LEARNING POSTERIOR TEACHERS FOR ON-POLICY SELF-DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation addresses the sparse supervision of reinforcement learning with verifiable rewards (RLVR) by providing dense token-level supervision from a privileged teacher. However, privileged information alone does not ensure reliable supervision, and teacher accuracy alone does not specify a suitable distribution for distillation. We identify three desirable teacher properties for distillation: correctness, minimal shift, and preference preservation. The correctness-conditioned posterior unifies these properties by restricting the initial student's prior to trajectories with correct final answers. We introduce Flow Teacher Distillation (FlowTD), which trains a privileged teacher to approximate this posterior using GFlowNet trajectory balance. Its probability-matching objective uses outcome verification to supervise correctness and explicitly trains the teacher to preserve the student's relative probabilities among correct trajectories. Our analyses show that FlowTD learns teachers with stronger correctness discrimination than prompt-only teachers, alongside smaller distributional shifts and better preference preservation than SFT- and RL-trained teachers. Across four mathematical reasoning benchmarks, FlowTD improves average accuracy over the strongest baselines by 4.7 and 3.9 points on Qwen2.5-3B and Qwen3-8B-Base, respectively. It also achieves the highest pass@k among the compared teacher constructions on AIME 2025.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) trains reasoning models with sparse, trajectory-level supervision derived from outcome verification (Shao et al., 2024; DeepSeek-AI, 2025). Such rewards assess final outcomes without directly identifying which intermediate decisions contributed to success. In contrast, on-policy distillation (OPD) (Agarwal et al., 2024; Thinking Machines, 2025) provides dense token-level supervision on student rollouts using a teacher's next-token distributions. On-policy self-distillation applies this idea with the model itself as a teacher, conditioned on *privileged information* unavailable to the student. OPSD (Zhao et al., 2026) uses reference solutions as teacher context, while SDPO (Hübotter et al., 2026) uses environment feedback or successful attempts.

When the student is trained solely by distillation, correctness supervision comes through the teacher's next-token predictions. Privileged information changes these predictions, but these changes need not align with eventual task success (Nguyen et al., 2026). This motivates training the teacher with outcome supervision, making final-answer correctness an explicit learning objective. Distillation also transfers the teacher's relative preferences among correct reasoning paths. Yet improving teacher accuracy alone leaves open how probability should be allocated among these paths. We therefore ask: *what makes a good teacher distribution, and how can we learn it?*

We identify three desirable teacher properties for distillation: **(P1) Correctness**, supporting only trajectories with correct final answers; **(P2) Minimal shift**, minimizing departure from the initial student's distribution subject to correctness; and **(P3) Preference preservation**, retaining the student's relative probabilities among correct trajectories. These properties are unified by the *correctness-conditioned posterior*, obtained by conditioning the initial student's prior p_0 on a correct final answer: $q^\star(z \mid x, y) \propto p_0(z \mid x)\, \mathbf{1}[\mathrm{Ans}(z) = y]$. This posterior uniquely minimizes $\mathrm{KL}(q \| p_0)$ over distributions q supported on correct trajectories (Proposition 1, Figure 1). The token-level supervision

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.

Table 1: Qualitative comparison of teacher constructions across P1–P3. Relative alignment: ● stronger, ◐ intermediate, and ○ weaker.

Teacher	Teacher learning	(P1) Correctness	(P2) Minimal shift	(P3) Preference preservation
Prompt-only	No update	○	●	●
SFT-trained	Maximum likelihood	●	○	○
RL-trained	KL-regularized policy optimization	●	◐	◐
FlowTD	Trajectory balance	●	●	●

induced by this ideal posterior is also outcome-grounded: relative to the prior, it upweights tokens that increase continuation success probability under the initial student (Proposition 2).

To learn a privileged teacher with these properties, we introduce *Flow Teacher Distillation* (FlowTD), which uses GFlowNet trajectory balance (Malkin et al., 2022), following FlowRL (Zhu et al., 2026). Outcome verification supplies the correctness signal, while the initial student's problem-only prior specifies the relative probabilities of correct trajectories. Trajectory balance turns this target into a computable probability-matching objective on sampled trajectories. The objective penalizes deviations from the target even for correct trajectories, explicitly training the teacher to preserve the prior's relative preferences among them. We then freeze the teacher for token-level distillation on the student's own rollouts (Figure 3).

On mathematical reasoning, we compare prompt-only, SFT-trained, RL-trained, and FlowTD teachers under a shared student-side distillation procedure. FlowTD provides stronger correctness discrimination than prompt-only teachers. Compared with similarly accurate SFT- and RL-trained teachers, it exhibits smaller distributional shifts and better preserves the student's relative preferences among correct trajectories (Table 1). Its student achieves the highest average accuracy on both Qwen2.5-3B and Qwen3-8B-Base, exceeding the strongest alternative teacher by 4.7 and 3.88 percentage points, respectively (Tables 2 and 3). On AIME 2025, its advantage persists across sampling budgets up to $k = 32$ (Figure 5). These results support the three properties as teacher design criteria and FlowTD as an effective approach to learning teachers that approximate the ideal posterior.

