SELF-SUPERVISED ON-POLICY DISTILLATION FROM CONTRASTING REASONING TRAJECTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy RLVR trains reasoning models from groups of sampled attempts but assigns token updates using terminal rewards. A mixed group contains a richer signal: a correct attempt shows how the current policy can solve a prompt, while an incorrect attempt exposes prefixes it visits before failing. The two trajectories usually diverge, so the successful tokens cannot directly supervise the failed prefixes. We introduce *Self-Supervised On-Policy Distillation* (SSOPD), which uses a successful completion as privileged hindsight for a teacher at prefixes of a failed completion, then distills the teacher's local distribution into a student that sees only the prefix. This converts within-group outcome contrast into dense on-policy supervision without external solution traces. We connect the method to the policy's success-conditioned action distribution and show how the frontier weight tracks available correct–wrong pairs. Across AIME 2024, AIME 2025, and HMMT 2025, SSOPD improves over GRPO in all nine model-benchmark settings. On Qwen3-8B, it reaches a macro Avg@12 of 65.6, exceeding GRPO by 1.6 points and the solution-conditioned OPSD baseline by 0.8 points.

1 INTRODUCTION

Verifiable rewards have made it practical to improve reasoning models without annotating every intermediate step: sample several attempts, check their final answers, and reinforce successful completions (Wei et al., 2022; Cobbe et al., 2021; Hendrycks et al., 2021). GRPO-style RLVR methods (Shao et al., 2024; DeepSeek-AI et al., 2025; Yu et al., 2025) use these on-policy samples effectively, but the verifier's terminal outcome is still the only direct signal assigned to each trajectory. A group of attempts contains more structure than its list of rewards.

Consider a prompt for which the current policy produces both a correct and an incorrect completion. The correct attempt is evidence that the policy can solve this particular problem; the incorrect attempt reveals prefixes it actually visits on the way to failure. Together, they suggest a process-level update using only information already present in the on-policy group. The difficulty is that the two trajectories usually take different paths. Copying the successful tokens would supervise states on the successful path, not the failed prefixes where guidance is needed.

We introduce *Self-Supervised On-Policy Distillation* (SSOPD) to bridge these paths. At a prefix of a failed completion, a stop-gradient teacher receives a successful completion from the same sampled group as privileged hindsight. The student is trained at that same prefix without the hindsight, exactly as it will act at inference time. The teacher thereby transfers information from a successful branch to an on-policy failure state. No external solution trace or stronger model is needed to construct the witness.

The two sides of the pair face different budgets. The teacher must read a successful witness within a finite prompt context, while the student can learn from at most K prefixes of one failure. SSOPD therefore pairs the shortest correct completion with the longest wrong completion, then edits the failure's first K prefixes. A prompt-level frontier weight scales this update with the number of available correct–wrong pairs. Figure 1 illustrates the pipeline, and the selector ablation tests the pairing choice.

Our analysis starts from the policy's own success-conditioned next-token distribution, an exact process target implicit in binary outcome feedback. A local improvement bound separates the quality

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

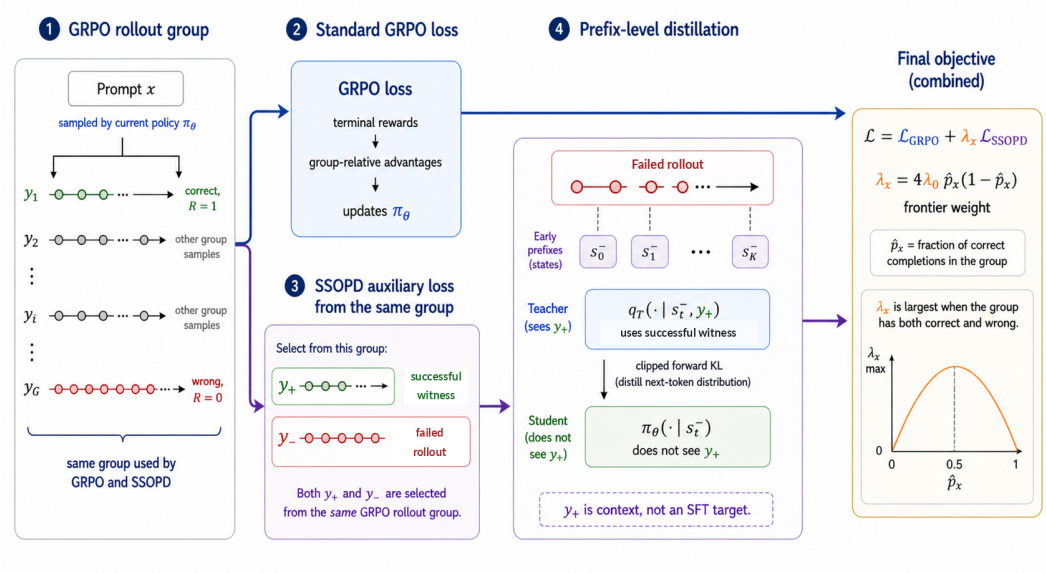

Figure 1: **SSOPD training pipeline.** A mixed on-policy group supplies a successful witness and a failed rollout. At prefixes of the failure, a teacher uses the witness as privileged hindsight to form local targets for a student that sees only the prompt and failed prefix. A frontier weight emphasizes groups containing both outcomes.

of the witness-conditioned teacher from the student's ability to learn its target. A group-level identity then connects frontier weighting to outcome-related branching. The length rule is the empirical choice used to instantiate this transfer.

We compare methods by the information available during training. SFT and OPSD use additional solution information; GRPO and SSOPD use only prompts, verified answers, and on-policy samples. SSOPD improves over GRPO in all nine model-benchmark settings across Qwen3-1.7B, 4B, and 8B on AIME 2024, AIME 2025, and HMMT 2025. At 8B, its macro Avg@12 reaches 65.6, exceeding GRPO by 1.6 points and the solution-conditioned OPSD baseline by 0.8 points.

