 1 U-OPSD: unsupervised on-policy self-distillation (one prompt)

Require: prompt x; student π_θ; detached teacher $\bar{\pi}$; rollouts G; threshold τ

1: sample $y^{(1)}, \ldots, y^{(G)} \overset{\text{i.i.d.}}{\sim} \bar{\pi}(\cdot \mid x)$ ▷ on-policy, training temperature
2: $a^{(g)} \leftarrow \text{Ans}(y^{(g)})$ for all g ▷ $\varnothing$ if unparsable
3: $\tilde{a}(x) \leftarrow$ plurality of $\{a^{(g)} : a^{(g)} \neq \varnothing\}$; $\quad c(x) \leftarrow \frac{1}{G}\sum_g \mathbb{1}[a^{(g)} = \tilde{a}(x)]$
4: $\mathcal{Y}_x^+ \leftarrow \{y^{(g)} : a^{(g)} = \tilde{a}(x)\}$; $\quad \mathcal{Y}_x^- \leftarrow \{y^{(g)} : a^{(g)} \notin \{\tilde{a}(x), \varnothing\}\}$
5: **if** $c(x) < \tau$ **or** $\mathcal{Y}_x^- = \emptyset$ **then**
6: **return** ▷ vote not trusted, or nothing to correct
7: **end if**
8: select $y^+ \in \mathcal{Y}_x^+$ and $\mathcal{B}_x^- \subseteq \mathcal{Y}_x^-$
9: minimize $\frac{1}{|\mathcal{B}_x^-|} \sum_{y^- \in \mathcal{B}_x^-} \sum_{t=1}^{|y^-|} D_\beta\big(\bar{\pi}(\cdot \mid x, y^+, y^-_{<t}) \,\|\, \pi_\theta(\cdot \mid x, y^-_{<t})\big)$

4 Experiments

4.1 Setup

Models and datasets. We adopt four variants of Qwen3 (Yang et al., 2025), including Qwen3-4B, Qwen3-8B, Qwen3-30B-A3B-Instruct-2507 and Qwen3-4B-Instruct-2507. For Qwen3-4B and Qwen3-8B, we train them in both non-thinking and thinking modes. In all experiments, the same checkpoint serves as both student and teacher. For the training set, we adopt a 30k subset of OpenThoughts (Guha et al., 2025) following OPSD; U-OPSD uses only the problem statements, never the solution field. We compare against three supervised baselines trained on the same prompts: SFT on their gold solutions, GRPO with binary outcome rewards verified against the gold answer, and OPSD, which conditions the teacher on the gold solution. All OPSD numbers we report are from our own rerun with the released code and hyperparameters, under conditions identical to the U-OPSD runs.

Evaluation and benchmarks. We adopt five competition-level math reasoning benchmarks: AIME24, AIME25, and HMMT25 at avg@12, and MATH500 (Hendrycks et al., 2021; Lightman et al., 2024) and AMC23 at avg@4. We follow OPSD's protocol for decoding using vLLM at temperature 1.0 with maximum generation length 38k, in the same reasoning mode the model was trained in.

Following OPSD's setup, we evaluate checkpoints every 25 steps up to 150 steps and report the best score for U-OPSD and OPSD, For GRPO, we report the peak performance within 500 training steps. For SFT, we train on the same number of samples as OPSD. For TTRL, RENT and Intuitor, we adopt

the same training setup as U-OPSD, i.e., training with the same rollout budget and the same prompt set.

Implementation details. We follow most of OPSD's training recipe (Zhao et al., 2026a): we use forward KL (β=0) over the full vocabulary with per-token pointwise clipping as the objective. We also adopt a teacher fixed to the initial policy rather than the running one to ensure fair comparison. Both choices are supported by ablations in Zhao et al. (2026a). We use LoRA of rank 64 (α=128) on all attention and MLP projections, a learning rate of 5×10^{-6}, gradient-norm clipping at 0.1, and sampling at temperature 1.1 with top-p 0.95 and top-k 20. All runs are trained for 150 steps with a checkpoint every 25.

For U-OPSD, we introduce two unique hyperparameters: for each prompt, we generate G=8 rollouts independently, filtered by confidence threshold τ=0.5 (default unless specified otherwise), where supervised OPSD instead draws a single rollout from each of 32 prompts per optimizer step. We also increase the maximum completion length from 1,024 to 4,096 tokens, as U-OPSD requires rollouts to reach a boxed final answer; Zhao et al. (2026a) report comparable performance between these two token budgets in their setting. For thinking-mode experiments, following OPSD, we keep the teacher in thinking mode and distill its behavior into a student in non-thinking mode, while evaluating the resulting model in thinking mode. For non-thinking experiments, both the teacher and student are in non-thinking mode during training, and evaluation is likewise conducted in non-thinking mode.

Table 1: Performance comparison on math reasoning benchmarks for Qwen3 models with non-thinking mode.

