SELD: SELECTIVE ON-POLICY DISTILLATION FOR MATHEMATICAL REASONING VIA SEMANTIC ERROR LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) trains students on their own solutions, but how should teacher supervision be allocated within failed reasoning? We investigate concentrating supervision near the first error that remains uncorrected and leads to the wrong answer. In a controlled placement study, a fixed supervision window centered on the diagnosed error outperforms the tested preceding, following, and random windows. Yet teacher–student disagreement need not identify that error. We propose **Selective On-Policy Distillation (SELD)**, which uses a frozen teacher and a reference solution to diagnose the first error and organize supervision around it. SELD emphasizes the diagnosed step and its predecessor, retains sparse earlier-prefix supervision, and allocates nonuniform weights within these regions while excluding the later continuation. On 159 expert-annotated incorrect solutions, semantic diagnosis raises first-error coverage from 44.7% with maximum-disagreement anchors to 67.9%, evaluating both anchors with SELD's predecessor-plus-anchor transition. Across four mathematical reasoning benchmarks and three training seeds, SELD improves Qwen3-1.7B macro-average response accuracy by 2.54 percentage points over full-token OPD, with gains also observed for a larger student and a smaller teacher. Repeated ablations support the error-step, prefix, and weighting choices.

1 INTRODUCTION

In on-policy distillation (OPD), a stronger teacher provides token-level knowledge distillation (KD) feedback on solutions generated by the student, called rollouts (Gu et al., 2023; Agarwal et al., 2024). This reduces the mismatch between training sequences and student outputs at inference (Kim & Rush, 2016; Lin et al., 2020). Incorrect rollouts pose a further question: which parts should receive supervision? A solution can contain valid initial reasoning, an error, and a continuation built on that error. Full-token KD weights these regions uniformly, while Correct-only KD discards the entire failed rollout. Neither uses the location where the reasoning goes wrong to organize supervision.

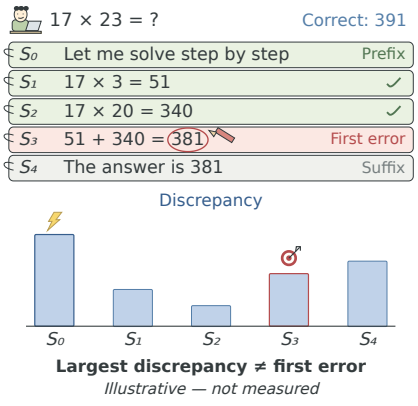

Figure 1: Illustrative example: disagreement peaks at harmless S_0; the first error occurs at S_3 and propagates to S_4.

We investigate *concentrating supervision near the first uncorrected error*. OPD already provides teacher distributions at erroneous positions; our question is whether making this neighborhood the supervision focus improves learning. In a fixed-window placement study, using the same three-step mask and offset weights, error-centered supervision achieves 35.49% Macro avg@12, compared with 32.89% before the error, 32.78% after it, and 34.31% at a random location (Section 4.3). These individual-run comparisons support the error neighborhood among the tested placements, while leaving room for complementary prefix supervision.

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

Exploiting this locality requires finding the error. Recent selective methods assess local teacher informativeness (Liu et al., 2026b), task relevance (Li et al., 2026a), or reasoning progress (Yang et al., 2026). Teacher–student disagreement alone need not locate a mathematical error: a harmless wording choice can yield a larger discrepancy (Figure 1). On 159 expert-annotated incorrect rollouts, semantic diagnosis covers the first-error step more often than maximum Jensen–Shannon divergence (JSD) selection when both use a predecessor-plus-anchor transition. This motivates examining the reasoning itself when placing supervision.

Selective On-Policy Distillation (SELD) implements this principle with a frozen teacher that uses a reference solution to diagnose the earliest error that remains uncorrected and leads to the wrong answer. The predicted step, or *anchor*, defines an *error-transition region* containing that step and its predecessor. SELD concentrates supervision there, retains sparse supervision on the earlier prefix, and excludes the later suffix. Distributional scores allocate weights within retained regions. The reference guides diagnosis; distillation targets are computed from the original student rollout.

Across four mathematical reasoning benchmarks and three training seeds, SELD improves Qwen3-1.7B macro-average response accuracy by 2.54 percentage points over full-token OPD and 2.42 points over Prefix-KD with a Qwen3-8B teacher. Gains over OPD also hold with a Qwen3-4B student and a smaller Qwen3-4B teacher. Repeated ablations test how anchor choice, prefix retention, error-step supervision, and token weighting realize the locality principle. Our contributions are:

- **First-error-localized supervision.** Fixed-window placement controls show that supervision near the diagnosed first error outperforms the tested preceding, following, and random windows.
- **SELD.** Semantic diagnosis locates the supervision focus more reliably than maximum disagreement in an annotated audit; SELD uses it to concentrate distillation on the error transition while retaining sparse prefix supervision.
- **Empirical validation.** Three-seed experiments across three student–teacher settings show gains over full-token OPD, with repeated ablations supporting the region and weighting choices.

2 Related Work

On-policy distillation and distributional selection. MiniLLM and Generalized Knowledge Distillation (GKD) provide teacher feedback on student-generated prefixes, including erroneous ones (Gu et al., 2023; Agarwal et al., 2024). MiniLLM uses reverse Kullback–Leibler (KL) divergence, while GKD supports several divergences. SELD studies how to allocate this feedback around the first error. Recent work examines teacher informativeness (Armandpour et al., 2026; Li et al., 2026b); Prefix Teach, Suffix Fade stops supervision at a drop in local teacher contrast (Liu et al., 2026b). SELD uses a semantic error location to define its supervision focus.

Task relevance and reasoning progress. Counterfactual Relevance for On-Policy Distillation (CROP) estimates task relevance using sensitivity to changed task conditions calibrated against meaning-preserving rewrites (Li et al., 2026a). Beyond Imitation filters rewards when teacher-agreement and reasoning-progress rankings conflict (Yang et al., 2026). SELD instead tests error-relative supervision placement. Our JSD controls assess disagreement-based localization; direct training comparisons with CROP and Beyond Imitation remain outside this evaluation.

First-error localization and prefix learning. Process supervision evaluates intermediate reasoning steps (Uesato et al., 2022; Lightman et al., 2023; Wang et al., 2024). Save the Good Prefix uses first-error localization to reward the preceding prefix and target later penalties during reinforcement learning (RL) (Liu et al., 2026a). SELD uses this location to concentrate distribution-matching loss on the error transition, retaining sparse prefix supervision. Its placement controls test the value of that concentration rather than introducing teacher feedback on erroneous tokens.

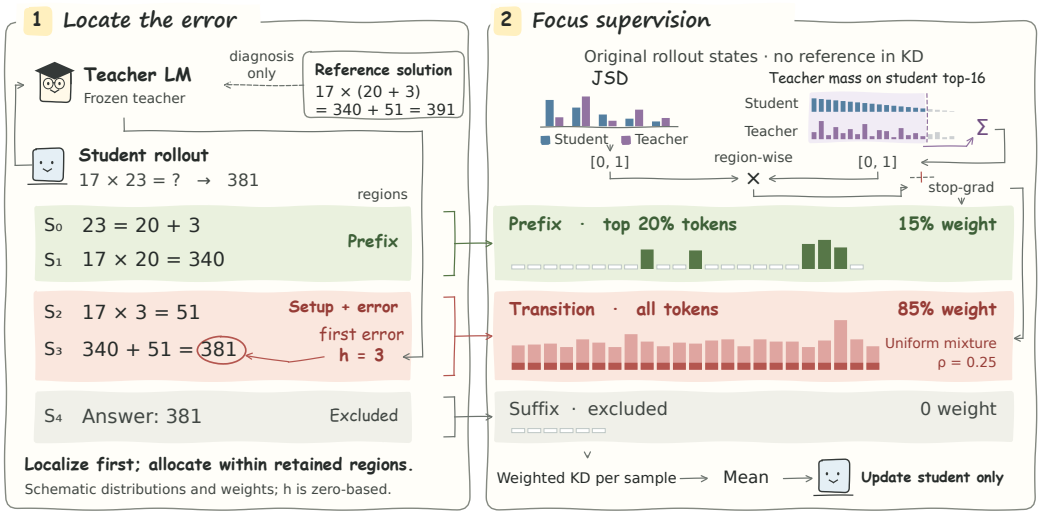

Figure 2: **Concentrating supervision around the error.** The frozen teacher language model (LM) diagnoses $h = 3$. The transition $S_2 \cup S_3$ receives 85% of the token weight, sparse earlier-prefix tokens receive 15%, and the suffix receives none. Dark bases show the transition's uniform weight floor. Per-sample KD losses are averaged to update the student. Steps and weights are illustrative.

3 SELD: Selective On-Policy Distillation

The fixed-window study tests where to concentrate supervision; SELD specifies how to use that focus. Its predecessor-plus-anchor transition accommodates one-step-late diagnoses, sparse prefix supervision retains earlier learning signal, and token weights refine allocation (Figure 2). Sections 4.4–4.6 evaluate these choices. Each training example provides a problem, reference answer, and reference solution; a frozen teacher supplies diagnosis and next-token distributions.

