BEYOND OUTPUT-LEVEL GUIDANCE: LEVERAGING INTERNAL EXPLORATION FOR LLM SELF-DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation (OPSD) provides dense token-level supervision for language model reasoning, but conditioning the self-teacher on privileged information may suppress the exploration needed by the student at inference time. By tracing middle-layer distributions throughout post-training from a interpretability perspective, we characterize how different training paradigms reshape the model's internal exploration–exploitation dynamics. In contrast to GRPO, OPSD progressively contracts latent exploration, and this contraction coincides with late-stage reasoning degradation. Motivated by this observation, we introduce *retained exploration*, a token-level signal that captures whether semantically diverse exploration persists across layers and identifies structurally important positions that shape subsequent reasoning trajectories. Based on this signal, we propose *Retained-Exploration-Guided Self-Distillation (RESD)*, which preserves exploration-supported teacher guidance while counteracting potentially over-concentrated preferences at low-retained-exploration positions. Across three models and five mathematical reasoning benchmarks, RESD consistently outperforms strong RLVR and self-distillation baselines.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR), exemplified by Group Relative Policy Optimization (GRPO) (Shao et al., 2024), has emerged as a dominant post-training paradigm for enhancing the reasoning capabilities of language models. However, in complex long-horizon reasoning, GRPO-style methods assign coarse sequence-level advantages even though tokens can play substantially different roles in determining the reasoning trajectory. To provide dense on-policy token-level supervision, On-policy Distillation (OPD) (Agarwal et al., 2024) introduces a stronger teacher model on student-generated trajectories. On-policy Self-Distillation (OPSD) (Zhao et al., 2026a) further removes the dependence on an external white-box teacher by using the same model conditioned on privileged information (PI), such as a verified solution, as the self-teacher.

Although OPSD has recently emerged as a promising post-training paradigm, prior work has shown that PI-conditioned supervision may introduce a train–inference mismatch and degrade the student's reasoning performance (Zhu et al., 2026; Kaur et al., 2026; Pan et al., 2026). Conditioning on a verified solution can narrow the reasoning alternatives considered by the self-teacher, whereas the student must still explore and revise candidate reasoning paths without access to such information at inference time. Uniformly imitating the PI-conditioned teacher at every token position, regardless of whether its internal states retain sufficient exploration, may therefore suppress the exploration required for difficult reasoning. As a result, it becomes a central issue to determine where and how the dense self-teacher supervision should be applied while preserving the student's internal ability to explore diverse reasoning paths.

Existing methods address this issue using teacher–student disagreement (Yang et al., 2026), predictive uncertainty (Zhang et al., 2026), or purification of privilege-induced distribution shifts (Shen et al., 2026; Pan et al., 2026) to augment outcome-level RLVR objectives with dense token-level self-distillation signals. However, these final-prediction-based signals cannot distinguish whether a teacher prediction is supported by diverse internal reasoning alternatives or instead arises from

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.

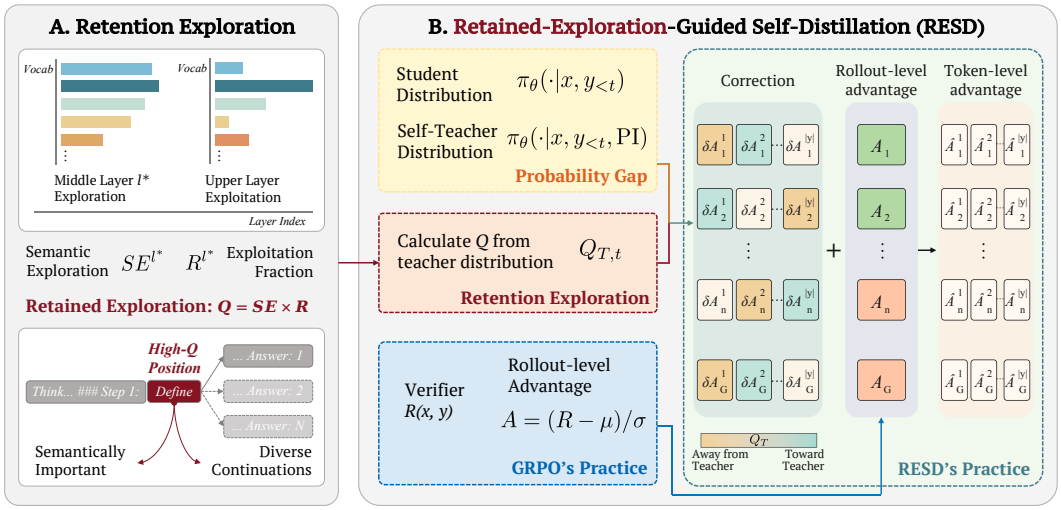


Figure 1: Overview of the RESD method.

a PI-conditioned state whose exploration has already substantially contracted. Such overly concentrated teacher states may in turn suppress the student's exploration. We therefore examine this distinction through internal-layer representations. Transformer-based language models exhibit an exploration–exploitation pattern across layers, with middle layers maintaining broader candidate reasoning states that provide a natural view of latent exploration before upper layers progressively integrate and prune them toward the final prediction (Ali et al., 2025; Tan et al., 2026). This perspective motivates studying how post-training reshapes internal exploration–exploitation dynamics and how these dynamics can inform the use of self-teacher supervision.

We begin by comparing the exploration–exploitation dynamics during GRPO and OPSD training. Using internal-layer predictive distributions, we identify a stable exploration layer and find that GRPO largely preserves latent exploration, whereas OPSD progressively contracts it. Since self-distillation is applied through token-level teacher distributions, we further introduce *retained exploration* to quantify whether individual teacher predictions are supported by semantically diverse reasoning alternatives that remain relevant through subsequent layers. Analyses along multiple dimensions show that high-retained-exploration positions concentrate on important reasoning operations and yield more diverse downstream continuations, indicating meaningful reasoning forks.

Building on these findings, we propose **Retained-Exploration-Guided Self-Distillation (RESD)**, which uses the teacher's internal exploration state to regulate dense token-level supervision. RESD guides the student toward the teacher at high-retained-exploration positions while applying reverse corrections at low-retained-exploration positions. This design retains exploration-supported teacher guidance while using reverse corrections to counteract potentially over-concentrated PI-conditioned preferences, thereby mitigating the exploration contraction associated with uniform self-distillation.

Experiments on multiple mathematical reasoning benchmarks and models show that our method consistently improves over strong RLVR and self-distillation baselines. Our main contributions are as follows:

- We systematically characterize the contrasting evolution of latent exploration under GRPO and OPSD, showing that OPSD's progressive exploration contraction is accompanied by late-stage teacher–student distribution mismatch and reasoning performance degradation.
- We introduce a retained-exploration signal derived from the self-teacher's internal states which is able to identify meaningful reasoning forks with obvious consequential downstream effects.
- We propose RESD, which regulates self-distillation using teacher-side retained exploration and achieves consistent reasoning improvements.

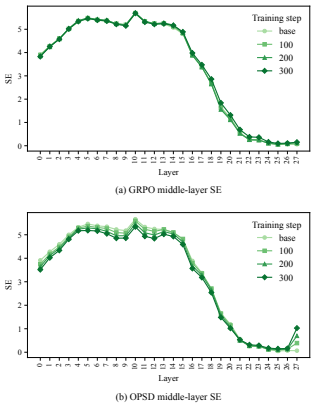


Figure 2: Layer-wise semantic exploration at different post-training checkpoints.

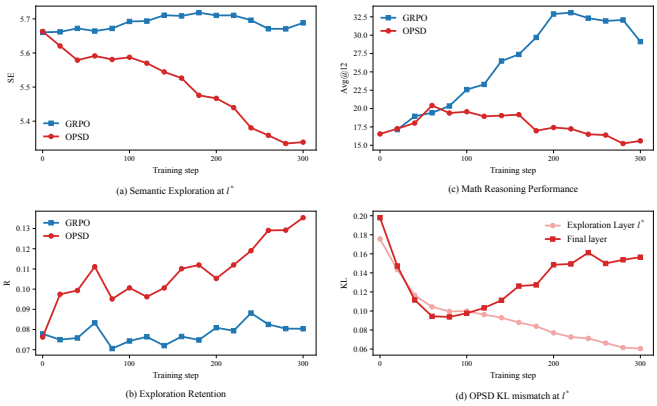


Figure 3: Latent exploration dynamics at the exploration layer during post-training, together with reasoning performance and OPSD teacher–student KL mismatch.

2 Latent Exploration Dynamics in Post-Training

2.1 Layer-wise Localization of Latent Exploration

During a single forward pass of a transformer-based language model, layers at different depths can exhibit distinct functional roles. Middle layers tend to preserve a broader set of candidate reasoning states, while upper layers progressively narrow this space toward the final prediction through exploitation and pruning (Tan et al., 2025; Ali et al., 2025).