	Method	AIME24	AIME25	HMMT25	MATH500	AMC23	Avg.
Qwen3-4B							
	Base	25.83	17.78	10.83	84.10	66.25	40.96
	+ SFT	26.67	19.72	13.06	84.85	69.38	42.73
w/ GT.	+ GRPO	25.00	22.50	15.00	86.20	80.62	45.86
	+ OPSD	32.22	20.83	**16.39**	85.75	76.25	46.29
	+ TTRL	25.00	20.83	11.94	83.60	67.50	41.77
	+ RENT	22.22	20.28	11.39	84.00	74.38	42.45
w/o GT.	+ Intuitor	23.89	20.00	11.67	83.70	70.62	41.98
	+ U-OPSD	**37.50**	**27.78**	14.44	**86.50**	**81.25**	**49.49**
Qwen3-8B							
	Base	27.50	23.33	13.61	84.05	69.38	43.57
	+ SFT	26.94	21.67	11.94	84.10	72.50	43.43
w/ GT.	+ GRPO	30.56	21.94	13.06	87.85	73.75	45.43
	+ OPSD	41.67	28.06	18.33	87.15	**85.00**	52.04
	+ TTRL	27.22	21.11	13.06	84.45	71.88	43.54
	+ RENT	28.33	21.67	10.83	84.00	70.00	42.97
w/o GT.	+ Intuitor	26.11	22.50	11.67	84.20	73.12	43.52
	+ U-OPSD	**45.56**	**34.72**	**18.61**	**89.55**	83.12	**54.31**

4.2 Main results

Non-thinking mode. Table 1 reports the results in non-thinking mode. U-OPSD achieves average scores of 49.49 and 54.31 on Qwen3-4B and Qwen3-8B, improving over the corresponding base models by 8.5% and 10.7%. Despite using no ground-truth solutions, U-OPSD outperforms all supervised baselines, including SFT, GRPO, and OPSD, exceeding OPSD by 3.2% and 2.3% at the two model scales. These gains are consistent across benchmarks: U-OPSD obtains the best result on four of the five benchmarks for both models. In contrast, the label-free RL baselines improve over the base models by at most 1.5% under the same rollout budget. This comparison highlights the advantage of using consensus-derived solutions as privileged teacher context for token-level distillation, rather than reducing them to scalar rewards.

Thinking mode. Table 2 presents the results in thinking mode. U-OPSD reaches 77.05 on Qwen3-4B and 77.99 on Qwen3-8B, improving over the base models by 2.2% and 1.9%, respectively. It matches or slightly exceeds supervised OPSD and consistently outperforms GRPO in average performance. The smaller gains relative to non-thinking mode suggest that consensus-based self-distillation is most

Table 2: Performance comparison on math reasoning benchmarks for Qwen3 models with thinking mode.

	Method	AIME24	AIME25	HMMT25	MATH500	AMC23	Avg.
Qwen3-4B							
	Base	74.17	64.72	45.56	94.80	95.00	74.85
w/ GT.	+ SFT	73.06	69.17	41.67	95.40	94.38	74.73
	+ GRPO	73.89	69.44	43.61	95.45	99.38	76.35
	+ OPSD	75.28	68.06	43.06	95.20	99.38	76.20
w/o GT.	+ TTRL	72.78	68.33	45.56	95.90	96.25	75.76
	+ RENT	74.72	65.83	43.06	95.25	99.38	75.65
	+ Intuitor	76.39	68.33	42.78	95.35	97.50	76.07
	+ U-OPSD	76.39	68.06	46.94	95.75	98.12	77.05
Qwen3-8B							
	Base	75.56	66.67	45.00	96.35	96.88	76.09
w/ GT.	+ SFT	76.39	69.72	43.89	95.80	95.00	76.16
	+ GRPO	76.94	69.17	47.78	95.70	95.00	76.92
	+ OPSD	80.83	69.72	46.67	95.75	96.88	77.97
w/o GT.	+ TTRL	77.22	68.61	46.94	95.75	96.25	76.95
	+ RENT	77.50	70.28	45.83	95.95	96.25	77.16
	+ Intuitor	76.94	70.28	44.17	96.20	95.62	76.64
	+ U-OPSD	76.94	71.39	47.50	96.00	98.12	77.99

Table 3: Performance comparison on math reasoning benchmarks for two instruction-tuned models, Qwen3-30B-A3B-Instruct-2507 and Qwen3-4B-Instruct-2507.

Model	Method	AIME24	AIME25	HMMT25	MATH500	AMC23	Avg.
Qwen3-30B-A3B -Instruct-2507	Base	**80.00**	63.33	43.33	96.33	95.83	75.77
	OPSD	78.89	61.11	47.78	**97.20**	96.67	76.33
	U-OPSD	75.56	**65.56**	**50.00**	97.00	**99.17**	**77.46**
Qwen3-4B -Instruct-2507	Base	66.67	53.33	27.78	93.87	93.33	67.00
	OPSD	62.22	52.22	**31.11**	**94.93**	**95.00**	67.10
	U-OPSD	**68.89**	**57.78**	28.89	94.20	94.17	**68.78**

effective when the base model is sufficiently reliable to produce informative consensus while retaining substantial room for improvement; the stronger thinking-mode models leave less such headroom.

Instruction-tuned models. Table 3 further evaluates U-OPSD on two instruction-tuned models. U-OPSD improves Qwen3-30B-A3B-Instruct-2507 from 75.77 to 77.46 and Qwen3-4B-Instruct-2507 from 67.00 to 68.78, achieving the highest average performance on both models. It also surpasses supervised OPSD by 1.1% and 1.7% on average, though on Qwen3-4B-Instruct-2507 OPSD leads on HMMT25, MATH500 and AMC23. The result on Qwen3-30B-A3B-Instruct-2507 demonstrates that the same training recipe transfers effectively to a substantially larger mixture-of-experts model without model-specific hyperparameter tuning.

Training dynamics across checkpoints are reported in Appendix B.1 (Figure 5). The curves show that U-OPSD's non-thinking gains persist throughout training, whereas all thinking-mode methods remain close to the base model.

4.3 Ablation and Analysis

Pseudo-label quality. On 64 training prompts under the full training configuration (G=8, 4096-token budget, τ=0.5), 96.3% of rollouts yield a parsable boxed answer, 94.0% of prompts clear the self-consistency threshold and receive a pseudo-label, and 86.7% of those pseudo-labels match the gold answer from data source (used in supervised methods). Fewer than 10% of valid rollouts disagree with their prompt's vote, so the distillation set is small and concentrated.