Our contributions are:

- We turn correct–wrong contrast within an on-policy group into dense supervision at failed prefixes through self-generated privileged hindsight.
- We connect this transfer to the success-conditioned policy, give a local improvement condition, and show that the frontier weight counts usable within-group pairs.
- We show consistent gains over information-matched GRPO across three model scales and three math benchmarks, with competitive performance against solution-conditioned distillation at 8B.

2 Background

We now fix notation and review the two ingredients used by SSOPD. GRPO supplies on-policy groups of attempts together with terminal verifier rewards. On-policy and student-distribution-aware distillation provide a way to construct dense next-token targets at model-visited prefixes (Agarwal et al., 2024; Lu & Thinking Machines Lab, 2025; Xu et al., 2025; Zhao et al., 2026). SSOPD combines these ideas while changing the source of privileged context: the context comes from another completion in the same group rather than from an external solution trace.

We use one notation throughout the background and method. For prompt x, a completion is $y = (a_0, \ldots, a_{\tau(y)-1})$, its prefix state is $s_t(y) = (x, y_{<t})$, and the verifier returns $R(y) \in \{0, 1\}$. The behavior policy that samples rollouts is π, the optimized student is π_θ, and the stop-gradient or frozen teacher is $\pi_{\bar{\theta}}$.

2.1 Group Relative Policy Optimization

Group Relative Policy Optimization (GRPO) is an RLVR algorithm introduced with DeepSeek-Math (Shao et al., 2024). For each prompt x, it samples a group

$$\mathcal{G}(x) = \{y_i\}_{i=1}^{G}, \qquad y_i \sim \pi(\cdot \mid x),$$

and computes binary rewards $r_i = R(y_i)$. Let

$$\bar{r} = \frac{1}{G}\sum_{j=1}^{G} r_j, \qquad \sigma_r^2 = \frac{1}{G}\sum_{j=1}^{G}(r_j - \bar{r})^2.$$

GRPO uses the group-normalized advantage

$$A_i = \frac{r_i - \bar{r}}{\sigma_r + \epsilon_r}, \tag{1}$$

where ϵ_r is a small stabilizer. For token $a_{i,t}$ at prefix $s_t(y_i)$, define the policy ratio

$$\omega_{i,t}(\theta) = \frac{\pi_\theta(a_{i,t} \mid s_t(y_i))}{\pi(a_{i,t} \mid s_t(y_i))}.$$

The denominator is the behavior, or old, policy that generated the sampled group. Up to the usual sign convention for minimization versus maximization, the token-averaged GRPO loss is

$$\mathcal{L}_{\mathrm{GRPO}}(x) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{\tau(y_i)}\sum_{t=0}^{\tau(y_i)-1}\left(-\min[\omega_{i,t}(\theta)A_i, \mathrm{clip}(\omega_{i,t}(\theta), 1-\epsilon, 1+\epsilon)A_i] + \beta\,\mathcal{K}_{i,t}\right), \tag{2}$$

where $\mathcal{K}_{i,t}$ denotes the reference-policy KL penalty at the same prefix. GRPO is a compact group-based RLVR objective: the group mean and variance turn terminal verifier outcomes into relative advantages, and the clipped ratio controls the size of the policy update. This makes it a natural on-policy backbone for adding auxiliary process-level supervision.

2.2 Privileged On-Policy Self-Distillation

Privileged on-policy self-distillation (Zhao et al., 2026) keeps the student update at an on-policy prefix while constructing a dense teacher distribution from additional training-time context. At a prefix state s, let z denote the privileged information. The teacher and student next-action distributions are

$$q_T(a \mid s, z) = \pi_{\bar{\theta}}(a \mid \mathrm{Hint}(s, z)), \qquad p_\theta(a \mid s) = \pi_\theta(a \mid s).$$

In our implementation, the local distillation loss is a pointwise-clipped forward-KL surrogate

$$\ell_{\mathrm{OPSD}}(s, z) = \sum_a \mathrm{clip}_{\tau_{\mathrm{clip}}}\left(q_T(a \mid s, z)\left[\log q_T(a \mid s, z) - \log p_\theta(a \mid s)\right]\right), \tag{3}$$

where $\mathrm{clip}_{\tau_{\mathrm{clip}}}(u) = \min(u, \tau_{\mathrm{clip}})$ caps each action-level summand before summing over the vocabulary. We use $\tau_{\mathrm{clip}} = 0.05$ unless otherwise stated. The clipping makes this an optimization surrogate rather than the exact KL. The prefix s comes from the behavior policy, and z is used only to construct the training-time teacher distribution.

3 Self-Supervised On-Policy Distillation

SSOPD uses a mixed GRPO group as a source of process supervision. A verified success supplies information about how the current policy can solve a prompt; a failed rollout supplies prefixes that the policy actually visits. These trajectories usually diverge, so the successful tokens cannot simply be copied onto the failed trajectory. Instead, a teacher uses the success as hindsight to make a local prediction at each failed prefix, and the student learns that prediction without seeing the hindsight.

3.1 The Process Target Hidden in Outcome Feedback

Fix a prompt x and let π be the behavior policy. A completion $y = (a_0, \ldots, a_{\tau(y)-1})$ receives a binary verifier reward $R(y) \in \{0, 1\}$, and its prefix state is $s_t(y) = (x, y_{<t})$. Write

$$V_R^\pi(s) = \Pr_\pi(R(Y) = 1 \mid s), \qquad Q_R^\pi(s, a) = \Pr_\pi(R(Y) = 1 \mid s, a).$$

At a prefix with $V_R^\pi(s) > 0$, the next-action distribution among successful continuations is

$$q^+(a \mid s) = \Pr_\pi(A = a \mid R(Y) = 1, s) = \frac{\pi(a \mid s) Q_R^\pi(s, a)}{V_R^\pi(s)}. \tag{4}$$

This target is induced by the policy's own successful behavior. It does not require an annotated reasoning step or an external teacher.