To characterize the reasoning states encoded at an intermediate layer, we project its hidden states into the model's vocabulary space using the unembedding matrix $W_U \in \mathbb{R}^{d_{\text{vocab}} \times d_{\text{model}}}$ (Belrose et al., 2023). Specifically, given the hidden state $h_t^l \in \mathbb{R}^{d_{\text{model}}}$ at token position t and layer l, the intermediate-layer predictive distribution is

$$p_t^l = \text{softmax}\left(W_U \text{Norm}\left(h_t^l\right)\right), \tag{1}$$

where Norm denotes the model's final normalization operator. Since p_t^l lies in the same vocabulary space as the final distribution, its top-K candidates provide vocabulary-level reasoning alternatives at position t (Tan et al., 2026).

Based on this intermediate distribution, a natural measurement of latent exploration is its entropy, $H_t^l = -\sum_v p_t^l(v) \log p_t^l(v)$. Prior work (Ali et al., 2025; Tan et al., 2025) has used H^l to characterize uncertainty and the richness of the exploration space. However, high entropy may reflect either genuinely different reasoning alternatives, which indicates meaningful exploration, or merely uncertainty among semantically similar candidates that does not correspond to distinct reasoning paths. To distinguish the former from the latter, we further incorporate the semantic diversity SD_t^l among the top-K candidates and define the layer-wise semantic exploration score as

$$SE_t^l = H_t^l \cdot SD_t^l, \qquad SD_t^l = \sum_{i \neq j} \overline{p_i} \cdot \overline{p_j} \cdot (1 - \cos(e_i, e_j)). \tag{2}$$

Here, H_t^l measures predictive uncertainty, while SD_t^l captures semantic diversity among the top-K candidates weighted by their re-normalized probabilities $\overline{p}$, where $e_i = W_U[i, :]$ is the unembedding vector corresponding to candidate token i. The product SE_t^l therefore measures whether an intermediate state is both uncertain and semantically diverse, i.e., how richly it encodes latent reasoning alternatives. This definition is similar to Zhao et al. (2026b), while we apply it from final logits to middle-layer distributions.

Based on this metric, we compare SE^l across training checkpoints under GRPO and OPSD. As shown in Figure 2, the layer-wise pattern remains remarkably stable despite substantial parameter updates. Middle layers consistently form a high-exploration region, followed by a sharp decline in upper layers. Moreover, the layer with the highest SE remains unchanged throughout training under both methods. Similar layer-wise patterns on Qwen3-4B are provided in Appendix D.3.

We therefore define the exploration layer l^* as the middle-range layer with the highest semantic exploration SE^l before post-training (layer 10 for Qwen3-1.7B as in Figure 2). The following analysis fixes l^* throughout training and examines how the richness of its latent exploration space, as well as its subsequent exploitation in upper layers, evolves under different post-training paradigms.

2.2 Evolution of Latent Exploration during Post-Training

At the fixed exploration layer l^*, SE^{l^*} characterizes the richness of latent reasoning alternatives, but does not indicate how much of this exploration is carried forward through upper layers. We therefore define an exploration retention metric R^{l^*} that jointly measures whether the candidates identified at l^* retain probability mass and semantic diversity during subsequent exploitation:

$$R_t^{l^*} = \frac{1}{L - l^*} \sum_{l=l^*+1}^{L} r_t^l \cdot \min\left(1, \frac{SD_t^l}{SD_t^{l^*}}\right), \qquad r_t^l = \sum_{j \in \text{topK}_t} p_t^l(j). \tag{3}$$

Here, r_t^l tracks the probability mass retained by the top-K candidates identified at l^*. Averaged across upper layers, $R_t^{l^*} \in [0, 1]$ therefore quantifies the extent to which exploration at l^* is retained during subsequent exploitation.

As shown in Figure 3, both SE and R remain relatively stable throughout GRPO training while the model's reasoning performance steadily improves, suggesting that GRPO enhances reasoning without substantially reshaping the model's latent exploration–exploitation dynamics. In contrast, OPSD exhibits markedly different dynamics. Early in training, SE decreases while R increases, indicating a narrower exploration space whose remaining alternatives are more strongly preserved through upper layers. Meanwhile, both exploration-layer and final-layer teacher–student KL decrease as reasoning performance improves. Later, SE continues to decline and the exploration-layer KL keeps decreasing, whereas the final-layer KL begins to increase. This growing mismatch between increasingly teacher-aligned internal exploration and the student's final prediction coincides with performance degradation, suggesting that continued PI-conditioned self-distillation progressively over-constrains latent exploration.

Overall, GRPO preserves relatively stable latent exploration but provides only coarse outcome-level supervision, whereas OPSD provides dense token-level guidance while progressively contracting the student's exploration space. These observations motivate regulating self-teacher supervision according to the internal exploration underlying each teacher prediction, which forms the basis of our method in Section 3.

3 Retained-Exploration-Guided Self-Distillation

3.1 Retained Exploration from the Privileged Teacher

In Section 2.2, we characterize training-level exploration dynamics using metrics averaged over tokens and rollouts. Since self-distillation operates through token-wise teacher distributions, we next characterize the internal exploration underlying token-level teacher predictions. For the privileged self-teacher at position t, we define its **retained exploration** as

$$Q_{T,t} = SE_{T,t}^{l^*} \cdot R_{T,t}^{l^*}. \tag{4}$$

Here, R_T measures how much of the semantic exploration represented by SE_T is retained through subsequent layers. A high $Q_{T,t}$ therefore requires both semantically diverse reasoning alternatives at the exploration layer and their continued retention in upper layers, filtering out both uncertainty among semantically similar candidates and transient exploration that is not carried forward. Thus, Q_T distinguishes teacher predictions supported by rich, retained exploration from those arising from more concentrated PI-conditioned states.

In this section, we evaluate whether Q_T identifies meaningful reasoning forks from two complementary aspects: (i) whether high-Q_T positions correspond to semantically important reasoning operations, and (ii) whether they lead to more diverse continuations. We compare Q with entropy H, semantic exploration SE, and $Q' = H \cdot R_{\text{w/o SD}}$, where $R_{\text{w/o SD}}$ retains only the probability-mass component, to isolate the contribution of each component in Q.

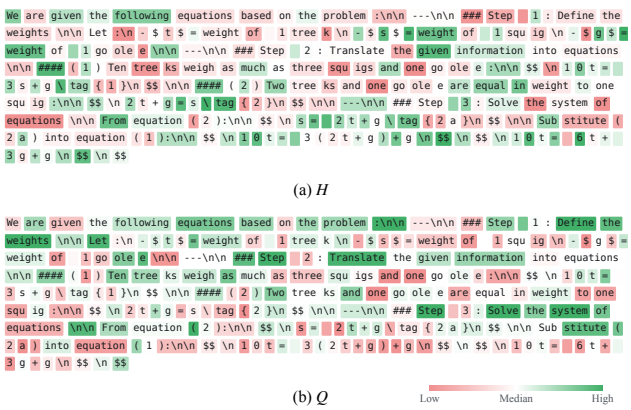


Figure 4: Token-level visualization of entropy H and retained exploration Q. High-Q regions frequently occur at important reasoning points (i.e., the beginning of each reasoning step).

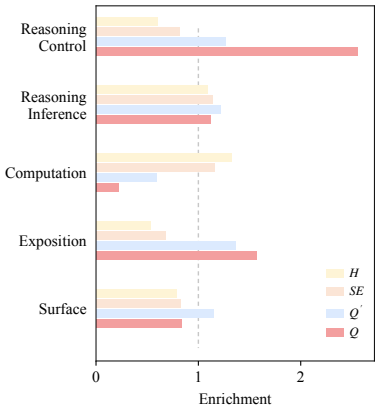


Figure 5: Semantic enrichment among top positions selected by different metrics.

3.1.1 Semantic Characteristics of High-Retained-Exploration Positions

We first examine whether high-Q_T tokens correspond to semantically important reasoning positions. Figure 4 visualizes H and Q_T along a mathematical reasoning trajectory part, where green and red indicate values above and below the median, respectively. The two metrics emphasize clearly different regions: high-Q_T positions frequently occur at reasoning-stage transitions and strategy selection, whereas high-H regions contain more local computation and formatting tokens. Full-trajectory visualizations for all four metrics are provided in Appendix D.1, which further illustrate these differences.

We further quantify this pattern by assigning each token in the trajectories to one of five semantic categories using an LLM judge: *reasoning control*, which selects, redirects, or verifies the reasoning process; *reasoning inference*, which advances intermediate reasoning; *computation*, which executes a selected operation; *exposition*, which restates or explains existing information; and *surface*, which serves syntactic or formatting functions. Full category definitions and the judging protocol are provided in Appendix A.1. For each metric M, we measure category C_k's enrichment as:

$$E_k(M) = \frac{P(C_k | t \in \text{Top5\%}(M))}{P(C_k)}, \qquad M \in \{H, SE, Q', Q\}. \tag{5}$$

As shown in Figure 5, Q shows the strongest enrichment for reasoning-control positions while substantially suppressing computation tokens. In contrast, H, SE, and Q' show much weaker selectivity toward reasoning control. Together with the qualitative visualization, these results indicate that combining semantic diversity with cross-layer retention makes Q_T more selective for semantically important reasoning positions.

