BEYOND POINTWISE SIGNALS: PROSPECTIVE SUPERVISION WEIGHTING IN ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) for large language models (LLMs) provides dense teacher supervision along student-generated trajectories, yet existing token-weighting methods judge each position by signals local to the current state. This overlooks a defining property of OPD: because the student generates its own trajectories, learning at one state changes which states, and hence which supervision, it encounters later. Through controlled single-state interventions, we show that updates with nearly identical immediate learning gains can have opposite downstream effects. We therefore introduce **Pro**spective Supervision **W**eighting (**PROW**), which scores each token by its immediate learning opportunity together with a prospective term estimating how the teacher-directed correction at that token shifts the student toward continuations offering more learning opportunity. PROW computes both terms from standard OPD rollouts and only reallocates a fixed per-token optimization budget, requiring no additional rollouts, reward models, or learned critics. Across reasoning benchmarks and model scales, PROW improves standard OPD and existing token-weighting baselines, while controlled interventions confirm that its prospective signal captures effects beyond local learning value.

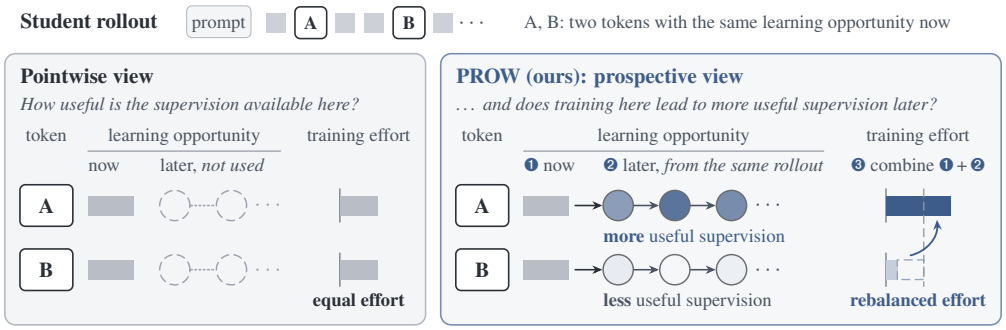


Figure 1: **Same learning opportunity now, different later.** In on-policy distillation the student writes its own continuation, so training at a token changes the tokens it writes afterwards. Tokens A and B from a student rollout have the same learning opportunity now: the student can learn equally much from the teacher at each. *Left:* a pointwise view scores each token by this alone and gives A and B equal training effort. *Right:* our method, **PROW**, takes a prospective view and also estimates their learning opportunity later from the tokens that follow in the same rollout, namely whether training at the token makes a path with more useful teacher supervision more likely. **PROW** combines both to rebalance training effort from B to A.

1 INTRODUCTION

Knowledge distillation transfers the behavior of a strong teacher to a smaller student, reducing the cost of deploying large language models (Hinton et al., 2015). For autoregressive models, conventional distillation is often performed on fixed reference or teacher-generated sequences (Kim & Rush, 2016; Wen et al., 2023), so the prefixes observed during training may differ from those induced by the student at inference time. On-policy distillation (OPD) mitigates this mismatch by querying

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

the teacher along student-generated trajectories (Gu et al., 2024; Agarwal et al., 2024), providing supervision directly at the states the student actually visits.

Recent work shows that not all on-policy supervision is equally valuable (Li et al., 2026b; Wang et al., 2026; Xie et al., 2026; Xu et al., 2026). This has motivated methods that selectively allocate optimization effort using signals such as teacher–student discrepancy, prefix-level mismatch, support compatibility, or local teachability (Xie et al., 2026; Wang et al., 2026; Li et al., 2026a). Despite their differences, these methods largely assess supervision using information available at, or accumulated up to, the current state. They therefore answer variants of the same question: *how useful is the supervision available here?* Recent trajectory-aware methods further suggest that future context can provide additional information beyond an isolated token (Jiang & Ferraro, 2026; Liu et al., 2026; Chen et al., 2026).

We argue that a further consequence of this sequential structure is that *learning at a state changes not only the prediction at that state, but also the supervision the student is likely to encounter next.* Because future prefixes are generated by the evolving student policy, a local update can redirect the subsequent trajectory and alter which learning opportunities become reachable. Consequently, two states with similar immediate learning value can have substantially different downstream consequences.

Figure 1 illustrates this distinction. Tokens A and B have the same immediate learning opportunity, yet learning at the two states can lead to continuations with very different future supervision. Figure 2 further confirms this effect through controlled interventions: states with similar immediate gains can induce widely different downstream changes. Together, these observations suggest that local learning value alone is insufficient for allocating supervision in OPD.

Motivated by this observation, we introduce **PROW** (**PRO**spective supervision **W**eighting), which combines immediate learning opportunity with the prospective effect of the same teacher-directed correction on downstream learning opportunities. The immediate term measures how much can be gained locally, while the prospective term measures whether the correction redirects the student toward continuations with more useful supervision. Both quantities are computed from the standard OPD rollout and used to reweight token-level supervision.

Across reasoning benchmarks, teacher–student scales, and training corpora, PROW improves standard OPD and existing token-weighting baselines. Controlled interventions further show that its prospective signal captures downstream consequences beyond immediate learning value, while ablations show that the two components provide complementary signals.

Accordingly, we make three contributions:

1. We identify a limitation of pointwise supervision signals in OPD: states with similar immediate learning opportunity can have substantially different downstream consequences because local updates change which supervision the student is likely to encounter next.
2. We introduce **PROW**, a prospective supervision weighting framework that combines immediate learning opportunity with the downstream learning opportunities made more or less reachable by the same teacher-directed correction.
3. We validate PROW across multiple teacher–student configurations, training corpora, and reasoning benchmarks, and provide controlled intervention evidence that its prospective signal captures supervision effects beyond local learning value.

2 Similar Local Gains, Different Downstream Supervision

A pointwise supervision signal implicitly assumes that the value of learning at a state can be assessed from its local effect. In OPD, however, an update can also change the continuation generated by the student and therefore the supervision encountered afterward. We first test whether this downstream effect is already determined by the immediate learning gain.

We conduct a controlled intervention study with Qwen3-4B $\rightarrow$ Qwen3-1.7B-Base on Competition-MATH. Starting from a common student checkpoint θ_0, we independently apply one teacher-supervised update at each of 300 student-visited states. For a state s_t, let θ_t^+ denote the student after

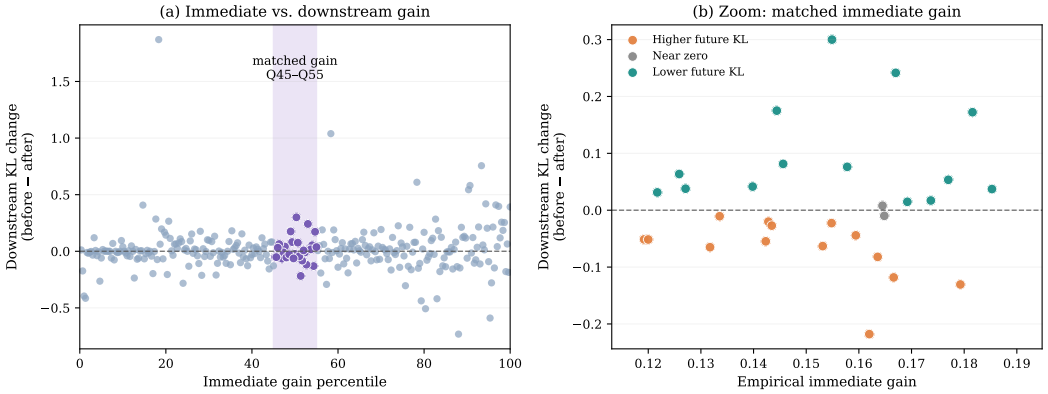


Figure 2: **Similar local gains can lead to different downstream supervision.** We independently apply one teacher-supervised update to 300 student-visited states from a common checkpoint. Immediate gain measures the local reduction in distillation loss, while downstream change measures the induced change in future state visitation using a frozen pre-update evaluator. **(a)** Downstream change is only weakly associated with immediate gain. **(b)** Even within the matched Q45-Q55 gain band, downstream effects remain widely dispersed and can have opposite signs. Thus, similar pointwise improvements can expose the student to substantially different supervision afterward.

this local update. We define its realized immediate gain as

$$\Delta_t^{\mathrm{imm}} = L(\theta_0; s_t) - L(\theta_t^+; s_t). \tag{1}$$

We then sample continuations from the same prefix before and after the update. To isolate changes in *which states are visited*, all future states are evaluated using the same frozen pre-update student and teacher. Letting F_t^{before} and F_t^{after} denote the corresponding mean future-state distillation losses, we measure

$$\Delta_t^{\mathrm{future}} = F_t^{\mathrm{before}} - F_t^{\mathrm{after}}. \tag{2}$$

We use $\Delta_t^{\mathrm{future}}$ to isolate the change in downstream state visitation induced by the local update.

Figure 2 reveals a clear mismatch between these two effects. Across all interventions, immediate gain is only weakly correlated with downstream change (Pearson $r = 0.115$; Spearman $\rho = 0.099$). More importantly, this heterogeneity persists even after controlling for local improvement. Within the Q45–Q55 band of matched immediate gains, downstream changes span $[-0.218, 0.300]$, with both positive and negative effects. Hence, states that appear nearly indistinguishable from a pointwise perspective can expose the student to markedly different supervision after training on them.

This observation motivates a prospective view of supervision: its value should depend not only on the immediate learning gain, but also on how the update changes the supervision encountered downstream. PROW captures both effects in a unified supervision score.