Proposition 3.1 (Success-conditioned policy)**.** *Suppose $p_x = \Pr_\pi(R(Y) = 1 \mid x) > 0$. At every reachable nonterminal prefix with positive success probability, use $q^+(\cdot \mid s)$ as the next-action policy. The resulting distribution over complete trajectories is*

$$\pi^+(y \mid x) = \frac{\pi(y \mid x) R(y)}{p_x}.$$

Thus q^+ is the token-level policy obtained by conditioning the original policy on eventual success.

The proposition identifies the process signal that a terminal verifier leaves implicit. The states of interest, however, are prefixes reached on failed on-policy rollouts. To make the difference precise, let $K_Y = \min(K, \tau(Y))$ and define the failure-prefix distribution for $p_x < 1$ by

$$d_\pi^-(s \mid x) = \frac{1}{1 - p_x} \mathbb{E}_{Y \sim \pi(\cdot|x)} \left[\mathbf{1}\{R(Y) = 0\} \frac{1}{K_Y} \sum_{t=0}^{K_Y - 1} \mathbf{1}\{s_t(Y) = s\} \right]. \tag{5}$$

This samples a failed completion and then one of its early prefixes. The ideal cross-trajectory objective would distill the success-conditioned target (4) at these failure states:

$$\mathcal{L}^*_{\text{cross}}(x) = \sum_{s : V_R^\pi(s) > 0} d_\pi^-(s \mid x) D_{\text{KL}}\big(q^+(\cdot \mid s) \big\| p_\theta(\cdot \mid s)\big). \tag{6}$$

The restriction to $V_R^\pi(s) > 0$ excludes prefixes from which the behavior policy has no successful continuation. This objective separates the two roles of the sampled group: successful continuations define the desired action distribution, while failed continuations define where to improve the policy. Neither distribution is directly available from a finite group at every failed prefix.

That mismatch is structural, not merely a lack of samples. Let $b(y^+, y^-)$ be the length of the longest common prefix of one correct and one wrong completion. Beyond this branching point,

$$s_t(y^-) \notin \{s_u(y^+) : 0 \le u < \tau(y^+)\} \qquad (t > b(y^+, y^-)).$$

Sequence imitation of y^+ therefore places its token targets on a path that the failed rollout no longer follows. More correct samples from the root do not, by themselves, supply the success-conditioned action distribution at a particular failed prefix. The key operation in SSOPD is to ask the teacher for a prediction *at that failed state*, while giving it the successful attempt as context. The student's target is the teacher's resulting distribution at $s_t(y^-)$, not any token from y^+ at its original position. This is how outcome contrast becomes supervision at states that ordinary successful-trace imitation misses.

The condition $V_R^\pi(s) > 0$ identifies the recoverable failure prefixes on which the ideal target and local analysis apply. This motivates the finite-group construction below: use a successful witness to guide the teacher across trajectories, and concentrate supervision on early prefixes of a failure.

3.2 Transferring Hindsight Across Trajectories

For each prompt, GRPO samples $\mathcal{G}(x) = \{y_i\}_{i=1}^G$ with $y_i \sim \pi(\cdot \mid x)$ and forms

$$\mathcal{C}(x) = \{y_i : R(y_i) = 1\}, \qquad \mathcal{W}(x) = \{y_i : R(y_i) = 0\}.$$

Consider a successful witness $z \in \mathcal{C}(x)$ and a prefix s of a failed rollout $w \in \mathcal{W}(x)$. The student sees only s; a stop-gradient teacher also sees z:

$$q_T(\cdot \mid s, z) = \pi_{\bar{\theta}}(\cdot \mid \mathrm{Hint}(s, z)), \qquad p_\theta(\cdot \mid s) = \pi_\theta(\cdot \mid s). \tag{7}$$

The witness is not a target sequence for the student. Its role is to help the teacher form a next-token distribution at a state visited by the behavior policy. Following the OPSD implementation in §2.2, the local training loss is the pointwise-clipped forward-KL surrogate

$$\ell_{\mathrm{OPSD}}(s, z) = \sum_a \mathrm{clip}_{\tau_{\mathrm{clip}}}\left(q_T(a \mid s, z)\left[\log q_T(a \mid s, z) - \log p_\theta(a \mid s)\right]\right). \tag{8}$$

The teacher supplies the extra information only during training. At inference, the student acts from the prompt and its own prefix.

The witness and the failed rollout play asymmetric roles in this transfer. Every candidate witness has already passed the verifier, so selection among successes concerns how economically to present that evidence to the teacher. The teacher must fit the problem and witness into its prompt window; a shorter witness spends less of that window on the reference trajectory. Every candidate failure has the same terminal label, while a longer one offers more observed prefix positions until the budget K is filled. In our runs, the teacher prompt is capped at 2048 tokens and $K = 1024$. These resource constraints motivate a simple length selector for each side:

$$y^+ = \arg\min_{y \in \mathcal{C}(x)} \tau(y), \qquad y^- = \arg\max_{y \in \mathcal{W}(x)} \tau(y). \tag{9}$$

The first rule conserves teacher context; the second maximizes the number $\min(K, \tau(y^-))$ of available supervised positions. Coverage saturates at K, so longer failures beyond that point are an inexpensive tie breaker rather than a claim that length measures error severity. This pair performs best among the tested selectors in §4.4.

Let $K_- = \min(K, \tau(y^-))$ and $s_t^- = (x, y_{<t}^-)$. The auxiliary loss uses the first K_- prefixes of the failed completion:

$$\mathcal{L}_{\mathrm{SSOPD}}(x) = \frac{1}{K_-} \sum_{t=0}^{K_- - 1} \ell_{\mathrm{OPSD}}(s_t^-, y^+). \tag{10}$$

The sampled failure determines where supervision is applied; the sampled success supplies the teacher's hindsight. Both come from the same on-policy group used by GRPO. Relative to (6), using one failed path replaces an expectation over failure prefixes with a hard selection, and the witness-conditioned teacher replaces the intractable success-conditioned distribution. These are the two approximations made by the implemented objective. The implementation averages the weighted local losses over all valid selected tokens in the minibatch; (10) displays the corresponding loss for one selected pair. The prefix budget concentrates the update on earlier decisions and reduces exposure to late states with little chance of recovery. Length selection and prefix truncation therefore serve distinct purposes: the former chooses a path from the group, while the latter fixes which of its states are supervised.