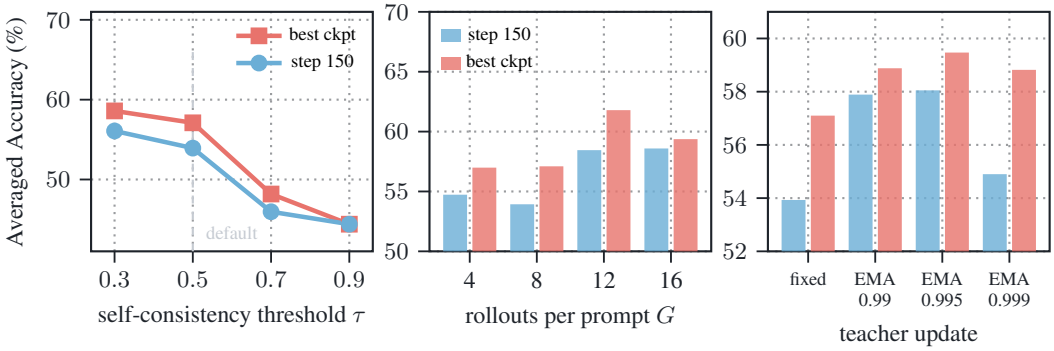

Figure 3: Configuration ablations, on Qwen3-8B non-thinking. The vertical axis is accuracy averaged over the five evaluation benchmarks. **Left:** self-consistency threshold τ, the fraction of valid rollouts that must agree for a prompt to receive a pseudo-label. **Middle:** rollouts per prompt G, at τ=0.5. **Right:** how the teacher is updated; "fixed" freezes it at the initial policy and is our default, while the EMA rows let it track the student at the given decay.

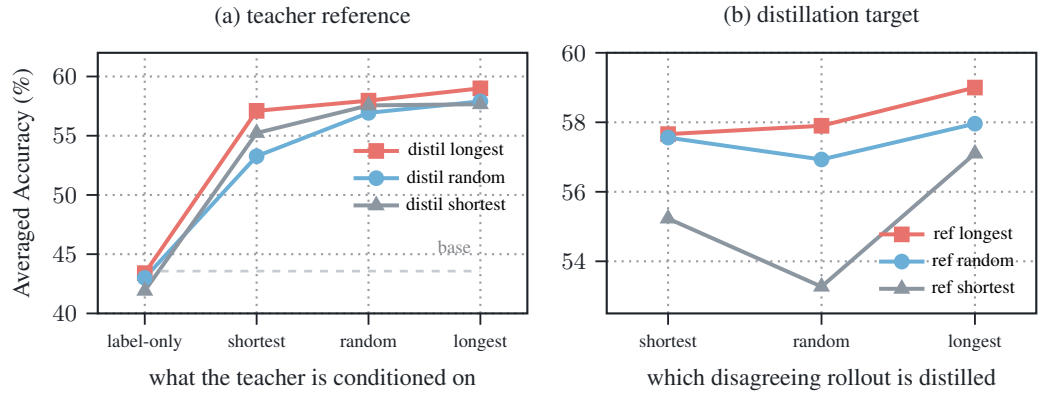

Figure 4: **Teacher reference against distillation target**, on Qwen3-8B non-thinking, G=8, τ=0.5, k=1; each point is the five-benchmark average. *(a)* What the teacher is conditioned on. *(b)* Which disagreeing rollout is distilled.

Self-consistency threshold. The threshold τ decides which prompts are allowed to supervise. Figure 3 (left) sweeps it and the ranking is monotone: the loosest threshold is best, at 58.59 for τ=0.3 against 57.10 for the default and 44.40 at τ=0.9, a 14.2% spread; the step-150 series in the figure is ordered the same way. The default τ=0.5 is therefore not the best setting in this sweep.

Rollouts per prompt. G, number of rollouts per prompt, controls both the resolution of the vote and the pool from which disagreeing rollouts are drawn, and it is the dominant cost term, since the sampling cost grows linearly in G. Figure 3 (middle) shows the return is real but soon saturating: G=4 and G=8 are indistinguishable at 56.99 and 57.10 at the best checkpoint, G=12 gains 4.7% over the default, and G=16 gives back half of that. We keep G=8 elsewhere because it is OPSD's own generation budget and makes the comparison in Tables 1 and 2 cost-matched, but G=12 is the better operating point when generation is not the constraint.

Teacher update. OPSD (Zhao et al., 2026a) freezes the teacher at the initial policy, which under LoRA means evaluating the base weights with the adapter disabled. We report the frozen teacher elsewhere because it is OPSD's setting and keeps the comparison in Tables 1 and 2 matched. Yet, Figure 3 (right) shows that the EMA settings offer more room for improvement: the best, at decay 0.995, gains 2.4% over the frozen default at the best checkpoint and 4.1% at step 150, and the two faster decays land within 0.2% of each other at step 150.

Teacher reference and distillation target. Figure 4 crosses which agreeing rollout the teacher is conditioned on against which disagreeing rollout is distilled. Both axes favour the longest rollout, though choosing the shortest or a random one also yields competitive performance: seven of the nine full-reference settings lie within 2.1% of the best on averaged benchmark accuracy, with random the weaker choice as a distillation target, at a column mean of 56.03 against 58.02 for longest. The best setting, at 59.00, conditions the teacher on the longest agreeing rollout and distils the longest disagreeing one. Yet, as (a) shows, further stripping the reference to the boxed pseudo-label costs 10.3% to 15.8% and puts all three label-only settings below the base model's 43.57. Effective distillation requires the teacher to condition on the longer, full reasoning trace rather than only the final answer.

