SPOT: SPARSE PROBING AND OUTCOME CALIBRATION FOR ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) provides dense teacher supervision on student-generated trajectories, but standard reverse-KL training can assign insufficient probability to other plausible continuations. Teacher entropy alone does not reveal whether uncertainty is concentrated among a few plausible next tokens or dispersed over a long probability tail, nor whether the student already represents those candidates well. Moreover, local teacher probabilities may not predict downstream success. We introduce *Sparse Probing and Outcome-calibrated Targets* OPD (`SPOT`), which addresses two coupled decisions, *where to probe* and *what to distill*, through an acquisition–exploration–exploitation procedure. During acquisition, a position-level score combines normalized teacher entropy, the probability mass captured by a small top-k candidate set, and student–teacher mismatch to allocate a limited probing budget. During exploration, `SPOT` evaluates teacher-proposed candidates through verifier-scored student continuations. During exploitation, these outcomes produce a closed-form, KL-regularized target that favors candidates with better downstream outcomes while remaining anchored to the teacher distribution. Extensive experiments across multiple student models and reasoning benchmarks demonstrate the effectiveness of `SPOT` in improving reasoning performance while balancing solution quality and coverage.

1 INTRODUCTION

Transferring the reasoning capabilities of large language models to smaller student models is an important goal in LLM post-training. Supervised fine-tuning and off-policy distillation leverage expert- or teacher-generated trajectories, but train the student on contexts that differ from those induced by its own predictions, leading to exposure bias and compounding errors at inference (Bengio et al., 2015; Ranzato et al., 2016). On-policy reinforcement learning (RL) instead optimizes student-generated rollouts, but reasoning-oriented RL typically relies on sequence-level or terminal verifier rewards, providing limited fine-grained credit assignment (Shao et al., 2024; Guo et al., 2025). On-policy distillation (OPD) combines these advantages: it trains on student-generated prefixes while retaining dense token-level teacher feedback (Agarwal et al., 2024; Lu & Lab, 2025).

Yet standard OPD typically minimizes reverse KL, whose mode-seeking behavior favors the teacher's dominant continuation but can assign insufficient probability to other plausible continuations, potentially limiting solution coverage. EOPD addresses this limitation by using teacher entropy as a trigger: at high-entropy positions, it augments reverse-KL training with a top-k approximation of forward KL to preserve plausible local alternatives (Minka, 2005; Jin et al., 2026). However, high entropy alone does not reveal whether the teacher's uncertainty is concentrated among a few plausible next tokens or dispersed over a long probability tail. Nor does it reveal whether the student already assigns sufficient probability to and similarly ranks those plausible tokens. Evaluating an alternative next token requires rolling out a student continuation and checking its final outcome, making such probing costly. An entropy threshold can flag uncertain positions, but by itself cannot prioritize where a limited probing budget will be most useful. The first challenge is therefore to allocate that budget to positions where the teacher assigns substantial probability to a small set of alternatives that the student does not yet represent well.

Selecting a candidate position does not yet determine *what to distill*. On student-generated prefixes, the teacher's local next-token probabilities need not predict downstream success: a token assigned

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

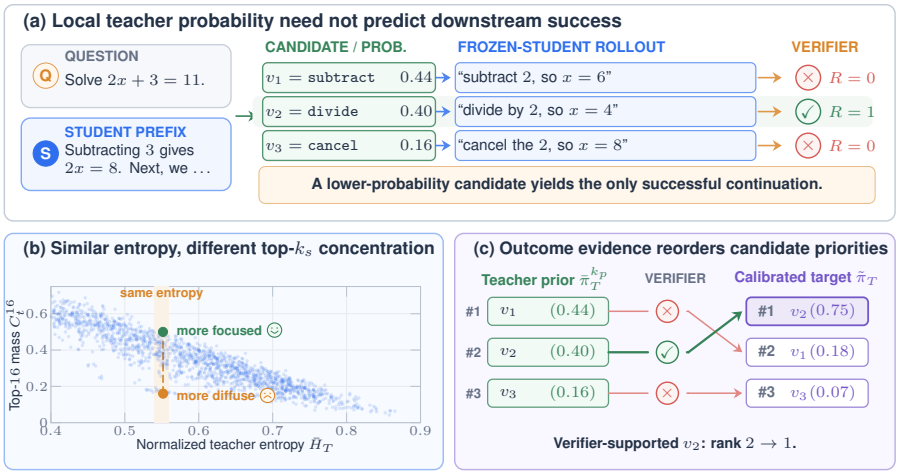


Figure 1: **Why SPOT probes before distilling.** (a) Similarly likely teacher candidates can yield different student-continuation outcomes. (b) Similar entropy can correspond to different top-k_s support concentration. (c) Outcome calibration promotes verifier-supported candidates.

higher probability by the teacher may yield an unsuccessful continuation under the current student policy, whereas a lower-probability alternative may yield a successful one (Li et al., 2026; Hou et al., 2026). We therefore treat the teacher distribution as a proposal prior over candidate branches, rather than as definitive evidence of downstream success, and calibrate it using the outcomes of student continuations. Thus, teacher uncertainty and student mismatch help prioritize where to acquire additional evidence, while verified downstream outcomes determine how that evidence should modify the supervision target.

These observations motivate **Sparse Probing and Outcome-calibrated Targets for on-policy distillation** (SPOT), which reframes uncertainty-aware distillation around two coupled decisions: *where to probe* and *what to distill*. SPOT addresses them through a three-stage acquisition–exploration–exploitation procedure. During acquisition, a lightweight position-level score s_t prioritizes positions that satisfy three conditions: the teacher assigns meaningful probability to multiple next tokens, most of the teacher's probability mass lies within a small top-k candidate set, and the student either underweights or differently ranks those candidates. Because these factors are multiplied, a low value on any one condition lowers the position's overall probing priority. The score therefore estimates *where additional evidence may be useful*; it does not identify which candidates yield successful student continuations, which is assessed only after the candidates are rolled out and verified. During exploration, at each selected position, SPOT appends each candidate from the teacher's top-k set in turn, rolls out a continuation under the student policy, and evaluates the completed continuation with a verifier. During exploitation, the verified continuation values produce a closed-form, KL-regularized target: candidates with better downstream outcomes receive more probability, while the target remains anchored to the teacher distribution. SPOT applies this additional local loss only at positions where at least one tested candidate receives positive verifier reward. This allows downstream evidence to modify the local token-level target rather than serving only as a trajectory-level score.

In summary, our contributions are threefold:

- **A two-decision formulation.** We formulate selective supervision in OPD as determining *where* additional outcome evidence is worth acquiring and *how* that evidence should be converted into a supervision target, separating position selection from target construction.
- **Sparse probing and outcome-calibrated targets.** We propose SPOT, which probes positions via teacher uncertainty, top-k mass, and student mismatch, then derives a closed-form target by reward-tilting the teacher distribution with verifier-scored student continuations.
- **Empirical validation.** Across three evaluated Qwen student scales and six mathematical reasoning benchmarks, SPOT achieves the highest macro Pass@8 in all three settings and the highest or second-highest macro Avg@8 among the compared methods. These results indicate stronger multi-sample solution coverage while maintaining competitive average accuracy.