3.1.2 Reasoning Branch Diversity at High-Retained-Exploration Positions

Semantic characteristics alone do not establish that high-Q_T positions function as actual reasoning forks. We therefore test whether resampling from these positions produces more diverse downstream trajectories and final answers. For each metric and rollout, we select the five highest-scoring positions and, from each preserved prefix $y_{<t}$, independently sample $N = 32$ suffix continuations from the student model.

Following Zhou et al. (2026), we measure diversity using *normalized edit distance*, which averages pairwise token-level Levenshtein distance normalized by suffix length, and *distinct-answer fraction*, the proportion of unique final answers among the N continuations. We also include a *Random* baseline that samples five positions uniformly from each rollout. Details on position selection, resampling, and metric computation are provided in Appendix A.2.

As shown in Table 1, positions selected by Q yield the highest continuation diversity under both metrics, whereas H, SE, and Q' remain close to or below the random baseline. Together with

Table 1: Continuation diversity when resampling from positions selected by different metrics.

Metric	H	SE	Q'	Q	Random
Normalized edit distance	0.6503	0.6605	0.6703	**0.7443**	0.6659
Distinct-answer fraction	0.1537	0.1614	0.1745	**0.1852**	0.1700

the semantic analysis above, this shows that high-retained-exploration positions are not only semantically important but also correspond to reasoning forks that support substantially more diverse reasoning trajectories and final answers. We therefore use $Q_{T,t}$ to construct the routing signal for token-level self-distillation in the following section.

3.2 Retained-Exploration-Guided Self-Distillation

Section 2.2 shows that uniform self-distillation is accompanied by progressive contraction of the student's latent exploration space and late-stage reasoning degradation. Based on this observation, we propose **Retained-Exploration-Guided Self-Distillation (RESD)**, which augments the sparse rollout-level RLVR advantage with dense token-level self-distillation corrections whose direction and strength are regulated by the retained exploration $Q_{T,t}$.

For each rollout i from a GRPO-style rollout group G, we first define a rollout-level base advantage

$$B^{(i)} = \begin{cases} +a_h, & \text{all-correct group,} \\ -a_h, & \text{all-wrong group,} \\ A_G^{(i)}, & \text{otherwise,} \end{cases} \tag{6}$$

where $A_G^{(i)}$ is the standard GRPO advantage. Mixed groups retain the original group-relative advantage, whereas all-correct and all-wrong groups are assigned small positive and negative base values $\pm a_h$, respectively, with a_h denoting the base advantage. This provides these groups with a rollout-level correctness signal before adding token-level self-distillation corrections.

We then extract the dense token-level self-distillation signal from the PI-conditioned teacher. For each sampled token y_t, we compute the teacher–student log-probability difference

$$d_t^{(i)} = \log p_T(y_t \mid x, y_{<t}, \mathrm{PI}) - \log p_S(y_t \mid x, y_{<t}). \tag{7}$$

The sign of $d_t^{(i)}$ indicates whether the teacher assigns stronger or weaker support to the sampled token than the student. We normalize and clip $d_t^{(i)}$ within each rollout and softly down-weight negligible differences, yielding the normalized correction signal $\tilde{d}_t^{(i)}$.

We use retained exploration to route the normalized teacher signal $\tilde{d}_t^{(i)}$. We normalize $Q_{T,t}$ within each rollout and map its distance from the ρ-quantile threshold to a routing coefficient $r_t \in (-1, 1)$. Positions above the threshold preserve the teacher-directed correction ($r_t > 0$), whereas positions below it reverse the correction ($r_t < 0$). The magnitude $|r_t|$ increases with the distance from the threshold.

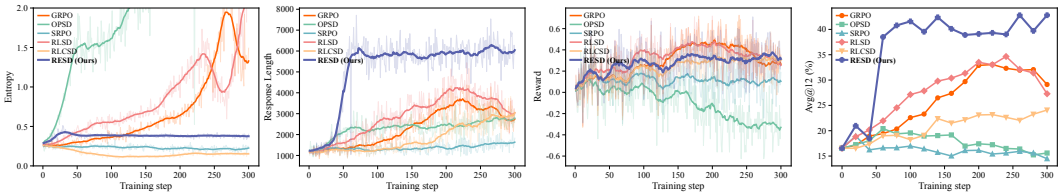
Combining the rollout-level reward signal with the routed self-distillation correction, we define the RESD advantage as

$$\hat{A}_t^{(i)} = B^{(i)} + \beta r_t^{(i)} \tilde{d}_t^{(i)}. \tag{8}$$

Here, $B^{(i)}$ supplies rollout-level reward supervision, while $\beta r_t^{(i)} \tilde{d}_t^{(i)}$ provides token-level corrections. High-Q_T predictions are supported by diverse reasoning alternatives retained across layers, motivating teacher-directed corrections. Low-Q_T predictions, in contrast, may reflect limited or weakly retained exploration, where an auxiliary reverse correction can counteract potentially over-concentrated teacher preferences and help preserve the student's exploration.

Finally, we replace the sequence-level advantage in the standard GRPO objective with the token-level advantage $\hat{A}_t^{(i)}$, yielding

$$\mathcal{L}_{\mathrm{RESD}} = -\mathbb{E}_{i,t}\left[\min\left(\omega_t^{(i)} \hat{A}_t^{(i)}, \mathrm{clip}\left(\omega_t^{(i)}, 1-\epsilon, 1+\epsilon\right) \hat{A}_t^{(i)}\right)\right], \tag{9}$$



(a) Entropy (b) Response Length (c) Reward (d) Accuracy

Figure 6: Training dynamics on Qwen3-1.7B. We report response entropy, response length, training reward, and the average Avg@12 over all test sets throughout post-training steps.

where $\omega_t^{(i)}$ denotes the importance sampling ratio. In practice, the retained-exploration correction is applied during the early stage of post-training to reshape the student's exploration dynamics, after which training continues with standard GRPO. Detailed RESD algorithm pesudocode and the corresponding hyperparameter settings are provided in Appendix B.1.

4 Experiments

4.1 Experimental Setup

Models and datasets. We experiment with models across different scales and model families: Qwen3-1.7B, Qwen3-4B (Team, 2025), and SmolLM3-3B (Bakouch et al., 2025). All models are trained on OpenThoughts-Math-30k (Guha et al., 2025), with ground-truth solutions provided as PI for the self-teacher construction. The test sets for evaluation are five challenging mathematical reasoning benchmarks: AMC23 (MAA, 2023), AIME24 (Zhang & Math-AI, 2024), AIME25 (Zhang & Math-AI, 2025), AIME26 (Zhang & Math-AI, 2026), and HMMT25 (Balunovic et al., 2025).

Baselines. We compare RESD with: (1) **GRPO** (Shao et al., 2024), a standard RLVR method using sparse rollout-level rewards; (2) **OPSD** (Zhao et al., 2026a), an on-policy self-distillation method providing dense token-level supervision; (3) **SRPO** (Li et al., 2026), which applies GRPO to successful rollouts and self-distillation to failed rollouts; and (4) **RLSD** (Yang et al., 2026) and **RLCSD** (Pan et al., 2026), two self-distillation methods that use token-level teacher signals to modulate rollout-level advantages.

Implementation details. All models are trained with LoRA using rank 64, scaling factor $\alpha = 128$, and all linear layers as adaptation targets. We use a learning rate of 5×10^{-6}, a rollout group size of 8, a batch size of 32, and train for 300 steps. The self-teacher is synchronized with the student every 20 training steps. Student rollouts are generated with extended thinking disabled, while self-teacher distributions are constructed with thinking enabled, following standard OPSD configuration (Zhao et al., 2026a). We report Avg@12 and Pass@12 on test sets as main evaluation metrics. Additional generation and implementation details are provided in Appendix B.2.

4.2 Main Results

Table 2 summarizes the reasoning performance of different post-training methods across five mathematical reasoning benchmarks. RESD achieves the best accuracy on 14 of 15 model–benchmark combinations, improving the average accuracy over the strongest baseline by 8.72 and 5.98 points on Qwen3-1.7B and Qwen3-4B, respectively, and by 10.56 points on the cross-family SmolLM3-3B. We also provide results on Pass@12 metric in Appendix D.2, where RESD also performs consistently better than all baselines. These results indicate that the substantial benefit of retained-exploration-guided supervision is not specific to a particular model scale or architecture.

Among the baselines, GRPO remains consistently strong across all three models, while the other ones show less consistent gains across model families. RLSD performs strongly on the Qwen3 models but transfers less effectively to SmolLM3-3B, while OPSD, SRPO, and RLCSD generally provide smaller improvements.

Figure 6 further compares training dynamics on Qwen3-1.7B, with similar trends on Qwen3-4B reported in Appendix D.3. During the first 60 steps, RESD improves accuracy and increases response

Table 2: Avg@12 results on mathematical reasoning benchmarks. *Avg.* denotes the average performance across all test sets. Best and second-best are highlighted in **bold** and underline.