3.1 Overview and Formulation

For an incorrect rollout, SELD uses the diagnosed error step to define the prefix and error-transition regions. At every supervised position, including within the error step, the target is the teacher's next-token distribution; the student's sampled token is not treated as a correct label. SELD selects tokens from the prefix, weights every transition token, and averages the resulting losses to update the student. Correct or truncated rollouts, and those without a valid error prediction, use full-token KD. Rollouts, predicted steps, token selections, and weights remain fixed during each update; gradients flow through the token losses.

Let π_s and π_t denote the student and frozen teacher. For a prompt x, the student samples an on-policy rollout $y = (y_1, \ldots, y_T) \sim \pi_s(\cdot \mid x)$. Write $\pi_s^{(\gamma)}$ and $\pi_t^{(\gamma)}$ for their next-token distributions with temperature γ. The base token loss is the symmetric Jensen–Shannon divergence at each visited state $(x, y_{<t})$,

$$d_t = \mathrm{JSD}\Big(\pi_s^{(\gamma)}(\cdot \mid x, y_{<t}) \,\|\, \pi_t^{(\gamma)}(\cdot \mid x, y_{<t})\Big), \tag{1}$$

where $\mathrm{JSD}(p\|q) = \frac{1}{2}\mathrm{KL}(p\|m) + \frac{1}{2}\mathrm{KL}(q\|m)$ and $m = (p+q)/2$. We use $\gamma = 1$. Standard on-policy KD weights all completion tokens equally,

$$\mathcal{L}_{\mathrm{std}}(y) = \frac{1}{T}\sum_{t=1}^{T} d_t. \tag{2}$$

3.2 Locating the Error and Defining Its Transition

First-error diagnosis. For a complete rollout whose final answer fails verification against the reference answer, divide the response into reasoning steps $\mathcal{S} = (S_0, \ldots, S_{J-1})$. Each step is a

contiguous span of tokens, and every response token belongs to exactly one step. Appendix C.2 specifies how step markers and paragraph boundaries define these spans.

The frozen teacher receives the problem, reference solution, and numbered steps. It predicts the step index h, counted from zero, of the earliest substantive error that remains uncorrected and leads to the wrong answer. The diagnosis accepts valid alternative solutions and exploration corrected before it is used. Missing, ambiguous, or out-of-range predictions use full-token KD; the prompt and checks appear in Appendix C.1.

The reference solution enters the diagnosis prompt only. For distillation, both models predict the next token from the original problem and student prefix $(x, y_{<t})$.

Defining the prefix and error transition. Given an accepted anchor h, define the eligible prefix and error-transition regions as

$$\mathcal{R}_P = \bigcup_{k=0}^{h-2} S_k, \qquad \mathcal{R}_E = S_{\max(0,h-1)} \cup S_h, \tag{3}$$

with $\mathcal{R}_P = \varnothing$ for $h < 2$. The predicted error step determines both regions. Including its predecessor keeps the true error in the transition when the prediction is one step late. We exclude the suffix after S_h because its predictions are conditioned on a history containing the diagnosed error. Section 4.4 measures how localization errors affect this partition.

In Figure 2, $h = 3$ yields prefix S_0–S_1, transition S_2–S_3, and excluded S_4. Whether this feedback favors a corrective continuation depends on the teacher's predictions under the visited student history; covering the error step alone does not guarantee correction.

3.3 Concentrating Supervision Around the Error

SELD selects a small fraction of prefix tokens and supervises every token in the error transition. The following scores determine which prefix tokens are selected and how tokens are weighted within each region.

Token scores within each region. We combine teacher–student disagreement d_t with a measure of overlap between their likely next tokens:

$$c_t = \sum_{v \in \mathrm{TopK}(\pi_s^{(\gamma)}(\cdot \mid x, y_{<t}))} \pi_t^{(\gamma)}(v \mid x, y_{<t}). \tag{4}$$

Here, TopK returns the student's K most probable next tokens, and c_t is the total probability the teacher assigns to those tokens. Combining c_t with d_t favors positions where the models disagree while sharing likely next tokens.

Within each nonempty region $r \in \{P, E\}$, independently scale d_t and c_t to $[0, 1]$ using clipped 5th–95th percentiles, yielding $\widetilde{d}_t^{(r)}$ and $\widetilde{c}_t^{(r)}$. A constant signal maps to one. Define

$$q_t^{(r)} = \mathrm{stopgrad}\left(\widetilde{d}_t^{(r)} \widetilde{c}_t^{(r)}\right). \tag{5}$$

The $\mathrm{stopgrad}$ operation holds these scores fixed during the student update. Section 4.6 compares the product with uniform weighting and alternatives using either signal alone.

Sparse prefix, dense transition. For a nonempty prefix, we retain a fraction r_P of its tokens. Let Ω_P contain the $M_P = \max(1, \lceil r_P |\mathcal{R}_P| \rceil)$ highest-scoring tokens. Supervision applies to these tokens and every token in $\mathcal{R}_E$. Selected prefix tokens receive weights that sum to one:

$$w_t^P = \frac{\mathbf{1}[t \in \Omega_P] \exp(q_t^{(P)}/\tau)}{\sum_{j \in \Omega_P} \exp(q_j^{(P)}/\tau)}. \tag{6}$$

Here, τ controls how strongly weights favor higher scores. For the transition, we mix uniform and score-based weights, using ρ as the uniform fraction:

$$w_t^E = \frac{\rho}{|\mathcal{R}_E|} + (1 - \rho) \frac{\exp(q_t^{(E)}/\tau)}{\sum_{j \in \mathcal{R}_E} \exp(q_j^{(E)}/\tau)}. \tag{7}$$

This gives every transition token a weight of at least $\rho/|\mathcal{R}_E|$, including tokens with low disagreement. Tokens in the suffix receive zero weight.

Balancing prefix and transition supervision. The localized loss assigns a fraction α of the total token weight to the prefix and $1-\alpha$ to the transition,

$$\mathcal{L}_{\text{loc}}(y) = \begin{cases} \alpha \sum_{t\in\mathcal{R}_P} w_t^P d_t + (1-\alpha) \sum_{t\in\mathcal{R}_E} w_t^E d_t, & |\mathcal{R}_P| > 0, \\ \sum_{t\in\mathcal{R}_E} w_t^E d_t, & |\mathcal{R}_P| = 0. \end{cases} \tag{8}$$

Weights within each region sum to one, so α sets their relative weighting independently of region length. An empty prefix transfers its weight to the transition. We use $K = 16$, retain 20% of prefix tokens ($r_P = 0.2$), assign the prefix 15% of the total token weight ($\alpha = 0.15$), and set $\rho = 0.25$ and $\tau = 1$ (Figure 2).

Training objective and fallback. For each rollout y_b in a global batch $\mathcal{B}$, define

$$\ell(y_b) = \begin{cases} \mathcal{L}_{\text{loc}}(y_b), & y_b \text{ is non-truncated, incorrect, and has a valid anchor,} \\ \mathcal{L}_{\text{std}}(y_b), & \text{otherwise.} \end{cases} \tag{9}$$

As at the bottom of Figure 2, average over the global batch:

$$\mathcal{L}_{\text{SELD}} = \frac{1}{|\mathcal{B}|} \sum_{y_b\in\mathcal{B}} \ell(y_b). \tag{10}$$

Only the student is updated. Each rollout contributes equally to the batch mean, with α controlling the prefix–transition balance within rollouts that use the localized loss.

4 Experiments

We evaluate SELD across model sizes, then test its design rationale: whether the first-error neighborhood is a useful supervision focus, how to locate it, and how to allocate supervision around it.

4.1 Experimental Setup

Models and data. Our main settings use Qwen3-1.7B and Qwen3-4B students with a frozen Qwen3-8B teacher (Yang et al., 2025). We also evaluate a Qwen3-1.7B student with a smaller Qwen3-4B teacher. All methods train on the 29,434 prompts in OPSD (Zhao et al., 2026)[1]. We verified that the training set and the four evaluation sets contain no overlapping problems. SELD uses the accompanying reference solutions for diagnosis only, not for its next-token distillation targets.

Baselines. **GKD/OPD** (Agarwal et al., 2024) supervises all rollout tokens, while **Correct-only KD** discards incorrect rollouts. **Prefix-KD** implements the release-point selection rule of Prefix Teach, Suffix Fade (Liu et al., 2026b) within our shared distillation framework, retaining a distributionally selected prefix of incorrect rollouts. We compare these methods in all three settings. For the 1.7B student with the 8B teacher, we additionally include **GRPO** (Shao et al., 2024; Guo et al., 2025) as a non-KD reference. The distillation methods share uniform token-level KD on verified-correct rollouts.

Training protocol. Within each setting, distillation methods share initialization, training data, optimizer family, and rollout-generation settings, using fresh rollouts from their evolving student policies and an effective batch size of 128. GRPO uses the separate configuration in Appendix B.3. The reported comparison is restricted to the first 100 optimizer steps. The main 1.7B setup uses a maximum training length of 8192 and learning rate 3×10^{-6}. Within each setting, all distillation methods and design controls use the same learning rate, warmup, and learning-rate schedule throughout the 100-step comparison window. Teacher diagnosis uses thinking and the full reference solution, with the same Qwen-recommended sampling settings as the primary audit configuration (Appendix C.1). The 4B-teacher setting uses the same SELD configuration as the main 8B-teacher experiment; only the frozen teacher is replaced (Appendix B.1). Control objectives are specified in Appendix C.5.