Additional objective ablations are deferred to Appendix B.2. Table 4 compares sampled-token, full-vocabulary, and top-k divergence computation and finds top-100 strongest on average; Table 5 compares divergence metrics and documents the non-termination failure under reverse KL.

5 LIMITATIONS

Scope. Our experiments cover one model family (Qwen3, at 4B and 8B and in both reasoning modes) and one domain (competition mathematics) with automatically checkable final answers. The voting mechanism requires an extractable, canonicalizable answer; extending U-OPSD to open-ended generation would require replacing exact-match voting with a softer consensus.

The size of the advantage is regime-dependent. The gain over supervised OPSD is large in non-thinking mode and less in thinking mode. U-OPSD exceeds OPSD by 3.2% and 2.3% at 4B and 8B in Table 1, and by 0.9% and 0.02% in Table 2; the improvement over the base model shrinks in the same way, from 8.5% and 10.7% to 2.2% and 1.9%. Thinking-mode base accuracy is already 74.9 and 76.1, and thinking-mode rollouts are far longer, so a 150-step budget covers many fewer completed votes per token spent. Readers should therefore treat consensus as a *stronger replacement* for gold solutions in the non-thinking regime we measured and as a *match* for them in the thinking regime, rather than as a uniform improvement on on-policy self-distillation.

Dependence on base-model competence. Majority-vote supervision reproduces whichever answer the base model already produces most often, so it is bounded by that answer's accuracy. Our probe measured 13.3% wrong pseudo-labels in-domain, and we did not measure training dynamics under deliberately corrupted votes. Mechanisms for detecting or down-weighting low-quality consensus (e.g., by vote margin) are natural next steps.

Variance and repeats. We mitigated variance by evaluating every checkpoint with 12 samples spanning five benchmarks, but seed-replicated error bars for every training setup are pending and will be added in a revision.

6 CONCLUSION

We showed that the ground-truth solution in on-policy self-distillation can be replaced by the model's own majority-vote consensus: the agreeing rollout serves as the teacher's reference, and distillation is applied only to self-inconsistent rollouts. With the OPSD recipe otherwise untouched, this label-free variant outperforms its supervised counterpart on the five-benchmark average, and a grid over two model sizes and both reasoning modes shows the advantage reproduces at both scales in non-thinking mode and a tie in thinking mode, where the base model is already strong enough that little headroom remains. The result suggests that, in the regime of on-policy self-distillation—a base model competent enough to vote well and fallible enough to have headroom—the binding constraint is not access to gold solutions but the machinery for surfacing and correcting the model's own inconsistencies, a machinery that requires no supervision.

AI USE STATEMENT

In this work, we used generative AI tools solely for language editing to improve the grammar, clarity, and readability of the manuscript. All AI-assisted edits were carefully reviewed to ensure that they preserved the intended meaning and technical accuracy. We take full responsibility for the final

content of this work, including any text, claims, or artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work studies unsupervised on-policy self-distillation for improving the mathematical reasoning ability of open-weight language models (Qwen3, 4B/8B). All training and evaluation use publicly released model checkpoints and publicly available, automatically-graded competition mathematics benchmarks (AIME24, AIME25, HMMT25, MATH500, AMC23); the method does not involve human subjects, crowdsourcing, or the collection of any new personal data.

REPRODUCIBILITY STATEMENT

We describe U-OPSD formally in Section 3 (pseudo-answer construction, the agreeing/disagreeing rollout partition, and the training objective), and the exact training and evaluation configuration – base models, benchmarks, sampling budget, and hyperparameters – in Section 4.1. Limitations relevant to reproducing our exact numbers (variance from a fixed evaluation sample size, pending seed repeats) are discussed in Section 5.

REFERENCES

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio César Teodoro Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, Vibhav Vineet, Yue Wu, Safoora Yousefi, and Guoqing Zheng. Phi-4-reasoning technical report, 2025. URL https://arxiv.org/abs/2504.21318.

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 *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. doi: 10.1109/ICCV48922.2021.00951.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning convertsweak language models to strong language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Wenkai Fang, Shunyu Liu, Yang Zhou, Kongcheng Zhang, Tongya Zheng, Kaixuan Chen, Mingli Song, and Dacheng Tao. SeRL: Self-play reinforcement learning for large language models with limited data. 2025. URL https://openreview.net/forum?id=ZF93vyH9He.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *ArXiv*, abs/2603.25562, 2026. URL https://api.semanticscholar.org/CorpusID:286791524.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1607–1616. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/furlanello18a.html.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alexandros G. Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models. 2025. URL https://arxiv.org/abs/2506.04178.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. Reinforced self-training (rest) for language modeling. 2023. URL https://arxiv.org/abs/2308.08998.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=7Bywt2mQsCe.

Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, et al. R-zero: Self-evolving reasoning llm from zero data. *arXiv preprint arXiv:2508.05004*, 2025.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. pages 1051–1068, December 2023. doi: 10.18653/v1/2023.emnlp-main.67. URL https://aclanthology.org/2023.emnlp-main.67/.

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. 2026. URL https://openreview.net/forum?id=QkfkxyRizZ.

Yiqiao Jin, Yiyang Wang, Lucheng Fu, Yijia Xiao, Yinyi Luo, Haoxin Liu, B. Aditya Prakash, Josiah Hester, Jindong Wang, and Srijan Kumar. UniSD: Towards a unified self-distillation framework for large language models. 2026. URL https://openreview.net/forum?id=ymCne2LfP8.