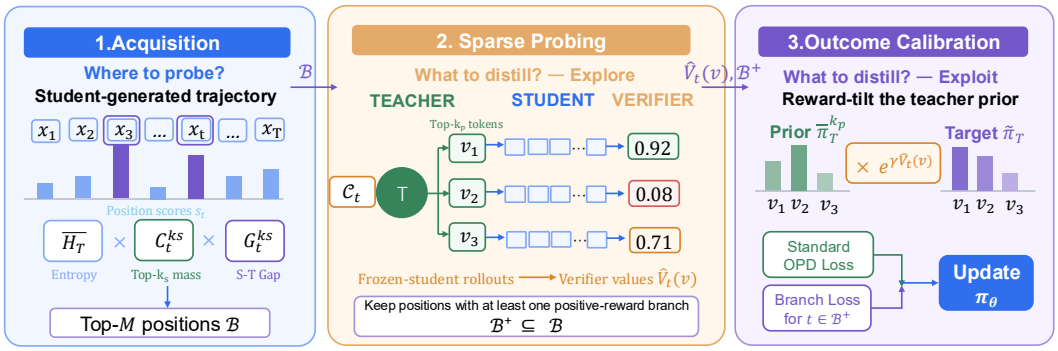


Figure 2: An overview of our SPOT framework for on-policy distillation.

2 Preliminaries

On-Policy Distillation. Let $\mathcal{D}$ denote the prompt distribution, and let π_θ and π_T be the student and teacher policies over a shared vocabulary $\mathcal{V}$. Given $q \sim \mathcal{D}$, OPD samples a student trajectory $x = (x_1, \ldots, x_T) \sim \pi_\theta(\cdot \mid q)$ and queries the teacher on each student-induced prefix $c_t = (q, x_{<t})$. At each prefix, it uses the reverse-KL loss $\mathcal{L}_t^{\mathrm{OPD}} := D_{\mathrm{KL}}(\pi_\theta(\cdot \mid c_t) \| \pi_T(\cdot \mid c_t))$. Averaging over the student trajectory gives

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{q \sim \mathcal{D},\, x \sim \pi_\theta(\cdot|q)} \left[\frac{1}{T} \sum_{t=1}^{T} \mathcal{L}_t^{\mathrm{OPD}} \right]. \tag{1}$$

In practice, this objective is estimated from tokens sampled by a frozen behavior policy $\pi_{\theta_{\mathrm{old}}}$ and optimized with PPO-style clipping. Querying the teacher on student-visited prefixes reduces state-distribution mismatch and provides dense token-level feedback. However, the mode-seeking reverse KL can under-cover plausible alternatives when the teacher is uncertain.

Entropy-Aware On-Policy Distillation. EOPD (Jin et al., 2026) augments OPD with mode-covering supervision at positions of high teacher entropy. Let $H_T(c_t) = \mathcal{H}(\pi_T(\cdot \mid c_t))$ denote the teacher entropy and $\bar{H}_T(c_t) = H_T(c_t)/\log|\mathcal{V}| \in [0,1]$ its vocabulary-normalized counterpart. EOPD gates supervision using H_T, whereas our acquisition score later uses the bounded, vocabulary-normalized entropy $\bar{H}_T$. For efficiency, let $S_t^k := \mathrm{TopK}_k(\pi_T(\cdot \mid c_t))$ denote the teacher's top-k candidate set. For $j \in \{T, \theta\}$, write $\bar{\pi}_j^k(v \mid c_t) := \pi_j(v \mid c_t) / \sum_{u \in S_t^k} \pi_j(u \mid c_t)$ for the restriction of π_j renormalized on this candidate set. EOPD approximates forward KL as

$$\mathcal{L}_t^{\mathrm{FKL}} := \sum_{v \in S_t^k} \bar{\pi}_T^k(v \mid c_t) \log \frac{\bar{\pi}_T^k(v \mid c_t)}{\pi_\theta(v \mid c_t)}. \tag{2}$$

Only the teacher target is truncated and renormalized; the student retains its full-vocabulary probabilities. With entropy threshold τ and forward-KL weight α, the per-token objective is

$$\mathcal{L}_t^{\mathrm{EOPD}} := \mathcal{L}_t^{\mathrm{OPD}} + \alpha\, \mathbb{I}[H_T(c_t) > \tau]\, \mathcal{L}_t^{\mathrm{FKL}}. \tag{3}$$

Here, π_j always denotes a full-vocabulary policy, while $\bar{\pi}_j^k$ denotes its top-k-renormalized shape—a distinction that separates probability mass from relative shape in our method. EOPD promotes coverage, but still uses a scalar entropy criterion to decide *where* to intervene and the uncalibrated teacher prior to determine *what* to distill; SPOT revisits both decisions.

3 Methodology

3.1 Overview

SPOT follows a three-stage acquisition–exploration–exploitation procedure that addresses two coupled decisions: *where to probe* and *what to distill*. As illustrated in Figure 2, acquisition uses a lightweight position-level score to prioritize where to acquire additional evidence; exploration estimates the downstream values of teacher-proposed next-token candidates through verifier-scored student continuations; and exploitation combines these values with the teacher probabilities to construct outcome-calibrated targets. In this way, acquisition allocates the probing budget, while exploration and exploitation determine how the acquired evidence modifies local supervision.

Algorithm 1 SPOT Training

Require: Student π_θ, teacher π_T, verifier R, prompt set $\mathcal{D}$, learning rate η
Require: Position budget M, top-k sizes k_s (scoring) and k_p (probing), inverse temperature γ, branch-loss weight β

for each training iteration **do**
  $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
  Rollout buffer $\mathcal{R} \leftarrow \emptyset$
  for each prompt q in a batch sampled from $\mathcal{D}$ **do**
    Roll out $x \sim \pi_{\theta_{\text{old}}}(\cdot \mid q)$
    Query the teacher $\pi_T(\cdot \mid c_t)$ at every prefix $c_t = (q, x_{<t})$
    Score valid positions by $s_t = \bar{H}_T(c_t) C_t^{k_s} G_t^{k_s}$ using Eqs. (4)–(5)
    Select sparse probing positions $\mathcal{B} = \text{Top-}M\big(\{s_t\}_{t=1}^{|x|}\big)$
    for each position $t \in \mathcal{B}$ and candidate $v \in S_t^{k_p}$ **do**
      Sample $y \sim \pi_{\theta_{\text{old}}}(\cdot \mid c_t, v)$ and estimate $\hat{V}_t(v)$ using $R(q, x_{<t}, v, y)$
    Retain positions with at least one positive-reward candidate as $\mathcal{B}^+ \subseteq \mathcal{B}$
    Form reward-tilted targets $\tilde{\pi}_T(\cdot \mid c_t)$ for $t \in \mathcal{B}^+$ using Eq. (8)
    Add the rollout and its supervision to $\mathcal{R}$
  for each gradient step on a mini-batch sampled from $\mathcal{R}$ **do**
    Update $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$ by minimizing

$$\mathcal{L} = \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}_t^{\text{OPD}} + \frac{\beta}{\max\{1, |\mathcal{B}^+|\}} \sum_{t \in \mathcal{B}^+} \mathcal{L}_t^{\text{Branch}}$$

    using Eq. (10)

3.2 SPOT: Sparse Probing and Outcome-Calibrated Targets

Position Acquisition: Where to Probe. Because probing next-token candidates requires extra rollouts, acquisition prioritizes positions where several teacher candidates are plausible, a small candidate set captures most teacher mass, and the student does not already represent those candidates well. Entropy alone cannot distinguish this pattern from uncertainty spread over a long tail. Recent analyses also suggest that OPD gains depend on student–teacher compatibility and genuinely new teacher information (Li et al., 2026). We therefore use student–teacher mismatch to signal potential correction needed, while verifier-scored rollouts test the utility of the selected candidates.

