# PAINT: ADAPTIVE PRIVILEGED-CONTEXT SELF-DISTILLATION FOR REASONING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Effective reasoning supervision must act on the states a model visits while offering richer guidance than a final-answer reward. We frame on-policy self-distillation as an information-allocation problem: training-only context re-scores student-generated prefixes, and its influence is set by how much information is exposed and where the resulting token targets are calibrated. We propose *PAINT* (*Privileged-context Adaptive INterpolated Training*), which makes these two choices at different resolutions. For mathematical reasoning, rollout-reference overlap determines how much of a verified solution the privileged scorer sees; a teacher/student entropy ratio then selects a sparse set of prefixes for small energy-space interpolation. Conditional mutual information explains the value of context exposure, and geometric interpolation controls the local target. Across three competition-level math benchmarks and three Qwen3 scales, PAINT improves over full-solution on-policy self-distillation. On Qwen3-8B, it raises macro Avg@12 by 2.1 points over that baseline and 2.9 points over GRPO.

## 1 INTRODUCTION

Reasoning post-training increasingly determines whether LLMs can solve difficult mathematical and scientific tasks, a capability made visible by chain-of-thought prompting and math-centered evaluation (Wei et al., 2022; Cobbe et al., 2021; Hendrycks et al., 2021). The central challenge is not just to reward correct final answers, but to provide supervision that is both *deployment-aligned* and *informative*: training should act on states the model actually visits at inference time, while still giving more guidance than a single sequence-level correctness bit.

Current recipes occupy different points in this trade-off. Reinforcement learning with verifiable rewards (RLVR), including GRPO-style methods (Shao et al., 2024; DeepSeek-AI et al., 2025; Yu et al., 2025), optimizes the desired outcome on on-policy samples, but its reward is sparse and often costly to estimate. Supervised fine-tuning (SFT) on curated reasoning traces (Guha et al., 2025; Xiaomi LLM-Core Team, 2026) is cheap and dense, but trains on reference prefixes rather than the model's own prefixes. Distillation supplies token-level targets, yet standard forms either remain off-policy or depend on an external stronger model (Hinton et al., 2015; Agarwal et al., 2024; Lu & Thinking Machines Lab, 2025).

Training-only context offers another route to dense supervision on student-generated trajectories. A model can score its own rollout prefixes again with information unavailable at deployment, such as a verified solution, and turn the difference between the two views into token-level guidance. Prior privileged on-policy self-distillation demonstrates this mechanism with a full reference solution (Zhao et al., 2026). The broader question is how to allocate the privileged information. Revealing everything can concentrate the target on one trace; applying the resulting target uniformly ignores how its information advantage varies from prefix to prefix. These choices arise whenever a model learns from a context-rich view of its own trajectories.

We formulate this as contextual re-scoring. Building on context distillation (Snell et al., 2022; Huang et al., 2022; Qi et al., 2025), the same student prefix is scored with and without training-only evidence. The expected uncertainty reduction from that evidence is conditional mutual information. This leads to two complementary controls: choose how much evidence the privileged view receives, then identify the prefixes where its extra information resolves the least uncertainty.



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.

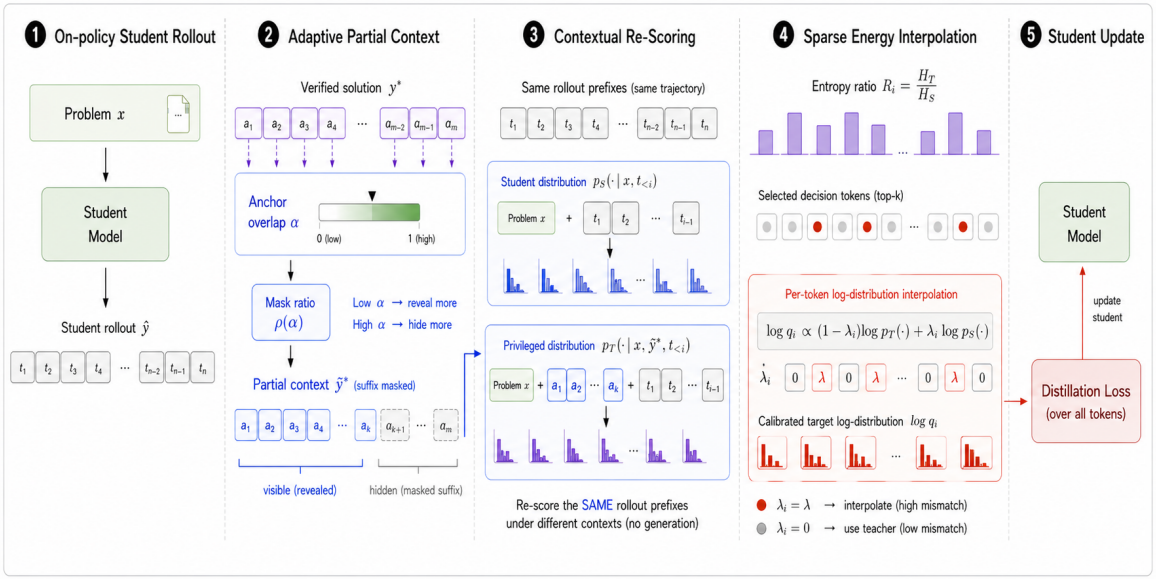



Figure 1: **PAINT training pipeline.** PAINT samples an on-policy rollout, uses rollout-reference overlap $\alpha$ to form a suffix-masked solution $\tilde{y}^\star$, re-scores the same prefixes with a fixed privileged view, and applies small energy interpolation only on entropy-mismatch positions.

As summarized in Figure 1, **P**rivileged-context **A**daptive **IN**terpolated **T**raining (PAINT) instantiates the first control with overlap-adaptive partial solutions: low-overlap rollouts receive more reference context, while high-overlap rollouts hide more. The second control ranks positions by the teacher/student entropy ratio and interpolates the privileged target toward the student only on selected positions. Together, they allocate privileged information across rollouts and token positions.

Across Qwen3-8B, 4B, and 1.7B, PAINT improves over the prior self-distillation baseline, with macro Avg@12 gains between 0.8 and 2.1 points. It also matches or exceeds GRPO across scales while training with one short rollout per problem rather than eight long rollouts.

Our main contributions are summarized below:

- We cast privileged on-policy distillation as an information-allocation problem: control the evidence available to the scorer and calibrate its targets where the information advantage is smallest.
- We propose PAINT, a two-resolution instantiation of this principle using overlap-adaptive context exposure and sparse teacher energy interpolation for verified math solutions.
- We evaluate PAINT on three competition-level math benchmarks and three model scales, showing gains over the prior on-policy self-distillation baseline and SFT, plus GRPO-level or better accuracy with fewer training rollouts.

## 2 Background

Reasoning post-training has a basic tension between *where the trajectory comes from* and *how informative the supervision is*. SFT and off-policy distillation provide dense token-level targets, but train on reference prefixes rather than the model's own prefixes. RLVR samples on-policy trajectories, but its correctness reward is sparse and sequence-level. On-policy distillation combines on-policy data with dense targets (Agarwal et al., 2024; Lu & Thinking Machines Lab, 2025; Xu et al., 2025), but usually assumes an external teacher. The prior OPSD framework (Zhao et al., 2026) resolves this tension by using the verified solution to construct a privileged training-time view of the same model, thereby combining on-policy rollouts with dense token-level supervision without introducing a separate teacher model.