[1] https://huggingface.co/datasets/siyanzhao/Openthoughts_math_30k_opsd

Table 1: Main results across three Qwen3 student–teacher settings (%). Values are three-seed means with ± sample standard deviations; bold marks column maxima within each setting. Macro averages benchmarks equally. MATH L3–5 denotes the MATH-500 subset. Per-seed results are in Appendix E.

Method	Benchmark accuracy (avg@12)				Macro	
	AIME-25	AIME-24	AMC-23	MATH L3–5	avg@12	pass@12
Qwen3-1.7B ← Qwen3-8B						
GKD / OPD	$10.28_{\pm 2.92}$	$11.95_{\pm 1.82}$	$47.40_{\pm 1.47}$	$66.65_{\pm 0.66}$	$34.07_{\pm 0.27}$	$58.33_{\pm 2.68}$
Correct-only KD	$8.15_{\pm 0.42}$	$10.83_{\pm 0.28}$	$47.16_{\pm 0.45}$	$65.68_{\pm 0.16}$	$32.96_{\pm 0.24}$	$58.66_{\pm 0.40}$
Prefix-KD	$8.80_{\pm 0.43}$	$12.87_{\pm 0.42}$	$48.25_{\pm 0.27}$	$66.84_{\pm 0.16}$	$34.19_{\pm 0.03}$	$58.28_{\pm 0.45}$
GRPO	$5.65_{\pm 0.42}$	$7.31_{\pm 0.43}$	$46.44_{\pm 0.46}$	$69.31_{\pm 0.14}$	$32.18_{\pm 0.10}$	$52.11_{\pm 0.38}$
SELD	$\mathbf{11.76}_{\pm 1.37}$	$\mathbf{13.34}_{\pm 0.48}$	$\mathbf{51.21}_{\pm 1.21}$	$\mathbf{70.13}_{\pm 0.50}$	$\mathbf{36.61}_{\pm 0.24}$	$\mathbf{61.90}_{\pm 1.90}$
Qwen3-4B ← Qwen3-8B						
GKD / OPD	$18.98_{\pm 0.42}$	$19.54_{\pm 0.42}$	$60.75_{\pm 0.45}$	$79.35_{\pm 0.16}$	$44.65_{\pm 0.09}$	$67.66_{\pm 0.67}$
Correct-only KD	$18.15_{\pm 0.42}$	$\mathbf{22.59}_{\pm 0.43}$	$62.38_{\pm 0.45}$	$78.15_{\pm 0.14}$	$45.32_{\pm 0.10}$	$\mathbf{69.21}_{\pm 0.65}$
Prefix-KD	$18.61_{\pm 0.28}$	$20.00_{\pm 0.28}$	$61.53_{\pm 0.46}$	$78.54_{\pm 0.12}$	$44.67_{\pm 0.06}$	$65.04_{\pm 0.39}$
SELD	$\mathbf{20.93}_{\pm 0.42}$	$21.11_{\pm 0.28}$	$\mathbf{63.47}_{\pm 0.46}$	$\mathbf{79.38}_{\pm 0.08}$	$\mathbf{46.22}_{\pm 0.21}$	$68.30_{\pm 0.45}$
Qwen3-1.7B ← Qwen3-4B						
GKD / OPD	$10.93_{\pm 0.42}$	$13.70_{\pm 0.43}$	$42.93_{\pm 0.37}$	$65.02_{\pm 0.10}$	$33.15_{\pm 0.13}$	$55.07_{\pm 0.47}$
Correct-only KD	$10.93_{\pm 0.42}$	$10.37_{\pm 0.42}$	$45.89_{\pm 0.46}$	$64.92_{\pm 0.12}$	$33.03_{\pm 0.10}$	$55.03_{\pm 0.67}$
Prefix-KD	$9.44_{\pm 0.28}$	$\mathbf{14.81}_{\pm 0.43}$	$45.11_{\pm 0.36}$	$\mathbf{65.49}_{\pm 0.12}$	$33.71_{\pm 0.03}$	$\mathbf{58.23}_{\pm 0.63}$
SELD	$\mathbf{11.85}_{\pm 0.42}$	$13.33_{\pm 0.74}$	$\mathbf{46.44}_{\pm 0.38}$	$65.06_{\pm 0.17}$	$\mathbf{34.17}_{\pm 0.43}$	$57.21_{\pm 0.95}$

Evaluation. We evaluate on **AIME-25**, **AIME-24**, **AMC-23**, and **MATH-500 L3–5**, with 30, 30, 46, and 367 problems, respectively. Each problem receives 12 sampled responses at temperature 0.7 and top-$p = 0.8$. We report **avg@12**, the mean response accuracy, and **pass@12**, the fraction of problems solved at least once (Chen et al., 2021). Macro averages weight the four benchmarks equally; **Macro avg@12 is the primary metric**.

During training, we evaluate steps 50 and 100 on a fixed validation set of 300 MATH Level 3–5 problems, excluding MATH-500 and screened against training and test sets. Each checkpoint generates 12 responses per problem. For each method and seed, we select the checkpoint with the highest validation avg@12 across all 3,600 responses and report every test metric from that checkpoint. Appendix B.2 details the protocol. The training-cost comparison uses step 50 for all methods.

We repeat all main methods in the three Qwen3 settings with seeds 42, 43, and 44, as well as the nine final-objective design controls in the main setting (Tables 1 and 3). Fixed-window placement controls use the same training and evaluation settings and report individual-run results separately (Table 4).

4.2 Main Results

In the main Qwen3-1.7B ← Qwen3-8B setting, SELD reaches $36.61 \pm 0.24\%$ Macro avg@12 across three seeds, compared with $34.07 \pm 0.27\%$ for OPD (Table 3). The mean gain is 2.54 percentage points, with SELD ahead in every seed. Mean Macro pass@12 also improves, from 58.33% to 61.90%. The same table groups the repeated design controls used to examine localization, supervision regions, and token weighting below.

Table 1 extends the three-seed comparison to all main baselines and model settings. SELD has the highest mean Macro avg@12 in each setting. With the 8B teacher, it exceeds Prefix-KD by 2.42 points for the 1.7B student and 1.55 points for the 4B student, with gains in every seed. For the 4B student, it also exceeds OPD by 1.57 points and Correct-only KD by 0.90 points. With the smaller 4B teacher, SELD improves over OPD by 1.02 points; its 0.46-point advantage over Prefix-KD is smaller, with a tie at the reported precision in seed 42 and gains in seeds 43 and 44. The advantage remains metric-dependent: Correct-only KD leads mean Macro pass@12 for the 4B student (69.21 versus 68.30), while Prefix-KD leads it with the 4B teacher (58.23 versus 57.21).

Table 2: Performance under a common budget of 25.61 GPU-hours for Qwen3-1.7B ← Qwen3-8B. Scores are three-seed means ± sample standard deviations (%). Cost details and individual scores appear in Appendices B.4 and E.

Method	Macro avg@12	Macro pass@12
GKD / OPD	34.85 ± 0.35	60.00 ± 1.20
Full-rollout Weighted OPD	35.55 ± 0.30	61.00 ± 1.00
SELD	$\mathbf{36.46} \pm 0.26$	$\mathbf{63.06} \pm 1.04$

Table 3: Design controls for Qwen3-1.7B ← Qwen3-8B. Metrics and three-seed statistics follow Table 1; bold marks column maxima. Controls are grouped by anchor, supervision region, and token allocation. Per-seed results are in Appendix E.