The on-policy construction matters for both sides of the pair. The witness is a solution that the behavior policy actually generated for this prompt, so its content reflects the model's current capabilities. The failed prefix is likewise a state encountered by that policy, so the student is trained under the ordinary prompt-prefix conditioning it will use at inference. In our experiments the teacher is a frozen base model; its extra information comes from the witness, while the student receives gradients from the auxiliary loss. As the behavior policy is refreshed during training, new groups supply new witnesses and failure states. SSOPD therefore uses an evolving source of privileged information without requiring a separate solution corpus for its own update. This keeps the comparison with GRPO matched in training information, though teacher and student scoring adds computation.

3.3 Local Improvement and the Reasoning Frontier

The ideal target in (4) is generally intractable at a failed prefix. SSOPD approximates it in two stages: the witness helps the teacher form a target, and distillation moves the student toward that target. Let $\mathrm{TV}(u, v) = \frac{1}{2}\sum_a |u(a) - v(a)|$ and define the teacher's witness-approximation error $\delta_T(s, z) = \mathrm{TV}(q^+(\cdot \mid s), q_T(\cdot \mid s, z))$.

Theorem 3.2 (Two-stage local transfer)**.** *Fix a prefix s with $V_R^\pi(s) > 0$ and a witness z. For $0 \le \eta \le 1$, change only the action distribution at s to*

$$\tilde{\pi}_\eta(a \mid s) = (1-\eta)\pi(a \mid s) + \eta p_\theta(a \mid s).$$

Then

$$V_R^{\tilde{\pi}_\eta}(s) - V_R^\pi(s) \ge \eta \left(\frac{\mathrm{Var}_{a\sim\pi(\cdot|s)}[Q_R^\pi(s,a)]}{V_R^\pi(s)} - \delta_T(s,z) - \sqrt{\tfrac{1}{2} D_{\mathrm{KL}}(q_T(\cdot \mid s,z) \| p_\theta(\cdot \mid s))} \right).$$

The bound gives a direct sufficient condition for local improvement: outcome-relevant branching exceeds the teacher's approximation error and the student's distillation error. It separates the two transfer stages of the method. The statement analyzes a local policy edit with exact KL; training uses the pointwise-clipped surrogate in (8). The proof is in Appendix B.

This sampled transfer is available only when a group contains both a witness and a failed rollout. Let $K_c = |\mathcal{C}(x)|$ and $\hat{p}_x = K_c/G$. SSOPD therefore uses the prompt-level weight

$$\lambda_x = \lambda_0 \cdot 4\hat{p}_x(1-\hat{p}_x) = \frac{4\lambda_0}{G^2} K_c(G-K_c), \tag{11}$$

treated as stop-gradient. It is proportional to the number of correct–wrong pairs in the sampled group and is zero when either side is missing.

Proposition 3.3 (Frontier weight and branching variance)**.** *For binary rewards and a finite horizon with terminal padding,*

$$p_x(1-p_x) = \sum_{t=0}^{H-1} \mathbb{E}_{S_t}\left[\mathrm{Var}_{a\sim\pi(\cdot|S_t)}[Q_R^\pi(S_t,a)]\right].$$

If the G rollouts are independent, then

$$\Pr(0 < K_c < G) = 1 - p_x^G - (1-p_x)^G,$$

and

$$\mathbb{E}[K_c(G-K_c)] = G(G-1)p_x(1-p_x), \qquad \mathbb{E}[\lambda_x \mid x] = 4\lambda_0 \frac{G-1}{G} p_x(1-p_x).$$

The empirical weight counts available contrasts; in expectation, it tracks the policy's outcome-related branching variance. It puts the auxiliary loss on the current reasoning frontier, where the policy both succeeds and fails on the same prompt. The two finite-group quantities in the proposition serve different purposes. The probability of a mixed group tells us whether the auxiliary update can be formed at all. The pair count distinguishes a group with one exceptional success from a group with several successes and failures. Although SSOPD uses only one selected pair to construct teacher targets, a group with more possible pairs provides stronger evidence that success and failure both occur under the current policy. The weight uses that evidence to scale the single-pair update.

Group size controls how often this opportunity appears. For example, with $G = 8$ and $p_x = 1/2$, the probability of a mixed group is $1 - 2^{1-G} \approx 0.992$; near either extreme of p_x, mixed groups become less frequent. The auxiliary signal therefore follows the moving boundary of the model's capabilities: a prompt begins to contribute after the policy first produces some successes, and its weight falls again as failures disappear. This behavior follows from the sampled outcomes alone and requires no separate difficulty model. All-correct and all-wrong groups receive no auxiliary distillation.

The final per-prompt objective is

$$\mathcal{L}(x) = \mathcal{L}_{\mathrm{GRPO}}(x) + \lambda_x \mathcal{L}_{\mathrm{SSOPD}}(x). \tag{12}$$

When the group is all-correct or all-wrong, no pair is instantiated and the auxiliary term is zero.

4 Experiments

Our experiments test whether Self-Supervised On-Policy Distillation (SSOPD) can turn the model's own on-policy attempts into useful process supervision. The central comparison is information-controlled. SFT and OPSD use additional solution information, while GRPO and SSOPD use only answer verification and samples from the current policy. We ask whether SSOPD can improve the non-privileged GRPO setting, and how close self-generated supervision can come to solution-conditioned distillation.

4.1 Experimental Setup

Models and data. We use the instruct-tuned Qwen3 family (Yang et al., 2025) at three scales: Qwen3-1.7B, Qwen3-4B, and Qwen3-8B. Training prompts come from the mathematical reasoning subset of OpenThoughts (Guha et al., 2025); for each scale, we sample up to 30K problems with verified final answers. Reference reasoning traces are used only by the baselines that explicitly require additional solution information, namely SFT and OPSD. Evaluation covers AIME 2024, AIME 2025, and HMMT 2025.