Junlong Ke, Zichen Wen, Weijia Li, Conghui He, and Linfeng Zhang. Respecting self-uncertainty in on-policy self-distillation for efficient llm reasoning. 2026. URL https://arxiv.org/abs/2605.13255.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and SeYoung Yun. Distillm: Towards streamlined distillation for large language models. volume abs/2402.03898, 2024. URL https://api.semanticscholar.org/CorpusID:267499832.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. DistiLLM-2: A contrastive approach boosts the distillation of LLMs. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 31044–31062. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/ko25a.html.

Pengyi Li, Matvey Skripkin, Alexander Zubrey, Andrey Kuznetsov, and Ivan Oseledets. Confidence is all you need: Few-shot rl fine-tuning of language models. 2025. URL https://arxiv.org/abs/2506.06395.

Yijiang Li, Xinjiang Wang, Lihe Yang, Litong Feng, Wayne Zhang, and Ying Gao. Diverse cotraining makes strong semi-supervised segmentor. pages 16009–16021, 2023. doi: 10.1109/ICCV51070.2023.01471.

Zhongzhi Li, Xuansheng Wu, Yijiang Li, Lijie Hu, and Ninghao Liu. Less is enough: synthesizing diverse data in feature space of llms. *arXiv e-prints*, pages arXiv–2602, 2026.

Yijun Liang, Hengguang Zhou, Ming Li, Lichen Li, Cho-Jui Hsieh, and Tianyi Zhou. Self-evolving visual questioner. 2026. URL https://arxiv.org/abs/2606.13929.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 39578–39601, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/aca97732e30bcf1303bc22ac3924fd16-Paper-Conference.pdf.

Bo Liu, Simon Yu, Zichen Liu, Leon Guertler, Penghui Qi, Daniel Balcells, Mickel Liu, Cheston Tan, Weiyan Shi, Min Lin, Wee Sun Lee, and Natasha Jaques. SPIRAL: Self-play on zero-sum games incentivizes reasoning via multi-agent multi-turn reinforcement learning. 2026a. URL https://openreview.net/forum?id=7Yayy5fNLg.

Haoran Liu, Yuwei Zhang, Xiyao Li, Bohan Lyu, and Jingbo Shang. Hero: Hindsight-enhanced reflection from environment observations for agentic self-distillation. 2026b. URL https://arxiv.org/abs/2606.11559.

Xiaogeng Liu, Xinyan Wang, Yingzi Ma, Yechao Zhang, and Chaowei Xiao. When are teacher tokens reliable? position-weighted on-policy self-distillation for reasoning. 2026c. URL https://arxiv.org/abs/2605.21606.

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.

MiniMax, :, Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, Chengjun Xiao, Chengyu Du, Chi Zhang, Chu Qiao, Chunhao Zhang, Chunhui Du, Congchao Guo, Da Chen, Deming Ding, Dianjun Sun, Dong Li, Enwei Jiao, Haigang Zhou, Haimo Zhang, Han Ding, Haohai Sun, Haoyu Feng, Huaiguang Cai, Haichao Zhu, Jian Sun, Jiaqi Zhuang, Jiaren Cai, Jiayuan Song, Jin Zhu, Jingyang Li, Jinhao Tian, Jinli Liu, Junhao Xu, Junjie Yan, Junteng Liu, Junxian He, Kaiyi Feng, Ke Yang, Kecheng Xiao, Le Han, Leyang Wang, Lianfei Yu, Liheng Feng, Lin Li, Lin Zheng, Linge Du, Lingyu Yang, Lunbin Zeng, Minghui Yu, Mingliang Tao, Mingyuan Chi, Mozhi Zhang, Mujie Lin, Nan Hu, Nongyu Di, Peng Gao, Pengfei Li, Pengyu Zhao, Qibing Ren, Qidi Xu, Qile Li, Qin Wang, Rong Tian, Ruitao Leng, Shaoxiang Chen, Shaoyu Chen, Shengmin Shi, Shitong Weng, Shuchang Guan, Shuqi Yu, Sichen Li, Songquan Zhu, Tengfei Li, Tianchi Cai, Tianrun Liang, Weiyu Cheng, Weize Kong, Wenkai Li, Xiancai Chen, Xiangjun Song, Xiao Luo, Xiao Su, Xiaobo Li, Xiaodong Han, Xinzhu Hou, Xuan Lu, Xun Zou, Xuyang Shen, Yan Gong, Yan Ma, Yang Wang, Yiqi Shi, Yiran Zhong, Yonghong Duan, Yongxiang Fu, Yongyi Hu, Yu Gao, Yuanxiang Fan, Yufeng Yang, Yuhao Li, Yulin Hu, Yunan Huang, Yunji Li, Yunzhi Xu, Yuxin Mao, Yuxuan Shi, Yuze Wenren, Zehan Li, Zelin Li, Zhanxu Tian, Zhengmao Zhu, Zhenhua Fan, Zhenzhen Wu, Zhichao Xu, Zhihang Yu, Zhiheng Lyu, Zhuo Jiang, Zibo Gao, Zijia Wu, Zijian Song, and Zijun Sun. Minimax-m1: Scaling test-time compute efficiently with lightning attention, 2025. URL https://arxiv.org/abs/2506.13585.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. 2026. URL https://arxiv.org/abs/2602.04942.

Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. Maximizing confidence alone improves reasoning. *arXiv preprint arXiv:2505.22660*, 2025.

Archiki Prasad, Weizhe Yuan, Richard Yuanzhe Pang, Jing Xu, Maryam Fazel-Zarandi, Mohit Bansal, Sainbayar Sukhbaatar, Jason E Weston, and Jane Yu. Self-consistency preference optimization. 2025. URL https://openreview.net/forum?id=94G4eL3RWi.

Hejian Sang, Yuanda Xu, Zhengze Zhou, Ran He, Zhipeng Wang, and Jiachen Sun. Crisp: Compressed reasoning via iterative self-policy distillation. 2026. URL https://arxiv.org/abs/2603.05433.