Method	**Benchmark accuracy (avg@12)**				**Macro**	
	AIME-25	AIME-24	AMC-23	MATH L3–5	avg@12	pass@12
GKD / OPD	$10.28_{\pm 2.92}$	$11.95_{\pm 1.82}$	$47.40_{\pm 1.47}$	$66.65_{\pm 0.66}$	$34.07_{\pm 0.27}$	$58.33_{\pm 2.68}$
SELD	$11.76_{\pm 1.37}$	$\mathbf{13.34}_{\pm 0.48}$	$\mathbf{51.21}_{\pm 1.21}$	$\mathbf{70.13}_{\pm 0.50}$	$\mathbf{36.61}_{\pm 0.24}$	$\mathbf{61.90}_{\pm 1.90}$
Anchor						
JSD step-max	$11.39_{\pm 1.27}$	$12.22_{\pm 0.74}$	$46.20_{\pm 0.73}$	$66.62_{\pm 0.57}$	$34.11_{\pm 0.30}$	$58.84_{\pm 0.63}$
Random	$9.72_{\pm 0.28}$	$11.57_{\pm 1.15}$	$47.46_{\pm 1.55}$	$66.37_{\pm 0.71}$	$33.78_{\pm 0.48}$	$57.95_{\pm 1.62}$
Region						
w/o anchor-step sup.	$9.26_{\pm 0.70}$	$12.68_{\pm 2.58}$	$46.68_{\pm 1.62}$	$66.93_{\pm 0.60}$	$33.89_{\pm 0.45}$	$58.07_{\pm 2.17}$
w/o prefix sup.	$\mathbf{12.04}_{\pm 0.58}$	$12.69_{\pm 0.16}$	$48.19_{\pm 1.99}$	$68.66_{\pm 1.98}$	$35.39_{\pm 1.04}$	$59.81_{\pm 2.09}$
Suffix-restored	$10.09_{\pm 1.70}$	$11.57_{\pm 0.80}$	$48.43_{\pm 2.03}$	$66.98_{\pm 0.30}$	$34.27_{\pm 0.33}$	$56.77_{\pm 1.20}$
Allocation						
Same-support Uniform Full-rollout Weighted OPD	$10.28_{\pm 0.48}$	$11.67_{\pm 1.47}$	$45.41_{\pm 0.91}$	$66.99_{\pm 0.25}$	$33.59_{\pm 0.33}$	$57.22_{\pm 1.64}$
	$11.20_{\pm 0.43}$	$12.87_{\pm 0.42}$	$48.61_{\pm 0.45}$	$68.35_{\pm 0.67}$	$35.26_{\pm 0.15}$	$60.39_{\pm 1.34}$
Disagreement-only	$11.94_{\pm 0.74}$	$12.50_{\pm 0.96}$	$48.07_{\pm 2.09}$	$68.37_{\pm 1.42}$	$35.22_{\pm 0.82}$	$59.03_{\pm 2.31}$
Compatibility-only	$10.46_{\pm 0.70}$	$12.04_{\pm 3.43}$	$49.21_{\pm 4.19}$	$68.79_{\pm 1.60}$	$35.13_{\pm 0.58}$	$58.08_{\pm 2.47}$

Performance at a matched compute budget. SELD costs 38.3% more than OPD over 50 steps (Appendix B.4). Under a common budget of 25.61 GPU-hours, however, it exceeds OPD and Full-rollout Weighted OPD by 1.61 and 0.91 points in Macro avg@12 (Table 2). Both Macro metrics improve over both baselines in every seed. This supports an advantage at the evaluated budget; broader accuracy–compute trade-offs remain untested.

4.3 Is the First-Error Neighborhood Specifically Useful?

We test placement with a simplified fixed-window objective. On incorrect rollouts, three consecutive steps receive weights $0.8, 2.0, 1.2$ in order; other tokens receive zero weight. The window covers the diagnosed error and its neighbors, the three steps immediately before it, the three immediately after it, or a random center and its neighbors. Correct rollouts retain full-token KD. Window shape and offset weights are fixed, but step lengths and boundary clipping can change token counts. Appendix C.3 gives details.

Centering supervision on the error improves Macro avg@12 by 2.60, 2.71, and 1.18 points over preceding, following, and random windows (Table 4). This supports the supervision focus among the tested placements. The study does not select SELD's exact two-step transition: localization audits motivate including the predecessor, and repeated region and weighting controls test the final allocation. In particular, restoring the whole suffix does not isolate the effect of adding just one successor step.

Table 4: **Where supervision is placed matters.** Fixed-window controls for Qwen3-1.7B ← Qwen3-8B. Offsets are relative to the diagnosed error (0). Individual-run Macro scores (%); window shape and offset weights are shared.

Window placement	Step offsets	avg@12	pass@12
Before error	$-3, -2, -1$	32.89	56.74
After error	$+1, +2, +3$	32.78	55.90
Random center	—	34.31	58.84
Centered on error	$-1, \mathbf{0}, +1$	**35.49**	**60.28**

4.4 Why Semantic Localization?

Semantic anchors improve downstream learning. With SELD's region and weighting rules fixed, semantic anchors improve Macro avg@12 by 2.50 points over JSD step-max and 2.83 over uniformly random anchors, with gains in every seed (Table 3). Random selection also changes region lengths, so this comparison does not match support-size distributions.

The selectors attend to different reasoning steps. We audit 159 independently annotated incorrect rollouts from the original Qwen3-1.7B student, sharing segmentation and gold first-error steps across selectors. Semantic diagnosis uses Qwen3-8B with thinking and the full reference. Exact first-error measures agreement with the gold step; Causal additionally includes precursors and direct consequences. Appendix D reports configurations and annotations (Table 12).

Semantic diagnosis increases exact first-error accuracy from 32.1% to 51.6% over JSD step-max (paired $p = 0.00045$). Its Causal rate rises to 83.6%, a 30.8-point gain (95% paired CI $[22.0, 39.6]$), while Harmless falls from 43.4% to 10.7% (Figure 3a). Other distributional selectors also trail semantic diagnosis (Appendix D).

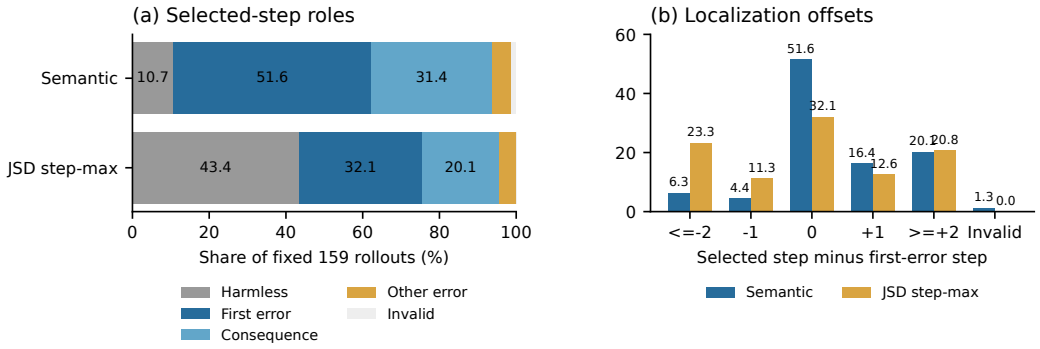

Figure 3: Roles and offsets on all 159 rollouts. Invalid outputs are shown separately; other error groups precursors and independent later errors. Offset tails are pooled.

Semantic anchors improve transition coverage. Under the same predecessor-plus-anchor rule, first-error coverage rises from 44.7% with JSD to 67.9% with semantic anchors, a 23.3-point gain (95% paired CI $[12.6, 34.0]$). Including the predecessor recovers 26 cases diagnosed one step late (Figure 3b; Appendix D.1).

On 60 recorded training responses with shared step boundaries, transition coverage among 57 clear-error cases is 12.3% for JSD, 47.4% for non-thinking diagnosis, and 52.6% for the thinking anchors used during training. Both semantic configurations lead in every stage; JSD selects step 0 on 39/60 responses. The semantic-mode difference remains unresolved ($p = 0.55$). This extends the comparison to training outputs without establishing corrective feedback (Appendix D, Table 11).

4.5 Which Regions Should Be Supervised?

We now test how SELD allocates supervision around the error: whether to include the anchor step, retain an earlier prefix, and supervise the suffix. Table 3 compares three-seed controls with diagnosis fixed.

Supervising the anchor step. Removing anchor-step supervision retains the selected prefix and predecessor, renormalizes the remaining coefficients, and uses full-token fallback for empty support (Appendix C.5). This lowers Macro avg@12 by 2.72 points on average, with SELD ahead on both Macro metrics in every seed. The result supports supervising the diagnosed anchor in addition to the retained preceding context. At the anchor, supervision uses teacher next-token distributions; the sampled erroneous tokens are not treated as correct labels.

Retaining prefix supervision. Removing prefix supervision retains the prefix in context but assigns all localized weight to the transition. Macro avg@12 falls by 1.21 points on average and in every seed. SELD gains 2.09 points in mean pass@12 but trails by 0.45 points in seed 44. Since earlier reasoning remains visible to both models, this result concerns the allocation of direct training signal to the prefix, rather than the availability of its conditioning context.

Allocating budget to the suffix. Suffix-restored transfers 20% of the transition budget to the full suffix while retaining prefix supervision. Macro avg@12 falls by 2.34 points on average, and SELD leads both Macro metrics in every seed. This favors concentrating the available budget on the transition; because suffix restoration reduces transition weight, the comparison does not isolate the effect of suffix supervision at fixed transition weight.

4.6 Does Within-Region Weighting Matter?

Same-support Uniform retains semantic anchors, token selection, and the 15%/85% regional budgets, replacing only within-region weights by uniform weights. Macro avg@12 falls by 3.02 points on average and in every seed (Table 3), supporting nonuniform allocation. Its score also falls below OPD (33.59 versus 34.07), so semantic region selection alone does not ensure a gain. Support is matched for a given rollout; training trajectories evolve separately.

Full-rollout Weighted OPD instead weights all tokens of incorrect, non-truncated rollouts, normalizing scores over the whole response and retaining uniform KD otherwise. It reaches 35.26 ± 0.15 Macro avg@12, exceeding OPD by 1.19 points. SELD improves this by 1.35 points, with gains in every seed. Mean Macro pass@12 rises from 60.39 to 61.90, although SELD trails in seed 43. This supports the additional value of SELD's region selection and allocation; the comparison does not isolate semantic anchors alone.