Compared post-training recipes. We evaluate four updates from the same base checkpoints, organized by the information available during training. **GRPO** (Shao et al., 2024) optimizes binary answer rewards with grouped rollouts and uses no reference solution. **SSOPD** keeps the same non-privileged setting, but adds our auxiliary distillation term: the model's own shortest correct completion serves as the successful witness, prefixes of the longest wrong completion are edited, and the auxiliary loss is weighted by the raw group success rate through $\hat{p}_x(1-\hat{p}_x)$. For privileged comparisons, **SFT** directly imitates reference reasoning traces, and **OPSD** (Zhao et al., 2026) scores student prefixes with a solution-conditioned teacher.

Training and evaluation protocol. Unless otherwise stated, SSOPD uses a fixed privileged teacher anchored at the initial checkpoint while the student policy is updated with LoRA (Hu et al., 2022). Teacher forward passes disable the trainable adapters, and the base coefficient of the SSOPD weight is set to $\lambda_0 = 0.5$. We report Avg@12 accuracy: for each problem, the model samples 12 completions and is credited according to the average verified correctness. Evaluation uses thinking mode, temperature 1.0, top-p 0.95, and a maximum generation length of 38k tokens. Additional training details appear in Appendix C.

The protocol is designed to isolate the value of reusing the sampled group itself. GRPO and SSOPD are information-matched: both update from prompts, on-policy completions, and verified final answers. Their difference is how much structure they extract from the same group. SFT and OPSD are included as privileged references because they use additional solution-side information, and therefore measure how close self-supervised group reuse can come to solution-conditioned supervision.

4.2 Main Results

Table 1 shows a consistent advantage for SSOPD within the information-matched comparison. It improves over GRPO on every benchmark at every scale, with macro-average gains of $+0.6$ on Qwen3-1.7B, $+0.3$ on Qwen3-4B, and $+1.6$ on Qwen3-8B. On-policy groups and verifier outcomes support a stronger update when their correct–wrong structure is used for prefix-level distillation.

The 8B result is especially encouraging. SSOPD reaches a macro Avg@12 of 65.6, above both GRPO (64.0) and the solution-conditioned OPSD baseline (64.8). It has the highest score in the table on AIME 2024 and HMMT 2025 and ties the highest score on AIME 2025. At this scale, self-generated witnesses provide enough guidance to compete with a teacher given reference solutions.

At Qwen3-1.7B and Qwen3-4B, OPSD remains stronger overall, while SSOPD still improves every GRPO benchmark score. This suggests a useful division of labor: external solutions help most when the policy has fewer strong successes of its own, whereas group-based supervision becomes more competitive as the policy strengthens. SFT trails the base model in macro average at all three scales.

The gains span all three benchmarks despite their different base accuracies. This breadth matters for the method: SSOPD can use a mixed group to refine prompts the model often solves and to learn from difficult prompts where a successful witness is rarer.

4.3 Ablation on the Frontier Weight

We ablate the auxiliary coefficient on HMMT 2025 using the Qwen3-4B setting. With the dynamic frontier weight $\lambda_x = \lambda_0 \cdot 4\hat{p}_x(1-\hat{p}_x)$, $\lambda_0 \in \{0.4, 0.5, 0.6, 0.7\}$ gives Avg@12 scores of 44.4, 45.0, 44.7, and 44.2, respectively. Using the same values as fixed global coefficients instead gives 44.2, 44.7, 44.4, and 43.9. The dynamic weight leads at every tested coefficient and peaks at $\lambda_0 = 0.5$, supporting the choice to concentrate distillation on mixed groups.

Table 1: Math reasoning Avg@12 after post-training Qwen3 backbones. Results are reported in percentage points. SFT and OPSD use additional solution information, while GRPO and SSOPD use only answer verification and on-policy samples. Evaluation uses thinking mode, temperature 1.0, top-p 0.95, and maximum length 38k.

Backbone / post-training recipe	AIME 2024	AIME 2025	HMMT 2025	Macro avg.
Qwen3-8B Instruct				
No update	75.8	65.6	43.9	61.8
No additional solution information				
+ GRPO	76.4	68.9	46.7	64.0
+ SSOPD (ours)	**78.6**	**70.8**	**47.5**	**65.6**
Uses additional solution information				
+ SFT	72.3	64.2	42.9	59.8
+ OPSD	77.8	**70.8**	45.8	64.8
Qwen3-4B Instruct				
No update	74.9	66.4	42.2	61.2
No additional solution information				
+ GRPO	75.6	68.1	44.4	62.7
+ SSOPD (ours)	75.8	**68.3**	45.0	63.0
Uses additional solution information				
+ SFT	70.2	62.3	43.4	58.6
+ OPSD	**76.4**	**68.3**	**46.1**	**63.6**
Qwen3-1.7B Instruct				
No update	51.5	36.7	23.1	37.1
No additional solution information				
+ GRPO	51.1	38.3	23.7	37.7
+ SSOPD (ours)	51.7	38.6	24.7	38.3
Uses additional solution information				
+ SFT	48.4	36.3	22.7	35.8
+ OPSD	**57.2**	**43.9**	**29.2**	**43.4**

Table 2: Ablation of selector choices on HMMT 2025. Each entry reports the best `Avg@12` across checkpoints.

	Wrong-side selector		
Correct-side selector	$\mathrm{Len}_{\max}$	$\mathrm{Len}_{\min}$	$\mathrm{AvgLogP}_{\max}$
$\mathrm{Len}_{\max}$	42.8	44.4	44.4
$\mathrm{Len}_{\min}$	**45.0**	44.2	44.2
$\mathrm{AvgLogP}_{\max}$	42.5	42.8	42.8
$\mathrm{AvgLogP}_{\min}$	42.8	42.8	42.8

4.4 Pair Selection

We ablate the choice of selectors by enumerating the Cartesian product between four correct-side selectors and three wrong-side selectors on HMMT 2025. We denote selecting the longest and shortest sequence as $\mathrm{Len}_{\max}$ and $\mathrm{Len}_{\min}$, respectively. $\mathrm{AvgLogP}_{\max}$ and $\mathrm{AvgLogP}_{\min}$ denote selecting the sequence with the maximum and minimum sequence-average log-probability. For each selector pair, we report the best `Avg@12` across checkpoints.