Our contributions are as follows:

1. **Teacher-distribution design and analysis.** We formulate correctness, minimal shift, and preference preservation as joint criteria, providing a common basis for designing and comparing distillation teachers. Our characterization of the correctness-conditioned posterior connects a minimum-KL teacher target with token-level supervision grounded in continuation success.
2. **FlowTD for posterior teacher learning.** We develop FlowTD, which uses GFlowNet trajectory balance to learn privileged teachers through explicit distribution matching. This makes preserving the initial student's relative preferences among correct trajectories part of teacher optimization, alongside correctness.
3. **Empirical analysis and validation.** Under a shared student-side distillation procedure, we show that comparable teacher accuracy does not imply comparable distillation performance. FlowTD exhibits a more favorable combination of the three teacher properties and achieves the highest student accuracy among the evaluated teacher constructions, outperforming direct RLVR.

2 A Principled Teacher Distribution

The teacher determines the next-token targets used for on-policy distillation. We seek a distribution that enforces correctness while preserving the prior's relative preferences among correct trajectories.

2.1 The Correctness Posterior

Let $\mathrm{supp}(q) = \{z : q(z) > 0\}$ and $\mathcal{C}(x, y) = \{z : \mathrm{Ans}(z) = y\}$. Write $Z(x, y) = \Pr_{z \sim p_0(\cdot|x)}[\mathrm{Ans}(z) = y]$ and assume $Z(x, y) > 0$.

Proposition 1 (I-projection onto the correct set; cf. Csiszár 1975)**.** *The unique minimizer of* $\mathrm{KL}(q\|p_0(\cdot \mid x))$ *over* $\mathcal{Q} = \{q : \mathrm{supp}(q) \subseteq \mathcal{C}(x, y)\}$ *is*

$$q^*(z \mid x, y) = p_0\big(z \mid x, \mathrm{Ans}(z) = y\big) = \frac{p_0(z \mid x)\,\mathbf{1}[\mathrm{Ans}(z) = y]}{Z(x, y)}. \tag{1}$$

Proof. For any $q \in \mathcal{Q}$ with finite $\mathrm{KL}(q\|p_0)$, we have $q \ll p_0$ and $\mathrm{KL}(q\|p_0) = \mathrm{KL}(q\|q^*) - \log Z(x, y)$. The first term vanishes only at $q = q^*$; distributions with infinite KL cannot minimize the objective since $-\log Z(x, y)$ is finite. □

The decomposition separates the shift required by correctness from any additional change in the teacher's distribution:

$$\mathrm{KL}\big(q\|p_0(\cdot \mid x)\big) = \underbrace{-\log Z(x, y)}_{\text{required by correctness}} + \underbrace{\mathrm{KL}(q\|q^*)}_{\text{additional shift}}. \tag{2}$$

2.2 Three Properties of the Posterior Teacher

(P1) Correctness. q^* assigns probability only to trajectories with verified correct final answers.

(P2) Minimal shift. Among distributions satisfying P1, q^* uniquely minimizes KL to p_0, with value $-\log Z(x, y)$ (Eq. (2)).

(P3) Preference preservation. For any two correct trajectories with positive prior probability, their probability ratio under q^* equals that under p_0. Thus, conditioning on correctness retains every correct trajectory in the prior's support and preserves their relative weights (Figure 2).

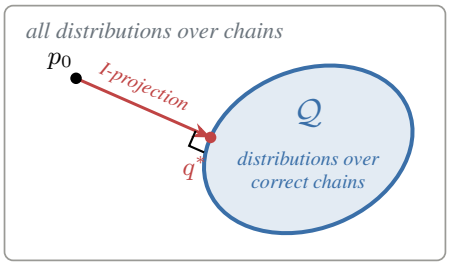

Figure 1: The correctness posterior q^* minimizes KL to p_0 among distributions supported on correct trajectories.

To examine the corresponding token-level predictions, let s be a prefix with $p_0(s \mid x) > 0$, and let sa denote s followed by token a. Define the *success-to-go* $h(s; x) = \Pr_{u \sim p_0(\cdot \mid x, s)}[\mathrm{Ans}(su) = y]$. Then $h(\varnothing; x) = Z(x, y)$ and $h(z; x) = \mathbf{1}[z \in \mathcal{C}(x, y)]$ for a completed trajectory z.

Proposition 2 (Next-token reweighting by success-to-go)**.** *For any such prefix with* $h(s; x) > 0$ *and token with* $p_0(a \mid x, s) > 0$*,*

$$\begin{aligned} q^*(a \mid x, s) &= p_0(a \mid x, s)\frac{h(sa; x)}{h(s; x)}, \\ \log \frac{q^*(a \mid x, s)}{p_0(a \mid x, s)} &= \log h(sa; x) - \log h(s; x). \end{aligned} \tag{3}$$

The log-ratio is $-\infty$ *when* $h(sa; x) = 0$*.*

Proof. Summing Eq. (1) over completed trajectories extending s gives $q^*(s \mid x, y) = p_0(s \mid x)h(s; x)/Z(x, y)$. Taking the ratio for sa and s yields the result. □

Relative to the prior, the ideal posterior upweights a token exactly when it increases continuation success probability under the base model. Along a correct trajectory, these log-ratios sum to $-\log Z(x, y)$. Our empirical analysis in Figure 6a examines outcome alignment at the trajectory level: the prompt teacher's scores provide near-chance correctness discrimination, while all three trained teachers separate correct and incorrect base-model responses substantially better.