The product score improves Macro avg@12 over Disagreement-only and Compatibility-only by 1.39 and 1.48 points, respectively, with gains in every seed. Mean pass@12 improves against both, although Disagreement-only is 0.36 points higher in seed 43. These controls change both prefix-token selection and within-region weights, supporting the combined score rather than isolating weighting alone. Appendix C.5 gives the objectives.

5 Conclusion

First-error location provides a useful basis for organizing distillation on failed reasoning. Fixed-window controls favor the error neighborhood over preceding, following, and random placements. SELD concentrates supervision on the diagnosed error transition with sparse prefix supervision and nonuniform weights. Three-seed gains over OPD hold across three model settings. The allocation controls show that locating the error and deciding how to weight the retained tokens are complementary design choices: semantic support with uniform weights does not recover the full method's benefit. The matched-budget comparison also supports an advantage after accounting for diagnosis cost at the evaluated budget. Future work should measure corrective teacher feedback at the selected positions and test broader compute budgets and generalization beyond the model family and mathematical tasks evaluated here.

AI USE STATEMENT

Generative AI tools assisted manuscript drafting and revision, feedback on experimental design, interpretation of results, and data reformatting. The authors commit to manually reviewing and verifying all AI-assisted content before submission and take full responsibility for the paper.

ETHICS STATEMENT

The primary 159-rollout first-error annotations in Appendix D were produced by three professionals with expertise in mathematical reasoning. They annotated model-generated mathematical solutions containing no personal or sensitive data; no human-subject study was conducted and no IRB approval was required. We use the publicly released OpenThoughts-math-30k / OPSD dataset (Zhao et al., 2026) and Qwen3 student/teacher checkpoints (Yang et al., 2025) under their release terms. SELD is a training-side modification for mathematical reasoning and does not introduce new personal-data collection, human-subject interaction, or capabilities beyond those of the underlying models.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.13649.

Mohammadreza Armandpour, Fatih Ilhan, David Harrison, Ajay Jaiswal, Duc N. M. Hoang, Fartash Faghri, Yizhe Zhang, Minsik Cho, and Mehrdad Farajtabar. Unmasking on-policy distillation: Where it helps, where it hurts, and why. *arXiv preprint arXiv:2605.10889*, 2026. URL https://arxiv.org/abs/2605.10889.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. URL https://arxiv.org/abs/2107.03374.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: On-policy distillation of large language models. *arXiv preprint arXiv:2306.08543*, 2023. URL https://arxiv.org/abs/2306.08543.

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

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, 2016. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

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

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

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023. URL https://arxiv.org/abs/2305.20050.

Alexander Lin, Jeremy Wohlwend, Howard Chen, and Tao Lei. Autoregressive knowledge distillation through imitation learning. *arXiv preprint arXiv:2009.07253*, 2020.

Haolin Liu, Dian Yu, Sidi Lu, Yujun Zhou, Rui Liu, Zhenwen Liang, Haitao Mi, Chen-Yu Wei, and Dong Yu. Save the good prefix: Precise error penalization via process-supervised RL to enhance LLM reasoning. *arXiv preprint arXiv:2601.18984*, 2026a. URL https://arxiv.org/abs/2601.18984.

Kaiyuan Liu, Ziyuan Zhuang, Yang Bai, Bing Wang, Rongxiang Weng, and Jieping Ye. Prefix teach, suffix fade: Local teachability collapse in strong-to-weak on-policy distillation. *arXiv preprint arXiv:2605.13643*, 2026b. URL https://arxiv.org/abs/2605.13643.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022. URL https://arxiv.org/abs/2211.14275.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://arxiv.org/abs/2312.08935.

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

Chen Yang, Haiyuan Wan, Rengrong Xiong, Yize Chen, and Danny H. K. Tsang. Beyond imitation: Filtering on-policy distillation by reasoning progress. *arXiv preprint arXiv:2608.19408*, 2026. URL https://arxiv.org/abs/2608.19408.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Zhiyuan Zeng, Xiaonan Li, Junqi Dai, Qinyuan Cheng, Xuanjing Huang, and Xipeng Qiu. Reasoning in flux: Enhancing large language models reasoning through uncertainty-aware adaptive guidance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2401–2416, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.131. URL https://aclanthology.org/2024.acl-long.131/.

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. URL https://arxiv.org/abs/2601.18734.

A LIMITATIONS

Scope and supervision requirements. SELD is evaluated on step-by-step mathematical reasoning with Qwen3 students. The DeepSeek results in Appendix F include baselines only and do not establish cross-family robustness for SELD. The method assumes complete reference solutions and step-local errors; reference-free training and tasks with less clearly localized failures remain untested.

Diagnosis reliability. The primary audit covers fixed, structured rollouts from the original Qwen3-1.7B student. The audit also evaluates training-time semantic anchors on 60 actual responses spanning a training run, but its annotated sampling frame and small stage-specific samples limit generalization. Coverage depends on annotation version, and the thinking-mode difference is not statistically resolved. Coverage of these training-time anchors does not by itself establish corrective teacher feedback (Appendix D, Table 11). The additional benefit of reference access is uncertain at the current sample size. Valid but incorrect anchors can still misplace supervision.

Strength of the experimental evidence. All main methods in the three Qwen3 settings and all nine final-objective design controls in the main setting are evaluated over three training seeds. These repetitions measure training-run variability within the tested family and 100-step training window; they do not establish cross-family robustness or longer-run performance. Fixed-window placement controls and additional DeepSeek baselines report individual runs. The placement controls match step-window shape and weights, not token counts, and do not compare all possible error locations. The 300-problem validation set covers MATH rather than the full range of test benchmarks, and lexical screening cannot rule out all semantic overlap. Semantic diagnosis adds training computation. The matched-budget comparison tests one budget in the main setting; it does not establish an advantage across budgets or model sizes.

Implementation and evaluation settings are in Appendix B; diagnosis, segmentation, and control objectives are in Appendix C. Appendix D contains the localization protocols and diagnostics, Appendix E the per-seed results, Appendix F the additional model-family baselines, and Appendix G the localization examples.

B IMPLEMENTATION DETAILS

B.1 TRAINING AND EVALUATION SETUP

Table 5 summarizes the main SELD run. All compared distillation methods use the same student model, training data, tokenizer, optimizer family, and evaluation protocol unless otherwise stated. Shared training and decoding settings are not repeated for each control. The shared smaller-teacher configuration is specified below; the anchor-step-removal control is specified in Appendix C.5.

Training runs on a single node with 8 NVIDIA A100 80GB GPUs. All main comparisons use an effective batch size of 128 and restrict reported checkpoints to the first 100 optimizer steps.

For each problem we sample 12 completions with temperature 0.7, top-p 0.8, and a budget of $16{,}384$ generated tokens. The evaluated datasets are AIME-25, AIME-24, AMC-23, and MATH-500 L3–5. We use zero-shot evaluation with the same long-reasoning system prompt across benchmarks; avg@12 is the primary metric and pass@12 is secondary. We report an unweighted macro average across benchmarks. Answers are evaluated using the Qwen Math symbolic-equivalence grader in EvalScope. Failed answer extractions are treated as incorrect. The training and evaluation problem sets were checked and are disjoint.

Fixed selection hyperparameters. We keep the prefix retention fraction $r_P = 0.2$, transition uniform fraction $\rho = 0.25$, and student support size $K = 16$ fixed. The configured prefix budget is $\alpha = 0.15$; the prefix-supervision ablation sets it to zero. Empty-region handling and the explicit normalization or budget changes in other controls follow their definitions in Appendix C.5. These fixed settings do not establish sensitivity or optimality across hyperparameter choices.

Table 5: Training details for the main SELD run.

Setting	Value
Student model	Qwen3-1.7B
Teacher model	Qwen3-8B
Training data	OPSD Hugging Face training set, 29,434 prompts
Reported training budget	First 100 optimizer steps
Max training length	8192 tokens
Precision	bfloat16
Optimizer	AdamW (β_1=0.9, β_2=0.999)
Learning rate	3×10^{-6}
Schedule / warmup	cosine / 3% warmup
Weight decay	0
Max gradient norm	1.0
Effective batch size	128
Checkpoint interval	50 steps
GKD on-policy ratio λ	1.0 (fully on-policy)
JSD mixture coefficient β	0.5
KD temperature	1.0
Student rollout temperature	1.0
Student rollout top-p	1.0
Student rollouts per prompt	1
Max student rollout length	8192 tokens
Semantic diagnosis	Qwen3-8B with thinking and complete reference solution
Diagnosis temperature / top-p	0.6 / 0.95
Diagnosis top-k / min-p	20 / 0
Diagnosis output limit	8192 tokens (thinking and step ID)
Valid-prefix selection	top 20% of prefix tokens; 15% coefficient weight
Error-transition region	predecessor step and diagnosed first-error step
Error-region uniform floor	25%
Within-support allocation	disagreement $\times$ student-support overlap
Post-error suffix	hard mask (weight 0)
Missing / unusable diagnosis	standard full-token KD

On-policy sampling. Each method generates one fresh response per prompt from its current student policy and uses each response once. The sampling policy is refreshed after every optimizer update. Methods share prompts and sampling settings but generate their own trajectories.