The shortest-correct / longest-wrong pair reaches 45.0, at least 0.6 points above every other pair in the sweep. Changing either selector lowers the best checkpoint score. This agrees with the method's asymmetric budgets: a concise witness leaves more teacher context for the problem and instructions, while an extended failure supplies the full early-prefix training window when available. The sweep tests the operational choice directly, rather than treating length as a measure of solution quality or error severity.

The two ablations probe complementary parts of the update. Selector changes alter the witness or failed prefixes; coefficient changes alter how strongly a mixed group contributes. Across selector cells,

the training recipe and binary reward criterion are held fixed, yet best-checkpoint scores span 42.5 to 45.0. This variation is consistent with the central mechanism: the group supplies more than correct and wrong labels, because the particular witness and visited prefixes determine the local supervision. The shortest-correct / longest-wrong pair leads the selector sweep, and dynamic weighting beats fixed weighting at every tested coefficient.

5 Related Work

GRPO-style RLVR methods (Shao et al., 2024; DeepSeek-AI et al., 2025; Yu et al., 2025) and recent reasoning post-training systems (Kimi Team, 2025; Mistral-AI, 2025) optimize verified answers from sampled trajectories. Their terminal rewards leave intermediate decisions without direct feedback. SSOPD uses the same on-policy groups and verifier, drawing dense targets from the contrast between a successful sample and failed prefixes of the same prompt.

Dense supervision for reasoning is often obtained through reference traces, process labels, or distillation. Supervised fine-tuning on curated mathematical traces (Guha et al., 2025; Xiaomi LLM-Core Team, 2026) and process supervision (Lightman et al., 2024; Zhang et al., 2025) provide more localized targets, but require reference states, annotations, or learned reward models. Distillation and self-improvement methods transfer teacher distributions, generated sequences, or behavior under richer context (Hinton et al., 2015; Kim & Rush, 2016; Sanh et al., 2019; Gu et al., 2024; Zelikman et al., 2022; Gulcehre et al., 2023; Snell et al., 2022; Huang et al., 2022). Closest to our setting, privileged on-policy self-distillation conditions a teacher on verified solution information (Zhao et al., 2026). SSOPD draws its privileged information from a successful completion in the same on-policy group and applies the resulting teacher distribution at prefixes of a failed completion. The supervision is therefore tied to both what the current policy can already solve and where it currently fails.

Filtering and imitating successful generations also reuses verifier outcomes (Zelikman et al., 2022; Gulcehre et al., 2023), but its targets stay on successful trajectories. After correct and wrong attempts branch, direct imitation supplies no target at the wrong attempt's later prefixes. SSOPD evaluates its witness-conditioned teacher at those prefixes instead.

This also separates SSOPD from solution-conditioned distillation (Zhao et al., 2026). In both methods the student is updated at an on-policy prefix and a teacher sees privileged context. The source and availability of that context differ. A reference solution is provided independently of the policy's sampled group; an SSOPD witness exists only when the current policy succeeds at least once in that group. The frontier weight handles this availability directly. Together, witness generation, failure-state visitation, and prompt-level gating make the training signal self-contained within the same outcome-labeled rollouts used for GRPO.

6 Conclusion

A mixed on-policy group contains more than a sparse reward signal. Its successful completion supplies prompt-specific hindsight, and its failed completion supplies states at which the policy needs guidance. SSOPD transfers that hindsight across trajectories: a privileged teacher forms local targets at failed prefixes, and the student learns to act there without the witness. This gives process-level supervision from the same samples and answer verification already used by GRPO.

The success-conditioned policy makes the latent process target explicit, while the local edit analysis separates teacher quality from student fitting. Frontier weighting concentrates the auxiliary loss where the sampled group contains both outcomes. On three math benchmarks and three Qwen3 scales, SSOPD improves over GRPO in every evaluated setting and exceeds the solution-conditioned OPSD baseline in macro Avg@12 at 8B.

Reliable answer verification and self-generated witnesses are the core ingredients. Future work can test witness and prefix selection in other verifiable domains. This paper aims to advance machine learning; we do not foresee specific negative societal consequences.

AI USE STATEMENT

Generative AI tools were used to improve the wording of the manuscript and to assist in checking the correctness of the proofs. The authors reviewed all AI-assisted suggestions, independently verified the final proofs, and take responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

The distillation objective, witness selection rule, and frontier weight are specified in Section 3. Experimental protocols and principal results are reported in Section 4; the complete training algorithm, proofs, implementation settings, and dataset details appear in Appendices A to D.

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.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. doi: 10.48550/arXiv.2110.14168. URL https://arxiv.org/abs/2110.14168.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. doi: 10.48550/arXiv.2501.12948. URL https://arxiv.org/abs/2501.12948.

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

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

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. Reinforced self-training (ReST) for language modeling. *arXiv preprint arXiv:2308.08998*, 2023. doi: 10.48550/arXiv.2308.08998. URL https://arxiv.org/abs/2308.08998.

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, 2015. URL https://arxiv.org/abs/1503.02531.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yukun Huang, Yanda Chen, Zhou Yu, and Kathleen McKeown. In-context learning distillation: Transferring few-shot learning ability of pre-trained language models. *arXiv preprint arXiv:2212.10670*, 2022. doi: 10.48550/arXiv.2212.10670. URL https://arxiv.org/abs/2212.10670.

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, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Kimi Team. Kimi K2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025. doi: 10.48550/arXiv.2507.20534. URL https://arxiv.org/abs/2507.20534.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation/.

Mistral-AI. Magistral. *arXiv preprint arXiv:2506.10910*, 2025. doi: 10.48550/arXiv.2506.10910. URL https://arxiv.org/abs/2506.10910.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. doi: 10.48550/arXiv.1910.01108. URL https://arxiv.org/abs/1910.01108.

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.