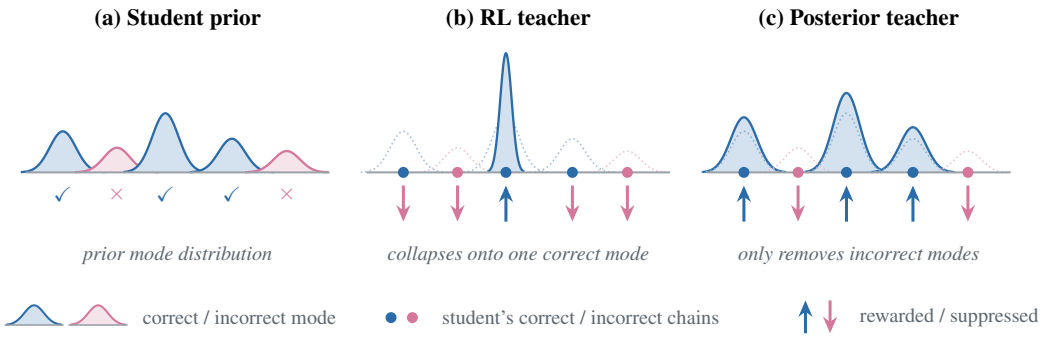

Figure 2: Illustration of probability allocation among reasoning trajectories. (a) The initial student assigns probability to correct and incorrect chains. (b) Reward maximization can concentrate probability on a few correct modes, reducing the relative weight of other correct chains sampled by the student. (c) The correctness posterior removes incorrect trajectories and preserves the prior's relative probabilities among correct ones.

3 Method: Flow Teacher Distillation

FlowTD (Figure 3) first learns a privileged teacher that approximates the posterior q^*, then freezes it for on-policy distillation. We hold the student-side objective fixed across teacher constructions.

3.1 Teacher: Posterior Matching by Flow Balance

GFlowNets (Bengio et al., 2023) train a policy to sample in proportion to an unnormalized terminal reward. We use this distribution-matching approach to fine-tune a teacher $q_{\mathrm{t}}(z \mid x, [\phi]; \theta_{\mathrm{t}})$ toward

$$R(z; x, y) = \underbrace{p_0(z \mid x)}_{\text{prior anchor}} \cdot \underbrace{\mathbf{1}[\mathrm{Ans}(z) = y]}_{\text{outcome verifier}}. \tag{4}$$

A bracketed argument marks privileged context available to the teacher alone; both the frozen reference p_0 and student p_{s} condition only on the problem. Normalizing R yields q^*: the verifier enforces correctness, while the prior preserves relative probabilities among correct trajectories and yields the minimal shift of Proposition 1.

Training objective. For numerical stability, we use the finite-penalty weight $R_\beta(z; x, y) = p_0(z \mid x) \exp[-\beta(1 - r(z; x, y))]$, where $r(z; x, y) = \mathbf{1}[\mathrm{Ans}(z) = y]$; its normalized distribution approaches q^* as $\beta \to \infty$. We fit the teacher and a learnable normalizer $Z_\psi(x)$ with the trajectory-balance loss (Malkin et al., 2022):

$$\mathcal{L}_{\mathrm{TB}}(\theta_{\mathrm{t}}, \psi) = \mathbb{E}_{z \sim \mu}\Big[\omega(z)\Big(\tfrac{1}{|z|}\big[\log Z_\psi(x) + \log q_{\mathrm{t}}(z \mid x, [\phi]) - \log R_\beta(z; x, y)\big]\Big)^2\Big]. \tag{5}$$

The factor $1/|z|$ scales the entire residual across trajectory lengths; $\omega(z)$ is a detached, clipped importance weight for trajectories reused from the rollout distribution μ. With positive weights and μ covering the target support, the population zero-residual solution satisfies $q_{\mathrm{t}} = R_\beta / Z_\psi$.

3.2 Student: Token-Level JSD Distillation

The student $p_{\mathrm{s}}(z \mid x; \theta_{\mathrm{s}})$ fine-tunes the same base model on x alone, without the teacher's feedback string, so deployment needs no privileged information. It draws K rollouts from its own policy; there we minimize the token-level Jensen–Shannon divergence to the frozen teacher:

$$\mathcal{L}_{\mathrm{distill}}(\theta_{\mathrm{s}}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\mathbb{E}_{z\sim p_{\mathrm{s}}(\cdot|x)}\Big[\sum_{t=1}^{|z|} m_t(z)\,\mathrm{JSD}\big(q_{\mathrm{t}}(\cdot \mid z_{<t}, x, [\phi]) \,\|\, p_{\mathrm{s}}(\cdot \mid z_{<t}, x; \theta_{\mathrm{s}})\big)\Big], \tag{6}$$

where the mask $m_t(z)$ excludes pad and post-EOS positions and

$$\mathrm{JSD}(q\|p) = \tfrac{1}{2}\,\mathrm{KL}(q\|\bar{m}) + \tfrac{1}{2}\,\mathrm{KL}(p\|\bar{m}), \qquad \bar{m} = \tfrac{1}{2}(q + p). \tag{7}$$