Model / Method	AMC23	AIME24	AIME25	AIME26	HMMT25	Avg.
Qwen3-1.7B	45.21	13.06	11.11	8.89	4.44	16.54
GRPO	70.63	33.06	26.39	21.67	15.83	33.52
OPSD	48.96	18.33	16.94	10.00	8.89	20.62
SRPO	50.00	15.28	13.61	9.44	6.39	18.94
RLSD	72.50	35.00	27.22	24.72	16.39	35.17
RLCSD	57.92	24.44	16.39	16.11	11.94	25.36
RESD (Ours)	**83.33**	**44.17**	**32.22**	**37.22**	**22.50**	**43.89**
Qwen3-4B	65.63	23.33	22.78	18.06	11.11	28.18
GRPO	89.58	47.78	41.39	42.50	22.22	48.69
OPSD	70.63	25.00	24.17	20.56	14.17	30.91
SRPO	70.83	25.83	21.94	18.61	14.72	30.39
RLSD	**91.46**	51.11	45.28	46.94	25.56	52.07
RLCSD	80.21	42.22	29.72	30.00	22.78	40.99
RESD (Ours)	90.00	**63.33**	**48.88**	**55.28**	**32.78**	**58.05**
SmolLM3-3B	47.50	10.28	11.11	5.56	5.00	15.89
GRPO	74.58	30.28	28.61	24.44	13.06	34.19
OPSD	66.67	30.56	23.33	20.56	13.89	31.00
SRPO	53.33	10.83	13.33	8.88	5.83	18.44
RLSD	70.83	28.89	21.39	16.94	12.50	30.11
RLCSD	72.92	26.67	26.11	23.06	10.56	31.86
RESD (Ours)	**82.08**	**45.83**	**35.56**	**38.33**	**21.94**	**44.75**

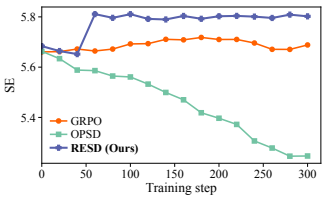


Figure 7: SE dynamics for GRPO, OPSD and RESD.

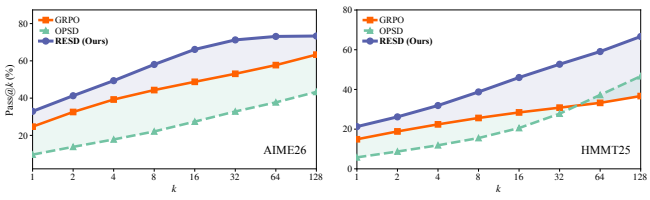


Figure 8: Pass@k curves on AIME26 and HMMT25 for GRPO, OPSD, and RESD, with $k \in \{1, 2, 4, 8, 16, 32, 64, 128\}$.

length while maintaining stable entropy. After the retained-exploration correction is disabled at step 60, both the performance advantage and increased response length are largely preserved, and RESD remains stable throughout later training. In contrast, GRPO and RLSD exhibit late-stage performance degradation, while OPSD shows rapidly increasing entropy and response length together with declining reward and accuracy. These results suggest that early retained-exploration-guided correction yields lasting performance gains in subsequent optimization without the instability observed under uniform self-distillation. We further analyze RESD's per-step computational cost and token-budget-aligned performance in Appendix E.

4.3 Discussions

Latent exploration and successful reasoning coverage. We examine how RESD affects latent exploration during post-training and whether this change is reflected in inference-time reasoning coverage. As shown in Figure 7, the semantic exploration of RESD increases early in training and remains at a higher level thereafter, showing that RESD effectively enhances and maintains latent exploration. We then evaluate Pass@k on the more challenging AIME26 and HMMT25 benchmarks in Figure 8. RESD consistently achieves higher Pass@k across sampling budgets, suggesting that its preserved latent exploration is associated with a broader set of reachable successful reasoning trajectories. Similar Pass@k trends on Qwen3-4B are reported in Appendix D.3.

Table 3: Core ablations and transferability of Q_T routing on Qwen3-1.7B. All results are Avg@12. Full RESD routes high-Q_T positions toward the teacher and low-Q_T positions away using the ρ-quantile threshold in Section 3.2.

Analysis	Configuration	AIME24	AIME25	HMMT25
A. RESD ablations	**RESD (full)**	**44.17**	**32.22**	**22.50**
	High-Q_T toward only	39.44	29.17	20.83
	High-Q_T toward + low-Q_T toward	38.33	29.72	20.28
	w/o base advantage a_h	38.33	30.83	21.67
	w/o Q_T routing	38.33	29.44	20.56
	Entropy routing (H instead of Q_T)	36.11	27.50	19.17
B. Q_T transfer	OPSD	18.33	16.94	8.89
	OPSD + Q_T routing	**20.83**	**18.05**	**9.16**

Core ablations and effectiveness of Q_T. Table 3 first isolates the role of low-Q_T routing while keeping the high-Q_T toward-teacher branch fixed. Full RESD, which routes low-Q_T positions away from the teacher, performs best; removing this correction degrades performance, while routing these positions toward the teacher leads to a further drop. This shows that RESD benefits from exploration-dependent correction directions rather than simply stronger teacher supervision. Ablating the base advantage a_h or Q_T-based routing also reduces performance, with a larger drop when Q_T is replaced by entropy, supporting the importance of retained exploration as the routing signal. Also, applying the same Q_T-based routing to OPSD improves all three benchmarks, indicating that its utility extends beyond the specific RESD formulation. Additional sensitivity analyses are provided in Appendix C.

5 Related Work

Self-distillation with RLVR. Recent methods combine dense self-teacher supervision with RLVR. SDPO converts rollout feedback into dense self-distillation (Hübotter et al., 2026), while SRPO applies GRPO to successful rollouts and SDPO-style distillation to failed ones, with entropy-aware target weighting (Li et al., 2026). RLSD uses verifier rewards to determine update direction and teacher–student probability gaps to scale token-level updates (Yang et al., 2026); RLCSD and Purified OPSD mitigate privilege-induced distribution shifts (Pan et al., 2026; Shen et al., 2026); and DASD routes tokens toward or away from the teacher using output entropy (Zhang et al., 2026). These methods mainly rely on rollout outcomes or output-level signals, whereas RESD regulates self-distillation based on whether teacher predictions internally retain rich reasoning alternatives.

Internal reasoning dynamics. Intermediate representations can reveal reasoning dynamics beyond final outputs. BuPO identifies a layer-wise transition from exploration to convergence (Tan et al., 2025), LED exploits richer intermediate-layer uncertainty for inference-time exploration (Tan et al., 2026), and SHEAR uses hidden-state differences between correct and incorrect trajectories for fine-grained credit assignment (Chen et al., 2026). RESD further studies how internal exploration evolves during post-training and uses self-teacher's retained exploration to regulate dense supervision.

6 Conclusion

We study post-training through internal exploration–exploitation dynamics and find that self-distillation progressively contracts latent exploration, unlike GRPO. We introduce retained exploration to measure whether the self-teacher's reasoning alternatives remain relevant through subsequent layers, and show that high-retained-exploration positions correspond to consequential reasoning forks. We propose RESD, which uses this signal to regulate token-level teacher guidance and consistently improves over strong RLVR and self-distillation baselines across multiple models and mathematical reasoning benchmarks.

AI USE STATEMENT

In this work, we used generative AI tools solely to edit the manuscript for language clarity and readability. All AI-assisted edits were manually reviewed by the authors and, where necessary, further revised before inclusion in the final manuscript. We did not use generative AI tools for any research tasks requiring mandatory disclosure under the ICLR 2027 AI Policy for Authors, or for any other part of the research process. We take responsibility for the final content of this work, including text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

We provide the methodological and experimental details required to reproduce our results throughout the paper and appendix. The formulation of retained exploration and the complete RESD objective are described in Section 3.2, with algorithmic details and hyperparameter settings provided in Appendix B.1. The models, datasets, baselines, training configuration, and evaluation protocol are specified in Section 4.1, with additional generation and implementation details reported in Appendix B.2. We additionally include the core source code for RESD and the scripts required to reproduce the main experiments in the supplementary materials.

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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Riccardo Ali, Francesco Caso, Christopher Irwin, and Pietro Liò. Entropy-lens: The information signature of transformer computations. *CoRR*, abs/2502.16570, 2025. doi: 10.48550/ARXIV.2502.16570. URL https://doi.org/10.48550/arXiv.2502.16570.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, Xuan-Son Nguyen, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3, 2025.

Mislav Balunovic, Jasper Dekoninck, Ivo Petrov, Nikola Jovanovic, and Martin T. Vechev. Matharena: Evaluating llms on uncontaminated math competitions. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/1d27c01ebd3e3aebe226b44fc970d803-Abstract-Datasets_and_Benchmarks_Track.html.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *CoRR*, abs/2303.08112, 2023. doi: 10.48550/ARXIV.2303.08112. URL https://doi.org/10.48550/arXiv.2303.08112.

Xinzhu Chen, Wei He, Huichuan Fan, Wenzhe Niu, Zhongxiang Sun, Xuanru Wang, Jiuchong Gao, Jinghua Hao, Renqing He, and Weijie Yu. Hidden states know where reasoning diverges: Credit assignment via span-level wasserstein distance. *CoRR*, abs/2604.23318, 2026. doi: 10.48550/ARXIV.2604.23318. URL https://doi.org/10.48550/arXiv.2604.23318.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

Etash Kumar 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 M. 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. *CoRR*, abs/2506.04178, 2025. doi: 10.48550/ARXIV.2506.04178. URL https://doi.org/10.48550/arXiv.2506.04178.