Sheikh Shafayat, Fahim Tajwar, Ruslan Salakhutdinov, Jeff Schneider, and Andrea Zanette. Can large reasoning models self-train? 2025. URL https://arxiv.org/abs/2505.21444.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. URL https://doi.org/10.48550/arXiv.2402.03300.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. 2026a. URL https://openreview.net/forum?id=qA6FgH0nnZ.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. RL's razor: Why online reinforcement learning forgets less. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=7HNRYT4V44.

Zhiquan Tan and Yinrong Hong. Self-supervised on-policy distillation for reasoning language models. 2026. URL https://arxiv.org/abs/2605.17497.

5 Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, Yean Cheng, Yifan An, Yilin Niu, Yuanhao Wen, Yushi Bai, Zhengxiao Du, Zihan Wang, Zilin Zhu, Bohan Zhang, Bosi Wen, Bowen Wu, Bowen Xu, Can Huang, Casey Zhao, Changpeng Cai, Chao Yu, Chen Li, Chendi Ge, Chenghua Huang, Chenhui Zhang, Chenxi Xu, Chenzheng Zhu, Chuang Li, Congfeng Yin, Daoyan Lin, Dayong Yang, Dazhi Jiang, Ding Ai, Erle Zhu, Fei Wang, Gengzheng Pan, Guo Wang, Hailong Sun, Haitao Li, Haiyang Li, Haiyi Hu, Hanyu Zhang, Hao Peng, Hao Tai, Haoke Zhang, Haoran Wang, Haoyu Yang, He Liu, He Zhao, Hongwei Liu, Hongxi Yan, Huan Liu, Huilong Chen, Ji Li,

Jiajing Zhao, Jiamin Ren, Jian Jiao, Jiani Zhao, Jianyang Yan, Jiaqi Wang, Jiayi Gui, Jiayue Zhao, Jie Liu, Jijie Li, Jing Li, Jing Lu, Jingsen Wang, Jingwei Yuan, Jingxuan Li, Jingzhao Du, Jinhua Du, Jinxin Liu, Junkai Zhi, Junli Gao, Ke Wang, Lekang Yang, Liang Xu, Lin Fan, Lindong Wu, Lintao Ding, Lu Wang, Man Zhang, Minghao Li, Minghuan Xu, Mingming Zhao, Mingshu Zhai, Pengfan Du, Qian Dong, Shangde Lei, Shangqing Tu, Shangtong Yang, Shaoyou Lu, Shijie Li, Shuang Li, Shuang-Li, Shuxun Yang, Sibo Yi, Tianshu Yu, Wei Tian, Weihan Wang, Wenbo Yu, Weng Lam Tam, Wenjie Liang, Wentao Liu, Xiao Wang, Xiaohan Jia, Xiaotao Gu, Xiaoying Ling, Xin Wang, Xing Fan, Xingru Pan, Xinyuan Zhang, Xinze Zhang, Xiuqing Fu, Xunkai Zhang, Yabo Xu, Yandong Wu, Yida Lu, Yidong Wang, Yilin Zhou, Yiming Pan, Ying Zhang, Yingli Wang, Yingru Li, Yinpei Su, Yipeng Geng, Yitong Zhu, Yongkun Yang, Yuhang Li, Yuhao Wu, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yuxuan Zhang, Zezhen Liu, Zhen Yang, Zhengda Zhou, Zhongpei Qiao, Zhuoer Feng, Zhuorui Liu, Zichen Zhang, Zihan Wang, Zijun Yao, Zikang Wang, Ziqiang Liu, Ziwei Chai, Zixuan Li, Zuodong Zhao, Wenguang Chen, Jidong Zhai, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models, 2025. URL https://arxiv.org/abs/2508.06471.

Rui Wang, Hongru Wang, Yi Chen, Boyang Xue, Tianqing Fang, Wenhao Yu, and Kam-Fai Wong. Demystifying on-policy distillation: Roles, pathologies, and regulations. 2026. URL https://arxiv.org/abs/2607.13399.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Lifu Tang, Xiaowei Lv, Haosheng Zou, Yongchao Deng, Shousheng Jia, and Xiangzheng Zhang. Light-r1: Curriculum SFT, DPO and RL for long COT from scratch and beyond. In Georg Rehm and Yunyao Li, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 318–327, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-288-6. doi: 10.18653/v1/2025.acl-industry.24. URL https://aclanthology.org/2025.acl-industry.24/.

Guande Wu, Huan Song, Yawei Wang, Qiaojing Yan, Yijun Tian, Lin Lee Cheong, and Panpan Xu. Sdrt: Enhance vision-language models by self-distillation with diverse reasoning traces. 2025. URL https://arxiv.org/abs/2503.01754.

Yecheng Wu, Song Han, and Hai Cai. Lightning opd: Efficient post-training for large reasoning models with offline on-policy distillation. 2026. URL https://arxiv.org/abs/2604.13010.

Wei Xiong, Hanning Zhang, Chenlu Ye, Lichang Chen, Nan Jiang, and Tong Zhang. Self-rewarding correction for mathematical reasoning. 2025. URL https://arxiv.org/abs/2502.19613.

Fangzhi Xu, Hang Yan, Chang Ma, Haiteng Zhao, Qiushi Sun, Kanzhi Cheng, Junxian He, Jun Liu, and Zhiyong Wu. Genius: A generalizable and purely unsupervised self-training framework for advanced reasoning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13153–13167, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.644. URL https://aclanthology.org/2025.acl-long.644/.

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.