Smaller-teacher configuration. The Qwen3-1.7B $\leftarrow$ Qwen3-4B experiment uses the same configuration as the main Qwen3-1.7B $\leftarrow$ Qwen3-8B experiment, replacing only the frozen teacher. The 4B teacher supplies both next-token distributions and reference-assisted step-ID diagnosis with thinking enabled. Diagnosis generation settings, student rollout and evaluation decoding, optimization, and regional allocation rules remain unchanged. Losses use the global per-sample average in Eq. 10; missing or unusable diagnoses fall back to full-token KD as in Eq. 9.

B.2 Validation Set and Checkpoint Selection

Construction and sampling. The validation set contains 300 problems from the original MATH test split, restricted to Levels 3–5. We first screen all 5,000 test problems against the complete MATH-500, AIME-24/25, AMC-23, and the 29,434 OPSD training problems. After external and internal screening, 2,690 Level 3–5 candidates remain. We proportionally sample a fixed 1,000-problem parent set over 21 subject-by-level strata, then sample the final 300 problems from that parent using the same stratification. Both stages use seed 20260924 and largest-remainder quota allocation. Neither stage selects problems using model scores. Table 6 reports the final distribution.

Table 6: Subject and difficulty distribution of the fixed validation set.

Subject	Level 3	Level 4	Level 5	Total
Algebra	18	20	24	62
Counting and Probability	9	9	12	30
Geometry	9	11	12	32
Intermediate Algebra	15	22	24	61
Number Theory	8	10	14	32
Prealgebra	19	17	17	53
Precalculus	9	9	12	30
Total	87	98	115	300

Overlap screening. Comparisons use problem statements, not generated solutions. We remove verbatim matches and matches after normalizing whitespace, math delimiters, and TeX sizing/spacing commands while preserving numbers, variables, operators, and case. We conservatively exclude pairs with normalized-character Indel similarity at least 85%, and apply an additional containment threshold of 80% over sets of unique contiguous five-token fragments to statements with at least 20 deduplicated tokens. These automatic lexical checks can remove distinct but similarly worded questions and do not prove the absence of semantic overlap.

The successive remaining counts are 5,000 (original test), 4,500 (excluding all MATH-500), 4,500 (after AIME/AMC exact and format matching), 4,499 (after training exact and format matching), 3,703 (character similarity), 3,689 (five-token containment), 3,541 (internal deduplication), and 2,690 (Levels 3–5), followed by the 1,000-problem parent and 300-problem final sample.

Evaluation inputs. Problems and reference solutions retain their original wording. Reference answers are extracted from the final boxed answer in each solution. Model inputs contain only the problem and the evaluation prompt; answers and reference solutions are used only for scoring.

Validation evaluation. Each checkpoint generates 12 responses per problem, for 3,600 responses in total, with temperature 0.7, top-p 0.8, top-k 20, and a maximum of 16,384 generated tokens. We use the Qwen3 template with thinking disabled and the same prompts and symbolic-equivalence grader as test evaluation. Validation avg@12 is the fraction of correct responses across all 3,600 generations, including truncated responses.

Checkpoint selection during training. During training, we evaluate the eligible step-50 and step-100 checkpoints on the fixed validation set. For each method and training seed, we select the checkpoint with the highest validation avg@12 and report all test metrics from that checkpoint. Test-benchmark scores are not used for checkpoint selection.

B.3 Compared method details

The current distillation comparisons share the prompts, tokenizer, optimizer family, learning rate, warmup, and learning-rate schedule within the 100-step comparison window, as well as the effective batch size of 128, rollout-sampling configuration, evaluation protocol, and correct-rollout supervision used by the main SELD run. The anchor-step-removal control retains the main SELD configuration, masking the anchor step and renormalizing the remaining coefficients on nonempty support, with full-token fallback for empty support. The main baselines (GKD/OPD, Correct-only KD, Prefix-KD, and GRPO) follow the descriptions in Section 4. Prefix-KD implements the distributional release-point method of Prefix Teach, Suffix Fade (Liu et al., 2026b); its selection settings and the separate GRPO configuration are specified below. The semantic diagnosis is described in Appendix C; the distributional audit signals are defined in Appendix D.2.

Prefix-KD configuration. Prefix-KD uses the four highest-probability student tokens as its candidate support and the gap between the two highest teacher log probabilities on this support as the local teacher-margin signal. Punkt sentence segmentation defines candidate boundaries. Supervision is truncated only when a downward change point yields a strict improvement in the Bayesian informa-

Table 7: Training cost for Qwen3-1.7B ← Qwen3-8B over 50 optimizer steps on eight A100-80GB GPUs. Evaluation is excluded; relative cost is normalized to OPD.

Method	GPU-hours	Relative cost
GKD / OPD	18.52	$1.00\times$
Prefix-KD	21.35	$1.15\times$
JSD step-max	16.53	$0.89\times$
Random	16.54	$0.89\times$
SELD	25.61	$1.38\times$

tion criterion (BIC) over the no-change model; otherwise, full-rollout supervision is retained. The selected boundary limits the KD loss on incorrect rollouts, while verified-correct rollouts receive full-token KD.

GRPO configuration. The GRPO baseline samples eight completions per prompt, normalizes rewards within each group, and uses a KL coefficient of 0.04 and a policy-ratio clipping parameter of 0.2. Its learning rate is 10^{-6} and its maximum completion length is 4,096 tokens. These settings differ from the distillation configuration above; GRPO is included as a non-KD reference, not as an isolated change to the distillation objective. Benchmark evaluation and checkpoint selection follow Section 4.1.

B.4 Compute and throughput

Table 7 reports training costs for Qwen3-1.7B ← Qwen3-8B over the first 50 optimizer steps, covering 6,400 rollouts. Each run uses eight A100-80GB GPUs and an effective batch size of 128, with no separate GPU for diagnosis. GPU-hours include all eight allocated GPUs and exclude evaluation. Relative cost is the ratio of GPU-hours to OPD.

SELD uses 38.3% more GPU-hours than OPD and 20.0% more than Prefix-KD. The anchor controls JSD step-max and Random use 16.53 and 16.54 GPU-hours, respectively, giving SELD approximately 55% higher cost than either control. These costs include semantic diagnosis for SELD.

For the measured SELD, JSD step-max, and Random runs, step-50 Macro avg@12 is 36.71%, 33.09%, and 33.67%, respectively. Their mean optimizer-step times are 230.51, 148.80, and 148.82 seconds. These run-level measurements are distinct from the three-seed accuracy summaries in Tables 1 and 3.

Comparison at a common budget. Table 2 compares OPD, Full-rollout Weighted OPD, and SELD under a budget of 25.61 GPU-hours. The reported cost records are 25.41 GPU-hours over 77 steps for OPD, 25.55 over 73 steps for Full-rollout Weighted OPD, and 25.61 over 50 steps for SELD. These are run-level cost records, not averages across the three accuracy seeds; the SELD record is the seed-42 run reported above. Cost accounting includes rollout generation, teacher scoring, diagnosis where applicable, and student updates on the allocated GPUs, excluding evaluation. Table 23 gives the individual accuracy results. This comparison uses a compute budget rather than the fixed checkpoint window of the main comparison and tests one budget in the main model setting.

Diagnosis and fallback. Among SELD's 1,710 incorrect, non-truncated rollouts, 1,464 receive an accepted anchor (85.61%). Acceptance indicates a usable step ID, not verified localization correctness. The remaining 246 rollouts fall back to full-token KD, accounting for 3.84% of all 6,400 rollouts. Truncated rollouts also use full-token KD and account for 1.95% of the total. JSD step-max and Random have no missing-anchor cases in this window; their truncation fallback rates are 0.047% and 0.063%, respectively. Separate diagnosis token costs are unavailable. JSD step-max and Random require no additional semantic-diagnosis generation.

C METHOD PIPELINE DETAILS

The pipeline first segments the sampled response and obtains a diagnosis when the response is incorrect and non-truncated. It computes teacher and student distributions on the original rollout states, constructs region weights, and averages normalized per-sample losses. Correct or truncated responses and unusable diagnoses use full-token KD. The reference solution enters the diagnosis prompt only. The following subsections specify diagnosis, segmentation, and the interventions used by the design controls.

C.1 DIRECT FIRST-ERROR DIAGNOSIS

For each verified-incorrect, non-truncated rollout, the teacher receives the problem, complete reference solution, reference final answer, and numbered student reasoning steps. Online training diagnosis uses the same thinking-enabled, full-reference configuration as the audit: temperature 0.6, top-$p = 0.95$, top-$k = 20$, min-$p = 0$, and an output limit of $8{,}192$ tokens. This limit includes thinking content and the final step ID, but excludes the input prompt. The instruction emphasizes the earliest *uncorrected* substantive error on the causal path to the wrong answer, permits valid methods that differ from the reference, and treats Step 0 as a valid candidate. This augmented context is used only to generate the discrete step ID; the reference solution is excluded from all student and teacher forward passes used to compute d_t and c_t.