Charlie Snell, Dan Klein, and Ruiqi Zhong. Learning by distilling context. *arXiv preprint arXiv:2209.15189*, 2022. doi: 10.48550/arXiv.2209.15189. URL https://arxiv.org/abs/2209.15189.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Xiaomi LLM-Core Team. MiMo-V2-Flash technical report, 2026. URL https://arxiv.org/abs/2601.02780.

Wenda Xu, Rujun Han, Zifeng Wang, Long T. Le, Dhruv Madeka, Lei Li, William Yang Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=EgJhwYR2tB.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. doi: 10.48550/arXiv.2503.14476. URL https://arxiv.org/abs/2503.14476.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15476–15488, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/639a9a172c044fbb64175b5fad42e9a5-Abstract-Conference.html.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. *arXiv preprint arXiv:2501.07301*, 2025. doi: 10.48550/arXiv.2501.07301. URL https://arxiv.org/abs/2501.07301.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-Distilled Reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026. doi: 10.48550/arXiv.2601.18734. URL https://arxiv.org/abs/2601.18734.

Algorithm 1 SSOPD auxiliary update

Require: Prompt minibatch $\mathcal{B}$, behavior policy π, student policy π_θ, stop-gradient teacher $\pi_{\bar{\theta}}$, verifier R, group size G, prefix budget K, base coefficient λ_0, clipping threshold τ_{clip}

for all $x \in \mathcal{B}$ **do**
  Sample $\mathcal{G}(x) = \{y_i\}_{i=1}^{G}$, with $y_i \sim \pi(\cdot \mid x)$
  Compute rewards $r_i \leftarrow R(y_i)$ and the GRPO loss $\mathcal{L}_{\text{GRPO}}(x)$
  $\mathcal{C}(x) \leftarrow \{y_i : r_i = 1\}$, $\mathcal{W}(x) \leftarrow \{y_i : r_i = 0\}$
  if $|\mathcal{C}(x)| > 0$ and $|\mathcal{W}(x)| > 0$ **then**
    $y^+ \leftarrow \arg\min_{y \in \mathcal{C}(x)} \tau(y)$
    $y^- \leftarrow \arg\max_{y \in \mathcal{W}(x)} \tau(y)$
    $K_- \leftarrow \min(K, \tau(y^-))$, $\hat{p}_x \leftarrow |\mathcal{C}(x)|/G$
    $\lambda_x \leftarrow \lambda_0 \cdot 4\hat{p}_x(1 - \hat{p}_x)$ ▷ stop-gradient
    for $t = 0, \ldots, K_- - 1$ **do**
      $s_t^- \leftarrow (x, y_{<t}^-)$
      $q_T(\cdot \mid s_t^-, y^+) \leftarrow \pi_{\bar{\theta}}(\cdot \mid \text{Hint}(s_t^-, y^+))$
      $p_\theta(\cdot \mid s_t^-) \leftarrow \pi_\theta(\cdot \mid s_t^-)$
    $\mathcal{L}_{\text{SSOPD}}(x) \leftarrow \frac{1}{K_-} \sum_{t=0}^{K_- - 1} \ell_{\text{OPSD}}(s_t^-, y^+)$
  else
    $\lambda_x \leftarrow 0$, $\mathcal{L}_{\text{SSOPD}}(x) \leftarrow 0$
Update θ using $\frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} [\mathcal{L}_{\text{GRPO}}(x) + \lambda_x \mathcal{L}_{\text{SSOPD}}(x)]$

A Training Algorithm

Algorithm 1 summarizes the SSOPD update using the same notation as §3. A prompt x produces a GRPO group $\mathcal{G}(x)$ under the behavior policy π. If the group contains both correct and wrong completions, the shortest correct completion y^+ is used as privileged hindsight, and the first K_- prefixes of the longest wrong completion y^- are edited by distillation.

B Detailed Proofs

Proof of Proposition 3.1. Write $h(s) = V_R^\pi(s)$. Since appending action a to an autoregressive prefix gives the next state sa, we have $Q_R^\pi(s, a) = h(sa)$. Bayes' rule gives

$$q^+(a \mid s) = \frac{\pi(a \mid s) h(sa)}{h(s)}.$$

Along a complete trajectory $y = (a_0, \ldots, a_{\tau(y)-1})$, these value ratios telescope:

$$\pi^+(y \mid x) = \prod_{t=0}^{\tau(y)-1} \frac{\pi(a_t \mid s_t(y)) h(s_{t+1}(y))}{h(s_t(y))} = \pi(y \mid x) \frac{h(s_{\tau(y)}(y))}{h(x)}.$$

At termination, $h(s_{\tau(y)}(y)) = R(y)$, while $h(x) = p_x$. This proves the identity. □

Proof of Theorem 3.2. Only the action distribution at s changes, so subsequent action values remain $Q_R^\pi(s, a)$. Abbreviate $V = V_R^\pi(s)$ and $Q(a) = Q_R^\pi(s, a)$. The improvement is

$$V_R^{\tilde{\pi}_\eta}(s) - V = \eta \left(\mathbb{E}_{a \sim p_\theta(\cdot \mid s)} Q(a) - V \right).$$

By (4),

$$\mathbb{E}_{a \sim q^+(\cdot \mid s)} Q(a) = \frac{\mathbb{E}_{a \sim \pi(\cdot \mid s)} Q(a)^2}{V} = V + \frac{\text{Var}_{a \sim \pi(\cdot \mid s)}[Q(a)]}{V}.$$

Since $0 \le Q(a) \le 1$, the difference in its expectation under two action distributions is at most their total variation distance. Add and subtract the expectations under q^+ and q_T in the first display, then use the triangle inequality and Pinsker's inequality:

$$\begin{aligned} \mathbb{E}_{p_\theta} Q &\ge \mathbb{E}_{q^+} Q - \text{TV}(q^+, q_T) - \text{TV}(q_T, p_\theta) \\ &\ge \mathbb{E}_{q^+} Q - \delta_T(s, z) - \sqrt{\tfrac{1}{2} D_{\text{KL}}(q_T \| p_\theta)}. \end{aligned}$$