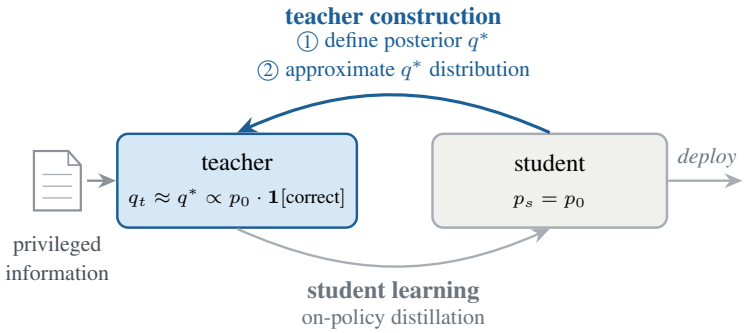

Figure 3: Our teacher construction. We define the teacher's target as the correctness-conditioned posterior q^* obtained from the student prior p_0, and train the teacher to approximate it. The student is then distilled on its own rollouts under a token-level on-policy distillation objective.

JSD is symmetric and bounded, providing a stable discrepancy between teacher and student next-token distributions.

Algorithm 1 summarizes the training procedures of FlowTD.

Algorithm 1 Flow Teacher Distillation.

Require: problems $\mathcal{D} = \{(x, y)\}$, base model p_0, privileged context ϕ, steps U, E, penalty β, rollouts K

initialize $\theta_t, \theta_s \leftarrow p_0$ and the partition head $\log Z_\psi$
Stage 1: teacher construction
for $u = 1$ to U **do**
  draw $(x, y) \sim \mathcal{D}$ and chains $z \sim \mu(\cdot \mid x, \phi)$
  score $\log p_0(z \mid x)$ with a frozen reference pass
  $\log R_\beta \leftarrow \log p_0(z \mid x) - \beta\big(1 - r(z; x, y)\big)$
  update (θ_t, ψ) on the trajectory-balance loss Eq. (5)
end for
freeze θ_t
Stage 2: student distillation
for $e = 1$ to E **do**
  draw $(x, y) \sim \mathcal{D}$ and rollouts $z^{(1:K)} \sim p_s(\cdot \mid x; \theta_s)$
  update θ_s on the token-level Jensen–Shannon objective Eq. (6)
end for
return $p_s(\cdot \mid x; \theta_s)$

4 Experiments

4.1 Experimental Setup

Models and datasets. We experiment with Qwen2.5-3B (Yang et al., 2024) and Qwen3-8B-Base (Yang et al., 2025). We evaluate mathematical reasoning on OlympiadBench (He et al., 2024), AMC 2023, AIME 2024, and AIME 2025. For the 3B experiments, we use 8,523 level-3–5 problems from MATH (Hendrycks et al., 2021), following the data processing of SimpleRL-Zoo (Zeng et al., 2025), and perform full-parameter fine-tuning of both teacher and student. For Qwen3-8B-Base, we use 10,000 mathematical reasoning problems from OpenThoughts-114k (Guha et al., 2025), with data selection detailed in Appendix A. During privileged teacher training, the teacher receives a reference solution and is instructed to reason independently. The student receives only the problem; correctness is determined by final-answer verification.

Baselines. We compare FlowTD with four alternative teacher constructions and direct RLVR. **OPSD** (Zhao et al., 2026) uses a frozen model conditioned on privileged information, while

Table 2: Greedy accuracy (%) on Qwen2.5-3B. Self-distillation rows report student results; SFT/RL teacher denote the teacher used. Best results are bolded.

Method	OlympiadBench	AMC 2023	AIME 2024	AIME 2025	AVG
References					
Qwen2.5-3B	20.1	30.0	0.0	0.0	12.5
GRPO	25.0	35.0	3.3	0.0	15.8
Self-distillation					
OPSD	21.8	35.0	3.3	0.0	15.0
SDPO	19.0	40.0	3.3	3.3	16.4
SFT teacher	24.1	35.0	3.3	0.0	15.6
RL teacher	**25.9**	32.5	3.3	3.3	16.3
FlowTD	25.4	**42.5**	**10.0**	**6.7**	**21.1**

SDPO (Hübotter et al., 2026) uses an exponential moving average (EMA) of the student as its privileged teacher. We also implement two trained-teacher baselines: an **SFT teacher** trained on reference solutions, and an **RL teacher** optimized with GRPO (Shao et al., 2024) using verifier rewards and KL regularization to a frozen, problem-only base reference. Both are frozen during distillation, as is the FlowTD teacher. Our controlled implementations of all distillation methods use the same student-side JSD objective (Eq. (6)), allowing us to compare different teacher constructions within a shared distillation procedure. As a non-distillation baseline, **GRPO** applies RLVR directly to the problem-only model.