The core diagnosis instruction is:

Teacher diagnosis prompt

> Reason through the problem, reference solution, and numbered student steps. Find the earliest uncorrected substantive mathematical error that remains on the causal path to the wrong final answer. Ignore an exploratory mistake that is explicitly repaired before use; do not select a later consequence of an earlier error. A different valid method is allowed. After thinking, end with exactly `FINAL_STEP: <integer>`.

We accept a single in-range integer from the required final field. Missing, ambiguous, negative, or out-of-range IDs trigger full-token KD under Eqs. 2 and 9. The returned index identifies the step used to construct the supervision regions.

C.2 STEP SEGMENTATION RULES

Reasoning steps are contiguous token spans covering the full response. Candidate boundaries are then collected in the following priority order.

1. **Major markers.** If present, SELD uses explicit reasoning markers such as Markdown headings, `Step`, `step`, `Case`, or `case`, optionally preceded by separator lines. These boundaries are accepted with a minimum spacing of 4 tokens.

2. **Numbered lists.** If no major markers are found, SELD considers numbered-list boundaries of the form `1.`, `1)`, etc. To avoid treating formula lists or option lists as reasoning steps, the list must begin with 1, the first boundary must occur within the first $\max(128, 0.2T)$ tokens, and the numbered sequence must contain a consecutive run of at least three items. Adjacent numbered spans longer than 512 tokens, or a final tail longer than 1024 tokens, cause the numbered-list fallback to be rejected.

3. **Paragraphs.** If neither major markers nor reliable numbered lists are available, SELD uses paragraph boundaries, defined as two or more newlines followed by non-whitespace text. Paragraph boundaries require a minimum spacing of 15 tokens.

Chat-template special tokens that mark the boundaries between reasoning and the final-solution block are also candidate boundaries. After candidates are merged and spacing constraints are applied, the resulting ordered spans are nonempty, contiguous, and cover all completion target tokens. The returned ID h directly selects S_h; Eq. 3 then constructs the prefix and error-transition regions.

C.3 Fixed-Window Placement Controls

The placement study in Section 4.3 uses Qwen3-1.7B with a Qwen3-8B teacher under the training and evaluation settings of Section 4.1. It studies a fixed-window objective rather than the final SELD objective. For an incorrect rollout with diagnosed anchor h, the centered window supervises steps $(h-1, h, h+1)$; the preceding and following windows use $(h-3, h-2, h-1)$ and $(h+1, h+2, h+3)$, respectively. The random control samples a center uniformly from the rollout's steps. The ordered window weights are $(0.8, 2.0, 1.2)$, nonexistent steps are omitted, and token losses are normalized by their total retained weight. Correct rollouts use full-token KD. The centered window includes the successor; SELD's final transition instead includes only the predecessor and anchor, with a separate sparse-prefix budget.

Table 8 reports individual-run benchmark accuracies; these scores are not three-seed averages. The window shape and weights are shared across placements, but different step lengths and boundary positions can change the number of supervised tokens. Methods generate fresh rollouts from their own evolving policies. The comparison therefore tests the training effect of each placement rule, not corrections on identical fixed trajectories.

Table 8: Fixed-window placement: individual-run avg@12 (%). Macro pass@12 is reported in Table 4.

Placement	AIME-25	AIME-24	AMC-23	MATH L3–5	Macro
Before error	8.89	11.11	45.65	65.92	32.89
After error	10.00	10.56	46.01	64.53	32.78
Random center	10.00	12.50	48.91	65.83	34.31
Error-centered	10.56	12.50	50.18	68.70	35.49

C.4 Anchor Substitution Controls

Per-seed anchor-substitution results appear in Appendix E; this section specifies the selectors used in Section 4.4. JSD uses the maximum token-level score within each step, with ties resolved by ascending step ID. Random samples uniformly over all segmented steps, including the first and last, and fixes the anchor for the update. Both selectors use the same step segmentation and region construction as Semantic, without querying its reference-assisted diagnosis. All variants retain Product weighting, the 20% prefix selection rate, the 15% prefix budget, and the 25% transition uniform floor.

C.5 Region and Weighting Controls

These controls use the Qwen3-1.7B ← Qwen3-8B student–teacher pair and the shared evaluation protocol. Same-support Uniform and Suffix-restored retain the routing in Eq. 9 and the global per-sample batch average in Eq. 10.

Removing prefix supervision. This control preserves the complete rollout and the contexts used for student and teacher predictions; it changes only direct KD supervision. It sets the prefix coefficient α to zero and assigns the entire localized-sample loss budget to the transition. The localized loss is

$$\mathcal{L}_{\mathrm{noP}}(y) = \sum_{t \in \mathcal{R}_E} \left[\frac{0.25}{|\mathcal{R}_E|} + 0.75 \frac{\exp(q_t^{(E)}/\tau)}{\sum_{j \in \mathcal{R}_E} \exp(q_j^{(E)}/\tau)} \right] d_t, \qquad \tau = 1. \tag{11}$$

Transition weights sum to one, replacing the usual 85% transition budget by 100%; no further division by the full rollout length is applied. Prefix and suffix tokens receive zero direct loss weight on this route, while prefix tokens remain in the conditioning context. Semantic diagnosis, transition construction, full-token fallback, and the global per-sample batch average are unchanged.

Removing anchor-step supervision. The anchor-step-removal control retains SELD's reference-assisted diagnosis, prefix token selection, standard full-token fallback cases, and global per-sample batch average. It additionally uses full-token fallback if masking the anchor leaves empty support.

On the localized route, it removes supervision from the diagnosed anchor step and renormalizes the remaining token coefficients. Let h_i be the diagnosed step for rollout i, and let c_{it} denote the full SELD token coefficient in Eq. 8, including its regional budget. Define the retained coefficient mass and normalized coefficients as

$$Z_i = \sum_t c_{it}\mathbf{1}[t \notin S_{h_i}], \qquad \bar{c}_{it} = \frac{c_{it}\mathbf{1}[t \notin S_{h_i}]}{Z_i} \quad (Z_i > 0). \tag{12}$$

The ablated localized loss is

$$\mathcal{L}_{\text{loc}}^{\text{SPO}}(y_i) = \begin{cases} \sum_t \bar{c}_{it} d_{it}, & Z_i > 0, \\ \mathcal{L}_{\text{std}}(y_i), & Z_i = 0. \end{cases} \tag{13}$$

For nonempty retained support, coefficients sum to one, preserving the per-sample loss normalization. Renormalization preserves the relative coefficients of the selected prefix and predecessor tokens but rescales their absolute weights; it does not retain their original regional masses. The suffix remains masked when retained support is nonempty. Coefficients and the normalizer are held fixed during the update. Correct, truncated, and missing-anchor rollouts follow the same full-token fallback as SELD. If masking the anchor leaves no retained coefficient mass ($Z_i = 0$, including a valid anchor at the first step), the sample falls back to full-token KD. It remains in the global batch denominator and contributes the standard normalized loss. The batch objective is

$$\mathcal{L}_{\text{SPO}} = \frac{1}{|\mathcal{B}|} \sum_{y_i \in \mathcal{B}} \begin{cases} \mathcal{L}_{\text{loc}}^{\text{SPO}}(y_i), & y_i \text{ is non-truncated, incorrect, and has a valid anchor,} \\ \mathcal{L}_{\text{std}}(y_i), & \text{otherwise.} \end{cases} \tag{14}$$

Thus, nonempty localized samples remove anchor-step supervision with unit coefficient mass, while empty-support samples receive normalized full-token KD. No samples are dropped and no separate branch means are introduced. The three-seed results are reported in Table 15.

Single-signal selection and weighting. Compatibility-only and Disagreement-only replace the Product score in Eq. 5 throughout the token-selection and weighting procedure. For each nonempty region $r \in \{P, E\}$, define

$$q_{t,C}^{(r)} = \text{stopgrad}\left(\widetilde{c}_t^{(r)}\right), \qquad q_{t,D}^{(r)} = \text{stopgrad}\left(\widetilde{d}_t^{(r)}\right). \tag{15}$$

Each signal uses the same region-wise clipped 5th–95th percentile normalization as SELD, including mapping a constant signal to one. For variant $a \in \{C, D\}$, the prefix set $\Omega_{P,a}$ contains the $M_P = \max(1, \lceil 0.2|\mathcal{R}_P| \rceil)$ tokens with the largest $q_{t,a}^{(P)}$, using the same tie-handling rule as SELD. We substitute $q_{t,a}^{(P)}$ and $\Omega_{P,a}$ into Eq. 6, and $q_{t,a}^{(E)}$ into Eq. 7, with $\tau = 1$ and $\rho = 0.25$. Thus, the single signal determines both which prefix tokens are retained and the weights within the retained prefix and dense transition.

Both variants retain reference-assisted semantic diagnosis, the prefix/transition partition, supervision of every transition token, suffix masking, the 15/85% regional budgets, and the empty-prefix rule. Compatibility uses the same student top-K support with $K = 16$. The token loss remains JSD d_t in both variants; Compatibility-only removes disagreement from the selection/weighting score, not from the distillation loss. Routing and batch reduction follow Eqs. 9–10. For a given rollout and anchor, the eligible regions and prefix retention count are unchanged, but the selected prefix-token identities can differ from Product SELD. These comparisons therefore test the score used for both selection and weighting; Same-support Uniform below isolates uniform versus nonuniform weighting while preserving Product-based prefix selection.