For each training example $(x, y^\star) \in \mathcal{S}$, let $z$ denote a privileged context derived from the verified solution. This framework defines two views:

$$q_{\text{base}}(\cdot \mid x) \triangleq p_\theta(\cdot \mid x), \qquad q_{\text{aug}}(\cdot \mid x, z) \triangleq p_\phi(\cdot \mid x, z).$$

The base view is the deployment-time policy; the augmented view sees the same problem plus privileged solution context and is used only during training. In this background section, we take $z = y^\star$ to describe the full-solution version. The concrete method in §3 instead uses the masked trace $z = \tilde{y}^\star$ and fixes $\phi = \theta_0$.

The baseline samples $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_{|\hat{y}|}) \sim q_{\text{base}}(\cdot \mid x)$ and evaluates both views on the same sampled prefixes. In the full-solution background form used to define the baseline, define

$$p_T^n(\cdot) \triangleq q_{\text{aug}}(\cdot \mid x, y^\star, \hat{y}_{<n}), \qquad p_S^n(\cdot) \triangleq q_{\text{base}}(\cdot \mid x, \hat{y}_{<n}).$$

The rollout discrepancy is then

$$\mathcal{D}_{\text{roll}}(\hat{y}; x, y^\star) \triangleq \frac{1}{|\hat{y}|} \sum_{n=1}^{|\hat{y}|} D\Big(p_T^n \,\Big\|\, p_S^n\Big), \tag{1}$$

where $D$ may be instantiated as forward KL, reverse KL, or a Jensen-Shannon-type divergence. The full objective is

$$\mathcal{J}_{\text{OPSD}}(\theta) = \mathbb{E}_{(x,y^\star)\sim\mathcal{S}} \mathbb{E}_{\hat{y}\sim q_{\text{base}}(\cdot|x)} \Big[\mathcal{D}_{\text{roll}}(\hat{y}; x, y^\star)\Big]. \tag{2}$$

The essential property is that supervision is dense and token-level, while the visited states still come from the policy being improved.

Our main implementation uses full-vocabulary forward KL with pointwise clipping. Given teacher targets $p_T^n$ and student distributions $p_S^n$ on the same sampled prefixes, define

$$D_{\text{fKL}}^{\text{clip}}(p_T, p_S; \hat{y}) = \frac{1}{|\hat{y}|} \sum_{n=1}^{|\hat{y}|} \sum_{v\in\mathcal{V}} \min\big(p_T^n(v)\big[\log p_T^n(v) - \log p_S^n(v)\big], \tau\big), \tag{3}$$

where $\tau$ is a pointwise clipping threshold.

## 3 METHOD

PAINT allocates training-only information while preserving on-policy student trajectories. A privileged view re-scores prefixes sampled by the deployment policy using additional evidence. Two controls determine how that evidence shapes learning: context exposure is chosen for each rollout, and target calibration is chosen for each token position. We first formalize contextual re-scoring, then instantiate these controls with adaptive partial-solution conditioning and sparse interpolation.

### 3.1 A THEORETICAL LENS: CONTEXTUAL RE-SCORING

Once the deployment policy samples a rollout $\hat{y}$, each prefix $\hat{y}_{<n}$ is evaluated under the ordinary prompt and a privileged prompt containing training-only evidence. Both views score the same visited state. The extra context changes the next-token distribution at that state, so it can supply dense guidance without requiring an alternative teacher trajectory. PAINT decides how much evidence the scorer receives and where the re-scored distribution is calibrated before distillation.

To quantify what privileged context can contribute at a student prefix $c = (x, \hat{y}_{<n})$, consider an ideal joint distribution over a desired next token $A$ and solution context $Z$. Its ordinary and privileged conditionals are $q_0(a \mid c) = \Pr(A = a \mid c)$ and $q_1(a \mid c, z) = \Pr(A = a \mid c, Z = z)$. The following identity measures the expected uncertainty removed by revealing $Z$.

**Proposition 3.1** (Information advantage of privileged context)**.** *For any such joint distribution with* $H(A \mid c) > 0$*,*

$$H(A \mid c) - \mathbb{E}_{z\sim\Pr(Z|c)} H(A \mid c, Z = z) = I(A; Z \mid c),$$

*and hence*

$$1 - \mathbb{E}_{z\sim\Pr(Z|c)}\left[\frac{H(A \mid c, Z = z)}{H(A \mid c)}\right] = \frac{I(A; Z \mid c)}{H(A \mid c)}.$$

*If* $Z_{\text{part}}$ *is a deterministic masking of* $Z_{\text{full}}$*, then* $I(A; Z_{\text{part}} \mid c) \le I(A; Z_{\text{full}} \mid c)$.

**Deployment prompt: student view**

Problem: How many ordered pairs of positive integers $(a, b)$ satisfy $\frac{1}{a} + \frac{1}{b} = \frac{1}{6}$?

Please reason step by step, and put your final answer within \boxed{}.

**Privileged scoring prompt: fixed teacher view**

Problem: How many ordered pairs of positive integers $(a, b)$ satisfy $\frac{1}{a} + \frac{1}{b} = \frac{1}{6}$?

Here is a part of reference solution to this problem:
=== Reference Solution Begin ===
Rewrite as $ab = 6a + 6b$, hence $(a-6)(b-6) = 36$. ...(omitted)...
=== Reference Solution End ===

**After reading the reference solution above, derive the same final answer in your own words.**

Please reason step by step, and put your final answer within \boxed{}.
[student rollout prefix being scored]

Figure 2: **Prompt views for privileged re-scoring.** The fixed teacher sees a masked reference solution and scores the same student-produced prefixes, yielding token-level targets without generating a separate trajectory.

The proof is in Appendix D.1. Revealing more of the reference supplies more information about the next action; masking regulates that exposure. The entropy ratio measures the fraction of uncertainty left at a sampled prefix and guides local calibration. PAINT operationalizes these two views with the current student and a frozen privileged scorer.

Although the abstract objective permits both views to share the current parameters, our main instantiation uses a fixed privileged teacher with $\phi = \theta_0$ and updates only the student parameters $\theta$. Under the re-scoring view, fixing the teacher keeps the privileged energy landscape stable and avoids having the target drift toward the student during training.

### 3.2 Overlap-Adaptive Partial Solution Conditioning

Once the privileged teacher is fixed, the next question is how much of the solution it should see. A full reference can concentrate the target on one trace; hiding too much can remove the guidance that makes privileged scoring useful. We reveal more when the rollout shares little with the reference and hide more when it already follows its structure. This design has two parts: how much of the solution to hide, and which part to hide.