Training and evaluation details. For each method, we select the checkpoint with the best validation performance. The main tables report final-answer accuracy under greedy decoding, with AVG denoting the unweighted mean across the four benchmarks. We additionally evaluate pass@k, the probability of at least one correct answer in k samples, on AIME 2025 with Qwen3-8B-Base (Figure 5). Training and evaluation protocols are provided in Appendix A.

4.2 Main Results

Table 3: Greedy accuracy (%) on Qwen3-8B-Base. Self-distillation rows report student results; SFT/RL teacher denote the teacher used. Best results are bolded, including ties.

Method	OlympiadBench	AMC 2023	AIME 2024	AIME 2025	AVG
References					
Qwen3-8B-Base	30.0	47.5	16.7	10.0	26.05
GRPO	**38.8**	60.0	13.3	6.7	29.70
Self-distillation					
OPSD	34.3	52.5	10.0	10.0	26.71
SDPO	32.4	45.0	**20.0**	10.0	26.85
SFT teacher	36.8	57.5	6.7	10.0	27.75
RL teacher	**38.8**	57.5	10.0	13.3	29.90
FlowTD	36.0	**62.5**	**20.0**	**16.7**	**33.78**

Overall performance across models. To evaluate FlowTD's effectiveness in improving student reasoning, we compare it with alternative teacher constructions and GRPO (Tables 2 and 3). On Qwen2.5-3B, students distilled from SFT- and RL-trained teachers outperform OPSD in average accuracy but do not surpass SDPO. On Qwen3-8B-Base, both trained-teacher baselines outperform OPSD and SDPO. Nevertheless, across both models, these alternative self-distillation methods remain below or close to direct GRPO. In contrast, FlowTD achieves the highest average student accuracy on both models, exceeding the strongest distillation baseline by 4.7 and 3.88 percentage points and direct GRPO by 5.3 and 4.08 points, respectively. These results demonstrate that FlowTD's teacher-learning approach translates into superior downstream distillation performance across both models.

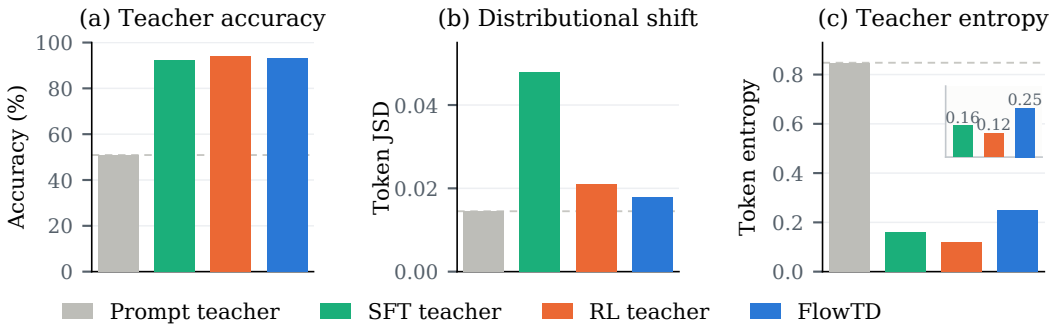

Figure 4: Teacher accuracy and distributional analysis on Qwen2.5-3B. (a) Teacher accuracy. (b) Distributional shift from the initial student (token JSD). (c) Teacher token-level diversity (token entropy). FlowTD achieves comparable accuracy with smaller distributional shifts and higher entropy than SFT and RL teachers.

Pass@k across sampling budgets. Figure 5 compares the distilled students on AIME 2025. OPSD largely plateaus beyond $k = 8$, whereas students distilled from trained teachers continue to improve. FlowTD leads among these teacher constructions at every evaluated k, with a widening advantage over SFT- and RL-teacher distillation at larger budgets. These results indicate improved problem-solving coverage beyond single-sample accuracy.

4.3 Teacher Accuracy and Distributional Analysis

To understand FlowTD's distillation advantage, we examine how different teacher constructions balance three teacher properties: correctness, distributional shift, and token-level diversity. Figure 4 compares the fixed prompt-only teacher used in OPSD with SFT, RL, and FlowTD teachers, all based on Qwen2.5-3B and evaluated on the same 500 problems at $T = 1.0$.

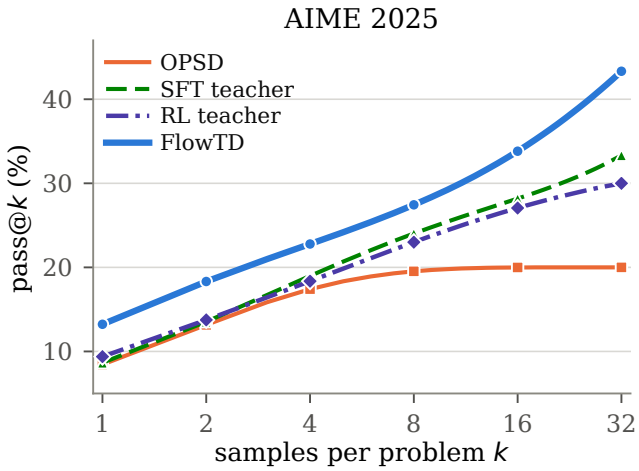

Figure 5: Pass@k on AIME 2025 with Qwen3-8B-Base, estimated from $n = 32$ samples per problem.