3 Preliminaries

Setup and notation. Let $\mathcal{D}$ denote the prompt distribution, π_θ the student policy, and π_{T} a fixed teacher policy over a shared vocabulary $\mathcal{V}$. For a prompt $x \sim \mathcal{D}$, the student generates $y = (y_1, \ldots, y_T) \sim \pi_\theta(\cdot \mid x)$, where T denotes the number of valid response tokens and $s_t := (x, y_{<t})$ is the student-visited prefix. For a candidate next token $a \in \mathcal{V}$, we write

$$p_t(a) := \pi_\theta(a \mid s_t), \qquad q_t(a) := \pi_{\mathrm{T}}(a \mid s_t). \tag{3}$$

Top-K local policy space. To avoid full-vocabulary computation at every visited prefix, we evaluate the local distillation quantities on the student's top-K support,

$$\mathcal{A}_t := \mathrm{TopK}_K(p_t). \tag{4}$$

For any fixed support $\mathcal{S}_t \subseteq \mathcal{V}$, let

$$p_t^{\mathcal{S}}(a) := \frac{p_t(a)}{\sum_{b \in \mathcal{S}_t} p_t(b)}, \qquad q_t^{\mathcal{S}}(a) := \frac{q_t(a)}{\sum_{b \in \mathcal{S}_t} q_t(b)}, \qquad a \in \mathcal{S}_t. \tag{5}$$

The rollout itself is still sampled from the full student policy π_θ; Top-K restricts the local distributions used for distillation and scoring.

On-policy distillation. OPD trains the student on its own rollouts while querying the teacher at the resulting student-visited prefixes. On the local policy space above, the token-level reverse-KL loss is

$$\ell_t^{\mathrm{OPD}} \coloneqq D_{\mathrm{KL}}\big(p_t^{\mathcal{A}} \| q_t^{\mathcal{A}}\big)\,, \tag{6}$$

and the length-normalized objective is

$$\mathcal{L}_{\mathrm{OPD}}(\theta) \coloneqq \mathbb{E}_{\substack{x\sim\mathcal{D}\\ y\sim\pi_\theta(\cdot|x)}}\left[\frac{1}{T}\sum_{t=1}^{T}\ell_t^{\mathrm{OPD}}\right]. \tag{7}$$

As in standard OPD, sampled prefixes are treated as fixed when differentiating each token-level loss; gradients are not propagated through the rollout process.

4 METHOD

Overview. PROW values supervision at a student-visited state through two complementary effects: the immediate benefit of learning at the current state and the prospective effect of the same correction on the supervision encountered afterward. Concretely, for each position t in a student-generated trajectory $y = (y_1, \ldots, y_T)$, we define $S_t = G_t + \lambda \widehat{D}_t$, where G_t measures the immediate learning opportunity and $\widehat{D}_t$ estimates the prospective learning effect. Section 4.1 develops the local term G_t, Section 4.2 derives the prospective term $\widehat{D}_t$, and Section 4.3 converts their combination into token-level training weights. PROW uses these scores only to redistribute optimization effort across sampled positions, leaving the underlying OPD loss unchanged.

Restricted supports. For efficient scoring, we use two restricted supports tailored to the two terms. The immediate opportunity is evaluated on the student's high-probability actions, $\mathcal{A}_t \coloneqq \mathrm{TopK}_K(p_t)$. For the prospective term, we additionally retain teacher-favored actions that may become more reachable after correction, $\mathcal{B}_t \coloneqq \mathrm{TopK}_K(p_t) \cup \mathrm{TopK}_K(q_t)$. For either $\mathcal{S}_t \in \{\mathcal{A}_t, \mathcal{B}_t\}$, let

$$p_t^{\mathcal{S}}(a) = \frac{p_t(a)}{\sum_{b\in\mathcal{S}_t} p_t(b)}, \qquad q_t^{\mathcal{S}}(a) = \frac{q_t(a)}{\sum_{b\in\mathcal{S}_t} q_t(b)}, \qquad a \in \mathcal{S}_t. \tag{8}$$

These restrictions affect only scoring; trajectories are still sampled from the full student policy.

To analyze both immediate and prospective effects under the same correction, we perturb the student distribution toward the teacher:

$$p_{t,\eta}^{\mathcal{S}}(a) \propto p_t^{\mathcal{S}}(a)\exp\left(\eta \log \frac{q_t^{\mathcal{S}}(a)}{p_t^{\mathcal{S}}(a)}\right), \qquad a \in \mathcal{S}_t. \tag{9}$$

This path satisfies $p_{t,0}^{\mathcal{S}} = p_t^{\mathcal{S}}$ and $p_{t,1}^{\mathcal{S}} = q_t^{\mathcal{S}}$. We use its first-order behavior at $\eta = 0$, treating the selected supports as fixed.

4.1 IMMEDIATE LEARNING OPPORTUNITY

We first quantify the immediate value of supervision at s_t. Under the teacher-directed perturbation in Eq. (9), a locally useful state is one for which moving the student toward the teacher rapidly reduces the current student–teacher mismatch. We therefore define the immediate learning opportunity as the first-order reduction in reverse KL at $\eta = 0$:

$$G_t = -\left.\frac{d}{d\eta} D_{\mathrm{KL}}\left(p_{t,\eta}^{\mathcal{A}} \| q_t^{\mathcal{A}}\right)\right|_{\eta=0} = \mathrm{Var}_{a\sim p_t^{\mathcal{A}}}\left[\log \frac{q_t^{\mathcal{A}}(a)}{p_t^{\mathcal{A}}(a)}\right]. \tag{10}$$

The variance form gives a direct interpretation of G_t. It is small when student and teacher have similar relative preferences over the student's high-probability actions, and large when the teacher would substantially redistribute probability among them. Thus, G_t captures the strength of the local teacher-directed correction, with larger values indicating greater immediate learning opportunity. Appendix B.1 gives the derivation.

4.2 PROSPECTIVE LEARNING EFFECT

The immediate term G_t measures how much can be learned at the current state. To go beyond this pointwise view, we ask a complementary question: *does the same teacher-directed correction move the student toward continuations with better learning opportunities?* We construct the prospective term to measure this effect.

Downstream learning opportunity. The correction at s_t changes which next-action branch the student is likely to follow. We therefore first assign each candidate action $a \in \mathcal{B}_t$ a value describing the learning opportunity available along its downstream continuation.

For an action a at a future state s_j, define $w_j(a) \coloneqq \mathbf{1}\{a \in \mathcal{B}_j\} \min\{1, q_j(a)/p_j(a)\}$. Since $p_j(a)w_j(a) = \mathbf{1}\{a \in \mathcal{B}_j\} \min\{p_j(a), q_j(a)\}$ under student sampling, w_j discounts transitions with little probability mass shared by student and teacher. We define the downstream opportunity of branch a as

$$U_t(a) = \mathbb{E}_{\tau \sim \pi_\theta(\cdot|s_t,a)} \left[\sum_{k=1}^{T-t} \gamma^{k-1} \left(\prod_{j=t+1}^{t+k-1} w_j(y_j) \right) \big(G_{t+k} - G_t\big) \right]. \tag{11}$$

Thus, $U_t(a)$ measures whether the continuation reached through a offers more or less learning opportunity than the current state, while discounting weakly shared transitions along that continuation. Because $U_t(a)$ is conditioned on already taking a, the overlap of the current transition is handled separately below. Appendix D shows that the root-centered term $G_{t+k} - G_t$ is the unique shift-invariant linear contrast under the fixed downstream weights.

From branch opportunity to prospective effect. Having assigned a downstream opportunity to each next-action branch, we now ask how the teacher-directed correction redistributes probability across these branches. The desired prospective effect is the first-order change in their expected downstream opportunity, with $w_t(a)$ retaining the shared mass of the current transition.

Proposition 1 (Directional prospective effect)**.** *On $\mathcal{B}_t$, let $r_t(a) \coloneqq \log(q_t^{\mathcal{B}}(a)/p_t^{\mathcal{B}}(a))$, $\bar{r}_t \coloneqq \mathbb{E}_{a\sim p_t^{\mathcal{B}}}[r_t(a)]$, and $Z_t^{\mathcal{B}} \coloneqq \sum_{a\in\mathcal{B}_t} p_t(a)$. Define*

$$\Phi_t(\eta) = \gamma Z_t^{\mathcal{B}} \mathbb{E}_{a\sim p_{t,\eta}^{\mathcal{B}}} \left[w_t(a) U_t(a) \right].$$

Holding $U_t(a)$ and $w_t(a)$ fixed under the local perturbation,

$$D_t \coloneqq \frac{d}{d\eta} \Phi_t(\eta) \Big|_{\eta=0} = \gamma \sum_{a\in\mathcal{B}_t} \min\{p_t(a), q_t(a)\} \big(r_t(a) - \bar{r}_t\big) U_t(a). \tag{12}$$

Hence, D_t is positive when the correction reallocates probability toward branches with greater downstream learning opportunity, or away from branches with lower opportunity. Together, G_t and D_t capture the immediate and prospective effects of the same teacher-directed correction. The proof is given in Appendix B.3.

Single-rollout estimate. Since evaluating D_t requires multiple actions and continuations, PROW reuses the sampled action y_t and its observed suffix, replacing $U_t(y_t)$ with its single-trajectory estimate $\widehat{U}_t(y_t)$ and using $\widehat{D}_t = \gamma w_t(r_t(y_t) - \bar{r}_t)\widehat{U}_t(y_t)$. When $y_t \notin \mathcal{B}_t$, we set $\widehat{D}_t = 0$.