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 Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

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. *CoRR*, abs/2601.20802, 2026. doi: 10.48550/ARXIV.2601.20802. URL https://doi.org/10.48550/arXiv.2601.20802.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking on-policy self-distillation for thinking models. *CoRR*, abs/2607.05184, 2026. doi: 10.48550/ARXIV.2607.05184. URL https://doi.org/10.48550/arXiv.2607.05184.

Gengsheng Li, Tianyu Yang, Junfeng Fang, Mingyang Song, Mao Zheng, Haiyun Guo, Dan Zhang, Jinqiao Wang, and Tat-Seng Chua. Unifying group-relative and self-distillation policy optimization via sample routing. *CoRR*, abs/2604.02288, 2026. doi: 10.48550/ARXIV.2604.02288. URL https://doi.org/10.48550/arXiv.2604.02288.

MAA. AMC 2023. https://huggingface.co/datasets/math-ai/amc23, 2023. Dataset compiled from the 2023 American Mathematics Competitions.

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. RLCSD: reinforcement learning with contrastive on-policy self-distillation. *CoRR*, abs/2606.11709, 2026. doi: 10.48550/ARXIV.2606.11709. URL https://doi.org/10.48550/arXiv.2606.11709.

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

Zhanming Shen, Jintao Tong, Shaotian Yan, Chen Shen, Hao Chen, Wentao Ye, Xiaomeng Hu, Rui Miao, Haobo Wang, Junbo Zhao, Gang Chen, and Jieping Ye. Purified OPSD: on-policy self-distillation without losing how to think. *CoRR*, abs/2607.02234, 2026. doi: 10.48550/ARXIV.2607.02234. URL https://doi.org/10.48550/arXiv.2607.02234.

Wenhui Tan, Fiorenzo Parascandolo, Enver Sangineto, Jianzhong Ju, Zhenbo Luo, Qian Cao, Rita Cucchiara, Ruihua Song, and Jian Luan. Restoring exploration after post-training: Latent exploration decoding for large reasoning models. *CoRR*, abs/2602.01698, 2026. doi: 10.48550/ARXIV.2602.01698. URL https://doi.org/10.48550/arXiv.2602.01698.

Yuqiao Tan, Minzheng Wang, Shizhu He, Huanxuan Liao, Chengfeng Zhao, Qiunan Lu, Tian Liang, Jun Zhao, and Kang Liu. Bottom-up policy optimization: Your language model policy secretly contains internal policies. *CoRR*, abs/2512.19673, 2025. doi: 10.48550/ARXIV.2512.19673. URL https://doi.org/10.48550/arXiv.2512.19673.

Qwen Team. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled RLVR. *CoRR*, abs/2604.03128, 2026. doi: 10.48550/ARXIV.2604.03128. URL https://doi.org/10.48550/arXiv.2604.03128.

Hongbin Zhang, Chaozheng Wang, Kehai Chen, Youcheng Pan, Yang Xiang, Jinpeng Wang, and Min Zhang. Tailoring teaching to aptitude: Direction-adaptive self-distillation for LLM reasoning. *CoRR*, abs/2605.22263, 2026. doi: 10.48550/ARXIV.2605.22263. URL https://doi.org/10.48550/arXiv.2605.22263.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2024, 2024.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2025, 2025.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2026, 2026.

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. *CoRR*, abs/2601.18734, 2026a. doi: 10.48550/ARXIV.2601.18734. URL https://doi.org/10.48550/arXiv.2601.18734.

Ziqi Zhao, Zhaochun Ren, Jiahong Zou, Liu Yang, Zhiwei Xu, Xuri Ge, Zhumin Chen, Xinyu Ma, Daiting Shi, Shuaiqiang Wang, Dawei Yin, and Xin Xin. Reinforced efficient reasoning via semantically diverse exploration. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 47994–48007. Association for Computational Linguistics, 2026b. doi: 10.18653/V1/2026.ACL-LONG.2216. URL https://doi.org/10.18653/v1/2026.acl-long.2216.

Felix Zhou, Anay Mehrotra, and Quanquan C. Liu. Reasoning with sampling: Cutting at decision points. *CoRR*, abs/2605.30327, 2026. doi: 10.48550/ARXIV.2605.30327. URL https://doi.org/10.48550/arXiv.2605.30327.

Siqi Zhu, Xuyan Ye, Hongyu Lu, Weiye Shi, and Ge Liu. The many faces of on-policy distillation: Pitfalls, mechanisms, and fixes. *CoRR*, abs/2605.11182, 2026. doi: 10.48550/ARXIV.2605.11182. URL https://doi.org/10.48550/arXiv.2605.11182.

A Detailed Configurations on Latent-Exploration Analysis

A.1 LLM-Judge Configurations on Semantic Token Categorization

We analyze mathematical reasoning trajectories generated by Qwen3-1.7B on MATH-500 (Hendrycks et al., 2021), retaining problem groups that contain both correct and incorrect responses. We use Deepseek-V4-Flash (DeepSeek-AI, 2026) as the token-category judge with temperature set to 0.0. Each annotation uses a local context containing up to 48 tokens on either side of the target token. The complete annotation prompt is provided in Figure 9.

For each metric, we select the top $\lceil 0.05L \rceil$ positions within each trajectory of length L. We pool the selected positions across trajectories to compute category proportions.

Prompt for Semantic Token Categorization

User Prompt:
You are annotating the functional role of a single target token in a mathematical reasoning trace.
You will receive:
- the original math problem;
- a local context from the model's reasoning trace;
- one target token marked as `<TARGET>...</TARGET>`.

Classify the target token into exactly one of the following categories according to its role in this specific context, not merely according to its lexical form.

1. **`reasoning_control`**
   The token is part of a reasoning action that determines, initiates, redirects, or checks what the model should do next. This includes introducing a new subgoal, selecting a strategy or representation, transitioning between reasoning stages, backtracking, reconsidering, or verification.
2. **`reasoning_inference`**
   The token is part of a semantic deduction, logical relation, or intermediate conclusion that advances the solution.
3. **`computation`**
   The token participates mainly in executing an already chosen mathematical operation, including arithmetic, algebraic manipulation, substitution, numerical values, or mathematical expressions.
4. **`exposition`**
   The token mainly restates the problem, describes given information, introduces notation, or provides explanatory/narrative text without introducing a new reasoning decision or inference.
5. **`surface`**
   The token serves mainly formatting, syntactic, or markup purposes and has little independent reasoning content.

Important rules:
- Judge the role of the target token together with the semantic phrase containing it.
- A word inside a step heading such as "Compute the dot product" should be `reasoning_control` if it states the next reasoning subgoal.
- The formatting portion of the same heading, such as "`###`", "Step 3", or punctuation, should be `surface`.
- A mathematical symbol or number used in an actual calculation should normally be `computation`; pure LaTeX/formatting syntax should be `surface`.
- Merely mentioning information from the problem is `exposition`; using that information to derive a new conclusion is `reasoning_inference`.
- Do not infer importance from whether the reasoning is correct or incorrect.
- Do not judge whether the token is useful overall. Judge only its functional role.

Return only one category name, with no JSON, confidence, reason, punctuation, or extra text.

Problem: `<PROBLEM>`
{`problem`}
`</PROBLEM>`

Target token raw text: {`target_text`}

Reasoning context: `<CONTEXT>`
{`context_with_target`}
`</CONTEXT>`

Figure 9: The instruction prompt used for semantic token categorization.

A.2 Configurations on Reasoning Diversity Analysis

We use the same trajectory pool as in Appendix A.1 and all metrics are evaluated on the same selected trajectories. For each trajectory, we rank positions by each metric and select five fork

Table 4: Default hyperparameters for RESD. The exploration layer l^* is identified before post-training following Section 2.1, rather than treated as a tunable scalar hyperparameter.

Hyperparameter	Value	Description
ρ	0.20	Routing quantile
β	0.30	Token-level correction weight
a_h	0.25	Base advantage for zero-variance groups
$s_{\min}$	0.05	Gap normalization floor
C	6.0	Normalized gap clipping bound
ε	10^{-6}	Numerical stability constant
T_{corr}	60	Correction horizon (steps)

positions from its top-5% positions in descending score order, requiring any two selected positions to be more than 8 tokens apart. The *Random* baseline selects five positions from the entire trajectory with the same spacing constraint.

At each selected position t, we fix the original problem prompt x and response prefix $y_{<t}$, and sample $N = 32$ suffixes starting at position t. While position selection uses the privileged teacher's metrics, continuation sampling uses the student prompt without the reference solution. Generation settings are identical to those used for evaluation. Each complete continuation consists of the fixed response prefix followed by a resampled suffix.

We measure continuation diversity using normalized edit distance and distinct-answer fraction. For sampled suffixes s_i and s_j, we compute token-level Levenshtein distance and normalize it by their average length. We then average over all $\binom{N}{2}$ suffix pairs:

$$D_{\text{edit}}(t) = \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} \frac{d_{\text{Lev}}(s_i, s_j)}{(|s_i| + |s_j|)/2}. \tag{10}$$

The distance is computed only over the sampled suffixes, excluding the shared prefix. A pair of empty suffixes is assigned a distance of zero.