Comparable accuracy with smaller distributional shifts. Despite access to privileged information, the prompt teacher is substantially less accurate than all three trained teachers and delivers modest distillation gains (Table 2), suggesting that weak teacher accuracy can limit the effectiveness of its supervision. SFT, RL, and FlowTD achieve similarly high teacher accuracy (Figure 4a). Yet the most accurate teacher, RL, does not produce the strongest student, indicating that higher teacher accuracy alone does not guarantee better distillation.

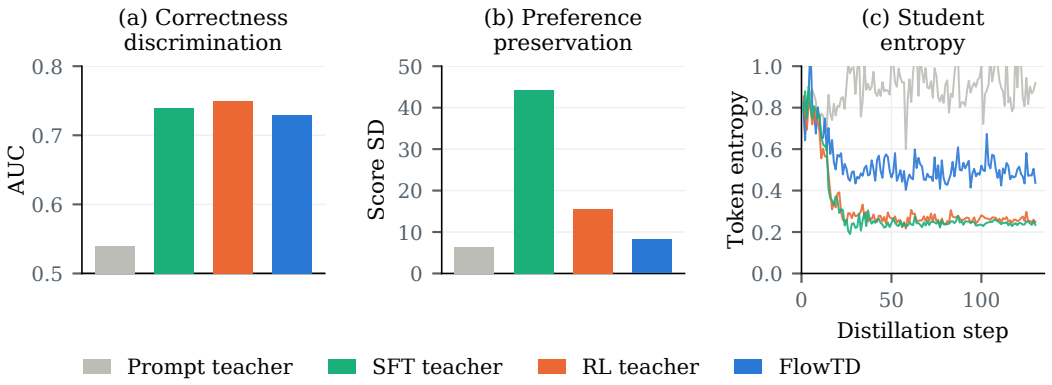

Figure 6: Teacher supervision and student dynamics on Qwen2.5-3B. Panels (a–b) score the same 800 frozen-base trajectories (100 problems, 8 samples each). (a) Within-problem correctness AUC using teacher/base log-ratio scores. (b) Standard deviation of sequence-summed log-ratios among each problem's correct trajectories. (c) Student token entropy during distillation. FlowTD combines comparable correctness discrimination with better preference preservation and higher student entropy than SFT and RL teachers.

The SFT teacher shows the largest distributional shift from the initial student, while the RL teacher stays substantially closer. FlowTD achieves the smallest shift among the trained teachers (Figure 4b).

Teacher entropy analysis. All trained teachers have lower token entropy than the prompt-only teacher, but FlowTD retains substantially higher entropy than SFT and RL (Figure 4c). At comparable teacher accuracy, this indicates that FlowTD retains greater token-level diversity.

Among the similarly accurate trained teachers, FlowTD's smaller distributional shifts and greater token-level diversity accompany the strongest student performance (Table 2). These results support jointly considering correctness and distributional properties when learning teachers for distillation.

4.4 Teacher Supervision and Student Dynamics

We compare teacher supervision on a shared set of frozen-base trajectories and track student token entropy during distillation on Qwen2.5-3B (Figure 6).

Correctness discrimination. Within-problem AUC measures whether teacher-to-base log-ratio scores rank correct trajectories above incorrect ones. The prompt teacher performs close to the random-ranking baseline of 0.5, indicating weak correctness discrimination despite privileged information. Teacher training substantially improves this ability, with SFT, RL, and FlowTD achieving comparable discrimination (Figure 6a).

Preference preservation among correct trajectories. The ideal posterior preserves the prior's relative probabilities among correct trajectories, yielding identical sequence-level teacher-to-base log-ratios within each problem. We therefore measure the standard deviation of these summed log-ratios (Appendix A.4). SFT shows the largest variation, indicating the greatest distortion of the prior's relative preferences. RL reduces this distortion, while FlowTD's explicit distribution-matching approach achieves the strongest preference preservation among the trained teachers.

Student entropy dynamics. We track student token entropy to examine how token-level diversity evolves during distillation. Students distilled from SFT and RL teachers undergo a sharp entropy decline, whereas FlowTD's student stabilizes at a substantially higher level (Figure 6c). Together with Figure 4c, these results suggest an association between teacher and student token-level diversity. FlowTD retains greater diversity in both.

Compared with SFT and RL teachers, FlowTD provides comparable correctness discrimination while better preserving prior preferences among correct trajectories, and its student retains greater

token-level diversity. Together with the teacher accuracy and distributional shift, these findings support jointly targeting correctness, minimal shift, and preference preservation in teacher design and support FlowTD's effectiveness in approximating the ideal posterior for distillation.