4.3 PROSPECTIVE SUPERVISION WEIGHTING

For an update, let $\mathcal{C}$ denote the set of N valid token positions. Given supervision scores $\{S_i\}_{i\in\mathcal{C}}$, PROW obtains mean-one token weights by solving

$$\max_{\{v_i\}} \quad \frac{1}{N} \sum_{i\in\mathcal{C}} v_i S_i \quad \text{s.t.} \quad \frac{1}{N} \sum_{i\in\mathcal{C}} v_i = 1, \quad D_{\mathrm{KL}}(v/N \| u_N) \le \epsilon, \quad v_{\min} \le v_i \le v_{\max}. \tag{13}$$

Its solution has the clipped Gibbs form $v_i^\star = \mathrm{clip}(c \exp(S_i/\tau), v_{\min}, v_{\max})$, with c and τ determined by the constraints; see Appendix C for the derivation. The detached weights are then applied to the original OPD objective, $\mathcal{L}_{\mathrm{train}} = \dfrac{1}{N} \sum_{i\in\mathcal{C}} v_i^\star \ell_i^{\mathrm{OPD}}$.

5 EXPERIMENTS

We evaluate PROW through three research questions:

- **RQ1:** Does PROW consistently improve on-policy distillation over standard and token-weighted OPD baselines across teacher–student scales and training corpora?
- **RQ2:** Does the prospective signal capture downstream consequences of supervision that are not explained by local learning value?
- **RQ3:** Do the gains from PROW transfer beyond the mathematical reasoning benchmarks used in the main evaluation?

5.1 EXPERIMENTAL SETUP

Teacher–student configurations. The main comparison considers three Qwen3 teacher–student pairs (Yang et al., 2025): Qwen3-14B → Qwen3-4B, Qwen3-4B → Qwen3-1.7B, and Qwen3-8B → Qwen3-4B-Base. The teacher is frozen throughout distillation. Within each setting, all methods share the student initialization, training data, rollout and evaluation protocols, and training step; they differ only in how supervision is allocated across student-visited positions. Additional Qwen3, Qwen3.5, and Llama results are reported in Appendix F.

Training corpora. We use Competition-MATH (Hendrycks et al., 2021) and DAPO-Math-17K (Yu et al., 2025). Each teacher–student pair is trained on both corpora, yielding six settings. For Competition-MATH, distillation and evaluation use its disjoint training and test splits, respectively.

Baselines. We compare PROW with three token-level OPD baselines. **OPD** uniformly weights valid response tokens (Agarwal et al., 2024); **TA-OPD** prioritizes locally teachable tokens (Wang et al., 2026); and **IW-OPD** uses sequential prefix-level importance weighting (Xie et al., 2026). PROW retains the same OPD loss but reallocates supervision across positions. For each baseline, we use the hyperparameter configuration reported in the corresponding paper, unless otherwise specified.

Evaluation benchmarks and metrics. We evaluate on Competition-MATH (Hendrycks et al., 2021), MATH-500 (Lightman et al., 2023), AIME 2024, AIME 2025, and AMC 2023 (Mathematical Association of America, 2026). For each problem, we sample eight responses under the same decoding protocol and report **Avg@8**, the mean correctness across samples, and **Pass@8**, the fraction of problems with at least one correct response. GPQA-Diamond (Rein et al., 2024) and MMLU-Pro (Wang et al., 2024) are reserved for the cross-domain study in Section 5.4.

5.2 MAIN RESULTS

Table 1 reports the results for RQ1. Each block compares methods under the same teacher–student pair and training corpus. Every entry reports *Avg@8 / Pass@8* (%); higher is better for both metrics.

RQ1: Overall effectiveness. Across the 30 benchmark–setting pairs in Table 1, PROW attains the highest or tied-highest Avg@8 in 23 cases and the highest or tied-highest Pass@8 in 24 cases. Among these, 22 Avg@8 results and 18 Pass@8 results are strict improvements over all compared token-weighting baselines. The gains occur across both student scales and both training corpora rather than being concentrated in a single configuration. Performance on AMC and AIME is more mixed, with another token-weighting baseline matching or outperforming PROW in several settings. We further analyze the training dynamics underlying the main results in Figure 7 in the appendix.

5.3 MECHANISM VALIDATION

RQ2: Mechanism validation. We test two implications of the proposed mechanism. First, conditioned on similar immediate learning opportunity G_t, does the prospective score $\widehat{D}_t$ distinguish updates with different downstream consequences? Second, does this additional information complement G_t when allocating supervision during training? Appendix G.2 separately verifies that the single-rollout prospective estimate preserves the ranking obtained from a multi-rollout reference, while Appendix G.3 compares the prospective signal with an existing local teachability criterion.

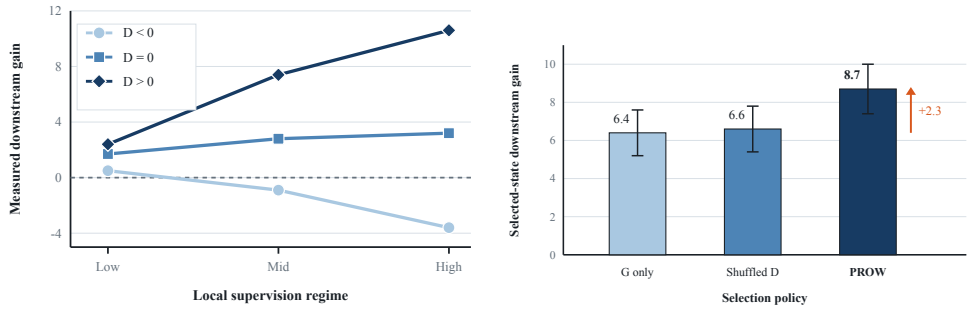
Table 1: Main results across teacher–student configurations and training corpora. Each entry reports **Avg@8 / Pass@8** (%). Within each setting, the best result is **boldfaced** and the second-best is underlined. Student and teacher performance is included for reference as lower and upper bounds.

Method	Compe-MATH	AMC23	AIME24	AIME25	MATH-500
Teacher: Qwen3-14B → **Student:** Qwen3-4B **Train:** DAPO-Math					
OPD	66.70 / 76.60	69.07 / 90.00	23.75 / 50.00	20.42 / **50.00**	67.20 / 75.80
IW-OPD	67.20 / 76.20	67.82 / 92.50	26.67 / **53.34**	**25.42** / 43.34	67.50 / 76.00
TA-OPD	67.70 / 76.00	66.57 / **95.00**	27.09 / **53.34**	21.67 / 43.34	67.90 / 76.40
PROW	**70.00 / 77.60**	**71.25** / 92.50	**29.17 / 53.34**	**25.42 / 50.00**	**70.00 / 77.40**
Student init. (lower ref.)	68.00 / 76.00	67.00 / 90.00	20.00 / 37.00	23.00 / 40.00	67.00 / 75.00
Teacher (upper ref.)	72.00 / 76.00	73.00 / 90.00	30.00 / 50.00	25.00 / 43.00	70.00 / 75.40
Teacher: Qwen3-14B → **Student:** Qwen3-4B **Train:** Compe-MATH					
OPD	67.20 / 76.40	66.25 / 95.00	22.92 / 36.67	20.84 / **46.67**	67.20 / 74.60
IW-OPD	67.80 / 75.80	67.50 / 92.50	25.00 / 50.00	24.17 / **46.67**	67.20 / 76.00
TA-OPD	67.60 / 75.40	66.88 / 95.00	26.25 / 43.34	20.00 / **46.67**	67.30 / 76.00
PROW	**70.20 / 78.80**	**69.07 / 97.50**	**27.09 / 53.34**	**26.25 / 46.67**	**68.90 / 77.60**
Student init. (lower ref.)	68.00 / 76.00	67.00 / 90.00	20.00 / 37.00	23.00 / 40.00	67.00 / 75.00
Teacher (upper ref.)	72.00 / 76.00	73.00 / 90.00	30.00 / 50.00	25.00 / 43.00	70.00 / 75.40
Teacher: Qwen3-4B → **Student:** Qwen3-1.7B **Train:** DAPO-Math					
OPD	61.00 / 72.20	49.69 / 87.50	17.09 / 40.00	14.17 / 36.67	60.50 / 73.40
IW-OPD	60.80 / 73.80	50.94 / 85.00	17.50 / 40.00	15.84 / 36.67	61.00 / 73.60
TA-OPD	60.70 / 74.40	50.94 / 87.50	**18.75** / 40.00	16.67 / 36.67	60.50 / 72.60
PROW	**63.40 / 74.60**	**54.38 / 90.00**	18.34 / **43.34**	**17.92 / 40.00**	**64.00 / 75.40**
Student init. (lower ref.)	58.00 / 73.00	41.00 / 72.50	15.00 / 33.00	11.00 / 27.00	58.00 / 70.00
Teacher (upper ref.)	68.00 / 76.00	67.00 / 90.00	20.00 / 37.00	23.00 / 40.00	67.00 / 75.00
Teacher: Qwen3-4B → **Student:** Qwen3-1.7B **Train:** Compe-MATH					
OPD	59.20 / 71.00	47.82 / 80.00	17.92 / 36.67	17.09 / 36.67	60.50 / 70.80
IW-OPD	60.30 / 73.60	50.32 / 85.00	18.34 / **40.00**	16.67 / 33.34	60.80 / 73.60
TA-OPD	59.60 / 73.80	51.88 / **87.50**	**19.59** / 36.67	17.09 / 40.00	61.20 / 73.40
PROW	**62.80 / 74.80**	**52.19** / 85.00	18.34 / **40.00**	**18.34 / 46.67**	**62.80 / 75.00**
Student init. (lower ref.)	58.00 / 73.00	41.00 / 72.50	15.00 / 33.00	11.00 / 27.00	58.00 / 70.00
Teacher (upper ref.)	68.00 / 76.00	67.00 / 90.00	20.00 / 37.00	23.00 / 40.00	67.00 / 75.00
Teacher: Qwen3-8B → **Student:** Qwen3-4B-Base **Train:** DAPO-Math					
OPD	63.70 / 73.60	54.38 / 85.00	18.34 / 36.67	16.67 / 33.34	63.80 / 74.20
IW-OPD	63.90 / 74.00	54.07 / 82.50	17.50 / 30.00	17.50 / 40.00	63.80 / 74.60
TA-OPD	**64.80** / 72.80	54.69 / 87.50	16.25 / **43.34**	**19.17 / 43.34**	**64.90** / 75.20
PROW	64.30 / **75.20**	**57.82 / 92.50**	**20.84** / 40.00	18.75 / 40.00	64.70 / **75.40**
Student init. (lower ref.)	33.00 / 70.00	27.00 / 62.50	7.00 / 27.00	2.50 / 23.00	33.00 / 68.00
Teacher (upper ref.)	68.00 / 77.00	69.00 / 92.50	26.00 / 50.00	21.00 / 40.00	68.00 / 76.00
Teacher: Qwen3-8B → **Student:** Qwen3-4B-Base **Train:** Compe-MATH					
OPD	63.80 / 74.00	56.88 / **90.00**	15.00 / 30.00	15.00 / 36.67	63.30 / 73.20
IW-OPD	63.20 / 74.00	55.94 / 87.50	18.34 / 33.34	17.50 / 40.00	63.60 / 74.20
TA-OPD	63.80 / **74.80**	**58.44** / 87.50	**20.00 / 40.00**	17.50 / **43.34**	63.50 / 74.00
PROW	**63.90 / 74.80**	57.82 / 87.50	19.17 / **40.00**	**21.67** / 40.00	**64.10 / 74.60**
Student init. (lower ref.)	33.00 / 70.00	27.00 / 62.50	7.00 / 27.00	2.50 / 23.00	33.00 / 68.00
Teacher (upper ref.)	68.00 / 77.00	69.00 / 92.50	26.00 / 50.00	21.00 / 40.00	68.00 / 76.00