For distinct-answer fraction, we extract the answer from the last `\boxed{}` expression in each complete continuation and merge mathematically equivalent answers using `math-verify`. Let $\mathcal{A}_t$ denote the resulting set of answer classes. We define

$$D_{\text{ans}}(t) = \frac{|\mathcal{A}_t|}{N}. \tag{11}$$

Continuations without an extractable answer are assigned to a single `INVALID` class and remain included in the denominator. For both diversity measures, we first average over the selected positions within each trajectory and then average equally across trajectories.

We further assess the stability of normalized edit distance under suffix subsampling. We randomly select 20 fork positions and, at each position, repeatedly sample 16 of the 32 available suffixes without replacement. We recompute D_{edit} for each of 100 repetitions. Across the selected positions, the mean coefficient of variation, defined as the standard deviation of the subsample estimates divided by their mean, is 6.45%. The mean absolute difference between the average subsample estimate and the full-32 estimate is 0.0025. These results indicate that the edit-distance estimates are relatively stable under suffix sampling at the examined positions, supporting its use as a reliable metric for quantifying downstream continuation diversity.

B Implementation Details on RESD

B.1 Algorithm Pseudocode and Hyperparameter Settings

We provide detailed pseudocode for RESD in Algorithm 1, with the default hyperparameters summarized in Table 4. We further conduct ablations and sensitivity analyses on a_h, ρ, l^*, and the retained-exploration metric Q_T in Appendix C.

Algorithm 1 Retained-Exploration-Guided Self-Distillation (RESD)

Require: student policy π_θ and behavior policy $\pi_{\theta_{\text{old}}}$; privileged information PI; rollout count G; exploration layer l^*; routing quantile ρ; coefficients β, a_h; normalization floor $s_{\min}$; clipping bound C; correction horizon T_{corr}

for each training batch **do**
  Sample rollout groups $\{y^{(i)}\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$ without privileged information
  Compute verifier rewards and standard GRPO advantages $\{A_G^{(i)}\}_{i=1}^{G}$
  if $s \le T_{\text{corr}}$ **then**
    Set $B^{(i)} \leftarrow A_G^{(i)}$; replace with $+a_h / -a_h$ for all-correct/all-wrong groups
    for each rollout i **do**
      Compute token-wise student and self-teacher probabilities:
      $p_{S,t}^{(i)} \leftarrow p_S(y_t^{(i)} \mid x, y_{<t}^{(i)}),\ p_{T,t}^{(i)} \leftarrow p_T(y_t^{(i)} \mid x, y_{<t}^{(i)}, \text{PI})$
      Compute the teacher correction signal: $\boldsymbol{d_t^{(i)}} \leftarrow \log p_{T,t}^{(i)} - \log p_{S,t}^{(i)}$
      $s_d^{(i)} \leftarrow \max(\text{median}_t\, |d_t^{(i)}|, s_{\min}),\ \bar{d}_t^{(i)} \leftarrow \text{clip}(d_t^{(i)}/s_d^{(i)}, -C, C)$
      $\boldsymbol{\tilde{d}_t^{(i)}} \leftarrow \sigma(|\bar{d}_t^{(i)}| - 1)\bar{d}_t^{(i)}$ ▷ normalize, clip, and suppress small gaps
      Compute retained-exploration routing: $\boldsymbol{Q_{T,t}^{(i)}} \leftarrow SE_{T,t}^{l^*} R_{T,t}^{l^*}$
      $\bar{Q}_{T,t}^{(i)} \leftarrow \log(1 + Q_{T,t}^{(i)}/(\text{median}_t\, Q_{T,t}^{(i)} + \varepsilon)),\ \tau^{(i)} \leftarrow \text{Quantile}_\rho(\{\bar{Q}_{T,t}^{(i)}\}_t)$
      $\boldsymbol{r_t^{(i)}} \leftarrow \tanh((\bar{Q}_{T,t}^{(i)} - \tau^{(i)})/(\text{MAD}_t(\bar{Q}_{T,t}^{(i)}) + \varepsilon))$ ▷ $r_t > 0$: toward teacher; $r_t < 0$: away
      Combine teacher guidance for token-level advantage:
      $\hat{A}_t^{(i)} \leftarrow B^{(i)} + \beta r_t^{(i)} \tilde{d}_t^{(i)}$
    end for
  else
    Use standard GRPO advantages: $\hat{A}_t^{(i)} \leftarrow A_G^{(i)}$ for all i, t
  end if
  Compute importance sampling ratios $\omega_t^{(i)} \leftarrow \pi_\theta(y_t^{(i)} \mid x, y_{<t}^{(i)})/\pi_{\theta_{\text{old}}}(y_t^{(i)} \mid x, y_{<t}^{(i)})$
  Update θ using the clipped GRPO objective in Eq. (9) with $\hat{A}_t^{(i)}$
end for
return updated policy π_θ

B.2 Implementation Details

Training and generation details. We implement RESD and other post-training baselines using verl and generate rollouts with vLLM. Following OPSD (Zhao et al., 2026a), we adopt the same student and privileged-teacher prompt templates (Figure 10) and generation parameter settings for training and evaluation (Table 5). Teacher signals are computed on student-generated trajectories through teacher-forced forward passes under the privileged prompt. The teacher remains frozen between synchronizations and is updated by copying the student parameters. All teacher signals and the resulting token-level advantages are detached from gradient computation.

Table 5: Generation settings for student rollouts during training and evaluation.

Parameter	Training	Evaluation
Temperature	1.1	1.0
Top-p	0.95	0.8
Maximum response length	16,384	16,384
Extended thinking (student)	Disabled	Disabled

For retained-exploration computation, we use $K = 20$ semantic candidates. The exploration-layer indices are 10, 14, and 10 for Qwen3-1.7B, Qwen3-4B, and SmolLM3-3B, respectively. Cross-layer retention tracks the same candidate set selected at the exploration layer. In the routing computation, the median used to rescale Q is computed over its strictly positive values, and MAD denotes the mean absolute deviation from the mean.

For evaluation, we extract the final answer from the last `\boxed{}` expression and assess correctness using `math-verify`; responses without an extractable answer are counted as incorrect. Avg@12 averages correctness over twelve sampled responses per problem, whereas Pass@12 mea-

sures the fraction of problems with at least one correct response among the twelve samples. Reported averages across benchmarks assign equal weight to each benchmark.

Student and Privileged-Teacher Prompts

Student Prompt:
Problem: {problem}
Please reason step by step, and put your final answer within \boxed{}.

Privileged-Teacher Prompt:
Problem: {problem}
Here is a reference solution to this problem:
=== Reference Solution Begin ===
{solution}
=== Reference Solution End ===
After reading the reference solution above, make sure you truly understand the reasoning behind each step—do not copy or paraphrase it. Now, using your own words and independent reasoning, derive the same final answer to the problem above. Think step by step, explore different approaches, and don't be afraid to backtrack or reconsider if something doesn't work out:
Please reason step by step, and put your final answer within \boxed{}.

Figure 10: Student and privileged-teacher prompts used during training. The reference solution is available only to the teacher.

Baseline implementation details. All baseline methods use the common post-training configuration described above, including the optimizer, rollout, batch, and generation settings. Method-specific hyperparameters and implementation choices follow the corresponding official implementations when available, and otherwise the original papers. DASD (Zhang et al., 2026) was not included in the quantitative comparison because no official implementation was publicly available at the time of submission, and our paper-based implementation exhibited unstable training. We therefore could not obtain results that we considered sufficiently reliable for comparison.

C Further Ablations on RESD

We further analyze additional design choices of RESD on Qwen3-1.7B beyond those examined in Table 3. Table 6 studies the base advantage a_h, routing quantile ρ, choice of internal layer, and construction of the retained-exploration signal.

Base advantage. Assigning a small signed base advantage to zero-variance rollout groups is beneficial, with $a_h = 0.25$ achieving the best performance across all three benchmarks. Removing this base advantage leads to a moderate drop, while further increasing a_h provides no additional gain. Nevertheless, all tested settings remain substantially stronger than the corresponding baseline results in Table 2, indicating that RESD is relatively robust to the exact choice of a_h.

Routing quantile. The routing quantile ρ controls the fraction of low-Q_T positions receiving away-from-teacher corrections. The default $\rho = 0.20$ consistently performs best, while increasing ρ to 0.50 or 0.80 causes substantial degradation. In particular, $\rho = 0.80$ sharply reduces performance across all three benchmarks. This indicates that reverse correction is most effective when localized to a relatively small low-retained-exploration region, while toward-teacher guidance remains the dominant supervision signal.

Choice of internal layer. We compute retained exploration at the identified exploration layer l^* by default and compare it with an early layer, an upper layer, and the final layer. For the final-layer variant, the retention term is set to 1, reducing the signal to semantic exploration at the output layer. None of these alternatives matches the performance obtained at l^*, supporting the use of the exploration layer for constructing the routing signal rather than an arbitrary intermediate or output layer.

Table 6: Ablations and sensitivity analyses of RESD on Qwen3-1.7B. All results are Avg@12. Bold rows denote the default RESD configuration.