Substitution yields the bound. □

Proof of Proposition 3.3. Let S_t be the random prefix at depth t, padded after termination by an absorbing terminal state, and set $M_t = V_R^\pi(S_t) = \Pr_\pi(R = 1 \mid S_t)$. This is a martingale: conditional on $S_t = s$,

$$\mathbb{E}[M_{t+1} \mid S_t = s] = \sum_a \pi(a \mid s) Q_R^\pi(s, a) = V_R^\pi(s) = M_t.$$

At the root $M_0 = p_x$, and at the terminal padded state $M_H = R$. Orthogonality of martingale increments therefore gives

$$p_x(1 - p_x) = \mathrm{Var}(R \mid x) = \sum_{t=0}^{H-1} \mathbb{E}[(M_{t+1} - M_t)^2 \mid x].$$

Conditioning on $S_t = s$, the summand is $\mathrm{Var}_{a \sim \pi(\cdot|s)}[Q_R^\pi(s, a)]$. Taking expectation over S_t proves the branching-variance identity.

For independent rollouts, $K_c \sim \mathrm{Binomial}(G, p_x)$. Counting pairs of distinct indices with a correct first member and a wrong second member gives

$$\mathbb{E}[K_c(G - K_c)] = G(G - 1)p_x(1 - p_x).$$

The group is mixed unless every rollout is correct or every rollout is wrong, giving $\Pr(0 < K_c < G) = 1 - p_x^G - (1 - p_x)^G$. Finally, $\lambda_x = 4\lambda_0 K_c(G - K_c)/G^2$. Taking expectations proves the coefficient identity. □

C Implementation Details

The auxiliary term is computed from the same sampled group used by GRPO; no additional rollout policy is introduced. It does require teacher and student scoring at the selected prefixes, so matching the information budget does not imply matching compute cost. In the notation of §3, π is the old policy that generates the group, π_θ is the LoRA-updated student, and $\pi_{\bar{\theta}}$ is the fixed teacher used only for target construction. Teacher forward passes are run with stop-gradient. The student receives gradients from both $\mathcal{L}_{\mathrm{GRPO}}$ and $\mathcal{L}_{\mathrm{SSOPD}}$. For a selected wrong completion y^-, we evaluate the auxiliary loss on prefix states $s_t^- = (x, y_{<t}^-)$ for $t < K_-$.

To construct $\mathrm{Hint}(s_t^-, y^+)$, we decode the problem and selected correct completion. The Qwen3 user message begins with "Problem: [problem]" and introduces the completion with "Here is a reference solution to this problem:" between the markers "=== Reference Solution Begin ===" and "=== Reference Solution End ===". It then asks the teacher to understand each step without copying or paraphrasing the solution, derive the same final answer in its own words, explore alternatives and backtrack if needed, and put the final answer in a box. The chat template enables thinking mode and adds the generation prompt. Teacher logits are computed by appending the selected wrong completion to this teacher prompt and scoring its next tokens; student logits use the ordinary problem prompt with that same wrong completion. The teacher prompt is tokenized with a maximum of 2048 tokens, using the tokenizer's truncation rule.

The prompt-level weight uses the raw group success rate

$$\hat{p}_x = \frac{|\mathcal{C}(x)|}{G}, \qquad \lambda_x = \lambda_0 \cdot 4\hat{p}_x(1 - \hat{p}_x),$$

and λ_x is treated as stop-gradient. No extra correction is applied to this estimate. If the group is all-correct or all-wrong, then $\hat{p}_x(1 - \hat{p}_x) = 0$, so the auxiliary term is automatically inactive. In code, the selected-pair losses, including their $4\hat{p}_x(1 - \hat{p}_x)$ factors, are summed over valid selected tokens and divided by the total number of valid selected tokens in the minibatch. Thus (12) is a per-prompt description of the update rather than an exact minibatch reduction when selected failures have different lengths.

Unless otherwise stated, the GRPO+SSOPD runs train LoRA adapters with learning rate 5×10^{-6} and effective batch size 32 and experiments are run on A100 or H100 GPUs. We sample $G = 8$ completions per prompt, which keeps each prompt group on one process and makes the correct–wrong

pair selection local. The LoRA rank is $r = 64$ with $\alpha = 128$; the adapted modules are q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, and down_proj.

The GRPO rollout configuration uses binary correctness rewards, group reward scaling, KL coefficient $\beta = 0$, sampling temperature 1.2, maximum prompt length 2048, and maximum rollout length 16000. The SSOPD regularizer uses $\lambda_0 = 0.5$, the fixed-teacher setting obtained by disabling trainable adapters during teacher scoring, and the shortest-correct / longest-wrong selection rule. We apply the regularizer to at most the first $K = 1024$ generated positions of y^-, compute full-vocabulary teacher targets in chunks of 16384 vocabulary entries, and clip each pointwise distillation summand at 0.05. Checkpoints are saved every 50 steps; the HMMT 2025 sweep reports checkpoints up to step 250.

Evaluation follows the Qwen3 recommended sampling setup with thinking mode enabled: temperature 1.0, top-p 0.95, top-k -1, min-p 0.0, no presence penalty, maximum generation length $38{,}912$, and 12 samples per prompt.

D Dataset Sources and Terms

Table 3 links the dataset cards and summarizes how each asset is used. SSOPD uses training prompts and verified final answers from the OpenThoughts math subset; reference traces are used only for the SFT and OPSD baselines. The AIME and HMMT sets are used only for evaluation. We report the terms stated on the dataset cards without inferring a separate license for a derived dataset from its upstream project.

Table 3: Data sources, uses, and stated terms. Dataset names link to their public source cards.

Dataset	**Use**	**Stated terms**
OpenThoughts math subset	Training prompts; reference traces for privileged baselines	The derived card has no separate license field. The upstream OpenThoughts project describes its releases as Apache-2.0.
AIME 2024	Evaluation only	No explicit license field on the dataset card.
AIME 2025	Evaluation only	No explicit license field on the dataset card.
HMMT February 2025	Evaluation only	CC BY-NC-SA 4.0, as stated on the MathArena dataset card.