Intervention study. We conduct a controlled intervention study using a Qwen3-4B teacher and a Qwen3-1.7B student. From 1,000 student-visited prefixes, we compute G_t and $\widehat{D}_t$ and independently apply a single reverse-KL update at each state. Starting from the same pre-update parameters θ_0, this intervention produces updated parameters θ_1. We then sample $M = 8$ continuations before and after the intervention and measure downstream distillation loss over a horizon of $H = 4096$ tokens. For $j \in \{0, 1\}$, define

$$C_j(s_t) = \frac{1}{M} \sum_{m=1}^{M} \sum_{k=1}^{H} \gamma^k D_{\mathrm{KL}}\Big(p_{\theta_j}(\cdot \mid s_{t+k}^{(m,j)}) \,\|\, q(\cdot \mid s_{t+k}^{(m,j)})\Big), \quad \widehat{\Delta}_t^{\,\mathrm{down}} = C_0(s_t) - C_1(s_t). \tag{14}$$

Positive $\widehat{\Delta}_t^{\,\mathrm{down}}$ therefore indicates that supervising s_t improves the distillation loss over subsequently visited states. To isolate information beyond immediate learning opportunity, we compare this measured outcome across $\widehat{D}_t$ groups within matched G_t strata.



(a) Downstream response by local value. (b) State-selection ablation.

Figure 3: **Mechanism validation of the prospective supervision signal.** (a) evaluates whether the prospective score separates the measured consequences of locally comparable updates. (b) compares PROW with local-only selection and a within-G_t shuffled-$\widehat{D}_t$ control.

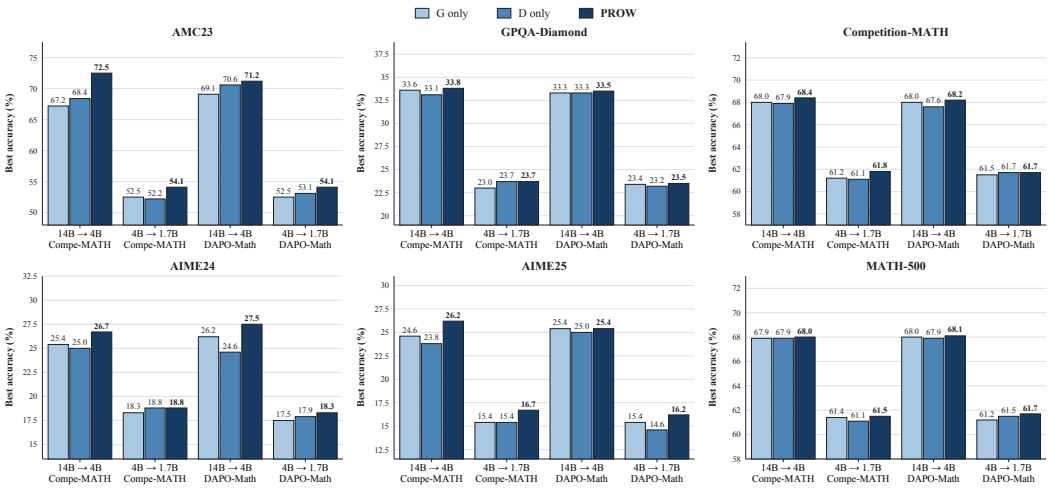


Figure 4: **Immediate and prospective supervision provide complementary signals.** We compare local-only, prospective-only, and combined PROW weighting across model scales, training corpora, and evaluation benchmarks.

Figure 3(a) shows little separation by $\widehat{D}_t$ at low G_t, but a clear ordering as local opportunity increases. At high G_t, the measured downstream gain is -3.6, 3.2, and 10.6 for $\widehat{D}_t < 0$, $\widehat{D}_t = 0$, and $\widehat{D}_t > 0$, respectively. Thus, states with similar immediate learning opportunity can induce substantially different, and even opposite, downstream consequences.

Figure 3(b) further shows that incorporating the prospective signal increases selected-state gain from 6.4 under G_t-only selection to 8.7, a paired improvement of 2.3 (95% CI: $[0.8, 3.8]$). Shuffling $\widehat{D}_t$ within matched-G_t strata reduces the gain to 6.6, indicating that the improvement depends on state-specific prospective information rather than its marginal distribution. Together, these results show that $\widehat{D}_t$ captures downstream heterogeneity not explained by G_t alone. Appendix G.3 further shows that the prospective ranking is largely distinct from that induced by TA-OPD on the same states.

Component ablation. We next test whether this additional information improves supervision allocation during full training. Figure 4 compares local-only, prospective-only, and combined PROW weighting across six benchmarks, two teacher–student configurations, and two training corpora. PROW obtains the highest or tied-highest point estimate in all 24 settings, including 20 strict improvements. Neither component consistently dominates the other, whereas their combination provides the strongest and most consistent performance. This supports the intended decomposition of

Table 2: Cross-domain generalization across teacher–student configurations. All models in this table are trained on DAPO-Math-17K. We report **Avg@8**, **Pass@k** for GPQA-Diamond, **Pass@k** for MMLU-Pro, where $k = 8$ for GPQA-Diamond and $k = 1$ for MMLU-Pro. Best results within each setting are boldfaced and second-best results are underlined.

Teacher → Student	Metric	GPQA-Diamond					MMLU-Pro				
		Base	OPD	IW-OPD	TA-OPD	PROW	Base	OPD	IW-OPD	TA-OPD	PROW
Qwen3-4B → Qwen3-1.7B	*Avg@8*	20.00	23.20	23.40	23.90	**25.60**	–	–	–	–	–
	Pass@k	64.00	65.70	<u>67.60</u>	67.30	**69.70**	36.80	46.00	<u>46.30</u>	43.20	**46.50**
Qwen3-8B → Qwen3-4B-Base	*Avg@8*	5.00	<u>25.80</u>	24.20	**27.50**	**27.50**	–	–	–	–	–
	Pass@k	35.00	76.80	72.20	<u>78.30</u>	**80.80**	50.20	57.40	<u>57.90</u>	**58.30**	**58.30**
Qwen3-14B → Qwen3-4B	*Avg@8*	30.00	30.60	32.60	<u>33.00</u>	**33.50**	–	–	–	–	–
	Pass@k	60.00	<u>65.20</u>	63.10	64.60	**65.70**	58.00	61.70	<u>62.20</u>	61.80	**62.50**

PROW: G_t identifies states with strong immediate learning opportunity, while $\widehat{D}_t$ further distinguishes them according to their downstream consequences. Appendix G.2 shows that the same prospective ordering is preserved when $\widehat{D}_t$ is estimated from a single continuation rather than a multi-rollout reference. Hyperparameter sensitivity is reported in Appendix G.1.

5.4 Cross-Domain Generalization

RQ3: Cross-domain generalization. We test whether PROW transfers beyond the mathematical reasoning benchmarks in Table 1. We evaluate on GPQA-Diamond (Rein et al., 2024) and MMLU-Pro (Wang et al., 2024); neither benchmark is used as a distillation corpus.