Starting from the normalized teacher entropy $\bar{H}_T(c_t)$, let k_s denote the number of teacher candidates used to compute the position-level acquisition score. We first measure the teacher mass captured by its top-k_s candidate set:

$$C_t^{k_s} = \sum_{v \in S_t^{k_s}} \pi_T(v \mid c_t). \tag{4}$$

The quantity $C_t^{k_s}$ is the teacher probability mass captured by its top-k_s candidates. A large value means that a small candidate set represents most of the teacher distribution. Combined with high entropy, this favors *positions with multiple plausible candidates* in that set over positions whose uncertainty is dispersed across a long probability tail.

Teacher-side structure, however, does not imply that the resulting supervision is useful to a student that may already represent the same alternatives. We thus measure the student–teacher gap on the same candidate set:

$$G_t^{k_s} = \lambda_{\text{mass}}(1 - A_t^{k_s}) + \lambda_{\text{shape}} D_{\text{JS}}(\bar{\pi}_T^{k_s} \,\|\, \bar{\pi}_{\theta_{\text{old}}}^{k_s}), \tag{5}$$

where $A_t^{k_s} = \sum_{v \in S_t^{k_s}} \pi_{\theta_{\text{old}}}(v \mid c_t)$ and $\lambda_{\text{mass}} + \lambda_{\text{shape}} = 1$. Here D_{JS} denotes the Jensen–Shannon divergence normalized to $[0, 1]$; the two terms capture mass under-coverage and relative-shape mismatch, respectively.

The final acquisition score $s_t = \bar{H}_T(c_t) C_t^{k_s} G_t^{k_s}$ acts as a soft conjunction: it is high only when the teacher presents multiple plausible candidates, the top-k_s set captures most teacher proba-

bility mass, and the student either underweights or differently ranks those candidates. After masking special, padding, whitespace-only, and punctuation-only tokens, we allocate probing to $\mathcal{B} = \text{Top-}M(\{s_t\}_{t=1}^{|x|})$. Consequently, the rollout budget is reserved for compact teacher alternatives that the current student neither adequately covers nor matches.

Sparse Probing and Outcome Calibration: What to Distill. Acquisition determines where to probe, but local teacher probabilities need not reliably indicate downstream success on deeper student-generated prefixes (Li et al., 2026): a higher-probability candidate may fail while a lower-probability one succeeds (Hou et al., 2026). We therefore treat the teacher distribution as a proposal prior and calibrate it with continuation outcomes.

To obtain outcome-grounded evidence, we probe the teacher's top-k_p candidates at each selected position $t \in \mathcal{B}$. For each $v \in S_t^{k_p}$, we append v to the prefix c_t and independently sample N_p continuations $y^{(n)} \sim \pi_{\theta_{\text{old}}}(\cdot \mid c_t, v)$, $n = 1, \ldots, N_p$, from the frozen behavior policy. We define the student-executable value of v as

$$\hat{V}_t(v) = \mathbb{E}_{y \sim \pi_{\theta_{\text{old}}}(\cdot \mid c_t, v)}[R(q, x_{<t}, v, y)], \tag{6}$$

which measures whether the current student can turn a teacher-proposed local alternative into a high-reward trajectory. To avoid imposing additional supervision where exploration finds no viable alternative, we retain only positions for which at least one candidate has a positive estimated continuation value; the resulting set is denoted by $\mathcal{B}^+ \subseteq \mathcal{B}$. Because probing evaluates at most $M \times k_p \times N_p$ candidate continuations per trajectory, its overhead is explicitly controlled by the position, candidate, and per-candidate sampling budgets.

To convert the probed outcomes into actionable supervision, we construct *outcome-calibrated targets* at each retained position $t \in \mathcal{B}^+$. Let $\Delta(S_t^{k_p}) = \{\rho_t : \rho_t(v) \geq 0, \sum_{v \in S_t^{k_p}} \rho_t(v) = 1\}$ denote the probability simplex over the teacher candidate set. Rather than collapsing onto the empirically best candidate, we maximize expected downstream value within a KL trust region around the renormalized teacher prior:

$$\max_{\rho_t \in \Delta(S_t^{k_p})} \sum_{v \in S_t^{k_p}} \rho_t(v) \hat{V}_t(v) \quad \text{s.t.} \quad D_{\text{KL}}\Big(\rho_t \,\|\, \bar{\pi}_T^{k_p}(\cdot \mid c_t)\Big) \leq \epsilon. \tag{7}$$

The derivation is deferred to Appendix A. For the corresponding dual parameter $\gamma > 0$, which acts as an inverse temperature, the problem admits the closed-form optimizer

$$\tilde{\pi}_T(v \mid c_t) = \frac{\bar{\pi}_T^{k_p}(v \mid c_t) \exp\Big(\gamma \hat{V}_t(v)\Big)}{\sum_{u \in S_t^{k_p}} \bar{\pi}_T^{k_p}(u \mid c_t) \exp\Big(\gamma \hat{V}_t(u)\Big)}. \tag{8}$$

In log space, equation 8 reads $\log \tilde{\pi}_T(v \mid c_t) = \log \bar{\pi}_T^{k_p}(v \mid c_t) + \gamma \hat{V}_t(v) - \log Z_t$, exposing each target log-probability as *teacher log-probability plus an outcome-grounded continuation-value bonus*, up to the normalizer Z_t. Subtracting this expression for any $u, v \in S_t^{k_p}$ yields the pairwise log-odds decomposition

$$\log \frac{\tilde{\pi}_T(v \mid c_t)}{\tilde{\pi}_T(u \mid c_t)} = \log \frac{\bar{\pi}_T^{k_p}(v \mid c_t)}{\bar{\pi}_T^{k_p}(u \mid c_t)} + \gamma\Big(\hat{V}_t(v) - \hat{V}_t(u)\Big). \tag{9}$$

This identity makes the division of roles explicit: the teacher supplies the prior odds, while the verifier adds a correction proportional to relative downstream value. Conditional on $t \in \mathcal{B}^+$, only value differences affect the target, making it invariant to uniform reward shifts. Together with the $\mathcal{B}^+$ gate, outcome feedback first tests absolute viability and then corrects relative preference. As $\gamma \to 0$, the target recovers the teacher prior; larger γ increasingly favors candidates with higher estimated continuation values.

Finally, we augment OPD with outcome-calibrated local supervision at verifier-supported positions:

$$\mathcal{L} = \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}_t^{\text{OPD}} + \frac{\beta}{\max\{1, |\mathcal{B}^+|\}} \sum_{t \in \mathcal{B}^+} \mathcal{L}_t^{\text{Branch}}, \tag{10}$$

where $\mathcal{L}_t^{\text{Branch}} = -\sum_{v \in S_t^{k_p}} \tilde{\pi}_T(v \mid c_t) \log \pi_\theta(v \mid c_t)$. The denominator averages this auxiliary loss over retained positions, and β controls its strength relative to trajectory-wide OPD. When $\mathcal{B}^+ = \varnothing$, the branch term vanishes and the objective reduces to standard OPD.

4 EXPERIMENTS

For a rigorous, controlled comparison of SPOT, we follow the setup of recent OPD work (Jin et al., 2026). Further details appear in Appendix B.

4.1 EXPERIMENTAL SETTINGS

Models and Training Data. We use Qwen3-8B, with thinking mode disabled, as the common teacher for all distillation methods, and Qwen3-0.6B-Base, Qwen3-1.7B-Base, and Qwen3-4B-Base as students (Yang et al., 2025). The 0.6B and 1.7B students are trained on MATH (Hendrycks et al., 2021); the 4B student uses the more challenging DAPO dataset (Yu et al., 2025).