We first address the amount of masking. Hiding more of the reference removes trace-specific information along with some useful guidance. To study this trade-off, consider a mixture of two targets. Let $\mathcal{X}$ be the finite set of valid continuations and let $\rho \in [0, 1)$ denote a continuous mixing weight. Let $q_0, q_1 \in \Delta(\mathcal{X})$ represent less and more hidden context, respectively. We interpolate them by

$$q_\rho(\tau) = (1 - \rho)q_0(\tau) + \rho q_1(\tau).$$

For student parameters $\theta$, define the uncentered gradient second moment

$$M_2(q, \theta) = \mathbb{E}_{\tau \sim q}\left[\|\nabla_\theta \log p_\theta(\tau)\|^2\right],$$

and let

$$v_0 = M_2(q_0, \theta), \qquad \epsilon(\theta) = M_2(q_1, \theta) - M_2(q_0, \theta).$$

Here $\epsilon$ is the change in uncentered gradient second moment between the two targets, and the entropy of $q_\rho$ measures diversity along the mixture path. By linearity of expectation,

$$M_2(q_\rho, \theta) = v_0 + \rho\epsilon.$$

**Theorem 3.2** (Monotone mask ratio under a mixture model)**.** *Assume $q_\rho(\tau) > 0$ on the relevant support and define the Shannon entropy*

$$H(\rho) = -\sum_{\tau \in \mathcal{X}} q_\rho(\tau) \log q_\rho(\tau).$$

*Assume $q_0 \neq q_1$. For coefficients $\beta, \lambda > 0$, consider the objective*

$$\mathcal{L}(\rho, \epsilon) = \beta H(\rho) - \lambda(v_0 + \rho\epsilon).$$

*Whenever an interior maximizer $\rho^\star(\epsilon) \in (0, 1)$ exists, its differentiable branch is a strictly decreasing function of the second-moment difference:*

$$\frac{d\rho^\star}{d\epsilon} < 0.$$

A detailed proof is given in Theorem D.2. The surrogate favors revealing more context when the cost of masking rises. In PAINT, rollout-reference overlap provides a readily computed signal for scheduling that context. After sampling $\hat{y} \sim q_{\text{base}}(\cdot \mid x)$, the trainer extracts a set of reference anchors $\mathcal{A}(y^\star)$, including the boxed answer, formula-like spans, and informative numbers, after normalization of common mathematical notation. It then computes a recall-style overlap score

$$\alpha(x, \hat{y}, y^\star) = \frac{1}{|\mathcal{A}(y^\star)|} \sum_{a \in \mathcal{A}(y^\star)} \mathbf{1}\{a \subseteq \text{Norm}(\hat{y})\}, \qquad \alpha \in [0, 1]. \tag{4}$$

If no anchors are found, the implementation sets $\alpha = 0$. The exact final-answer verifier is still computed for logging and analysis, but the teacher schedule is driven by this continuous overlap score rather than by a hard correct/incorrect branch. Higher overlap means that the rollout already reproduces more of the reference's mathematical structure, so it is safe to hide more. We therefore use the affine schedule

$$\rho(\alpha) = \rho_{\text{wrong}} + (\rho_{\text{correct}} - \rho_{\text{wrong}})\, \alpha, \qquad \rho_{\text{correct}} > \rho_{\text{wrong}}. \tag{5}$$

This schedule puts more solution context where overlap is low and hides more where the rollout already follows the reference. Appendix C.1 evaluates its two endpoints.

We use a contiguous suffix mask to preserve the setup and high-level plan often found near the start of a reference solution. For a fixed masking budget $k$, we study which placement retains the most information about the teacher's next action. Let

$$M_t = y^\star_{t:t+k-1}, \qquad C_t = (y^\star_{<t}, y^\star_{t+k:T}),$$

and quantify this trade-off by

$$\Gamma(t) = \mathbb{E}[I(A_n; M_t \mid x, \hat{y}_{<n}, C_t)],$$

the expected privileged information about the teacher's next action that is available only in the hidden block. Under the explicit causal information asymmetry assumption below, suffix masking minimizes this information-loss criterion.

**Theorem 3.3** (Suffix placement under information asymmetry)**.** *Assume later mask positions leave weakly more informative visible context for predicting the teacher's next action than earlier mask positions. Then $\Gamma(t)$ is non-increasing in $t$. Hence the minimum expected information loss is attained at the largest feasible mask position,*

$$t^\star = T - k + 1,$$

*which masks the suffix of length $k$.*

Appendix D proves the conditional placement result and studies an ideal posterior-mixture target for hidden completions. The mask-placement ablation in Appendix C.2 supports suffix placement in our setting. Combining the ratio schedule with suffix placement gives the privileged trace used in practice:

$$\tilde{y}^\star = M_{\text{tail}}(y^\star; \rho(\alpha)). \tag{6}$$

In the main runs, $M_{\text{tail}}$ removes a character-level suffix of the reference solution and replaces it with an omission marker. The code also supports prefix, middle, and punctuation-aware suffix masking variants, but the default setting used in the reported experiments is suffix masking.

### 3.3 SPARSE TEACHER ENERGY INTERPOLATION

After deciding what the teacher sees, we select where to calibrate its token distribution. Privileged context is meant to clarify a next step. If the teacher remains more uncertain than the student on a sampled prefix, that position is a natural candidate for a small correction toward the student's distribution. PAINT reserves a sparse budget for the strongest such relative uncertainty mismatches and leaves the teacher target intact elsewhere. Given the masked teacher prompt, define the two next-token distributions on the same student prefix:

$$p_S^n(\cdot) \triangleq q_{\text{base}}(\cdot \mid x, \hat{y}_{<n}), \qquad p_T^n(\cdot) \triangleq q_{\text{aug}}(\cdot \mid x, \tilde{y}^\star, \hat{y}_{<n}). \tag{7}$$

The trainer applies a small interpolation step in token-energy space at positions selected by a sparse budget. Let $z_S^n$ and $z_T^n$ be the raw student and teacher logits for prefix $\hat{y}_{<n}$, and let $p_S^n = \mathrm{softmax}(z_S^n/T)$ and $p_T^n = \mathrm{softmax}(z_T^n/T)$ be the temperature-scaled distributions. Define token energies

$$E_S^n(v) = -\log p_S^n(v), \qquad E_T^n(v) = -\log p_T^n(v).$$

Let

$$H_S^n = -\sum_{v\in\mathcal{V}} p_S^n(v)\log p_S^n(v), \qquad H_T^n = -\sum_{v\in\mathcal{V}} p_T^n(v)\log p_T^n(v), \tag{8}$$

and compare their uncertainty through the dimensionless ratio

$$R_n = \frac{H_T^n}{H_S^n}. \tag{9}$$

In the ideal conditional model of Proposition 3.1, this ratio is the fraction of uncertainty remaining after seeing the solution context. Its complement, $1 - R_n$, is the realized normalized entropy reduction, whose expectation is normalized conditional mutual information. We therefore prioritize the largest ratios: positions where privileged context yielded the least relative reduction. This also accounts for how uncertain the student already is, unlike teacher entropy alone. Let $\mathcal{J}$ be the valid generated-token positions, and let $\mathcal{I}$ contain the $\lceil f_{\text{top}}|\mathcal{J}|\rceil$ largest ratios within each sample, with arbitrary tie-breaking. Appendix C.3 evaluates this sparse choice. At selected positions, we move the teacher energy a small fraction toward the student energy:

$$\tilde{E}_T^n(v) = \begin{cases} (1-k_{\text{base}})E_T^n(v) + k_{\text{base}}E_S^n(v), & n \in \mathcal{I}, \\ E_T^n(v), & n \notin \mathcal{I}. \end{cases} \tag{10}$$

The reported runs use a constant $k_{\text{base}}$ and do not use an entropy-ratio scale term; the entropy ratio only determines which positions receive interpolation. Finally,

$$\tilde{p}_T^n(v) = \frac{\exp(-\tilde{E}_T^n(v))}{\sum_{u\in\mathcal{V}}\exp(-\tilde{E}_T^n(u))}. \tag{11}$$

Equivalently, on selected positions,

$$\tilde{p}_T^n(v) \propto \big(p_T^n(v)\big)^{1-k_{\text{base}}}\big(p_S^n(v)\big)^{k_{\text{base}}}.$$

This geometric form makes the calibration easy to interpret. For any two tokens $u, v$ at a selected position, normalization cancels in their odds ratio:

$$\log\frac{\tilde{p}_T^n(v)}{\tilde{p}_T^n(u)} = (1-k_{\text{base}})\log\frac{p_T^n(v)}{p_T^n(u)} + k_{\text{base}}\log\frac{p_S^n(v)}{p_S^n(u)}.$$

Thus interpolation moves each pairwise log preference only a small fraction toward the student's preference. At unselected positions the teacher's odds remain exactly as scored under the partial solution. The sparse budget therefore controls how often calibration occurs, while $k_{\text{base}}$ controls its strength at those positions.