Results. As shown in Table 2, PROW achieves the highest or tied-highest GPQA-Diamond Avg@8 and Pass@8 across all three reported DAPO-Math-17K settings. On MMLU-Pro, PROW is highest in two settings and tied for the highest result in the remaining setting. These results indicate that the supervision allocation learned from mathematical reasoning trajectories can also improve performance on evaluation domains not used for distillation.

6 Discussion

Our results suggest that supervision allocation in on-policy distillation should account for downstream consequences, not only local learning value. Because the student generates its own future prefixes, a correction at one state can change not only the current prediction, but also the supervision encountered later along the trajectory. Our intervention results show that states with similar immediate gains can nevertheless produce substantially different downstream effects, highlighting a limitation of purely pointwise allocation criteria.

PROW addresses this issue by combining immediate learning opportunity with a prospective signal that estimates how a teacher-directed correction reshapes future supervision. The two signals capture complementary aspects of supervision value: the local term reflects how useful the current correction is, while the prospective term distinguishes states according to the downstream opportunities made more or less reachable by that correction. The ablation results support the use of both components rather than relying on either signal alone.

PROW requires no additional rollouts, reward models, or learned critics, and can therefore be integrated into standard OPD pipelines with limited changes. Its prospective estimate, however, is computed from a single sampled rollout and restricted Top-K supports, which may introduce estimation noise and omit low-probability actions outside the retained support. Although our experiments cover multiple model scales and model families, broader validation at larger scales, with more diverse architectures, and on open-ended generation domains remains an important direction for future work.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, improving the clarity and readability of the manuscript, and providing limited assistance with code development and debugging. All AI-assisted text, code, and suggestions were reviewed and verified by the authors. The research methodology, experimental design, analysis, interpretation of results, and scientific conclusions were determined by the authors. The authors take full responsibility for the final content of this work, including any material produced with the assistance of generative AI.

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.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Chishui Chen, Yaoyou Fan, Te Sun, Yi Yang, Chenghao Sun, Delin Mao, Hongbo Qiao, Zuowei Zhang, Junxi Wang, Chenxing Sun, Yangen Hu, Lu Pan, Xuyang Liu, and Linfeng Zhang. Look ahead before you distill: Future trajectory validation of teacher guidance for agentic on-policy distillation. *arXiv preprint arXiv:2608.01953*, 2026. URL https://arxiv.org/abs/2608.01953.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8ac015d409635f196f9e3e9dcfb9a94e-Abstract-Conference.html.

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 *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

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

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

Gyeongman Kim, Doohyuk Jang, and Eunho Yang. PromptKD: Distilling student-friendly knowledge for generative language models via prompt tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6266–6282. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.364. URL https://aclanthology.org/2024.findings-emnlp.364/.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

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

Enhan Li, Junhao He, and Hongyang Du. CROP: Task relevance via counterfactuals for selective on-policy distillation. *arXiv preprint arXiv:2608.13387*, 2026a. URL https://arxiv.org/abs/2608.13387.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026b. 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. *arXiv preprint arXiv:2305.20050*, 2023. URL https://arxiv.org/abs/2305.20050.

Yanjiang Liu, Jie Lou, Xinyan Guan, Yuqiu Ji, Hongyu Lin, Ben He, Xianpei Han, Le Sun, Xing Yu, and Yaojie Lu. Your teacher can't help you here: Combating supervision fidelity decay in on-policy distillation. *arXiv preprint arXiv:2605.30833*, 2026. URL https://arxiv.org/abs/2605.30833.

Mathematical Association of America. American mathematics competitions. MAA Student Programs, 2026. URL https://maa.org/student-programs/amc/. Accessed 2026.

Meta AI. Llama-3.2-3B-Instruct model card. Hugging Face, 2024. URL https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct. Accessed 2026.

Qwen Team. Qwen3.5-9B model card. Hugging Face, 2026. URL https://huggingface.co/Qwen/Qwen3.5-9B. Accessed 2026.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635. PMLR, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

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

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3018. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract.html.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. f-divergence minimization for sequence-level knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10817–10834. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.605. URL https://aclanthology.org/2023.acl-long.605/.

Yan Xie, Sijie Zhu, Tiansheng Wen, Bo Chen, and Yifei Wang. On the position bias of on-policy distillation. *arXiv preprint arXiv:2606.22600*, 2026. URL https://arxiv.org/abs/2606.22600.

Xingrun Xing, Haoqing Wang, Boyan Gao, Ziheng Li, and Yehui Tang. Trust region on-policy distillation. *arXiv preprint arXiv:2606.01249*, 2026. URL https://arxiv.org/abs/2606.01249.

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

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. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

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

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Gao, 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, Weinan Dai, 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. *arXiv preprint arXiv:2503.14476*, 2025. URL https://arxiv.org/abs/2503.14476.

Qihuang Zhong, Liang Ding, Li Shen, Juhua Liu, Bo Du, and Dacheng Tao. Revisiting knowledge distillation for autoregressive language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10900–10913. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.587. URL https://aclanthology.org/2024.acl-long.587/.

A Related Work

Knowledge distillation for autoregressive language models. Knowledge distillation (KD) transfers predictive behavior from a teacher to a smaller student, originally through matching soft output distributions (Hinton et al., 2015). For autoregressive models, sequence-level distillation trains the student on teacher-generated sequences (Kim & Rush, 2016), while subsequent work has studied alternative divergence objectives and decompositions tailored to generative modeling (Wen et al., 2023; Gu et al., 2024). Recent LLM distillation methods further improve the objective, data mixture, or teacher signal, including skew-divergence and adaptive off-policy training (Ko et al., 2024), student-friendly teacher adaptation (Kim et al., 2024), and token-dependent teaching strategies (Zhong et al., 2024). These approaches primarily concern what distribution should be distilled or how teacher knowledge should be presented. PROW is orthogonal to these choices: it retains the underlying OPD loss and instead studies how supervision should be allocated across states visited by the student.

On-policy distillation. A central difficulty in autoregressive learning is the mismatch between training prefixes and those induced by the model at generation time, closely related to exposure bias in sequential prediction (Bengio et al., 2015; Ross et al., 2011). On-policy distillation addresses this mismatch by evaluating teacher supervision on student-generated trajectories. MiniLLM develops reverse-KL distillation for generative LMs (Gu et al., 2024), while GKD provides a general framework that combines student-generated data with alternative distillation divergences (Agarwal et al., 2024). Subsequent analyses have examined when OPD succeeds, emphasizing teacher–student compatibility and the concentration of useful supervision on shared high-probability tokens (Li et al., 2026b). Other work improves reliability under large student–teacher mismatch through trust-region or hybrid on/off-policy supervision (Xing et al., 2026), or reduces long-horizon cost by detecting prefix drift and truncating unreliable continuations (Yang et al., 2026). On-policy self-distillation further replaces the external teacher with a privileged view of the student model, extending the same student-trajectory principle to self-improvement settings. Our focus is complementary: given a standard OPD rollout and teacher, we ask how the value of supervision at one visited state depends on the supervision that the same correction makes reachable later.

Selective and token-weighted supervision in OPD. Recent work has established that dense OPD supervision is highly non-uniform across token positions. Token-importance analyses use student uncertainty and teacher–student disagreement to identify informative positions (Xu et al., 2026). IW-OPD attributes the degradation of later-token supervision to accumulated student–teacher discrepancy and derives position-dependent importance weights from this prefix-level mismatch (Xie et al., 2026). TA-OPD instead distinguishes learnable from incompatible disagreement through local student–teacher support compatibility and selects positions according to token teachability (Wang et al., 2026). Related analyses emphasize shared-support alignment and the dependence of OPD on student–teacher compatibility (Li et al., 2026b), while CROP adds task relevance through counterfactual sensitivity to the input (Li et al., 2026a). These criteria capture complementary notions of uncertainty, discrepancy, compatibility, or relevance, but assess a state through information available at that state, accumulated before it, or obtained from alternative contexts. PROW addresses a different axis: whether learning at the current state changes access to useful supervision downstream.

Trajectory-aware and lookahead supervision. Several recent methods move beyond strictly local token diagnostics by incorporating information from nearby or longer-horizon trajectory structure. Supervision Fidelity Decay characterizes the deterioration of teacher supervision along long student trajectories and introduces a lookahead signal that evaluates candidate tokens through the teacher confidence induced at the next step (Liu et al., 2026). Trajectory-aware OPD uses near-future information to distinguish genuine reasoning divergences from local surface-form mismatches and distributes guidance across short future segments (Jiang & Ferraro, 2026). Prune-OPD similarly uses prefix-level compatibility to detect trajectory drift and reduce supervision after unreliable regions (Yang et al., 2026). These works demonstrate that trajectory structure can provide information not available from an isolated token. PROW takes a complementary view of nonlocal supervision: rather than using future context only to assess the current signal, it asks how the teacher-directed correction itself changes which future learning opportunities become reachable. This leads to a prospective score that couples downstream learning opportunity with the first-order change in action reachability under the same correction.

Relation to sequential credit assignment. The dependence of future states on current predictions is fundamental to sequential decision making and imitation learning (Ross et al., 2011; Bengio et al., 2015). PROW shares this concern with state-distribution shift, but its objective is different from reward-based temporal credit assignment. The downstream quantity is defined entirely from teacher–student supervision along the OPD trajectory, rather than from task rewards or a learned value function. Accordingly, PROW can be viewed as introducing a prospective notion of supervision value within distillation: a state is valuable not only for the correction available there, but also for how that correction redistributes access to subsequent teacher supervision.