Baselines. We compare SPOT with four baselines:

- **KD** (Hinton et al., 2015; Kim & Rush, 2016): off-policy distillation using forward KL and cross-entropy on teacher-generated data.
- **OPD** (Agarwal et al., 2024; Li et al., 2026): on-policy training on student rollouts with per-token reverse-KL supervision.
- **GRPO** (Shao et al., 2024): outcome-based RL using group-relative advantages from verifiable rewards, without teacher supervision.
- **EOPD** (Jin et al., 2026): OPD augmented with a top-k forward-KL term at positions selected by a fixed teacher-entropy threshold.

Evaluation. We evaluate all models zero-shot with the same prompt template and answer verifier on MATH-500 (Hendrycks et al., 2021; Lightman et al., 2023), AIME 2024 (Zhang & Math-AI, 2024), AIME 2025 (Zhang & Math-AI, 2025), AMC 2023 (Yang et al., 2024), Minerva Math (Lewkowycz et al., 2022), and HMMT 2025 (Balunović et al., 2026). We sample eight responses per problem with temperature 1.0, top-p sampling with $p = 0.8$, and a maximum response length of 8192 tokens. **Avg@8** is the mean response accuracy, and **Pass@8** is the fraction of problems solved at least once. We report per-benchmark scores and their unweighted macro-average across the six benchmarks.

4.2 MAIN RESULTS

Consistent gains across student scales and training sets. As shown in Table 1, SPOT achieves the best macro Pass@8 at all three evaluated student scales and the best or second-best macro Avg@8. Relative to standard OPD, SPOT improves macro Avg@8 by 0.47–1.48 points and macro Pass@8 by 4.55–5.28 points. Relative to EOPD, the closest uncertainty-aware OPD baseline, SPOT improves macro Avg@8 by 0.29–0.68 points and macro Pass@8 by 2.49–3.19 points. These gains persist from 0.6B to 4B students and across MATH and DAPO-Math-14k training, demonstrating that the benefit is not tied to a particular capacity or training regime.

Broader multi-sample coverage without sacrificing average accuracy. Avg@8 estimates per-sample correctness, whereas Pass@8 measures whether eight samples reach at least one correct solution. The markedly larger Pass@8 gains, together with preserved or improved Avg@8, reveal a favorable coverage–quality profile: SPOT places meaningful probability on viable alternatives without diluting an individual attempt. This directly matches SPOT's goal of improving coverage without sacrificing average accuracy. The comparison with EOPD is consistent with our motivating distinction: teacher entropy alone cannot determine whether uncertainty is concentrated in a compact candidate set, whether the student already represents those candidates, or which candidates lead to successful student continuations. SPOT addresses these gaps by using normalized teacher entropy, top-k_s probability mass, and student–teacher mismatch to allocate the probing budget, then using verifier-scored continuation values to construct outcome-calibrated local targets. SPOT therefore selectively expands coverage toward useful alternative reasoning modes rather than stylistic or erroneous variation.

Table 1: Main results on six mathematical reasoning benchmarks. *Avg.* is the unweighted mean over the six benchmarks, computed before rounding. **Bold** and underlined entries denote the best and second-best results within each student setting, respectively.

Method	Avg@8							Pass@8						
	MATH500	AMC23	Minerva	HMMT	AIME24	AIME25	Avg.	MATH500	AMC23	Minerva	HMMT	AIME24	AIME25	Avg.
Student Model: ***Qwen3-0.6B-Base*** (Training Data: MATH)														
KD	47.80	23.43	14.34	0.21	2.08	0.83	14.78	69.60	52.50	30.88	1.67	6.67	6.67	28.00
GRPO	**53.33**	28.13	**16.41**	1.04	4.58	0.83	**17.39**	74.40	55.00	32.72	3.33	10.00	10.00	30.91
OPD	50.10	24.69	16.04	0.42	2.50	1.25	15.83	73.20	57.50	31.25	1.67	10.00	6.67	30.05
EOPD	50.50	27.81	15.95	1.46	4.17	1.25	16.86	75.00	55.00	33.46	5.00	**13.33**	6.67	31.41
SPOT (Ours)	50.50	**28.44**	15.99	**1.88**	**5.00**	**2.08**	17.31	**76.40**	**60.00**	**34.56**	**6.67**	**13.33**	**16.67**	**34.60**
Student Model: ***Qwen3-1.7B-Base*** (Training Data: MATH)														
KD	62.85	37.81	27.07	1.25	10.42	3.33	23.79	84.20	70.00	44.12	3.33	20.00	16.67	39.72
GRPO	67.00	38.13	28.17	1.88	9.17	5.42	24.96	84.00	72.50	**48.16**	**6.67**	20.00	16.67	41.33
OPD	67.03	39.06	27.62	1.46	8.33	6.25	24.96	84.80	70.00	47.06	3.33	20.00	16.67	40.31
EOPD	**67.68**	**40.31**	27.90	1.67	10.83	**7.50**	25.98	83.60	**75.00**	44.49	**6.67**	26.67	20.00	42.74
SPOT (Ours)	67.43	39.69	**28.45**	**2.08**	**12.50**	**7.50**	**26.27**	**87.80**	**75.00**	47.43	**6.67**	**33.33**	**23.33**	**45.59**
Student Model: ***Qwen3-4B-Base*** (Training Data: DAPO-Math-14k)														
KD	74.73	48.13	34.65	3.13	12.50	12.08	30.87	92.20	80.00	51.47	13.33	26.67	23.33	47.83
GRPO	79.18	53.13	**39.84**	2.71	14.17	13.33	33.73	90.00	80.00	52.21	11.67	26.67	26.67	47.87
OPD	**80.38**	55.94	37.64	5.83	**17.92**	16.25	35.66	93.00	80.00	51.10	13.33	30.00	30.00	49.57
EOPD	79.73	55.63	36.08	7.08	17.50	16.67	35.45	93.20	82.50	51.84	16.67	**33.33**	33.33	51.81
SPOT (Ours)	79.88	**56.25**	37.73	**7.50**	17.50	**17.92**	**36.13**	**93.40**	**85.00**	**54.04**	**23.33**	**33.33**	**36.67**	**54.30**

Table 2: **Branch-acquisition score ablation.** Avg@8 / Pass@8 (%) for Qwen3-0.6B-Base trained on MATH. *Macro* denotes the unweighted mean; best results are bold and FULL is shaded.

Acquisition score	Avg@8 / Pass@8				
	MATH500	AMC23	AIME24	AIME25	Macro
`H`: $\bar{H}_T$	50.50 / 76.20	25.00 / 60.00	1.67 / 3.33	0.42 / 3.33	19.40 / 35.72
`HC`: $\bar{H}_T C_t^{k_s}$	50.30 / 73.20	27.50 / 57.50	1.67 / 3.33	0.00 / 0.00	19.87 / 33.51
`HG`: $\bar{H}_T G_t^{k_s}$	50.00 / 74.20	28.13 / 55.00	2.08 / 10.00	1.25 / 10.00	20.37 / 37.30
`HC-Mass`: $\bar{H}_T C_t^{k_s}(1 - A_t^{k_s})$	**50.75** / 75.80	27.81 / **62.50**	2.08 / 6.67	1.25 / 6.67	20.47 / 37.91
`Full`: $\bar{H}_T C_t^{k_s} G_t^{k_s}$	50.50 / **76.40**	**28.44** / 60.00	**5.00 / 13.33**	**2.08 / 16.67**	**21.51 / 41.60**

4.3 ABLATION STUDY

Effect of the Branch-Acquisition Score. Because verifier-based probing is costly, `SPOT` applies it only at the top-M positions. We ablate normalized teacher entropy H, top-k_s teacher mass C, and the student–teacher gap G, which combines mass undercoverage and JS shape mismatch. We compare H, HC, HG, HC-Mass, and Full, where HC-Mass retains only the mass-undercoverage component of G. All variants otherwise share the same experimental protocol.