**Proposition 3.4** (Geometric calibration contracts the teacher–student KL)**.** *For strictly positive token distributions $p_T, p_S$ and $k \in [0,1]$, let $q_k(v) \propto p_T(v)^{1-k}p_S(v)^k$. Then $q_k$ uniquely minimizes*

$$(1-k)D_{\text{KL}}(q\,\|\,p_T) + kD_{\text{KL}}(q\,\|\,p_S)$$

*over token distributions $q$. Moreover, $D_{\text{KL}}(q_k\,\|\,p_S)$ is non-increasing in $k$, with*

$$\frac{d}{dk}D_{\text{KL}}(q_k\,\|\,p_S) = -(1-k)\,\mathrm{Var}_{q_k}\left[\log\frac{p_S(v)}{p_T(v)}\right] \le 0.$$

The proof is in Appendix D.3. Thus each selected interpolation step is the exact KL barycenter of teacher and student distributions and contracts their forward-KL discrepancy. The entropy ratio determines *where* this contraction is applied; the proposition describes *what* it does at those positions.

The two controls operate at different resolutions. The masked solution is chosen once per sampled rollout and changes the privileged distribution at every prefix. The entropy ratio is computed after re-scoring and changes only the selected token targets. A partial reference can therefore guide most of a rollout directly, while positions with little relative uncertainty reduction receive a modest student-aware correction. This ordering connects the context-level information advantage to the token-level KL calibration without treating every teacher prediction as equally informative.

Substituting the interpolated teacher targets into Eq. 3 gives the rollout loss used in the main experiments:

$$\widetilde{\mathcal{D}_{\text{roll}}}(\hat{y}; x, \tilde{y}^{\star}) = D_{\text{fKL}}^{\text{clip}}\left(\{\tilde{p}_T^n\}_{n=1}^{|\hat{y}|}, \{p_S^n\}_{n=1}^{|\hat{y}|}; \hat{y}\right). \tag{12}$$

The complete update, including overlap-adaptive masking and sparse teacher energy interpolation, is summarized in Appendix A.

**Information allocation.** PAINT adapts the information supplied to the scorer, then adapts how its distribution supervises the student. A base policy supplies visited prefixes; an exposure operator selects the privileged evidence available when those prefixes are re-scored. A rollout-level relevance signal controls exposure, and a token-level uncertainty signal assigns the calibration budget. Same-prefix scoring, normalized uncertainty, and geometric interpolation operate on predictive distributions independently of the evidence format. For mathematical reasoning, anchor overlap controls a suffix-masked verified solution. The two levels of adaptation give the scorer enough context to guide a rollout while concentrating calibration at prefixes where that context resolves the least uncertainty.

# 4 Experiments

We instantiate PAINT with verified mathematical solutions and test information allocation across rollouts and token positions. PAINT and full-solution on-policy self-distillation share data, backbones, rollout budgets, and the fixed privileged teacher. We also compare against supervised imitation and reward-based RL across model scales and competition-level math benchmarks.

## 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 data comes from the mathematical reasoning subset of OpenThoughts (Guha et al., 2025); for each scale, we sample up to 30K problem-solution pairs with verified final answers and reasoning traces. Evaluation covers AIME 2024, AIME 2025, and HMMT 2025.

**Compared post-training recipes.** We evaluate four updates from the same base checkpoints. **SFT** directly imitates the reference reasoning traces and represents an off-policy token-level baseline. **GRPO** (Shao et al., 2024) optimizes binary answer rewards with grouped rollouts; its maximum generation length is 16k. **OPSD** is the closest prior on-policy self-distillation baseline (Zhao et al., 2026): it scores each student rollout with a privileged full-solution view. **PAINT** allocates the visible solution context across rollouts and calibrates the resulting targets at selected positions, using the same rollout-and-score protocol for a direct comparison.

**Training and evaluation protocol.** Unless otherwise stated, PAINT 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. PAINT conditions the teacher on an overlap-adaptive partial reference solution with $\rho_{\text{wrong}} = 0.30$ and $\rho_{\text{correct}} = 0.40$, then interpolates teacher logits toward student logits with $k_{\text{base}} = 0.03$ on the top $f_{\text{top}} = 0.03$ fraction of generated-token positions ranked by the teacher/student entropy ratio. Additional training details appear in Appendix B.

Table 1: Math reasoning Avg@12 after post-training Qwen3 backbones. Evaluation uses thinking mode, temperature 1.0, top-p 0.95, and maximum length 38k. PAINT and OPSD use the same training split, rollout budget, and fixed privileged teacher; PAINT adds only adaptive partial-solution conditioning and sparse teacher interpolation.

| 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 |
| + SFT | 72.3 | 64.2 | 42.9 | 59.8 |
| + GRPO | 76.4 | 68.9 | 46.7 | 64.0 |
| + OPSD (prior) | 77.8 | 70.8 | 45.8 | 64.8 |
| + PAINT (ours) | **79.2** | **72.2** | **49.4** | **66.9** |
| *Qwen3-4B Instruct* | | | | |
| No update | 74.9 | 66.4 | 42.2 | 61.2 |
| + SFT | 70.2 | 62.3 | 43.4 | 58.6 |
| + GRPO | 75.6 | 68.1 | 44.4 | 62.7 |
| + OPSD (prior) | 76.4 | 68.3 | 46.1 | 63.6 |
| + PAINT (ours) | **76.9** | **69.7** | **46.7** | **64.4** |
| *Qwen3-1.7B Instruct* | | | | |
| No update | 51.5 | 36.7 | 23.1 | 37.1 |
| + SFT | 48.4 | 36.3 | 22.7 | 35.8 |
| + GRPO | 51.1 | 38.3 | 23.7 | 37.7 |
| + OPSD (prior) | 57.2 | 43.9 | 29.2 | 43.4 |
| + PAINT (ours) | **59.7** | **44.2** | **29.7** | **44.5** |

### 4.2 Main Results

Table 1 shows that PAINT improves over the prior on-policy self-distillation baseline at every model scale: +2.1 points on Qwen3-8B, +0.8 on Qwen3-4B, and +1.1 on Qwen3-1.7B. Since OPSD already beats SFT by turning reference solutions into dense on-policy supervisory distributions, these gains suggest that controlling privileged context and concentrating calibration on entropy-mismatch positions matters beyond the basic re-scoring scaffold.

PAINT also matches or exceeds GRPO across all three scales using one training rollout of up to 1024 tokens per problem, compared with eight GRPO rollouts of up to 16k tokens each. PAINT checkpoints were evaluated through 100 steps; GRPO used a longer training horizon.

SFT scores lower than the base checkpoint on average at all three scales. In contrast, OPSD and PAINT supervise the student's own prefixes. PAINT further improves this setup by controlling the visible solution context and calibrating a small subset of targets.

Appendix C isolates the two PAINT components on Qwen3-1.7B. Moderate adaptive suffix masking outperforms fixed full-solution conditioning and heavier masks, suffix placement beats prefix, middle, and random contiguous masks, and sparse interpolation is strongest when applied to a small top-$f_{\text{top}}$ budget rather than every token. These ablations support the intended mechanisms.