Ablation	Variant	AIME24	AIME25	HMMT25
A. Base advantage a_h	$a_h = 0.00$	38.33	30.83	21.67
	$\boldsymbol{a_h = 0.25}$	**44.17**	**32.22**	**22.50**
	$a_h = 0.50$	38.89	29.17	20.56
	$a_h = 1.00$	38.33	29.17	21.11
B. Routing quantile ρ	$\boldsymbol{\rho = 0.20}$	**44.17**	**32.22**	**22.50**
	$\rho = 0.50$	32.22	22.50	14.72
	$\rho = 0.80$	11.11	6.39	2.78
C. Choice of internal layer	Layer 2	37.50	29.73	20.83
	Layer 24	39.16	28.05	20.56
	Final layer	38.06	29.17	20.83
	Exploration layer $\boldsymbol{l^*}$	**44.17**	**32.22**	**22.50**
D. Retained-exploration metric	H	36.11	27.50	19.17
	SE	35.83	28.61	19.17
	Q'	36.11	29.17	19.72
	$\boldsymbol{Q_T}$	**44.17**	**32.22**	**22.50**

Exploration related metrics. We further replace Q_T with the alternative internal signals considered in Section 3.1. Using entropy H, semantic exploration SE, or Q' consistently yields substantially lower performance than the full retained-exploration metric. Since these alternatives remove different components of Q_T, the results show that both semantic diversity and cross-layer retention contribute to an effective supervision-routing signal.

D Supplementary Experimental Results

D.1 Token-level Visualization under Different Metrics

We analyze whether high-Q positions in reasoning trajectories correspond to important reasoning points in Section 3.1, and provide token-level visualizations of H and Q over the same trajectory segment in Figure 4. Here, we provide full-trajectory visualizations for all four metrics, including entropy H, semantic exploration SE, Q', and retained exploration Q in Figure 15. We follow the same visualization settings as in the main text, where green and red indicate metric values above and below the trajectory median, respectively.

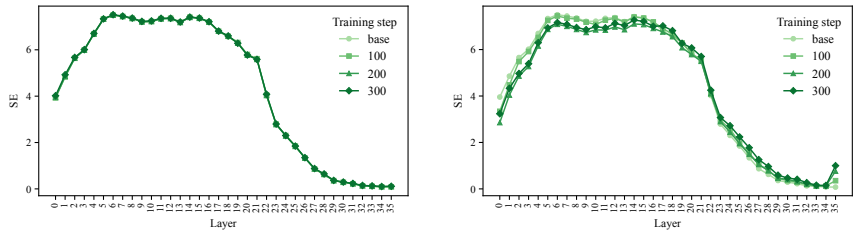
The full-trajectory visualizations further support the observations in Section 3.1. High-H positions are broadly distributed throughout the trajectory and frequently appear within local computation and exposition, while incorporating semantic diversity in SE only partially reduces such diffuse activations. Q' further incorporates cross-layer retention and exhibits greater selectivity, but still retains substantial activation on local reasoning tokens. In contrast, high-Q positions are more concentrated around reasoning-stage transitions and reasoning-control regions, consistent with the quantitative semantic-enrichment and continuation-diversity results in the main text. These comparisons qualitatively illustrate the complementary roles of semantic diversity and cross-layer retention in filtering generic uncertainty and isolating consequential reasoning positions.

D.2 Pass@12 Performance of RESD and Other Baselines

Consistent with the Avg@12 results in the main text, RESD also achieves the strongest overall Pass@12 performance across all three models as shown in Table 7. On Qwen3-1.7B and Qwen3-4B, the improvements are particularly pronounced on the more challenging AIME26 and HMMT25 benchmarks, indicating that RESD increases the coverage of successful reasoning trajectories rather than only improving average sampled accuracy. The same overall advantage is preserved on the cross-family SmolLM3-3B, where RESD achieves the highest average Pass@12 among all com-

Table 7: Pass@12 results on mathematical reasoning benchmarks. *Avg.* denotes the average performance across all test sets. Best and second-best results are highlighted in **bold** and underline.

Model / Method	AMC23	AIME24	AIME25	AIME26	HMMT25	Avg.
Qwen3-1.7B	85.00	36.67	20.00	20.00	10.00	34.33
GRPO	92.50	63.33	53.33	46.67	30.00	57.17
OPSD	87.50	50.00	40.00	30.00	26.67	46.83
SRPO	85.00	46.67	40.00	33.33	23.33	45.67
RLSD	**95.00**	66.67	46.67	50.00	30.00	57.67
RLCSD	**95.00**	56.67	43.33	40.00	33.33	53.67
RESD (Ours)	**95.00**	**76.67**	**60.00**	**70.00**	**50.00**	**70.33**
Qwen3-4B	90.00	50.00	46.67	43.33	23.33	50.67
GRPO	97.50	76.67	73.33	63.33	43.33	70.83
OPSD	97.50	56.67	56.67	56.67	30.00	59.50
SRPO	95.00	53.33	46.67	46.67	26.67	53.67
RLSD	97.50	80.00	70.00	73.33	50.00	74.17
RLCSD	97.50	76.67	60.00	66.67	50.00	70.17
RESD (Ours)	**100.00**	**83.33**	**76.67**	**80.00**	**60.00**	**80.00**
SmolLM3-3B	87.50	26.67	30.00	16.67	16.67	35.50
GRPO	**97.50**	63.33	56.67	60.00	33.33	62.17
OPSD	95.00	63.33	56.67	56.67	30.00	60.33
SRPO	90.00	36.67	36.67	33.33	23.33	44.00
RLSD	95.00	60.00	53.33	53.33	**40.00**	60.33
RLCSD	**97.50**	63.33	56.67	66.67	33.33	63.50
RESD (Ours)	95.00	**80.00**	**60.00**	**70.00**	**40.00**	**69.00**



(a) GRPO middle-layer SE (b) OPSD middle-layer SE

Figure 11: Qwen3-4B's layer-wise semantic exploration at different post-training checkpoints under GRPO and OPSD.

pared methods. These results further support that RESD improves both average reasoning accuracy and inference-time reasoning coverage.

D.3 Additional Analyses on Qwen3-4B

We provide additional analyses on Qwen3-4B to examine whether the exploration dynamics and training behavior observed on Qwen3-1.7B remain consistent at a larger model scale. Figure 11 shows the layer-wise semantic exploration (SE) of GRPO and OPSD at different post-training checkpoints. Similar to the observations on Qwen3-1.7B (Figure 2), the overall layer-wise structure remains stable throughout training, with a high-exploration region in the middle layers followed by a sharp reduction toward the upper layers. Under GRPO, the SE values at different checkpoints largely overlap, indicating that outcome-level optimization preserves the model's internal exploration structure. In contrast, OPSD progressively suppresses SE within the middle-layer exploration region while leaving the location of this region largely unchanged. This further supports that the exploration contraction induced by uniform self-distillation is not specific to the smaller model.

Figure 12 further compares the training dynamics of different baselines and RESD on Qwen3-4B. The overall trends are consistent with those observed on Qwen3-1.7B (Figure 6). OPSD ex-

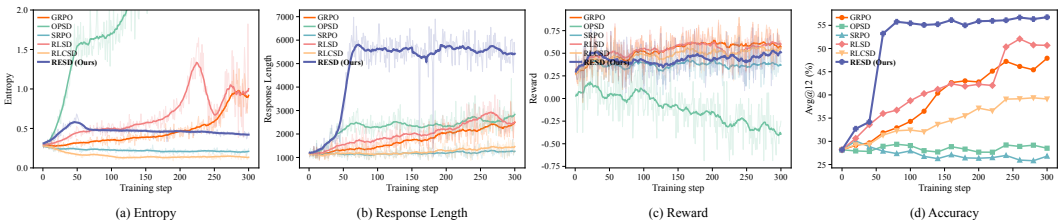


Figure 12: Training dynamics on Qwen3-4B. We report response entropy, response length, training reward, and the average Avg@12 accuracy over the five evaluation benchmarks throughout post-training.

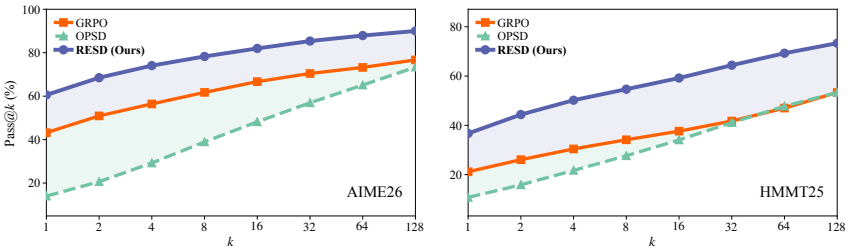


Figure 13: Pass@k curves on AIME26 and HMMT25 for GRPO, OPSD, and RESD on Qwen3-4B, with $k \in \{1, 2, 4, 8, 16, 32, 64, 128\}$.

hibits a rapid increase in response entropy together with deteriorating reward and limited improvement in reasoning accuracy, indicating unstable optimization under uniform self-distillation. GRPO and RLSD achieve substantial performance improvements but remain below RESD. In contrast, RESD undergoes a clear early-stage transition: its response length increases substantially during the retained-exploration correction stage, while its entropy remains controlled and the reasoning accuracy rapidly improves. After the retained-exploration correction is disabled at step 60, the higher response length and performance are largely maintained throughout the remaining GRPO training.