Table 2 shows that `Full` achieves the best macro Avg@8/Pass@8 (21.51/41.60), with its larger advantage in Pass@8 suggesting improved solution coverage. `HG` underperforms `Full` on both metrics across all benchmarks, supporting $C_t^{k_s}$ as a reliable filter against diffuse teacher uncertainty. `HC`, which omits the student–teacher gap, also consistently underperforms `Full`, supporting the inclusion of $G_t^{k_s}$. `HC-Mass` is the strongest partial variant but remains below `Full`, suggesting that mass undercoverage and JS shape mismatch capture complementary student deficits. Overall, the results support the multiplicative score as a soft conjunction of teacher ambiguity.

Effect of Verifier-Guided Calibration. Verifier-guided calibration converts the teacher prior into an outcome-aware branch target. To isolate this component, we remove reward tilting and positive-reward gating while retaining FULL acquisition and the branch-distillation objective. This ablation therefore uses the uncalibrated teacher proposal prior at all probed positions. All other training and evaluation settings match the Qwen3-1.7B-Base main experiment. Table 3 shows that the full configuration improves macro Avg@8/Pass@8 by 3.21/7.38 points. Pass@8 increases on all four benchmarks, with the largest gain on AIME24 (+13.33). On MATH500, Avg@8 changes by only +0.13 while Pass@8 rises by +2.00, suggesting that verifier-guided calibration primarily improves solution coverage in this setting.

Table 3: **Verifier-guidance ablation.** Avg@8 / Pass@8 (%) for Qwen3-1.7B-Base trained on MATH. *Macro* denotes the unweighted mean; best results are bold and the full model is shaded.

	Avg@8 / Pass@8				
Variant	MATH500	AMC23	AIME24	AIME25	Macro
w/o verifier guidance	67.30 / 85.80	34.06 / 67.50	8.33 / 20.00	4.58 / 16.67	28.57 / 47.49
SPOT (Full)	**67.43 / 87.80**	**39.69 / 75.00**	**12.50 / 33.33**	**7.50 / 23.33**	**31.78 / 54.87**

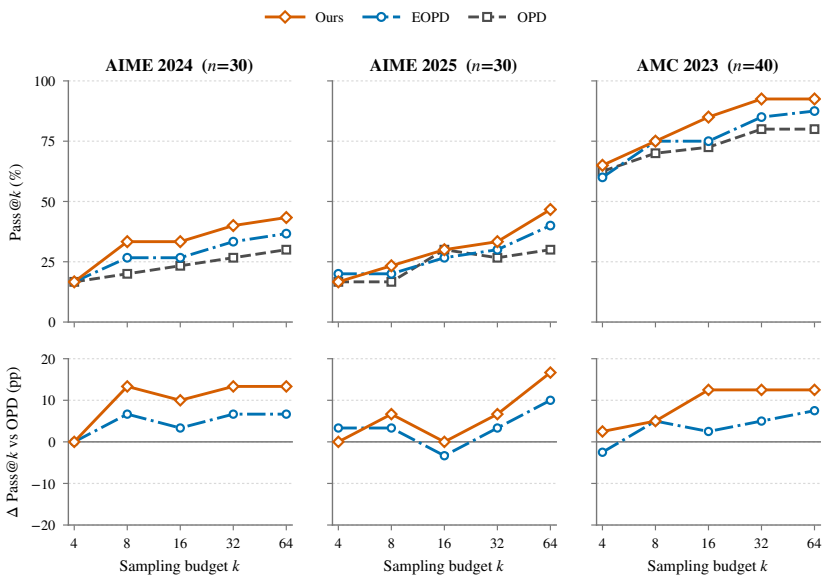


Figure 3: Pass@k and its gain over OPD on AIME 2024, AIME 2025, and AMC 2023. Each point is read from the corresponding summary of a separately generated k-sample run.

Scaling with the Evaluation-Time Sampling Budget. To test whether SPOT improves coverage beyond a fixed Pass@8 protocol, we vary $k \in \{4, 8, 16, 32, 64\}$ on AIME24, AIME25, and AMC23, holding decoding fixed. Here, k denotes responses per problem. Figure 3 reveals a budget-dependent separation. The methods are close at $k = 4$, with EOPD leading on AIME25, whereas SPOT ranks first or ties from $k = 8$ onward. At $k = 64$, SPOT retains 12.50–16.67-point gains over OPD, showing that its coverage advantage extends beyond Pass@8 and remains visible as the sampling budget grows.

Effect of Branch-Distillation Weight. The coefficient β trades off trajectory-wide OPD against outcome-calibrated branch supervision. Among the tested values in Figure 4, $\beta = 0.1$ leads the three competition benchmarks and ties $\beta = 0.5$ on MATH500. The competition benchmarks show greater observed sensitivity, with their mean Pass@8 lowest at $\beta = 1.0$. This pattern is consistent with branch feedback serving as a calibrated correction to trajectory-wide supervision: weaker weighting may underuse local outcome signals, whereas stronger weighting may overemphasize sparse branch targets. We therefore use $\beta = 0.1$ as a practical default.

Out-of-Domain Generalization. Although trained exclusively on MATH, SPOT generalizes beyond the training distribution, with its clearest gains on deliberative reasoning while remaining competitive on broader knowledge and instruction following (Appendix C.1, Table 9). This pattern suggests that outcome-calibrated branch supervision transfers reusable decision behavior rather than merely fitting math-specific solution templates. The non-uniform gains further indicate that branch calibration complements, rather than replaces, token-level uncertainty transfer and domain-aligned supervision.

Consistency Across Model Families. The Llama results (Appendix C.2, Table 10) preserve the main Qwen pattern: SPOT leads both macro metrics, with a clearer advantage in Pass@8 than Avg@8. Together with the Qwen results across model scales and training sets, this supports the

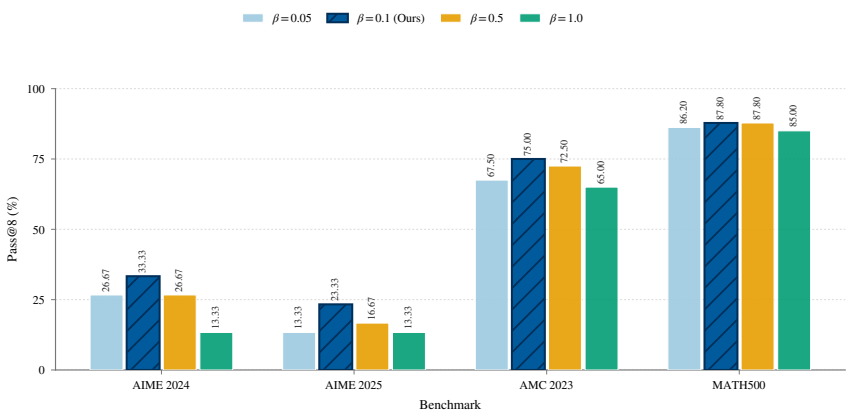


Figure 4: **Sensitivity to the branch-distillation weight.** Pass@8 of Qwen3-1.7B-Base for $\beta \in \{0.05, 0.1, 0.5, 1.0\}$. The main configuration $\beta = 0.1$ is marked as Ours.

portability of outcome-calibrated branch supervision beyond a single architecture and suggests that its primary benefit is consistently broader coverage of viable reasoning paths.