B Mathematical Derivations

B.1 Teacher-Directed Path and Local Learning Opportunity

All quantities in this derivation are restricted to the student Top-K support $\mathcal{A}_t$ from Eq. (4). For compactness, write

$$p(a) = p_t^{\mathcal{A}}(a), \qquad q(a) = q_t^{\mathcal{A}}(a), \qquad \ell(a) = \log \frac{q(a)}{p(a)}.$$

Define the normalized teacher-directed path

$$p_\eta(a) = \frac{p(a)e^{\eta \ell(a)}}{Z(\eta)}, \qquad Z(\eta) = \sum_{b \in \mathcal{A}_t} p(b)e^{\eta \ell(b)}. \tag{15}$$

Because $Z(0) = Z(1) = 1$, we have $p_0 = p$ and $p_1 = q$. Let $\mu(\eta) = \mathbb{E}_{p_\eta}[\ell]$. Standard log-partition identities give

$$\frac{d}{d\eta} \log Z(\eta) = \mu(\eta), \qquad \mu'(\eta) = \mathrm{Var}_{p_\eta}[\ell]. \tag{16}$$

Moreover,

$$\log \frac{p_\eta(a)}{q(a)} = (\eta - 1)\ell(a) - \log Z(\eta),$$

so

$$D_{\mathrm{KL}}(p_\eta \| q) = (\eta - 1)\mu(\eta) - \log Z(\eta). \tag{17}$$

Differentiating and using Eq. (16),

$$\frac{d}{d\eta} D_{\mathrm{KL}}(p_\eta \| q) = (\eta - 1)\,\mathrm{Var}_{p_\eta}[\ell]. \tag{18}$$

Therefore, at the current student policy,

$$-\left.\frac{d}{d\eta} D_{\mathrm{KL}}(p_\eta \| q)\right|_{\eta=0} = \mathrm{Var}_{a \sim p_t^{\mathcal{A}}}\left[\log \frac{q_t^{\mathcal{A}}(a)}{p_t^{\mathcal{A}}(a)}\right] = G_t, \tag{19}$$

which is Eq. (10). In particular, $G_t \geq 0$.

B.2 Top-K Shared Continuation Mass

Fix a prefix s and let $\mathcal{B} = \mathrm{TopK}_K(p) \cup \mathrm{TopK}_K(q)$. For an action a sampled from the student, define

$$w(a) = \mathbf{1}\{a \in \mathcal{B}\} \min\left\{1, \frac{q(a)}{p(a)}\right\}. \tag{20}$$

Then

$$p(a)w(a) = \mathbf{1}\{a \in \mathcal{B}\} \min\{p(a), q(a)\}. \tag{21}$$

Thus the factor used in the main text is exactly the student-sampling acceptance probability associated with the student–teacher common mass retained on the Top-K union. Summing Eq. (21) over the retained support gives

$$\mathbb{E}_{A \sim p}[w(A)] = \sum_{a \in \mathcal{B}} \min\{p(a), q(a)\} \leq 1. \tag{22}$$

Hence the continuation factor is computed entirely from the Top-K union; no full-vocabulary common-mass quantity is needed.

B.3 Directional Derivation of the Prospective Effect

We prove Proposition 1 under the same teacher-directed perturbation used for the immediate term. For brevity, write

$$p(a) = p_t^{\mathcal{B}}(a), \qquad q(a) = q_t^{\mathcal{B}}(a), \qquad r(a) = \log \frac{q(a)}{p(a)}.$$

On the fixed support $\mathcal{B}_t$, the perturbation is

$$p_\eta(a) = \frac{p(a)e^{\eta r(a)}}{\mathcal{Z}(\eta)}, \qquad \mathcal{Z}(\eta) = \sum_{b \in \mathcal{B}_t} p(b)e^{\eta r(b)}. \tag{23}$$

Its log-probability derivative is

$$\frac{\partial}{\partial \eta} \log p_\eta(a) = r(a) - \mathbb{E}_{b \sim p_\eta}[r(b)]. \tag{24}$$

At $\eta = 0$, since $p_0 = p_t^{\mathcal{B}}$,

$$\frac{\partial}{\partial \eta} \log p_\eta(a)\Big|_{\eta=0} = r_t(a) - \bar{r}_t. \tag{25}$$

Now consider the shared-mass-weighted downstream opportunity

$$\Phi_t(\eta) = \gamma Z_t^{\mathcal{B}} \sum_{a \in \mathcal{B}_t} p_\eta(a)\, w_t(a)\, U_t(a), \qquad Z_t^{\mathcal{B}} = \sum_{a \in \mathcal{B}_t} p_t(a). \tag{26}$$

Holding $w_t(a)$ and $U_t(a)$ fixed under the local perturbation and using $\partial_\eta p_\eta(a) = p_\eta(a)\partial_\eta \log p_\eta(a)$, we obtain

$$\begin{aligned}
\frac{d}{d\eta}\Phi_t(\eta)\Big|_{\eta=0} &= \gamma Z_t^{\mathcal{B}} \sum_{a \in \mathcal{B}_t} p_t^{\mathcal{B}}(a)\, w_t(a) \big(r_t(a) - \bar{r}_t\big)\, U_t(a) \\
&= \gamma \sum_{a \in \mathcal{B}_t} p_t(a)\, w_t(a) \big(r_t(a) - \bar{r}_t\big)\, U_t(a) \\
&= \gamma \sum_{a \in \mathcal{B}_t} \min\{p_t(a), q_t(a)\} \big(r_t(a) - \bar{r}_t\big)\, U_t(a) = D_t.
\end{aligned} \tag{27}$$

The second equality uses $Z_t^{\mathcal{B}} p_t^{\mathcal{B}}(a) = p_t(a)$ on $\mathcal{B}_t$, while the third uses $p_t(a)w_t(a) = \min\{p_t(a), q_t(a)\}$. This proves Proposition 1.

Single-rollout form. Direct evaluation of D_t requires $U_t(a)$ for multiple actions and continuations. Under the standard student rollout, we replace $U_t(y_t)$ by its single-trajectory estimate $\widehat{U}_t(y_t)$, yielding

$$\widehat{D}_t = \gamma w_t\big(r_t(y_t) - \bar{r}_t\big)\widehat{U}_t(y_t). \tag{28}$$

When $y_t \notin \mathcal{B}_t$, we set $\widehat{D}_t = 0$.

C Budgeted Allocation

Let $S = (S_1, \ldots, S_N)$ denote the supervision scores in one optimizer update and set $\pi_i = v_i/N$. Equation (13) is equivalent to

$$\begin{aligned}
\max_{\pi \in \Delta_N} \quad & \sum_{i=1}^{N} \pi_i S_i \\
\text{s.t.} \quad & D_{\mathrm{KL}}(\pi \| u_N) \le \epsilon, \qquad \ell \le \pi_i \le u,
\end{aligned} \tag{29}$$

where $\ell = v_{\min}/N$ and $u = v_{\max}/N$.

For an active KL constraint, let $\tau > 0$ be its multiplier, ν the normalization multiplier, and let $\alpha_i, \beta_i \geq 0$ be the lower- and upper-bound multipliers. The KKT stationarity condition is

$$S_i - \tau\big(\log(N\pi_i) + 1\big) + \nu + \alpha_i - \beta_i = 0. \tag{30}$$

For every coordinate strictly inside the box, complementary slackness gives $\alpha_i = \beta_i = 0$, hence

$$\pi_i = C \exp(S_i/\tau)$$

for a common normalizing constant $C > 0$. Coordinates whose unconstrained value lies outside the box are fixed at the corresponding boundary. Thus the solution has the clipped Gibbs form

$$\boxed{\pi_i^\star = [C \exp(S_i/\tau)]_\ell^u}, \qquad [z]_\ell^u \coloneqq \min\{u, \max\{\ell, z\}\}. \tag{31}$$

The constants C and τ are chosen so that $\sum_i \pi_i^\star = 1$ and, when the KL constraint is active, $D_{\mathrm{KL}}(\pi^\star \| u_N) = \epsilon$. The allocation used in training is then $v_i^\star = N\pi_i^\star$. Without active box constraints, Eq. (31) reduces to the usual Gibbs allocation $\pi_i^\star \propto \exp(S_i/\tau)$.

D Why the Downstream Opportunity Is Root-Centered

For a fixed sampled suffix, define the nonnegative coefficient

$$\alpha_{t,k} \coloneqq \gamma^{k-1} \prod_{j=t+1}^{t+k-1} w_j, \qquad k = 1, \ldots, T-t. \tag{32}$$

The corresponding downstream opportunity can then be written as

$$H_t = \sum_{k=1}^{T-t} \alpha_{t,k}\big(G_{t+k} - G_t\big). \tag{33}$$

The subtraction of G_t is not an additional fitted term; it is forced by a natural invariance of a relative future-opportunity score.

Proposition 2 (Characterization of the root-centered contrast)**.** *Fix the coefficients $\{\alpha_{t,k}\}_{k=1}^{T-t}$ and consider any linear summary*

$$C_t = b_t G_t + \sum_{k=1}^{T-t} \alpha_{t,k} G_{t+k}, \tag{34}$$

where only the coefficient b_t of the current opportunity remains to be chosen. The following are equivalent:

1. *C_t is unchanged when the same constant is added to every current and future G value;*
2. *$C_t = 0$ whenever all current and future opportunities are equal;*
3. *$b_t = -\sum_{k=1}^{T-t} \alpha_{t,k}$.*

Under any of these conditions, the unique resulting linear summary is $C_t = H_t$ in Eq. (33)*.*

Proof. Adding a common constant c to all G values changes Eq. (34) by

$$c\left(b_t + \sum_{k=1}^{T-t} \alpha_{t,k}\right).$$

This vanishes for every c if and only if $b_t = -\sum_k \alpha_{t,k}$. The same condition is necessary and sufficient for C_t to vanish when all opportunities take a common value. Substituting this coefficient into Eq. (34) gives exactly Eq. (33). □

At $t = T$, the sum is empty and $H_T = 0$ by definition. The proposition only characterizes linear summaries with the fixed downstream coefficients above; it does not impose a claim on arbitrary nonlinear notions of future utility.