### 4.3 Reading the Gains Across Scales and Components

The gains over OPSD are present in every benchmark–backbone cell of Table 1, although their sizes vary. On Qwen3-8B, PAINT adds 1.4 points on each AIME set and 3.6 points on HMMT 2025. On Qwen3-4B, the largest gain is 1.4 points on AIME 2025; on Qwen3-1.7B, it is 2.5 points on AIME 2024. Thus the macro improvement is not driven by the same benchmark at every scale. The consistent direction across all nine comparisons is the more useful pattern: changing the privileged target helps even when the full-solution OPSD baseline is already strong.

The Qwen3-1.7B AIME 2024 ablations show how the two controls contribute in sequence. Full-solution OPSD scores 57.2. Replacing the full reference with the selected adaptive suffix mask, without interpolation, raises the score to 58.6. Adding sparse interpolation at $k_{\text{base}} = f_{\text{top}} = 0.03$ raises it to 59.7. These comparisons use the same backbone and evaluation task, so they connect the main result to concrete changes in the teacher target. The mask-placement sweep further shows that which part of the reference remains visible matters: the suffix mask scores 58.6, compared with

57.8 for a middle mask and 56.1 for a prefix mask. Keeping the solution's opening setup visible is therefore a useful design choice in this setting.

The interpolation sweep makes the role of sparsity concrete: with the same adaptive mask and $k_{\text{base}} = 0.03$, applying interpolation everywhere scores 58.9, while selecting the top 3% of positions scores 59.7. PAINT thus retains the privileged target on most tokens and calibrates it where the relative uncertainty signal is strongest.

## 5 Related Work

Reasoning post-training includes RLVR with verifiable rewards (Shao et al., 2024; DeepSeek-AI et al., 2025; Kimi Team, 2025; Yu et al., 2025; Mistral-AI, 2025), process supervision (Lightman et al., 2024; Zhang et al., 2025), and SFT or self-improvement over curated and self-generated traces (Guha et al., 2025; Xiaomi LLM-Core Team, 2026; Zelikman et al., 2022; Gulcehre et al., 2023; Wang et al., 2023; Yuan et al., 2024). These approaches provide useful signals but differ in whether supervision is on-policy, dense, or externally annotated. RLVR directly optimizes final-answer correctness but can be statistically inefficient when rewards are sparse; process supervision gives finer feedback but requires additional labels or learned reward models; and imitation-style training depends strongly on the quality and length of reference traces. PAINT targets the complementary regime where verified answers are available but step labels and stronger external teachers are not. In this regime, the main question is how to turn a verified solution into useful token-level feedback without moving training off the student's own trajectories.

PAINT is closest to distillation and privileged-context self-training. Classical and sequence-level distillation transfer teacher distributions or generated sequences (Hinton et al., 2015; Kim & Rush, 2016; Sanh et al., 2019), while on-policy distillation supervises student-sampled trajectories (Agarwal et al., 2024; Gu et al., 2024; Xu et al., 2025; Lu & Thinking Machines Lab, 2025). Context distillation compresses richer-context behavior into a deployment-time model (Snell et al., 2022; Huang et al., 2022; Qi et al., 2025), and privileged on-policy self-distillation applies this idea to verified reasoning solutions (Zhao et al., 2026; Hübotter et al., 2026; Shenfeld et al., 2026). PAINT focuses on a further question shared by these settings: how much training-only evidence should the scorer see, and at which visited states should its distribution be calibrated?

PAINT realizes this principle in math reasoning with overlap-adaptive partial solutions and sparse energy interpolation. The student's rollout determines how much of the verified solution the scorer sees; the resulting teacher distribution is calibrated at selected prefixes. The training examples, answer verifier, and student-generated states remain fixed across the direct comparison with full-solution self-distillation, isolating information allocation as the change.

Methods trained on trajectories generated with privileged context may change both the visited states and the supervision. PAINT holds the student prefix fixed while varying the evidence used to score it. This separates the information available to the scorer from the trajectory on which the student is trained.

## 6 Conclusion

Training with privileged information raises an allocation problem: how much evidence should shape each student trajectory, and where should its token-level guidance be calibrated? PAINT solves both while keeping the visited states on-policy. With verified math solutions, it combines overlap-adaptive context exposure and sparse energy interpolation. Across three benchmarks and three Qwen3 scales, PAINT improves over full-solution on-policy self-distillation and matches or exceeds GRPO with fewer, shorter training rollouts. On Qwen3-8B, it gains 2.1 macro Avg@12 points over OPSD and 2.9 over GRPO. Information allocation is a central design principle for privileged on-policy distillation.

## 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 training objective, adaptive masking rule, and sparse interpolation procedure are specified in Section 3. Experimental protocols and principal results are reported in Section 4; the complete training algorithm, implementation settings, and ablation studies appear in Appendices A to C.

## 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.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026. doi: 10.48550/arXiv.2601.20802. URL https://arxiv.org/abs/2601.20802.

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.

Siyuan Qi, Bangcheng Yang, Kailin Jiang, Xiaobo Wang, Jiaqi Li, Yifan Zhong, Yaodong Yang, and Zilong Zheng. In-context editing: Learning knowledge from self-induced distributions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=w6rHCuN3YG.

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.

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

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.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-Instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.754. URL https://aclanthology.org/2023.acl-long.754/.

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.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E. Weston. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57905–57923. PMLR, 2024. URL https://proceedings.mlr.press/v235/yuan24d.html.

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** PAINT: complete adaptive privileged re-scoring recipe used in our main experiments

**Require:** Reasoning dataset $\mathcal{S} = \{(x_i, y_i^\star)\}_{i=1}^N$, initialized model $p_{\theta_0}$, student policy $q_{\text{base}}$, fixed privileged teacher $q_{\text{aug}}$, anchor extractor $\mathcal{A}$, mask ratios $\rho_{\text{correct}} > \rho_{\text{wrong}}$, sparse interpolation parameters $(k_{\text{base}}, f_{\text{top}})$, clipping threshold $\tau$

**while** not converged **do**
  Sample a minibatch $\mathcal{B} \subset \mathcal{S}$
  **for all** $(x, y^\star) \in \mathcal{B}$ **do**
    Sample one student rollout $\hat{y} \sim q_{\text{base}}(\cdot \mid x)$
    Set $\alpha \leftarrow 0$ if $\mathcal{A}(y^\star) = \varnothing$; otherwise set $\alpha \leftarrow |\{a \in \mathcal{A}(y^\star) : a \subseteq \text{Norm}(\hat{y})\}| / |\mathcal{A}(y^\star)|$
    Set mask ratio $\rho \leftarrow \rho_{\text{wrong}} + (\rho_{\text{correct}} - \rho_{\text{wrong}})\alpha$
    Build masked privileged trace $\tilde{y}^\star \leftarrow M_{\text{tail}}(y^\star; \rho)$
    **for** $n = 1, \ldots, |\hat{y}|$ **do**
      $p_S^n \leftarrow q_{\text{base}}(\cdot \mid x, \hat{y}_{<n})$ and $p_T^n \leftarrow q_{\text{aug}}(\cdot \mid x, \tilde{y}^\star, \hat{y}_{<n})$
      $E_S^n \leftarrow -\log p_S^n$ and $E_T^n \leftarrow -\log p_T^n$
      Compute $H_S^n$, $H_T^n$, and $R_n \leftarrow H_T^n / H_S^n$
    Let $\mathcal{I}$ be the top $f_{\text{top}}$ fraction of generated-token positions by $R_n$
    Set $\tilde{E}_T^n \leftarrow (1 - k_{\text{base}})E_T^n + k_{\text{base}}E_S^n$ for $n \in \mathcal{I}$, else $\tilde{E}_T^n \leftarrow E_T^n$
    $\tilde{p}_T^n(v) \leftarrow \exp(-\tilde{E}_T^n(v)) / \sum_{u \in \mathcal{V}} \exp(-\tilde{E}_T^n(u))$
    $\ell(x, y^\star) \leftarrow \widetilde{\mathcal{D}_{\text{roll}}}(\hat{y}; x, \tilde{y}^\star)$
  Update the student parameters underlying $q_{\text{base}}$ using $\frac{1}{|\mathcal{B}|} \sum_{(x,y^\star) \in \mathcal{B}} \ell(x, y^\star)$