Chenglin Yang, Lingxi Xie, Chi Su, and Alan L. Yuille. Snapshot distillation: Teacher-student optimization in one generation. 2018. URL https://arxiv.org/abs/1812.00123.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. 2026. URL https://arxiv.org/abs/2604.03128.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. 2026. URL https://arxiv.org/abs/2602.12275.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. LIMO: Less is more for reasoning. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=T2TZ0RY4Zk.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. 2024.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models. 2023. URL https://arxiv.org/abs/2308.01825.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? 2025. URL https://openreview.net/forum?id=4OsgYD7em5.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. Star: Bootstrapping reasoning with reasoning. 2022. URL https://arxiv.org/abs/2203.14465.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3712–3721, 2019. doi: 10.1109/ICCV.2019.00381.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Peilin Zhao, and Yatao Bian. Right question is already half the answer: Fully unsupervised llm reasoning incentivization. 2025. URL https://arxiv.org/abs/2504.05812.

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. 2025. URL https://arxiv.org/abs/2505.03335.

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*, 2026a.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to reason without external rewards. 2026b. URL https://openreview.net/forum?id=OU9nFEYR2M.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. TTRL: Test-time reinforcement learning. 2025. URL https://openreview.net/forum?id=VuVhgEiu20.

A Background Objectives

For completeness, we formalize the externally supervised objectives summarized in Section 3.1. These objectives are prior work and are not contributions of this paper.

A.1 Group Relative Policy Optimization

GRPO (Shao et al., 2024) optimizes the policy with a verifiable reward normalized within a group of sampled rollouts: for each problem it samples G rollouts $y^{(1)}, \dots, y^{(G)}$, scores them with $r^{(g)} = \mathbb{1}\big[\mathrm{Ans}(y^{(g)}) = a^\star\big]$, and normalizes within the group, $A^{(g)} = \big(r^{(g)} - \mathrm{mean}\{r^{(j)}\}_{j=1}^{G}\big) / \mathrm{std}\{r^{(j)}\}_{j=1}^{G}$, so that all tokens of a rollout share one sequence-level advantage. The policy is then updated on the clipped surrogate

$$\mathcal{L}_{\mathrm{GRPO}}(\theta) = -\,\mathbb{E}_{x\sim\mathcal{S}}\,\mathbb{E}_{\{y^{(g)}\}_{g=1}^{G}\sim\bar\pi(\cdot|x)}\,\frac{1}{G}\sum_{g=1}^{G}\frac{1}{|y^{(g)}|}\sum_{t=1}^{|y^{(g)}|}\min\Big(\rho_t^{(g)}A^{(g)},\ \mathrm{clip}\,\big(\rho_t^{(g)},\,1-\varepsilon,\,1+\varepsilon\big)A^{(g)}\Big), \tag{3}$$

where $\rho_t^{(g)} = \pi_\theta\big(y_t^{(g)} \mid x, y_{<t}^{(g)}\big) / \bar\pi\big(y_t^{(g)} \mid x, y_{<t}^{(g)}\big)$ is the importance ratio to the behaviour policy and ε is the clipping range. The supervision is the gold answer $a^\star$, and the signal is sparse: it is sequence-level, and vanishes whenever all G rewards coincide.

A.2 On-Policy Distillation

On-policy distillation (OPD) addresses this sparsity by retaining the model's own rollouts as training trajectories while replacing the scalar reward with dense supervision from a teacher's next-token distribution. The student is conditioned only on the problem x, matching inference time, while the teacher receives a possibly richer context c. Writing π_T for the teacher, these methods share the form

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_x\,\mathbb{E}_{y\sim\bar\pi(\cdot|x)}\sum_{t=1}^{|y|} D_\beta\big(\pi_T(\cdot \mid c,\, y_{<t})\,\big\|\,\pi_\theta(\cdot \mid x,\, y_{<t})\big), \tag{4}$$

which yields dense token-level supervision at states the student actually visits. Variants differ only in the choice of teacher π_T and the context c it receives. OPD (Gu et al., 2024; Agarwal et al., 2024; Lu and Lab, 2025) takes π_T to be a separate, typically stronger model with $c = x$. The supervision is dense, but it resides in an external teacher that must already surpass the model being trained.

A.3 On-Policy Self-Distillation

OPSD (Zhao et al., 2026a) takes a further step by constructing the teacher and student from the same language model under different conditioning contexts, removing the external teacher and supplying a *gold solution* in its place. The student policy observes only the problem statement x, matching inference time, while the teacher policy conditions on both the problem statement x and the GT solution $y^\star$:

$$\pi_S(\cdot \mid x) \triangleq \pi_\theta(\cdot \mid x), \qquad \pi_T(\cdot \mid x, y^\star) \triangleq \bar\pi(\cdot \mid x, y^\star).$$

Given a problem x paired with a GT solution $y^\star$, OPSD samples the full next-token distribution from both policies, then minimizes a token-level divergence between them:

$$\mathcal{L}_{\mathrm{OPSD}}(\theta) = \mathbb{E}_{(x,y^\star)\sim\mathcal{S}}\,\mathbb{E}_{y\sim\bar\pi(\cdot|x)}\sum_{t=1}^{|y|} D_\beta\big(\pi_T(\cdot \mid x,\, y^\star,\, y_{<t})\,\big\|\,\pi_S(\cdot \mid x,\, y_{<t})\big). \tag{5}$$

Because the teacher sees $y^\star$, its next-token distribution along the student's trajectory encodes what a model that already knows the solution would do at each position, and matching it distills solution-conditioned behavior into the solution-free policy. Following the strongest OPSD configuration, we use $\beta = 0$ (forward KL, teacher $\to$ student) with per-token point-wise clipping for stability. Teacher and student share parameters, yet OPSD still requires external supervision, i.e., the reference solution $y^\star$.