5 Related Work

Knowledge Distillation. Knowledge distillation transfers capabilities from a high-capacity teacher to a compact student by matching teacher predictive distributions or imitating teacher-generated sequences (Hinton et al., 2015; Kim & Rush, 2016). Recent work improves this paradigm through alternative divergence objectives, contrastive learning, and selective supervision based on sample difficulty or token importance (Ko et al., 2024; 2025; He et al., 2025; Guo et al., 2025). Despite their effectiveness, these methods train on contexts that are not sampled from the current student policy, creating exposure bias and compounding errors when the student conditions on its own predictions at inference time (Bengio et al., 2015; Ranzato et al., 2016).

On-Policy Distillation for Reasoning Models. On-policy distillation trains students on self-generated rollouts with dense token-level teacher feedback, mitigating exposure bias and providing finer credit assignment than outcome-only RL (Agarwal et al., 2024; Gu et al., 2026; Lu & Lab, 2025). Recent work has broadened OPD along four directions: teacher access and transfer scope, through reward extrapolation, black-box teachers, and multi-teacher distillation (Yang et al., 2026b; Ye et al., 2026a; Hou et al., 2026); adaptive supervision, through uncertainty- or disagreement-aware token selection, teachability modeling, and trajectory-aware guidance (Jin et al., 2026; Xu et al., 2026; Wang et al., 2026; Jiang & Ferraro, 2026); representation and training efficiency, through hidden-state alignment and offline teacher scoring (Yang et al., 2026a; Wu et al., 2026; Ziheng et al., 2026); and self-distillation, where privileged or auxiliary context induces internal teachers for knowledge internalization (Zhao et al., 2026; Ye et al., 2026b). Together, these advances have established OPD as a general post-training paradigm, with adoption in large-scale systems such as Qwen3 and DeepSeek-V4 (Yang et al., 2025; DeepSeek-AI et al., 2026).

6 Conclusion

Standard reverse-KL OPD can under-cover plausible alternatives, while teacher entropy alone cannot determine where limited probing is most useful or which candidates lead to successful student continuations. We introduced `SPOT`, which uses normalized teacher entropy, top-k_s probability mass, and student–teacher mismatch to allocate a sparse probing budget, then uses verifier-scored continuation values to construct KL-regularized, outcome-calibrated targets anchored to the teacher proposal prior. Across the evaluated student scales, training sets, reasoning benchmarks, and model families, `SPOT` achieves the highest macro Pass@8 in every setting and the highest or second-highest macro Avg@8, with controlled ablations supporting both selective acquisition and verifier guidance. Overall, our results support separating where to probe from what to distill: uncertainty and student mismatch guide probing, while downstream outcomes calibrate supervision.

AI USE STATEMENT

In this work, we used generative AI tools to improve the clarity, grammar, and presentation of the manuscript. We also used generative AI tools to assist with research execution, specifically in implementing, debugging, and refining parts of the research code. All AI-assisted text was reviewed and revised by the authors, and all AI-assisted code was inspected and tested by the authors. The authors take full responsibility for the final manuscript, implementation, and experimental results.

ETHICS STATEMENT

This work presents a methodological study of on-policy distillation for mathematical reasoning using existing language models and standard benchmarks. It does not involve human participants, private or sensitive personal data, or deployment in high-stakes settings. We are not aware of any specific ethical concerns arising from the experiments conducted in this study.

REPRODUCIBILITY STATEMENT

To support reproducibility, Algorithm 1 and Section 3.2 specify the complete SPOT procedure and training objective, while Appendix A provides the derivations and proofs for the outcome-calibrated target. Section 4.1 describes the models, training data, baselines, benchmarks, and evaluation protocol. Appendix B further reports the compute setup, training schedules and hyperparameters, chat templates, SPOT-specific acquisition and probing configurations, verifier implementation, decoding settings, and evaluation seed. All models, datasets, software frameworks, and evaluation tools used in our experiments are identified and cited. An anonymized implementation and the corresponding configuration files are provided in the supplementary material.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes, 2024. URL https://arxiv.org/abs/2306.13649.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating llms on uncontaminated math competitions, 2026. URL https://arxiv.org/abs/2505.23281.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks, 2015. URL https://arxiv.org/abs/1506.03099.

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.

Charles Goddard and Lucas Atkins. Distillkit: Flexible knowledge distillation for large language models, 2024. URL https://github.com/arcee-ai/distillkit.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: On-policy distillation of large language models, 2026. URL https://arxiv.org/abs/2306.08543.

Yiju Guo, Wenkai Yang, Zexu Sun, Ning Ding, Zhiyuan Liu, and Yankai Lin. Learning to focus: Causal attention distillation via gradient-guided token pruning, 2025. URL https://arxiv.org/abs/2506.07851.

Changyi He, Yifu Ding, Jinyang Guo, Ruihao Gong, Haotong Qin, and Xianglong Liu. DA-KD: Difficulty-aware knowledge distillation for efficient large language models. In *Forty-second International Conference on Machine Learning*, 2025.

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, 2021. URL https://arxiv.org/abs/2103.03874.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Wenjin Hou, Shangpin Peng, Weinong Wang, Zheng Ruan, Yue Zhang, Zhenglin Zhou, Mingqi Gao, Yifei Chen, Kaiqi Wang, Hongming Yang, Chengquan Zhang, Zhuotao Tian, Han Hu, Yi Yang, Fei Wu, and Hehe Fan. Uni-opd: Unifying on-policy distillation with a dual-perspective recipe, 2026. URL https://arxiv.org/abs/2605.03677.

Yuxuan Jiang and Francis Ferraro. Bridging reasoning trajectories in on-policy distillation via near-future guidance, 2026. URL https://arxiv.org/abs/2606.00305.

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, 2026. URL https://arxiv.org/abs/2603.07079.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation, 2016. URL https://arxiv.org/abs/1606.07947.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models, 2024. URL https://arxiv.org/abs/2402.03898.

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, 2025. URL https://arxiv.org/abs/2503.07067.

Hynek Kydlíček. Math-Verify: Math Verification Library. URL https://github.com/huggingface/math-verify.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models, 2022. URL https://arxiv.org/abs/2206.14858.

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, 2026. URL https://arxiv.org/abs/2604.13016.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. URL https://arxiv.org/abs/2305.20050.

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.

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

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks, 2016. URL https://arxiv.org/abs/1511.06732.

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

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.

Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. Not all disagreement is learnable: Token teachability in on-policy distillation, 2026. URL https://arxiv.org/abs/2605.26844.

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.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. Tip: Token importance in on-policy distillation, 2026. URL https://arxiv.org/abs/2604.14084.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

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.

Shenzhi Yang, Guangcheng Zhu, Bowen Song, Haobo Wang, Mingxuan Xia, Xing Zheng, Yingfan Ma, Zhongqi Chen, Weiqiang Wang, Junbo Zhao, and Gang Chen. Oprd: On-policy representation distillation, 2026a. URL https://arxiv.org/abs/2606.06021.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation, 2026b. URL https://arxiv.org/abs/2602.12125.

Tianzhu Ye, Li Dong, Zewen Chi, Xun Wu, Shaohan Huang, and Furu Wei. Black-box on-policy distillation of large language models, 2026a. URL https://arxiv.org/abs/2511.10643.

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

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.

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, 2026. URL https://arxiv.org/abs/2601.18734.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. Sglang: Efficient execution of structured language model programs, 2024. URL https://arxiv.org/abs/2312.07104.

Zhou Ziheng, Jiaqi Li, Huacong Tang, Ying Nian Wu, and Demetri Terzopoulos. Less is more: Early stopping rollout for on-policy distillation, 2026. URL https://arxiv.org/abs/2605.27028.