Table 3: Core training and evaluation settings used in our experiments.

Setting	Value
Optimizer	AdamW
Learning rate	5×10^{-6}
Max. prompt length	1,024
Max. training response length	4,096
Global prompt batch size	64
Training rollouts per prompt	4
PPO minibatch size	64 trajectories
Microbatch size / GPU	16 trajectories
Top-K	16
Competition-MATH epochs	3
DAPO-Math-17K epochs	2
PROW λ	1
Discount factor γ	0.995
Allocation KL budget ϵ	0.1
Training sampling	$T = 1.0$, top-$p = 1.0$
Evaluation sampling	$T = 0.7$, top-$p = 0.95$, 8 responses

E Experimental Details

Training setup. All experiments use the same on-policy distillation pipeline described in Section 5.1. We optimize the student with AdamW using a learning rate of 5×10^{-6}, Adam coefficients $(\beta_1, \beta_2) = (0.9, 0.95)$, zero weight decay, and a maximum gradient norm of 1.0. Models are trained in bfloat16 with fully sharded data parallelism (FSDP) and gradient checkpointing.

The maximum prompt length is 1,024 tokens. Prompts exceeding this limit after template rendering are filtered rather than truncated. Training responses are limited to 4,096 tokens. For PROW, OPD, and TA-OPD, the global prompt batch size is 64, with four student rollouts sampled per prompt. We use a PPO minibatch size of 64 trajectories and a per-GPU microbatch size of 16. Training rollouts are sampled with temperature 1.0 and top-$p = 1.0$.

Hardware. All experiments were run on $48\times$ NVIDIA B200 GPUs. We use distributed training with FSDP across the available devices.

Training data and duration. Competition-MATH uses its training split for distillation and its disjoint test split for evaluation. We train for three epochs on Competition-MATH and two epochs on DAPO-Math-17K. The additional model configurations in Appendix F are trained only on Competition-MATH and use the same training protocol.

PROW configuration. We use a Top-K size of $K = 16$ throughout the main experiments. The prospective coefficient and temporal discount are set to $\lambda = 1$ and $\gamma = 0.995$, respectively. The bounded allocation constrains the token weights to $[0.5, 2.0]$ and uses a KL budget of $\epsilon = 0.1$. Before allocation, the prospective term is stabilized using the implementation's $\tanh$ transformation with scale determined by the 0.99 quantile of its absolute values. Unless stated otherwise, these values are fixed across all teacher–student configurations and training corpora.

Evaluation protocol. For each evaluation problem, we sample eight responses with temperature 0.7 and top-$p = 0.95$. The same decoding settings are used for all compared methods. Training and rollout generation use seed 42.

F Results on Other Models

We further evaluate PROW on additional teacher–student configurations: Qwen3-8B $\rightarrow$ Qwen3-1.7B-Base (Yang et al., 2025), Qwen3.5-9B $\rightarrow$ Qwen3.5-4B (Qwen Team, 2026), and Llama-3.1-8B-Instruct $\rightarrow$ Llama-3.2-3B-Instruct (Grattafiori et al., 2024; Meta AI, 2024). All configurations are trained only on the Competition-MATH training split and evaluated on Competition-MATH,

Table 4: Results on additional teacher–student configurations trained on Competition-MATH. Each entry reports **Avg@8 / Pass@8** (%). Best results within each model pair are boldfaced and second-best results are underlined.

Setting	Dataset	Method	Compe-MATH	AMC23	AIME24	AIME25	MATH-500	GPQA-Diamond
Qwen3-8B → Qwen3-1.7B-Base	Compe-MATH	OPD	**55.30** / 70.40	38.13 / 77.50	11.67 / 23.34	8.75 / **23.34**	**55.10** / 68.20	19.80 / 69.00
		IW-OPD	53.30 / 70.00	34.07 / 75.00	7.50 / **26.67**	7.09 / 20.00	52.00 / 69.00	16.00 / 71.00
		TA-OPD	54.00 / **71.00**	39.38 / 72.50	10.84 / **26.67**	8.75 / **23.34**	53.80 / 69.00	19.40 / 68.70
		PROW	55.10 / 70.60	**40.94 / 80.00**	**14.17 / 26.67**	**9.17 / 23.34**	**55.10 / 70.80**	**23.20 / 73.70**
Qwen3.5-9B → Qwen3.5-4B	Compe-MATH	OPD	72.90 / 78.00	85.90 / **100.00**	**50.40 / 76.70**	37.50 / 60.00	73.00 / 77.60	47.50 / 86.90
		IW-OPD	73.00 / 77.40	86.60 / **100.00**	47.10 / 73.30	37.50 / 56.70	73.00 / 77.40	47.70 / 86.30
		TA-OPD	73.20 / 77.20	**87.20 / 100.00**	48.30 / 73.30	33.30 / 56.70	73.10 / **78.20**	48.20 / **90.40**
		PROW	**74.00 / 78.40**	**87.20 / 100.00**	**50.40 / 76.70**	**38.30 / 66.70**	**73.60 / 78.20**	**49.50** / 89.40
Llama-3.1-8B-Instruct → Llama-3.2-3B-Instruct	Compe-MATH	OPD	34.70 / 57.60	21.90 / 60.00	3.70 / 16.70	0.80 / **6.70**	33.70 / 58.40	19.50 / 67.20
		IW-OPD	**35.20** / 60.00	22.20 / 57.50	**5.00** / 16.70	1.70 / **6.70**	34.90 / 60.00	21.10 / 69.00
		TA-OPD	34.80 / 60.00	23.40 / **65.00**	3.70 / **20.00**	**2.10 / 6.70**	34.50 / **60.40**	21.10 / 68.20
		PROW	**35.20 / 61.20**	**24.10** / 62.50	4.60 / 16.70	1.70 / **6.70**	**35.60 / 60.40**	**21.40 / 71.20**

MATH-500, AMC 2023, AIME 2024, AIME 2025, and GPQA-Diamond. We use the same Avg@8 and Pass@8 metrics and evaluation protocol as in the main experiments.

The additional configurations show that the behavior of PROW is not restricted to the model scales used in the main comparison. On Qwen3-8B → Qwen3-1.7B-Base, PROW provides particularly clear improvements on AMC 2023, AIME 2024, and GPQA-Diamond. On Qwen3.5, PROW achieves the highest or tied-highest Avg@8 on all six benchmarks. The Llama results are more mixed, but PROW remains best/tied-best Avg@8 on four of six benchmarks và best/tied-best Pass@8 on four of six. These results suggest that the prospective signal transfers across student scales and model families.

G Additional Experiments

G.1 Hyperparameter Ablations

Ablation setup. We conduct one-dimensional sweeps over four PROW hyperparameters using Qwen3-4B → Qwen3-1.7B trained on Competition-MATH. The coefficient λ weights the prospective term in the supervision score,

$$S_t = G_t + \lambda \widehat{D}_t, \tag{35}$$

with $\lambda = 0$ reducing the score to the local component G_t and $\lambda = 1$ retaining the full downstream contribution. The support size K sets the size of the local support $\mathcal{A}_t = \mathrm{TopK}_K(p_t)$ and the reachability support $\mathcal{B}_t = \mathrm{TopK}_K(p_t) \cup \mathrm{TopK}_K(q_t)$. The KL budget ϵ controls the admissible deviation from uniform allocation through $D_{\mathrm{KL}}(\pi \| u_N) \le \epsilon$. The discount factor γ controls the temporal horizon of the downstream return, with smaller values attenuating contributions from more distant successor states.

Results. The λ sweep directly measures the effect of the prospective component. Compared with $\lambda = 0.1$, using $\lambda = 1$ raises the Avg@8 from 60.95 to 61.83 on Competition-MATH, from 61.48 to 61.73 on MATH-500, from 15.42 to 16.25 on AIME25, and from 23.04 to 23.61 on GPQA-Diamond. AIME24 reaches its highest value at $\lambda = 0.25$, whereas AMC23 is highest at $\lambda = 0.1$. The benefit of the downstream term is therefore visible on several benchmarks, although its optimal scaling is not universal.

No monotonic trend is observed for the support size K. The best results occur at $K = 16$ on Competition-MATH, GPQA-Diamond, and AMC23, at $K = 48$ on MATH-500, and at $K = 32$ on AIME24. The allocation-budget sweep favors $\epsilon = 0.10$ on Competition-MATH, MATH-500, GPQA-Diamond, and AMC23, while smaller budgets are better on the two AIME benchmarks. The clearest pattern is obtained for γ: $\gamma = 0.995$ exceeds the undiscounted setting $\gamma = 1$ on all six benchmarks and is the best setting on MATH-500, AIME24, AIME25, and AMC23. Within this

Table 5: Hyperparameter sensitivity for Qwen3-4B → Qwen3-1.7B trained on Competition-MATH. Each cell reports Avg@8 (%) at the shared evaluation step used for the sweep. Boldface denotes the highest value within each sweep.