We further evaluate inference-time reasoning coverage using Pass@k on the two challenging benchmarks AIME26 and HMMT25. As shown in Figure 13, RESD consistently achieves higher Pass@k than both GRPO and OPSD across all sampling budgets on Qwen3-4B. The advantage is already evident at small k and remains substantial as the sampling budget increases on both benchmarks. Together with the layer-wise and training-dynamics analyses above, these results show that the exploration-preserving behavior of RESD and its associated reasoning improvements consistently extend from Qwen3-1.7B to Qwen3-4B.

E Computational Overhead Analysis

We analyze the computational overhead of RESD compared with GRPO and OPSD from two complementary perspectives. First, we examine per-step training cost to assess the additional computation associated with RESD. Second, since RESD generates longer trajectories, we compare performance under a shared cumulative training-token budget to determine whether its gains can be explained solely by greater token consumption.

Per-step computational cost. Table 8 reports average training statistics over steps 11–300, excluding early warmup effects. Tokens per step count non-padding student prompt and response tokens, and seconds per 1K tokens normalize the measured training time by this count.

Although RESD takes longer per optimization step, it also generates substantially longer responses. Relative to GRPO, the increase in step time is considerably smaller than the increase in token volume, yielding a lower normalized cost on both models. OPSD, in contrast, processes fewer tokens per step but requires more time per 1K tokens. Thus, RESD achieves the lowest measured time per 1K student tokens despite its additional middle-layer computation. These results suggest that

Table 8: Average computational cost per training step over steps 11–300. Token counts include student prompts and responses. Values in parentheses for OPSD and RESD denote ratios relative to GRPO.

Model	Method	Step time (s) ↓	Resp. len	Tokens / step	Sec. / 1K tokens ↓
Qwen3-1.7B	GRPO	166.6	2372	634K	0.263
	OPSD	86.5 (0.52×)	2337 (0.99×)	78K (0.12×)	1.107 (4.21×)
	RESD	262.3 (1.57×)	5312 (2.24×)	1387K (2.19×)	**0.189 (0.72×)**
Qwen3-4B	GRPO	217.8	1768	480K	0.454
	OPSD	112.6 (0.52×)	2348 (1.33×)	79K (0.16×)	1.435 (3.16×)
	RESD	353.5 (1.62×)	5010 (2.83×)	1309K (2.73×)	**0.270 (0.59×)**

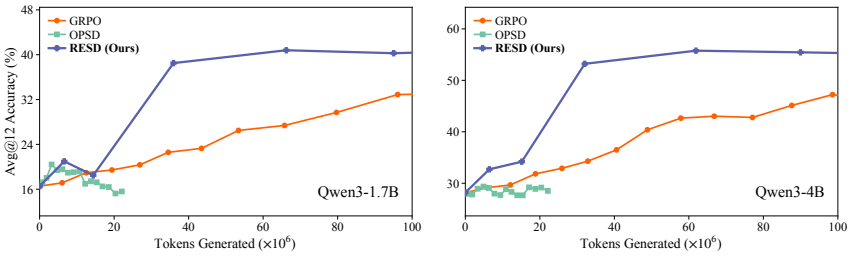


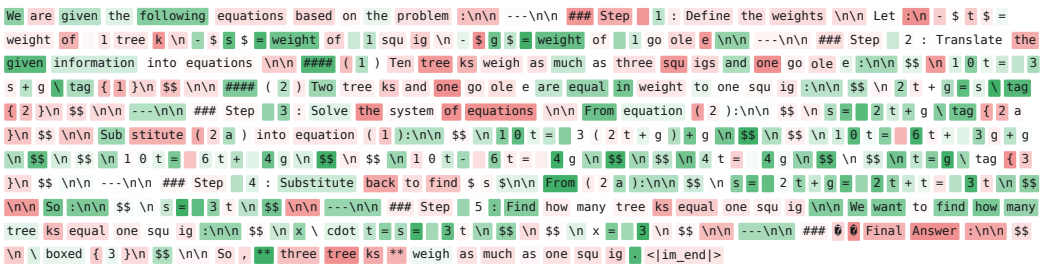
Figure 14: Mean Avg@12 accuracy across five mathematical reasoning benchmarks versus cumulative training-generated response tokens on Qwen3-1.7B (left) and Qwen3-4B (right). Markers denote evaluated checkpoints, with lines connecting adjacent observations. The horizontal axis shows the first 100M tokens; OPSD ends at its available budget of approximately 22M tokens.

its higher absolute step time is largely associated with longer generated trajectories, rather than a substantial increase in per-token computational cost.

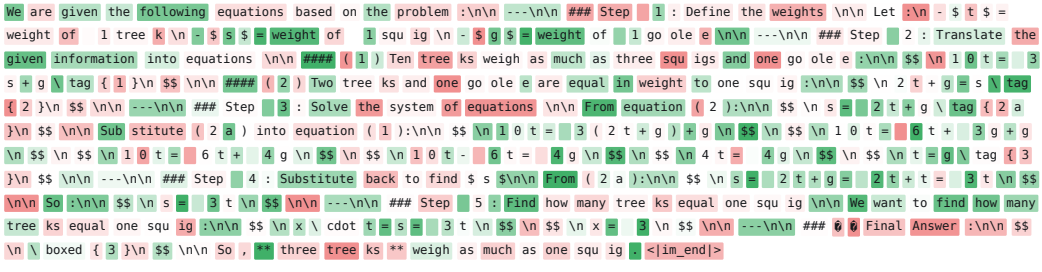
Performance under a training-token budget. The longer trajectories also raise a separate question: does RESD improve reasoning simply by consuming more tokens during training? To examine this possibility, Figure 14 compares mean Avg@12 accuracy across the five mathematical reasoning benchmarks against cumulative training-generated response tokens. For each evaluated checkpoint, we accumulate the response tokens generated from the start of training through that checkpoint. This aligns the performance curves by training-token consumption rather than optimization steps.

Across both models, RESD reaches substantially higher accuracy within roughly 40M generated tokens than GRPO achieves near 100M tokens. Its advantage is still clearly preserved when training-token consumption is taken into account. OPSD shows limited improvement over its available budget, whereas RESD develops a clear performance advantage as training progresses. These results indicate that RESD's gains are not solely a consequence of generating more training tokens, but reflect more effective learning from its rollouts.

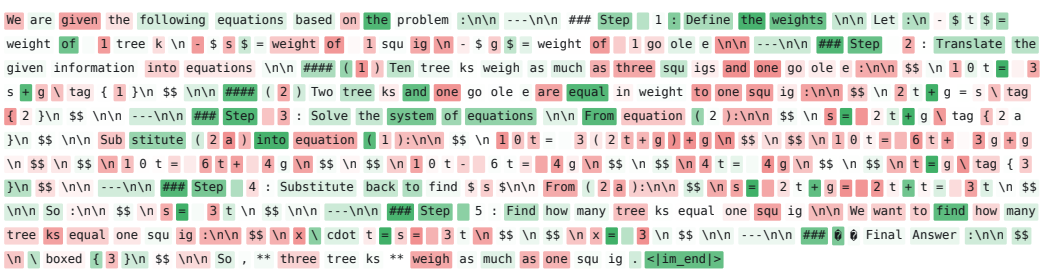
Together, the two analyses establish complementary aspects of RESD's effectiveness and efficiency. The per-step measurements show that its longer trajectories incur a less-than-proportional increase in training time, resulting in lower time per 1K tokens. The token-budget-aligned performance curves show that RESD also converts generated training tokens into reasoning improvements more effectively than GRPO. RESD therefore combines stronger reasoning performance with favorable training-token efficiency and measured throughput, despite its higher absolute cost per optimization step.



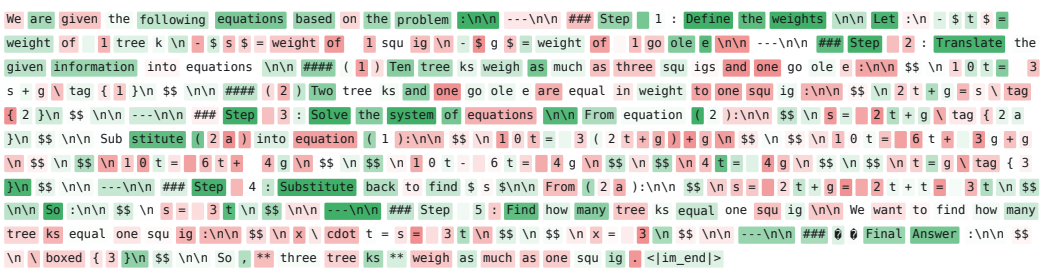
(a) Entropy H.



(b) Semantic exploration SE.



(c) Q'.



(d) Retained exploration Q.

Figure 15: Full-trajectory token-level visualizations of entropy H, semantic exploration SE, Q', and retained exploration Q.