# A Training Algorithm

Algorithm 1 summarizes the complete PAINT recipe. Each example contributes one on-policy rollout, one overlap-adaptive masked privileged prompt, one pair of student/teacher distributions per prefix, and one clipped forward-KL loss. Unless otherwise stated, all experiments use this fixed-teacher, suffix-masked, sparsely energy-interpolated full-vocabulary objective. For readability, the algorithm writes the interpolation in its equivalent normalized energy form; the implementation performs the same operation by interpolating logits before the final softmax.

The trainer also includes optional variants such as EMA teachers, sampled-token objectives, and reason-first teachers, but they are not part of the reported main recipe.

# B Experimental Details

Unless otherwise stated, all privileged on-policy distillation runs use the same post-training setting. We train LoRA adapters with learning rate $5 \times 10^{-6}$, effective batch size 32, rank $r = 64$, $\alpha = 128$, and target modules q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, and down_proj. Each update samples one base-policy completion per prompt with sampling temperature 1.1 and maximum completion length 1024, and scores the same prefixes with the fixed privileged teacher. For PAINT, we apply suffix masking with $\rho_{\text{wrong}} = 0.30$ and $\rho_{\text{correct}} = 0.40$, use sparse teacher-energy interpolation with $k_{\text{base}} = 0.03$ and $f_{\text{top}} = 0.03$, and clip the pointwise forward-KL term at $\tau = 0.06$. The vanilla prior self-distillation baseline removes these PAINT-specific adaptive masking and sparse interpolation components while keeping the same training budget and teacher anchoring. We train for 100 steps and evaluate checkpoints every 20 steps. 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.

# C Ablation Studies

All ablations in this section use Qwen3-1.7B and report AIME 2024 Avg@12. The first two sweeps isolate the masking design with sparse interpolation disabled; the final sweep fixes the best masking setting and ablates sparse interpolation.

### C.1 Mask-Ratio Selection

The adaptive masking schedule uses two endpoints: $\rho_{\text{wrong}}$, applied when the rollout has little overlap with the verified solution, and $\rho_{\text{correct}}$, applied when the rollout already matches more reference anchors. We report the sweep as $(\rho_{\text{correct}}, \rho_{\text{wrong}})$, with the actual per-example ratio computed as $\rho(\alpha) = \rho_{\text{wrong}} + (\rho_{\text{correct}} - \rho_{\text{wrong}})\alpha$. Table 2 evaluates this choice with suffix masking and no sparse interpolation.

Table 2: Ablation of the adaptive suffix-mask ratios with suffix masking and sparse interpolation disabled. Rows with equal endpoints correspond to a fixed mask ratio, while unequal endpoints use the overlap-adaptive schedule.

| $\rho_{\text{correct}}$ | $\rho_{\text{wrong}}$ | AIME 2024 |
|---|---|---|
| 0.00 | 0.00 | 57.2 |
| 0.25 | 0.25 | 57.8 |
| 0.50 | 0.50 | 56.9 |
| 0.75 | 0.75 | 55.8 |
| 0.50 | 0.25 | 58.1 |
| 0.50 | 0.30 | 57.8 |
| 0.40 | 0.30 | **58.6** |

The sweep supports the conservative adaptive schedule used in the main experiments. Revealing the full solution is strong but leaves little context control, while a large fixed hidden ratio can remove too much privileged information. Moderate adaptive masking performs best in this sweep: low-overlap rollouts still receive enough visible solution context for correction, whereas high-overlap rollouts hide slightly more of the suffix to avoid over-committing to a single reference trace.

### C.2 Mask Placement Strategy

We next fix the adaptive endpoints to the main setting, $(\rho_{\text{correct}}, \rho_{\text{wrong}}) = (0.40, 0.30)$, and ablate where the contiguous block is hidden. Table 3 compares prefix, middle, random contiguous-span, and suffix masking with sparse interpolation disabled.

Table 3: Mask-placement ablation with the default adaptive endpoints fixed and sparse interpolation disabled.

| Mask placement | AIME 2024 |
|---|---|
| Prefix mask | 56.1 |
| Middle mask | 57.8 |
| Random contiguous mask | 57.2 |
| Suffix mask | **58.6** |

Suffix masking is the strongest placement in this sweep. This is consistent with the visible-context argument in Theorem 3.3: early reference tokens often contain setup, decomposition, and high-level plan information, so preserving them gives the privileged scorer a coherent context while still hiding downstream execution details.

### C.3 Sparse Interpolation Strength and Budget

Finally, we fix suffix masking and the adaptive endpoints to $(\rho_{\text{correct}}, \rho_{\text{wrong}}) = (0.40, 0.30)$, then vary the interpolation strength $k_{\text{base}}$ and the selected-token fraction $f_{\text{top}}$. Table 4 reports the resulting AIME 2024 Avg@12 scores. The row $(k_{\text{base}}, f_{\text{top}}) = (0, 0)$ disables interpolation, while $f_{\text{top}} = 1.00$ applies interpolation at every generated-token position.

Sparse interpolation improves over the masking-only setting, but applying it everywhere is less effective. The best setting uses a moderate interpolation strength and a small selected-position

Table 4: Ablation of sparse energy interpolation with suffix masking and default adaptive endpoints fixed.

| $k_{\text{base}}$ | $f_{\text{top}}$ | AIME 2024 |
|---|---|---|
| 0.00 | 0.00 | 58.6 |
| 0.01 | 1.00 | 58.6 |
| 0.03 | 1.00 | 58.9 |
| 0.05 | 1.00 | 57.8 |
| 0.03 | 0.01 | 59.2 |
| 0.03 | 0.03 | **59.7** |
| 0.03 | 0.05 | 58.9 |

budget, supporting PAINT's design choice to calibrate only entropy-mismatch positions rather than all stylistic continuation positions.

# D Formal Analysis of Privileged Information, Adaptive Masking, Geometric Calibration, and Suffix Masking

**Proposition D.1** (Information advantage of privileged context)**.** *Fix a rollout prefix $c$ and any joint distribution over a desired next token $A$ and privileged solution context $Z$. If $H(A \mid c) > 0$, then*

$$H(A \mid c) - \mathbb{E}_{z\sim \Pr(Z|c)} H(A \mid c, Z = z) = I(A; Z \mid c),$$

*and*

$$1 - \mathbb{E}_{z\sim \Pr(Z|c)}\left[\frac{H(A \mid c, Z = z)}{H(A \mid c)}\right] = \frac{I(A; Z \mid c)}{H(A \mid c)}.$$

*If $Z_{\text{part}}$ is a deterministic function of $Z_{\text{full}}$, then*

$$I(A; Z_{\text{part}} \mid c) \le I(A; Z_{\text{full}} \mid c).$$

*Proof.* By the definition of conditional mutual information at the fixed prefix $c$,

$$\begin{aligned} I(A; Z \mid c) &= \mathbb{E}_{a,z|c}\left[\log \frac{\Pr(A = a \mid c, Z = z)}{\Pr(A = a \mid c)}\right] \\ &= H(A \mid c) - \mathbb{E}_{z|c} H(A \mid c, Z = z). \end{aligned}$$