A Derivation and Properties of Outcome-Calibrated Targets

Fix a retained position $t \in \mathcal{B}^+$. Write $\mathcal{S} = S_t^{k_p}$, $p(v) = \bar{\pi}_T^{k_p}(v \mid c_t)$, and $V(v)$ for the branch values supplied to calibration. The derivation applies to both expected values and their empirical estimates. Assume that $p(v) > 0$ on the finite candidate set $\mathcal{S}$, and let $\Delta(\mathcal{S})$ be its probability simplex.

A.1 Closed-Form Target and Trust-Region Interpretation

In the constrained problem of Eq. (7), suppose the KL constraint has a positive multiplier λ. Up to a constant, its Lagrangian maximizes

$$\max_{\rho \in \Delta(\mathcal{S})} \{\mathbb{E}_\rho[V] - \lambda D_{\mathrm{KL}}(\rho \| p)\}. \tag{11}$$

Adding a multiplier ν for $\sum_v \rho(v) = 1$, stationarity gives $V(v) - \lambda(\log(\rho(v)/p(v)) + 1) + \nu = 0$. With $\gamma = 1/\lambda$, normalization yields

$$\rho_\gamma(v) = \frac{p(v)e^{\gamma V(v)}}{Z(\gamma)}, \qquad Z(\gamma) = \sum_{u \in \mathcal{S}} p(u)e^{\gamma V(u)}. \tag{12}$$

The regularized objective is strictly concave in ρ, so this stationary point uniquely maximizes Eq. (11). For any $\gamma > 0$, it also uniquely solves the constrained problem with radius $\epsilon_t(\gamma) = D_{\mathrm{KL}}(\rho_\gamma \| p)$: a feasible distinct distribution with at least as high a value would contradict the unique regularized optimum. In our experiments, γ is fixed rather than the KL radius; the induced $\epsilon_t(\gamma)$ can differ across positions.

A.2 Properties of Outcome Calibration

Let $F(\gamma) = \log Z(\gamma)$. Direct differentiation gives $F'(\gamma) = \mathbb{E}_{\rho_\gamma}[V]$ and $F''(\gamma) = \mathrm{Var}_{\rho_\gamma}(V)$. Since $D_{\mathrm{KL}}(\rho_\gamma \| p) = \gamma F'(\gamma) - F(\gamma)$,

$$\frac{\mathrm{d}}{\mathrm{d}\gamma}\mathbb{E}_{\rho_\gamma}[V] = \mathrm{Var}_{\rho_\gamma}(V), \qquad \frac{\mathrm{d}}{\mathrm{d}\gamma}D_{\mathrm{KL}}(\rho_\gamma \| p) = \gamma\,\mathrm{Var}_{\rho_\gamma}(V). \tag{13}$$

For nonconstant V, increasing γ strictly increases both the expected branch value under the calibrated target and its distance from the teacher prior. It therefore determines the induced KL radius uniquely. As $\gamma \to 0$, $\rho_\gamma \to p$; as $\gamma \to \infty$, it concentrates on the highest-value candidates, preserving their relative teacher probabilities when several candidates tie.

The prior p is a feasible comparator in the regularized objective. Optimality of ρ_γ therefore gives

$$\mathbb{E}_{\rho_\gamma}[V] - \mathbb{E}_p[V] \geq \frac{1}{\gamma} D_{\mathrm{KL}}(\rho_\gamma \| p) \geq 0. \tag{14}$$

This concerns the values supplied to calibration; with sampled values, it does not guarantee improvement in the unknown expected continuation value or in the trained student's performance.

A.3 Finite-Sample Effects of Branch Probing

In our training configuration, one continuation is sampled per candidate and the verifier reward is binary. The single-sample estimate used to form the target thus has $V(v) \in \{0, 1\}$, although the expectation in Eq. (6) can lie anywhere in $[0, 1]$. Let $\mathcal{S}^+ = \{v : V(v) = 1\}$, and write $P_+ = \sum_{v \in \mathcal{S}^+} p(v)$ and $Q_+ = \sum_{v \in \mathcal{S}^+} \rho_\gamma(v)$. When $0 < P_+ < 1$, Eq. (12) gives

$$\frac{Q_+}{1 - Q_+} = e^\gamma \frac{P_+}{1 - P_+}. \tag{15}$$

Successful candidates receive the common multiplier e^γ, so their aggregate odds increase by that factor while teacher-relative probabilities within each outcome group remain unchanged. If every candidate succeeds, all multipliers coincide and $\rho_\gamma = p$.

To examine sampling variation, fix a probed position $t \in \mathcal{B}$ before the $\mathcal{B}^+$ gate. For each candidate v, let $R_v^{(n)} \in \{0, 1\}$ be independent verifier outcomes with mean $\mu_v = \mathbb{E}[R_v^{(n)}]$.

Set $\hat{\mu}_v = N_p^{-1} \sum_{n=1}^{N_p} R_v^{(n)}$, and let $\hat{\rho}_\gamma$ be the exponential tilt in Eq. (12) formed using $\hat{\mu}_v$. For distinct candidates $u, v \in \mathcal{S}$, its correction to the teacher log-odds is

$$\widehat{\Delta}_{v,u} := \log \frac{\hat{\rho}_\gamma(v)/\hat{\rho}_\gamma(u)}{p(v)/p(u)} = \gamma(\hat{\mu}_v - \hat{\mu}_u). \tag{16}$$

Independence across candidate rollouts then gives

$$\mathbb{E}[\widehat{\Delta}_{v,u}] = \gamma(\mu_v - \mu_u), \qquad \mathrm{Var}(\widehat{\Delta}_{v,u}) = \frac{\gamma^2}{N_p}\left[\mu_v(1-\mu_v) + \mu_u(1-\mu_u)\right]. \tag{17}$$

Thus the sampled log-odds correction is unbiased before gating, and its standard deviation decreases as $N_p^{-1/2}$. Increasing γ scales both the mean correction and its sampling noise. These statements do not imply an unbiased target probability distribution; conditioning on retention by the $\mathcal{B}^+$ gate also changes the sampling law.

B Implementation Details

Off-policy Training. Following (Jin et al., 2026), KD is implemented with DistillKit (Goddard & Atkins, 2024), with one teacher response sampled per training problem. Table 4 summarizes the off-policy configuration.

Table 4: Hyperparameters used for off-policy distillation.

Hyperparameter	**Value**
Learning rate	1×10^{-5}
LR scheduler type	cosine
Optimizer	AdamW
CE loss weight	0.5
Forward-KL loss weight	0.5
Training batch size	128
Training epoch	3
Cutoff length	4,096
Top-k (for FKL)	16

On-policy Training. We implement OPD, EOPD, GRPO, and `SPOT` with `verl` (Sheng et al., 2025) and generate rollouts asynchronously with `SGLang` (Zheng et al., 2024). All main runs use 4 $\times$ A800 GPUs, FSDP2, and bfloat16. Table 5 summarizes the principal training hyperparameters.

Table 5: Hyperparameters used for on-policy distillation and GRPO.

Hyperparameter	**OPD, EOPD, SPOT**	**GRPO**
Learning rate	3×10^{-6}	3×10^{-6}
LR scheduler type	cosine	cosine
Optimizer	AdamW	AdamW
Training batch size	128	128
Mini-batch size	32	32
Samples per prompt	1	8
Temperature	1.0	1.0
Top-p	1.0 (Qwen), 0.8 (Llama)	1.0 (Qwen), 0.8 (Llama)
Max prompt length	1,024	1,024
Max response length	4,096	4,096
Clipping ratio	0.2	0.2

All methods use AdamW, weight decay 0.01, gradient clipping at 1.0, and four mini-batch updates per rollout iteration; the distillation methods use no learning-rate warmup. EOPD applies a unit-weight top-16 forward-KL loss at teacher-entropy values of at least 0.8, while GRPO uses KL and entropy coefficients of 10^{-3} and 0, respectively.