(a) Downstream-score coefficient λ

Benchmark	0.1	0.25	0.5	1.0
Compe-MATH	60.95	61.60	61.43	**61.83**
MATH-500	61.48	61.33	61.38	**61.73**
AIME24	18.75	**20.00**	17.50	17.92
AIME25	15.42	**16.25**	**16.25**	**16.25**
AMC23	**52.50**	52.19	51.88	51.56
GPQA-Diamond	23.04	22.98	23.17	**23.61**

(b) Student-support size K

Benchmark	8	16	32	48	64
Compe-MATH	61.15	**61.35**	61.08	61.15	61.20
MATH-500	61.10	61.35	61.13	**61.40**	61.35
AIME24	17.92	18.33	**19.58**	18.75	17.50
AIME25	**17.50**	16.67	15.42	16.25	15.83
AMC23	50.94	**52.19**	50.63	51.56	51.25
GPQA-Diamond	22.85	**23.67**	22.92	23.17	22.92

(c) Allocation KL budget ϵ

Benchmark	0.01	0.02	0.05	0.10
Compe-MATH	61.03	61.10	60.65	**61.15**
MATH-500	60.88	60.85	60.93	**61.35**
AIME24	**17.92**	**17.92**	16.67	16.67
AIME25	15.00	13.75	**17.50**	13.75
AMC23	51.88	49.06	49.69	**55.00**
GPQA-Diamond	22.35	22.35	22.66	**22.73**

(d) Downstream discount factor γ

Benchmark	0.95	0.99	0.995	1.0
Compe-MATH	60.55	**61.05**	60.68	60.23
MATH-500	60.70	60.65	**61.08**	60.53
AIME24	15.42	15.83	**17.92**	15.83
AIME25	14.17	14.17	**15.42**	15.00
AMC23	48.44	47.81	**49.69**	48.44
GPQA-Diamond	**23.61**	22.73	22.35	22.03

sweep, a small amount of temporal discounting is therefore preferable to an undiscounted downstream return.

G.2 Reliability of Single-Rollout Prospective Scoring

PROW estimates prospective value from a single sampled continuation. We examine how closely this estimate preserves the ranking obtained from multiple continuations. Across 13 student checkpoints, we evaluate 128 Competition-MATH states per checkpoint and compare the single-rollout score $\widehat{D}_t^{(1)}$ with the 32-rollout reference $\overline{D}_t^{(32)} = \frac{1}{32}\sum_{r=1}^{32} \widehat{D}_t^{(r)}$. We evaluate both rank agreement and recovery of the top 20% of states under the reference supervision score $G_t + \overline{D}_t^{(32)}$.

Figure 5 shows strong agreement between single- and multi-rollout prospective estimates. The mean within-checkpoint Spearman correlation is 0.791 (95% CI: $[0.710, 0.859]$), increasing to 0.928 among states with nonzero single-rollout prospective scores, with 98.2% sign agreement. For state selection, $G_t + \widehat{D}_t^{(1)}$ recovers 76.9% of the reference top 20%, compared with 59.2% for G_t alone. The resulting 17.8 percentage-point improvement has a 95% CI of $[5.2, 31.7]$ and is positive at all 13 checkpoints. These results show that a single continuation provides a reliable ranking signal for prospective supervision and materially improves recovery of states preferred by the multi-rollout reference.

G.3 Prospective Supervision Beyond Local Teachability

We next examine whether the prospective component captures information beyond a local teachability criterion. Using the same student-visited states at each checkpoint, we compare the TA-OPD (Wang et al., 2026) score with the immediate component G_t, the single-rollout prospective component $\widehat{D}_t^{(1)}$, and its 32-rollout reference $\overline{D}_t^{(32)}$. Spearman correlations are computed within each checkpoint and averaged across the 13 checkpoints.

As shown in Figure 6, TA-OPD is strongly aligned with the immediate component G_t (mean $\rho = 0.776$), but much less correlated with the prospective components ($\rho = 0.177$ for $\widehat{D}_t^{(1)}$ and $\rho = 0.139$ for $\overline{D}_t^{(32)}$). The gap between TA–G and TA–$\overline{D}^{(32)}$ correlations is 0.637 (95% CI: $[0.399, 0.902]$) and remains positive at every checkpoint. These results indicate that prospective supervision induces a state ranking that is largely distinct from local teachability, supporting its role as a complementary signal rather than a reparameterization of existing local criteria.

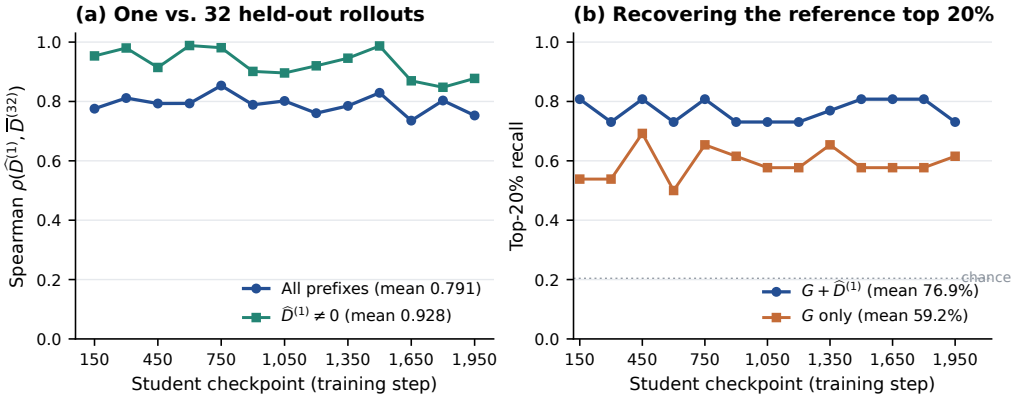


Figure 5: **Single-rollout prospective scores preserve multi-rollout rankings.** Across 13 student checkpoints, we compare the prospective score from one continuation, $\widehat{D}^{(1)}$, with a 32-rollout reference, $\overline{D}^{(32)}$. **(a)** Within-checkpoint Spearman agreement between the two estimates. **(b)** Recall of the reference top 20% under $G + \widehat{D}^{(1)}$ versus G alone. A single rollout preserves the prospective ordering and substantially improves recovery of states preferred by the multi-rollout reference.

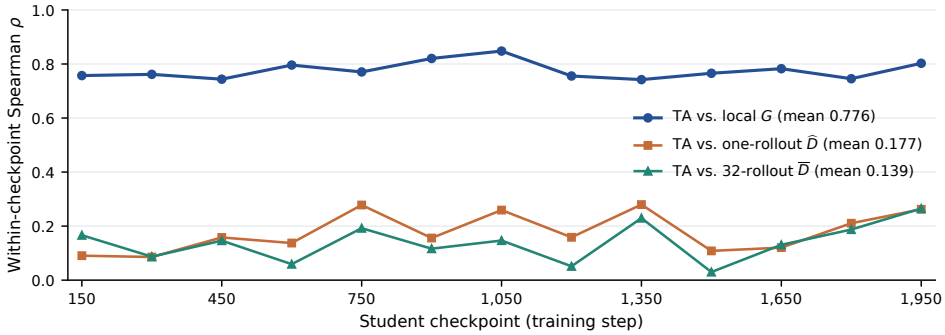


Figure 6: **Prospective supervision captures information distinct from local teachability.** We report within-checkpoint Spearman correlations between the TA-OPD score and the immediate component G, the single-rollout prospective component $\widehat{D}^{(1)}$, and the 32-rollout reference $\overline{D}^{(32)}$. TA-OPD closely tracks G but shows substantially weaker agreement with both prospective estimates, indicating that prospective supervision provides a complementary ranking signal.

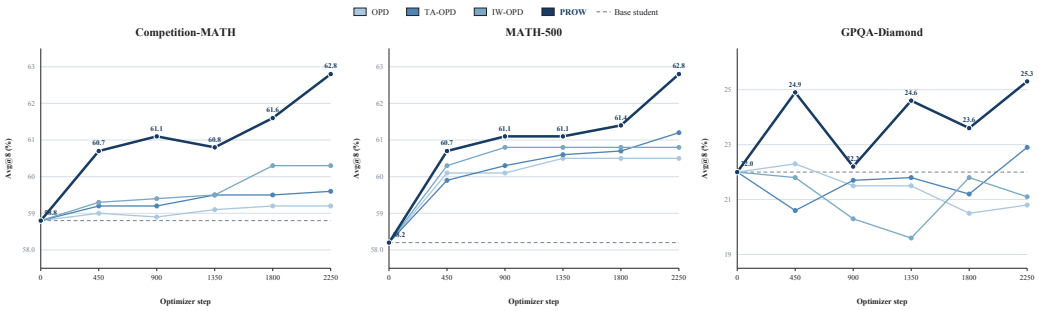


Figure 7: **Training dynamics of PROW and on-policy distillation baselines.** We report Avg@8 accuracy (%) at uniformly spaced checkpoints, evaluated every 450 optimizer steps, for the teacher (Qwen3-4B) and the student (Qwen3-1.7B) configuration trained on Competition-MATH. The dashed horizontal line denotes the shared student initialization, from which all methods start at step 0. Across all post-initialization checkpoints shown, PROW consistently outperforms OPD, TA-OPD, and IW-OPD on Competition-MATH, MATH-500, and the out-of-domain GPQA-Diamond benchmark. At 2,250 steps, PROW reaches 62.8% on both Competition-MATH and MATH-500 and 25.3% on GPQA-Diamond, demonstrating stronger in-domain learning and cross-domain generalization.