Dividing by the positive quantity $H(A \mid c)$ gives the normalized identity. The final inequality follows from the conditional data-processing inequality for the Markov chain $A \to Z_{\text{full}} \to Z_{\text{part}}$ at the fixed prefix $c$. □

**Theorem D.2** (Monotone mask ratio under a mixture model)**.** *Let $\mathcal{X}$ be a finite set of valid continuations and let $\rho \in [0, 1)$ denote a surrogate mixture weight for targets representing different amounts of hidden context. Let*

$$q_\rho(\tau) = (1 - \rho)q_0(\tau) + \rho q_1(\tau), \qquad \tau \in \mathcal{X},$$

*where $q_0, q_1 \in \Delta(\mathcal{X})$, $q_0 \ne q_1$, and $q_\rho(\tau) > 0$ on the relevant support. For student parameters $\theta$, define*

$$M_2(q, \theta) = \mathbb{E}_{\tau\sim q}\left[\|\nabla_\theta \log p_\theta(\tau)\|^2\right], \qquad v_0 = M_2(q_0, \theta), \qquad \epsilon(\theta) = M_2(q_1, \theta) - M_2(q_0, \theta).$$

*Let*

$$H(\rho) = -\sum_{\tau\in\mathcal{X}} q_\rho(\tau) \log q_\rho(\tau),$$

*and consider the objective*

$$\mathcal{L}(\rho, \epsilon) = \beta H(\rho) - \lambda(v_0 + \rho\epsilon), \qquad \beta, \lambda > 0.$$

*Whenever an interior maximizer $\rho^\star(\epsilon) \in (0, 1)$ exists, its differentiable branch is a strictly decreasing function of the second-moment difference:*

$$\frac{d\rho^\star}{d\epsilon} < 0.$$

*Proof.* By linearity of expectation under a convex combination of measures,

$$\begin{aligned} M_2(q_\rho, \theta) &= \mathbb{E}_{\tau \sim q_\rho}\left[\|\nabla_\theta \log p_\theta(\tau)\|^2\right] \\ &= (1-\rho) M_2(q_0, \theta) + \rho M_2(q_1, \theta) \\ &= (1-\rho) v_0 + \rho(v_0 + \epsilon) = v_0 + \rho\epsilon. \end{aligned}$$

Thus the objective can be written as

$$\mathcal{L}(\rho, \epsilon) = \beta H(\rho) - \lambda(v_0 + \rho\epsilon).$$

Next, because $q_\rho$ is affine in $\rho$, its Shannon entropy is strictly concave along this path. Differentiating twice gives

$$H''(\rho) = -\sum_{\tau \in \mathcal{X}} \frac{(q_1(\tau) - q_0(\tau))^2}{q_\rho(\tau)} < 0,$$

where strict negativity follows because $q_\rho(\tau) > 0$ on the relevant support and $q_0 \neq q_1$.

For an interior maximizer $\rho^\star(\epsilon)$, the first-order condition is

$$\frac{\partial \mathcal{L}}{\partial \rho} = \beta H'(\rho^\star) - \lambda\epsilon = 0.$$

Define

$$\mathcal{G}(\rho, \epsilon) = \beta H'(\rho) - \lambda\epsilon.$$

Since

$$\frac{\partial \mathcal{G}}{\partial \rho} = \beta H''(\rho^\star) \neq 0,$$

the implicit function theorem applies and yields

$$\frac{d\rho^\star}{d\epsilon} = -\frac{\partial \mathcal{G}/\partial \epsilon}{\partial \mathcal{G}/\partial \rho} = -\frac{-\lambda}{\beta H''(\rho^\star)} = \frac{\lambda}{\beta H''(\rho^\star)}.$$

Because $\lambda > 0$, $\beta > 0$, and $H''(\rho^\star) < 0$, the ratio is strictly negative:

$$\frac{d\rho^\star}{d\epsilon} < 0.$$

This proves the claim. □

**Proposition D.3** (Geometric calibration as a KL barycenter)**.** *Let $p_T$ and $p_S$ be strictly positive distributions on a finite vocabulary, and let $k \in [0, 1]$. Define*

$$q_k(v) = \frac{p_T(v)^{1-k} p_S(v)^k}{\sum_u p_T(u)^{1-k} p_S(u)^k}.$$

*Then $q_k$ is the unique minimizer over distributions $q$ of*

$$(1-k) D_{\mathrm{KL}}(q \,\|\, p_T) + k D_{\mathrm{KL}}(q \,\|\, p_S).$$

*Furthermore, $D_{\mathrm{KL}}(q_k \,\|\, p_S)$ is non-increasing in $k$, and*

$$\frac{d}{dk} D_{\mathrm{KL}}(q_k \,\|\, p_S) = -(1-k) \operatorname{Var}_{q_k}\left[\log \frac{p_S(v)}{p_T(v)}\right].$$

*Proof.* Write $L(v) = \log[p_S(v)/p_T(v)]$ and $\psi(k) = \log \sum_v p_T(v) e^{kL(v)}$. Then $\log q_k(v) = \log p_T(v) + kL(v) - \psi(k)$. For any distribution $q$,

$$\begin{aligned} &(1-k) D_{\mathrm{KL}}(q \,\|\, p_T) + k D_{\mathrm{KL}}(q \,\|\, p_S) \\ &\quad = D_{\mathrm{KL}}(q \,\|\, q_k) - \psi(k). \end{aligned}$$

Gibbs' inequality gives a unique minimum at $q = q_k$. Since $\psi'(k) = \mathbb{E}_{q_k}[L]$ and $\psi''(k) = \operatorname{Var}_{q_k}[L]$, we also have

$$D_{\mathrm{KL}}(q_k \,\|\, p_S) = (k-1)\psi'(k) - \psi(k).$$

Differentiation yields the stated non-positive derivative. It is strictly negative for $k < 1$ whenever $p_T \neq p_S$. □

**Proposition D.4** (Entropy smoothing)**.** *Let $y^\star = (y_1^\star, \dots, y_T^\star)$, let $M_t = y_{t:t+k-1}^\star$, let $C_t = (y_{<t}^\star, y_{t+k:T}^\star)$, and let $\Xi_t^n = (x, \hat{y}_{<n}, C_t)$. For*

$$q_{m,t}^n(a) = \Pr(A_n = a \mid \Xi_t^n, M_t = m), \qquad \bar{q}_t^n(a) = \sum_m \rho_t(m \mid \Xi_t^n) q_{m,t}^n(a),$$

*where $\rho_t(m \mid \Xi_t^n) = \Pr(M_t = m \mid \Xi_t^n)$,*

$$H(\bar{q}_t^n) \geq \mathbb{E}_{m \sim \rho_t(\cdot \mid \Xi_t^n)} \left[ H(q_{m,t}^n) \right].$$

*Equality holds if and only if $q_{m,t}^n$ is identical for all $m$ with positive posterior probability, up to events of zero probability.*

*Proof.* Let $f(z) = -z \log z$, with the usual convention $0 \log 0 = 0$. The function $f$ is concave on $[0, 1]$ and strictly concave on $(0, 1)$. Since

$$\bar{q}_t^n(a) = \sum_m \rho_t(m \mid \Xi_t^n) q_{m,t}^n(a),$$

Jensen's inequality gives, for each token $a$,

$$f(\bar{q}_t^n(a)) \geq \sum_m \rho_t(m \mid \Xi_t^n) f(q_{m,t}^n(a)).$$

Summing over $a \in \mathcal{V}$ yields the entropy inequality. The equality condition follows from the strict concavity of $f$ on the positive-probability coordinates: equality can hold only when the token probabilities agree across all posterior-supported hidden blocks. □