B Additional Experimental Results

B.1 Training Dynamics and Convergence

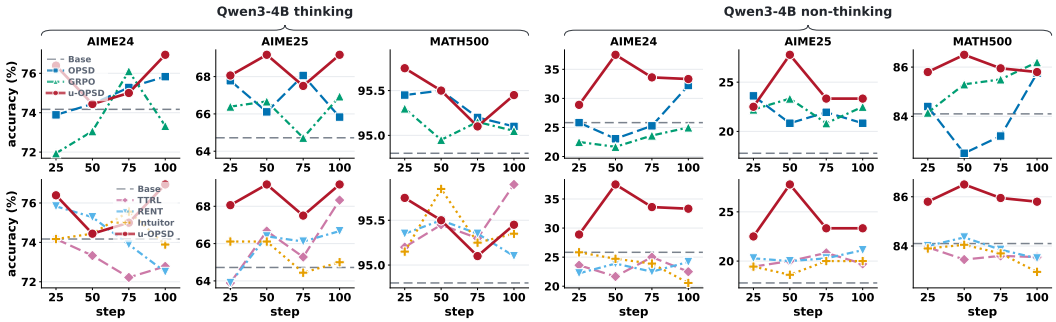

Figure 5: Training curves of the Qwen3-4B thinking (left) and non-thinking mode (right) runs on AIME24, AIME25 and MATH500. **Top:** U-OPSD against the supervised methods. **Bottom:** U-OPSD against the label-free methods.

Figure 5 plots the accuracy over checkpoints of the Qwen3-4B runs on AIME24, AIME25 and MATH500 in both modes. The two modes separate immediately. In thinking mode every method stays within about 2% of the base model and the ordering changes from checkpoint to checkpoint, with the whole MATH500 panel spanning less than 1%. The label-free baselines track the base model closely, as do U-OPSD and OPSD. In non-thinking mode U-OPSD is above every other method at every checkpoint on both AIME benchmarks, against the supervised and the label-free baselines alike; on MATH500 the other methods catch up to the same level by step 100.

B.2 Objective Ablations

B.2.1 Divergence computation strategy

Table 4: Comparison of divergence computation strategy: Full vocabulary is logit distillation over every token (Agarwal et al., 2024); sampled token evaluates the two policies only at the token the student drew (Lu and Lab, 2025); top-k rows truncate the teacher to its k largest entries. We report on Qwen3-8B non-thinking at the best checkpoint.

Variant		AIME24	AIME25	HMMT25	MATH500	AMC23	Avg.
Student token		29.44	19.72	12.78	84.70	70.62	43.45
Full-vocabulary		**53.89**	37.50	20.00	89.75	84.38	57.10
top-k	20	49.17	37.78	19.44	90.20	**88.12**	56.94
	50	48.06	34.44	**24.17**	89.45	83.12	55.85
	100	**53.89**	**43.33**	20.56	**91.65**	85.62	**59.01**
	200	51.67	40.00	22.50	90.15	86.25	58.11

Full-vocabulary logit distillation (Agarwal et al., 2024) evaluates D_β over the entire vocabulary at each position, giving a proper token-level divergence between the two policies. Sampled-token distillation (Lu and Lab, 2025) evaluates teacher and student log-probabilities only at the token the student sampled, and uses that term as a scalar advantage inside a policy-gradient objective. OPSD (Zhao et al., 2026a) reports the full-distribution objective to be the stronger of the two under gold supervision, and because U-OPSD changes the teacher's conditioning context rather than its objective, the open question is whether that preference survives once the privileged context is a pseudo-solution. Table 4 shows that it does, and by a wider margin: the full-distribution objective leads by 17.8% on AIME25 and 7.2% on HMMT25, against 2.0% and 2.7% under gold supervision. The student token-only variant is 13.7% below the default and 15.6% below the best setting, so it is not competitive under pseudo-label supervision. Yet these tend to be the extremes of one axis, so we

further test whether an intermediate solution suffices: restricting the divergence to the teacher's k largest entries, from k=20 upward. The best score on each of the five benchmarks falls mostly under top-100, and the full vocabulary exceeds none of them on average, tying top-100 on AIME24 and trailing elsewhere. The top-k truncation is therefore a practical optimization that reduces the cost of computing the divergence.

B.2.2 DIVERGENCE METRICS

Table 5: Divergence metrics D_β under U-OPSD, on Qwen3-8B non-thinking, longest-1, G=8, τ=0.5.

Objective	AIME24	AIME25	HMMT25	MATH500	AMC23	Avg.
Base	27.50	23.33	13.61	84.05	69.38	43.57
Forward KL $\mathrm{KL}(\pi_T \,\|\, \pi_S)$, β=0 (default)	**53.89**	**37.50**	**20.00**	**89.75**	**84.38**	**57.10**
Reverse KL $\mathrm{KL}(\pi_S \,\|\, \pi_T)$, β=1	*training diverges (not scored)*					
JSD (β=0.5)	29.44	21.94	12.22	83.70	69.38	43.34

OPSD uses forward KL with β=0 and Table 5 shows the preference towards forward KL is stronger under U-OPSD. Using symmetric Jensen–Shannon divergence (JSD) costs 13.8%, which puts it level with the untrained model, and reverse KL does not produce a converged model. The collapse under reverse KL is gradual and one-directional: generations average 2.7k characters at step 25, 12k at step 50, 76k at step 75, and 99k — the token ceiling — from step 125 on, while the fraction of rollouts that yield a parsable boxed answer falls from 99% to 33%. Inspection of the completions shows a loss of termination rather than cohesive reasoning (e.g., a phrase, a LaTeX command pair, or a nesting bracket repeated until the token budget runs out).