Training Schedule. Table 6 records the scale-specific data and effective schedules of the main on-policy runs.

Table 6: Training data and schedules. Iterations denote rollout iterations.

Student	Training set	Examples	Epochs	Iterations
Qwen3-0.6B-Base	MATH	7,500	3	174
Qwen3-1.7B-Base	MATH	7,500	3	174
Qwen3-4B-Base	DAPO-Math	14,116	2	220

Training examples are shuffled, no validation set is used, and prompts exceeding the listed length limit are filtered.

Chat Template. Following EOPD, the Qwen runs of OPD, EOPD, and `SPOT` use the Qwen3-8B teacher's non-thinking chat format. With thinking disabled, the Qwen3-Base tokenizer renders the same prompt: `<|im_start|>user\n{query}<|im_end|>\n<|im_start|>assistant\n<think>\n\n</think>\n\n`. GRPO instead uses the default Qwen3-Base generation prompt, which omits the empty thinking block: `<|im_start|>user\n{query}<|im_end|>\n<|im_start|>assistant\n`.

SPOT-specific Training. Table 7 summarizes the hyperparameters specific to `SPOT`; the acquisition, probing, gating, and loss are defined in Algorithm 1 and Section 3.2.

Table 7: Hyperparameters used for SPOT.

Hyperparameter	**Value**
Selected positions M	2
Scoring top-k k_s	16
Probing top-k k_p	4
$\lambda_{\text{mass}}, \lambda_{\text{shape}}$	$0.5, 0.5$
Reward-tilt coefficient γ	1.0
Branch-loss weight β	0.1
Probe rollouts per candidate	1
Training verifier	binary rule-based boxed-answer match

Evaluation Details. We use vLLM with bfloat16 precision and random seed 42, render one user message with the model's native chat template and thinking disabled, and append the instruction `Please reason step by step, and put your final answer within \boxed{}`. We extract the final balanced `\boxed{}` expression and use Math-Verify (Kydlíček) for symbolic equivalence; missing or unparsable answers are counted as incorrect.

Table 8: Sizes of the evaluation benchmarks.

Benchmark	**Problems**
MATH-500	500
AMC 2023	40
Minerva Math	272
HMMT 2025 (Feb.+Nov.)	60
AIME 2024	30
AIME 2025	30

Table 9: Out-of-domain results for Qwen3-1.7B-Base trained on MATH (all values in %). GPQA-Diamond uses eight samples per question; MMLU-Pro uses category-matched five-shot CoT demonstrations and reports Pass@1; and AlpacaEval 2.0 reports win rate (WR) and length-controlled win rate (LC-WR). **Bold** and underlined entries denote the best and second-best results, respectively. The SPOT column is shaded. AlpacaEval scores marked † use 803 (EOPD) and 800 (SPOT) valid judgments out of 805, respectively.

Benchmark	Metric	KD	GRPO	OPD	EOPD	**SPOT (Ours)**
GPQA-Diamond	Avg@8	21.10	19.26	25.00	27.21	**29.42**
	Pass@8	62.18	55.05	67.17	69.19	**80.81**
MMLU-Pro	Pass@1	38.13	41.86	41.23	**42.90**	42.26
AlpacaEval 2.0	LC-WR	24.10	21.86	27.08	**28.13**†	27.59†
	WR	27.58	25.71	31.18	33.23†	**34.63**†

C More Ablation Experiment Results

C.1 Out-of-Domain Generalization

Table 9 reports the complete out-of-domain results analyzed in Section 4.3. Together, these benchmarks separate deliberative reasoning, broad knowledge, and instruction following, providing a more diagnostic view of transfer than any single aggregate score.

The comparison reveals a structured transfer profile rather than uniform dominance. Relative to the strongest baseline on GPQA-Diamond, SPOT improves Avg@8 by 2.21 points and Pass@8 by 11.62 points, and it leads AlpacaEval WR by 1.40 points. On MMLU-Pro, SPOT ranks second at 42.26: it trails EOPD by only 0.64 points, while exceeding GRPO and OPD by 0.40 and 1.03 points, respectively. This near-best result indicates that branch calibration preserves broad subject-matter reasoning despite training only on MATH, while EOPD's edge suggests that token-level uncertainty matching remains especially useful for heterogeneous knowledge questions. On AlpacaEval, length control reverses the raw-WR ordering, indicating that part of the preference gain is associated with response length. Overall, outcome calibration transfers most clearly to deliberative reasoning, with method-dependent tradeoffs on broader forms of generalization.

C.2 Consistency Across Model Families

To evaluate the generality of our method beyond Qwen3, we repeat the comparison with a Llama-3.2-3B-Instruct student trained on MATH for three epochs and a Llama-3.1-8B-Instruct teacher for the distillation methods. All methods retain their native optimization protocols and are evaluated with eight samples under identical decoding. Because this setting differs from the Qwen experiments in architecture, checkpoint type, and student–teacher scale, Table 10 compares methods only within Llama; the experiment tests transfer of the relative method pattern, not absolute performance across families.

Table 10: **Results on the Llama model family.** *Macro avg.* is the unweighted mean over the four benchmarks, computed before rounding. **Bold** and underlined entries denote the best and second-best results, respectively. SPOT cells are shaded in blue for Avg@8 and green for Pass@8.

Benchmark	Metric	GRPO	OPD	EOPD	SPOT
MATH500	Avg@8	43.65	35.90	37.75	**43.78**
	Pass@8	61.60	63.60	65.20	**66.20**
AMC23	Avg@8	**24.06**	13.13	19.06	23.44
	Pass@8	47.50	45.00	47.50	**50.00**
AIME24	Avg@8	2.08	0.83	1.67	**2.50**
	Pass@8	**16.67**	6.67	13.33	**16.67**
AIME25	Avg@8	0.00	**0.42**	0.00	**0.42**
	Pass@8	0.00	**3.33**	0.00	**3.33**
Macro avg.	Avg@8	17.45	12.57	14.62	**17.53**
	Pass@8	31.44	29.65	31.51	**34.05**

Within-family comparison. Within the Llama setting, SPOT raises macro Avg@8 from 12.57 for OPD to 17.53 and macro Pass@8 from 29.65 to 34.05. Relative to the strongest baseline for each metric, however, the distinction is asymmetric: Avg@8 is essentially tied with GRPO (17.53 vs. 17.45), whereas Pass@8 exceeds EOPD by 2.54 points. The benchmark-level results show the same pattern. SPOT is best or tied-best in Pass@8 on all four benchmarks, while GRPO remains stronger in AMC23 Avg@8 and OPD ties SPOT on AIME25. Thus, the evidence supports a more consistent improvement in multi-sample solution coverage than in average per-sample accuracy, rather than uniform dominance on every benchmark.

Cross-family interpretation. This asymmetry mirrors the Qwen results: across the 0.6B and 1.7B students trained on MATH and the 4B student trained on DAPO, SPOT leads macro Pass@8 in every setting, while macro Avg@8 is best or nearly tied with the strongest baseline. Its recurrence across Base and Instruct checkpoints, two model families, multiple capacities, and two training-data regimes makes a Qwen-specific explanation less plausible. The aggregates are not directly comparable because the Llama panel contains four benchmarks whereas the Qwen panel contains six; moreover, architecture, capacity, initialization, teacher pairing, and training data are not independently controlled. We therefore view these results as evidence of portability across the evaluated configurations, not architecture invariance or a causal scaling law.