5 Related Work

On-policy distillation and self-distillation. On-policy distillation (OPD) trains the student to match the teacher's next-token predictions on trajectories sampled from the student's current policy (Agarwal et al., 2024; Thinking Machines, 2025). By providing guidance at prefixes the student actually visits, OPD addresses the training–inference distribution mismatch associated with learning only from fixed demonstrations. In self-distillation, these predictions can come from the model itself conditioned on additional context. OPSD (Zhao et al., 2026) uses reference solutions, SDPO (Hübotter et al., 2026) uses environment feedback or successful attempts, and SDFT (Shenfeld et al., 2026) applies demonstration-conditioned self-distillation to continual learning.

Privileged context, however, does not by itself ensure reliable supervision. Nguyen et al. (2026) distinguish context-induced likelihood changes from outcome-based action values. More broadly, prior work has explored improvements to both student updates and teacher learning. On the student side, RLSD (Yang et al., 2026) reweights verifier-based trajectory advantages at the token level using positive teacher–student probability ratios. For general OPD, Fu et al. (2026) study sampled-token supervision and use local distribution matching over the teacher's top-K tokens at each student prefix. On the teacher side, π-Distill (Penaloza et al., 2026) trains a privileged teacher with KL-regularized reinforcement learning, while variational policy distillation (VPD) (Li et al., 2026) trains a feedback-conditioned teacher through outcome-based preference optimization. Our work investigates what makes a teacher distribution suitable for on-policy distillation, focusing on correctness and the preservation of the initial student's relative preferences among correct reasoning paths. We compare teacher constructions under a shared student-side distillation procedure.

Posterior inference and distribution matching. Posterior inference defines target distributions by combining a prior with evidence about desired outcomes. Hu et al. (2024) use GFlowNets for amortized posterior inference in language models, while JEPO (Tang et al., 2025) studies posteriors over latent reasoning chains conditioned on the correct answer.

To fit such targets, GFlowNets (Bengio et al., 2023) learn to sample in proportion to unnormalized terminal weights, with trajectory balance (Malkin et al., 2022) providing a complete-trajectory training objective. FlowRL (Zhu et al., 2026) adapts trajectory balance to train reasoning policies toward reward-weighted distributions. FlowTD uses this distribution-matching approach to fit a privileged teacher to a correctness-conditioned target defined by the frozen initial student's problem-only prior. The learned teacher then provides next-token supervision for on-policy distillation.

6 Conclusion

We presented a framework for designing and learning teacher distributions for on-policy self-distillation. Rather than judging a teacher by accuracy alone, we identified correctness, minimal shift, and preference preservation as desirable properties and showed that the correctness-conditioned posterior unifies them. FlowTD approximates this target with GFlowNet trajectory balance, making probability allocation among correct reasoning paths an explicit part of teacher training. Our analyses show that FlowTD learns teachers with a more favorable balance of these properties than prompt-only, SFT-trained, and RL-trained teachers. Across mathematical reasoning benchmarks, its student achieves the highest average accuracy among the evaluated teacher constructions.

More broadly, these results suggest that the distribution a teacher represents, not only how often it answers correctly, shapes what a student learns from distillation. By specifying which responses to favor and how to weight alternative correct solutions, the correctness-conditioned posterior turns teacher construction from a heuristic choice into an explicit design target. FlowTD offers a principled way to learn such teachers, translating verifiable outcomes into dense token-level supervision while preserving the model's relative preferences among correct solutions.

Reproducibility statement

Sections 2 and 3 specify the teacher target, proofs, and training objectives. Appendix A records data sources, training configurations, checkpoint selection, and evaluation procedures, including the pass@k estimator and teacher-score definitions. The main tables report greedy accuracy from validation-selected checkpoints.

Ethics statement

This work uses existing pretrained models and public mathematical reasoning datasets and does not involve human participants. Final-answer verification does not establish the validity of every reasoning step; generated solutions can still contain incorrect or misleading reasoning.

AI use statement

We used generative AI tools, including OpenAI Codex and Claude, to assist with code development, manuscript writing, and literature checks. All core research ideas, theoretical derivations, experimental design, and algorithmic innovations are developed by the authors. AI-assisted code was manually reviewed by the authors, who take full responsibility for the content and results of this paper.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Yoshua Bengio, Salem Lahlou, Tristan Deleu, Edward J. Hu, Mo Tiwari, and Emmanuel Bengio. GFlowNet foundations. *Journal of Machine Learning Research*, 24(210):1–55, 2023.

Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026. URL https://arxiv.org/abs/2603.25562.

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.

Chaoqun He, Renjie Luo, Yuzhuo Bai, et al. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024. doi: 10.48550/arXiv.2402.14008. URL https://arxiv.org/abs/2402.14008.

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 *NeurIPS Datasets and Benchmarks Track*, 2021. doi: 10.48550/arXiv.2103.03874. URL https://arxiv.org/abs/2103.03874.

Edward J. Hu, Moksh Jain, Eric Elmoznino, Younesse Kaddar, Guillaume Lajoie, Yoshua Bengio, and Nikolay Malkin. Amortizing intractable inference in large language models. In *ICLR*, 2024.