**Lemma D.5** (Forward-KL decomposition)**.** *Under the setup of Proposition D.4, let $p_S^n(a) = q_{\text{base}}(a \mid x, \hat{y}_{<n})$ be the student distribution, with support containing the support of every $q_{m,t}^n$. Then*

$$\mathbb{E}_{m \sim \rho_t(\cdot \mid \Xi_t^n)} \left[ D_{\text{KL}}(q_{m,t}^n \,\|\, p_S^n) \right] = D_{\text{KL}}(\bar{q}_t^n \,\|\, p_S^n) + I(A_n; M_t \mid \Xi_t^n).$$

*Proof.* Expanding the left-hand side,

$$\begin{aligned}
&\sum_m \rho_t(m \mid \Xi_t^n) \sum_a q_{m,t}^n(a) \log \frac{q_{m,t}^n(a)}{p_S^n(a)} \\
&= \sum_m \rho_t(m \mid \Xi_t^n) \sum_a q_{m,t}^n(a) \log \frac{q_{m,t}^n(a)}{\bar{q}_t^n(a)} + \sum_m \rho_t(m \mid \Xi_t^n) \sum_a q_{m,t}^n(a) \log \frac{\bar{q}_t^n(a)}{p_S^n(a)}.
\end{aligned}$$

The first term is exactly $I(A_n; M_t \mid \Xi_t^n)$ under the joint distribution

$$\Pr(M_t = m, A_n = a \mid \Xi_t^n) = \rho_t(m \mid \Xi_t^n) q_{m,t}^n(a).$$

For the second term, use $\sum_m \rho_t(m \mid \Xi_t^n) q_{m,t}^n(a) = \bar{q}_t^n(a)$ to obtain

$$\sum_a \bar{q}_t^n(a) \log \frac{\bar{q}_t^n(a)}{p_S^n(a)} = D_{\text{KL}}(\bar{q}_t^n \,\|\, p_S^n).$$

This proves the identity. □

**Proposition D.6** (Rao–Blackwellized gradient variance)**.** *Under the setup of Proposition D.4, assume the target distributions are treated as fixed quantities, as in the frozen-teacher implementation. Define*

$$G_{m,t}^n(\theta) = \nabla_\theta D_{\text{KL}}(q_{m,t}^n \,\|\, p_S^n), \qquad \bar{G}_t^n(\theta) = \nabla_\theta D_{\text{KL}}(\bar{q}_t^n \,\|\, p_S^n).$$

*Then*

$$\mathbb{E}_{m \sim \rho_t(\cdot \mid \Xi_t^n)} \left[ G_{m,t}^n(\theta) \right] = \bar{G}_t^n(\theta).$$

*Moreover, over the joint randomness of examples, rollout prefixes, positions, and hidden blocks,*

$$\text{Cov}(G_{M_t,t}^n) = \text{Cov}(\bar{G}_t^n) + \mathbb{E}\big[\text{Cov}(G_{M_t,t}^n \mid \Xi_t^n)\big],$$

*so the covariance difference is positive semidefinite.*

*Proof.* Because the target is held fixed when differentiating,

$$G^n_{m,t}(\theta) = -\sum_a q^n_{m,t}(a)\nabla_\theta \log p^n_S(a).$$

Taking the posterior expectation over $m$ gives

$$\mathbb{E}_m[G^n_{m,t}(\theta) \mid \Xi^n_t] = -\sum_a \bar{q}^n_t(a)\nabla_\theta \log p^n_S(a) = \bar{G}^n_t(\theta).$$

Conditioning on the sampled prefixes and treating rollout sampling as fixed for the supervised distillation update, the covariance identity is the law of total covariance applied to the random vector $G^n_{M_t,t}$ with conditioning variable $\Xi^n_t$:

$$\mathrm{Cov}(G^n_{M_t,t}) = \mathrm{Cov}(\mathbb{E}[G^n_{M_t,t} \mid \Xi^n_t]) + \mathbb{E}[\mathrm{Cov}(G^n_{M_t,t} \mid \Xi^n_t)].$$

Substituting $\mathbb{E}[G^n_{M_t,t} \mid \Xi^n_t] = \bar{G}^n_t$ gives the stated result. Conditional covariance matrices are positive semidefinite, so the final term is positive semidefinite. □

**Lemma D.7** (Visible-context duality)**.** *Let $M_t = y^\star_{t:t+k-1}$ and $C_t = (y^\star_{<t}, y^\star_{t+k:T})$ for a fixed masking budget $k$, and define*

$$\Gamma(t) = \mathbb{E}_{(x,y^\star),\hat{y},n}\left[I(A_n; M_t \mid x, \hat{y}_{<n}, C_t)\right],$$

*where $n$ is sampled uniformly from the rollout positions. Then*

$$\Gamma(t) = K - \mathbb{E}_{(x,y^\star),\hat{y},n}\left[I(A_n; C_t \mid x, \hat{y}_{<n})\right],$$

*where*

$$K = \mathbb{E}_{(x,y^\star),\hat{y},n}\left[I(A_n; y^\star \mid x, \hat{y}_{<n})\right]$$

*does not depend on $t$.*

*Proof.* For a fixed $t$ and $k$, the pair $(C_t, M_t)$ is a deterministic partition of the same reference solution $y^\star$. By the chain rule for mutual information,

$$I(A_n; C_t, M_t \mid x, \hat{y}_{<n}) = I(A_n; C_t \mid x, \hat{y}_{<n}) + I(A_n; M_t \mid x, \hat{y}_{<n}, C_t).$$

Since $(C_t, M_t)$ and $y^\star$ carry the same information,

$$I(A_n; C_t, M_t \mid x, \hat{y}_{<n}) = I(A_n; y^\star \mid x, \hat{y}_{<n}).$$

Rearranging and taking expectations gives the result. The first term on the right is independent of the mask position because it conditions on the complete reference solution. □

**Assumption D.8** (Causal information asymmetry)**.** For a fixed masking budget $k$, later mask positions leave a weakly more informative visible context for predicting the teacher's next action: for every $t_1 < t_2$,

$$\mathbb{E}\left[I(A_n; C_{t_1} \mid x, \hat{y}_{<n})\right] \le \mathbb{E}\left[I(A_n; C_{t_2} \mid x, \hat{y}_{<n})\right]$$

holds.

**Theorem D.9** (Suffix placement under information asymmetry)**.** *Under Lemma D.7 and Assumption D.8, $\Gamma(t)$ is non-increasing in $t$, and the minimum expected information loss is attained at the largest feasible mask position,*

$$t^\star = T - k + 1,$$

*which masks the suffix of length $k$.*

*Proof.* Let $t_1 < t_2$. By Lemma D.7,

$$\Gamma(t_1) - \Gamma(t_2) = \mathbb{E}[I(A_n; C_{t_2} \mid x, \hat{y}_{<n})] - \mathbb{E}[I(A_n; C_{t_1} \mid x, \hat{y}_{<n})].$$

The assumed causal information asymmetry makes the right-hand side nonnegative, so $\Gamma(t_1) \ge \Gamma(t_2)$. Therefore $\Gamma(t)$ is non-increasing in $t$, and its minimum over $t \in \{1, \dots, T-k+1\}$ is achieved at $t = T - k + 1$. □