Jonas Hübotter, Frederike Lübeck, Lejs 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://arxiv.org/abs/2601.20802.

Yang Li, Erik Nijkamp, Semih Yavuz, and Shafiq Joty. Learning from language feedback via variational policy distillation. *arXiv preprint arXiv:2605.15113*, 2026.

Nikolay Malkin, Moksh Jain, Emmanuel Bengio, Chen Sun, and Yoshua Bengio. Trajectory balance: Improved credit assignment in GFlowNets. In *NeurIPS*, 2022.

Xuan-Phi Nguyen, Zeyu Leo Liu, Yang Li, Shrey Pandit, Yiran Zhao, Anurag Koul, and Shafiq Joty. Privileged likelihood is not automatically value: Three checks for token credit in on-policy self-distillation. *arXiv preprint arXiv:2608.09263*, 2026.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. *arXiv preprint arXiv:2602.04942*, 2026.

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

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*, 2026. URL https://arxiv.org/abs/2601.19897.

Yunhao Tang, Sid Wang, Lovish Madaan, and Rémi Munos. Beyond verifiable rewards: Scaling reinforcement learning for language models to unverifiable data. *arXiv preprint arXiv:2503.19618*, 2025. URL https://arxiv.org/abs/2503.19618.

Thinking Machines. On-policy distillation. Blog post, https://thinkingmachines.ai/blog/on-policy-distillation/, 2025.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

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

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. SimpleRL-Zoo: Investigating and taming zero reinforcement learning for open base models in the wild, 2025. URL https://arxiv.org/abs/2503.18892.

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.

Xuekai Zhu, Daixuan Cheng, Dinghuai Zhang, Hengli Li, Kaiyan Zhang, Che Jiang, Youbang Sun, Ermo Hua, Yuxin Zuo, Xingtai Lv, Qizheng Zhang, Lin Chen, Fanghao Shao, Bo Xue, Yunchong Song, Zhenjie Yang, Ganqu Cui, Ning Ding, Jianfeng Gao, Xiaodong Liu, et al. FlowRL: Matching reward distributions for LLM reasoning. In *ICLR*, 2026. arXiv:2509.15207.

A Additional Experiment Details

A.1 Hyperparameters

In the Qwen2.5-3B setting, both teacher and student use AdamW with learning rate 5×10^{-7} and gradient clipping 1.0. The configuration uses 128 prompts and 8 responses per prompt. Teacher sampling uses $T = 1.0$, top-$p = 0.95$, and up to 4,000 response tokens. Students also sample $K = 8$ responses per prompt.

A.2 Qwen3-8B-Base implementation

We use 10,000 problems from the `metadata` configuration of OpenThoughts-114k, selected with `domain=math` and `source=numina_math`. The privileged prompt contains the DeepSeek-R1 reference in `deepseek_solution`; for SFT, the supervised target is the same problem's `ground_truth_solution`.

Experiments use `verl` on four NVIDIA B200 GPUs. Teacher and student training each run for one epoch with a batch size of 64 prompts. Rollout-based training uses 8 responses per prompt and a configured maximum length of 6k tokens. The FlowTD teacher uses $\beta = 15$. Teacher prompts follow the SDPO template (Hübotter et al., 2026), with reference trajectories from the training dataset providing privileged context. SDPO follows the original EMA setting and uses dataset reference trajectories as feedback. The RL teacher uses a KL coefficient of 10^{-4} with a frozen, problem-only base reference, following the SimpleRL training setup (Zeng et al., 2025).

A.3 Checkpoint selection and evaluation

All methods are evaluated at the checkpoint with the best validation performance, using the same selection rule. The reported 8B RL-teacher baseline uses the validation-selected checkpoint at step 30. Main-table results use greedy decoding. For Figure 5, we sample $n = 32$ responses per AIME 2025 problem at temperature $T = 1.0$ from each Qwen3-8B-Base student. If c_i of the n responses to problem i are correct, we use the unbiased estimator

$$\widehat{\text{pass@}k} = \frac{1}{N} \sum_{i=1}^{N} \left[1 - \frac{\binom{n-c_i}{k}}{\binom{n}{k}} \right], \qquad 1 \le k \le n, \tag{8}$$

where N is the number of evaluation problems. Thus, estimates are computed per problem and then averaged across problems.

A.4 Teacher-score diagnostics

For Figure 6a–b, all teachers score the same 800 trajectories sampled from the frozen Qwen2.5-3B base model: 8 responses for each of 100 problems. For panel (b), the teacher-to-base log-ratio is summed over the entire response:

$$s_{q_\mathrm{t}}(z) = \sum_{t=1}^{|z|} \left[\log q_\mathrm{t}(z_t \mid x, [\phi], z_{<t}) - \log p_0(z_t \mid x, z_{<t}) \right]. \tag{9}$$

We measure the standard deviation of this score among correct responses within each problem. For the ideal posterior, $s_{q^*}(z) = -\log Z(x, y)$ for every correct response, so its within-problem standard deviation is zero.