Uniform weights on the selected support. Same-support Uniform preserves the semantic anchor, Product scores and tie handling, the top-20% prefix selection Ω_P, every transition token, and the suffix mask. Only the weights change: $w_t^P = \mathbf{1}[t \in \Omega_P]/|\Omega_P|$ on a nonempty prefix and $w_t^E = 1/|\mathcal{R}_E|$. Regional budgets remain 15% and 85%, with the entire budget assigned to the transition when the prefix is empty. This matches the selection procedure for a given rollout, not the realized rollouts across training runs.

Weighting the full rollout. Full-rollout Weighted OPD removes semantic diagnosis and region selection. On an incorrect, non-truncated rollout, it normalizes disagreement d_t and compatibility c_t over all T completion tokens using the clipped 5th–95th percentile rule underlying Eq. 5, including mapping a constant signal to one. Write $q_t^{(F)} = \mathrm{stopgrad}(\widetilde{d}_t^{(F)}\widetilde{c}_t^{(F)})$ for their product. Every completion token receives a mixture of uniform and score-based weight,

$$w_t^F = \frac{\rho}{T} + (1-\rho)\frac{\exp(q_t^{(F)}/\tau)}{\sum_{j=1}^{T}\exp(q_j^{(F)}/\tau)}, \qquad L_F(y) = \sum_{t=1}^{T} w_t^F d_t. \tag{16}$$

The score and mixture use the shared settings $K = 16$, $\rho = 0.25$, and $\tau = 1$. Correct or truncated rollouts use L_{std}; all other rollouts use L_F, without an anchor-validity condition. Weights remain fixed during each update, and normalized per-sample losses are averaged over the global batch as in Eq. 10. Checkpoints follow the same validation selection protocol as the other methods. This control changes support, score-normalization scope, and regional allocation relative to SELD.

Restoring the suffix. Let $\mathcal{R}_S = \bigcup_{k=h+1}^{J-1} S_k$, and let L_P, L_E, L_S denote normalized regional losses. Prefix selection and Product weighting are unchanged. Transition and suffix each normalize their Product scores within the region and use

$$w_t^r = \frac{\rho}{|\mathcal{R}_r|} + (1-\rho)\frac{\exp(q_t^{(r)}/\tau)}{\sum_{j\in\mathcal{R}_r}\exp(q_j^{(r)}/\tau)}, \quad r \in \{E, S\}, \quad \rho = 0.25, \quad \tau = 1. \tag{17}$$

Set $m_P = 0.15$ for a nonempty prefix and zero otherwise, and $\eta = 0.2$ for a nonempty suffix and zero otherwise. The localized loss is

$$L_{\mathrm{SR}} = m_P L_P + (1-m_P)(1-\eta)L_E + (1-m_P)\eta L_S. \tag{18}$$

Empty-region terms are omitted. The prefix/transition/suffix budgets are 15/68/17% when both prefix and suffix exist, 0/80/20% without a prefix, 15/85/0% without a suffix, and 0/100/0% when both are absent. Here η controls the suffix budget, whereas ρ controls uniform mixing within a region. Restoring the suffix also reduces transition weight, so this comparison tests budget allocation rather than the effect of adding suffix supervision at a fixed transition weight.

D Anchor-Quality Audit

The audit regenerates 300 distinct problems selected from the training-log candidate pool with the original Qwen3-1.7B model before SELD training. Selection retains 189 eligible problems from the original 300-problem sample and replaces 111 within the same source, original segmentation-kind, and training-quartile strata; eligibility requires the old completion to be non-truncated and at most 4096 tokens. Generation uses explicit numbered-step and final-answer headings, thinking disabled, temperature 1, top-p 1, and a 7168-token limit. These are newly generated completions from the original model, not completions recorded during SELD training. Of 170 non-truncated, verifier-incorrect outputs, 159 have a clear substantive first error. Two mathematics-reasoning domain experts independently annotate the earliest uncorrected substantive error on the causal path to the wrong answer; a third adjudicates disagreements (52 of the 170 candidate records). Annotators receive the problem, reference answer and solution, and fixed numbered reasoning, but not model identities or signal outputs. All evaluated signals share the 159 examples and frozen segmentation.

Agreement and scope. Recorded pre-adjudication first-error agreement is $118/170 = 69.41\%$ over all candidate solutions and $109/159 = 68.55\%$ on the final clear-error subset. Only aggregate pre-adjudication agreement is available for verification. A separate comparison over 391 rollout–step pairs gives nine-category role agreement of 88.49% and coarse-category agreement of 92.84%. Because 110 pairs are assigned directly by the first-error rule, these role-agreement rates are descriptive label-consistency measures, not independent human inter-annotator agreement estimates. They also do not measure agreement over all nine selector configurations in the current audit.

The primary semantic row in Table 12 uses Qwen3-8B with thinking enabled and the full reference solution. All four configurations share rollouts, segmentation, gold labels, top-$k = 20$, min-$p = 0$, and an $8{,}192$-token output limit, including thinking and the final step ID. Following Qwen

Table 9: Paired causal-selection differences on all 159 rollouts. Invalid outputs count as no causal selection; Adaptive-EMA also reflects its lower valid-anchor rate. Intervals and exact McNemar tests are unadjusted.

Semantic vs.	n	Δ causal (95% CI)	p
JSD	159	$+30.8$ $[22.0, 39.6]$	1.38×10^{-9}
Sampled-gap	159	$+25.2$ $[15.7, 34.6]$	1.65×10^{-6}
Adaptive-EMA	159	$+65.4$ $[56.6, 73.6]$	7.55×10^{-26}

Table 10: Complete role distribution on all 159 rollouts (percent). Roles and invalid outputs partition the full population.

Signal	**Format**	**Valid**	**Repaired**	**Precursor**	**First error**	**Conse-quence**	**Later error**	**Correction**	**Final restate**	**Invalid**
Semantic	1.3	5.0	3.8	0.6	51.6	31.4	4.4	0.6	0.0	1.3
JSD step-max	13.2	21.4	3.8	0.6	32.1	20.1	3.8	0.0	5.0	0.0
Sampled-gap	10.1	18.9	3.1	0.6	22.6	35.2	4.4	1.3	3.8	0.0
Adaptive-EMA	0.6	5.0	3.8	1.3	4.4	12.6	1.3	0.0	0.0	71.1

recommendations for each mode, thinking-enabled diagnosis uses temperature 0.6 and top-$p = 0.95$; thinking-disabled diagnosis uses temperature 0.7 and top-$p = 0.8$. Diagnosis uses random seed 20260921. All configurations use the same step-index acceptance rule; invalid outputs count as misses. The thinking-enabled full-reference condition uses the same diagnosis sampling settings as online training (Appendix C.1). JSD and sampled-gap use within-step maxima; Adaptive-EMA retains its frozen persistence rule.

All distributional signals are replayed on these structured completions and fixed step spans, using the same original Qwen3-1.7B model that generated them, before SELD training and Qwen3-8B as teacher. JSD is symmetric full-vocabulary divergence at temperature 1; sampled-gap is the teacher minus student log-probability of the generated token. Both select the step containing their largest token-level value. Adaptive-EMA applies its frozen smoothing, threshold, skip, and persistence rules. This controls the replay state across distributional signals but does not reconstruct localization performance across evolving training checkpoints or unstructured rollouts.

Selected-step role audit. The audit compares five distributional selectors with four semantic configurations on the same 159 expert-annotated rollouts. Identical rollout–step pairs have consistent labels across all nine configurations. Roles distinguish formatting, valid reasoning, repaired exploration, causal precursor, first causal error, direct consequence, independent later error, self-correction, and final-answer restatement.

We define *causal selection* as choosing a precursor, first causal error, or direct consequence, and *harmless selection* as choosing formatting, valid reasoning, repaired exploration, self-correction, or final-answer restatement. An independent later error belongs to neither group. All rates, including complete role distributions and support diagnostics, use all 159 rollouts. Invalid or abstaining predictions are localization misses and count as neither causal nor harmless; they remain in every denominator and are reported separately. Paired comparisons use 10,000 bootstrap resamples and exact two-sided McNemar tests, without multiple-comparison adjustment.

Paired localization and reference comparisons. Semantic exceeds JSD step-max in exact first-error accuracy by 19.5 points (95% CI $[9.4, 29.6]$; $p = 0.00045$) and in Causal by 30.8 points (95% CI $[22.0, 39.6]$; $p = 1.38 \times 10^{-9}$). Within thinking-enabled diagnosis, adding the reference changes Causal by 3.8 points (95% CI $[-1.9, 10.1]$; $p = 0.307$). The latter interval includes zero, so the observed reference gain is not statistically established.

Sampling actual training outputs. We use responses recorded during SELD training, without regenerating student rollouts. From 330 previously annotated training responses, we exclude truncated

Table 11: First-error-step coverage on the same 60 recorded training responses. All methods share step boundaries and use the anchor-plus-predecessor transition. Coverage denotes any token intersection with the human-labeled first-error step; denominators include all 57 clear-error responses, counting failed diagnoses as misses.

Stage (training step)	JSD step-max	Semantic, no thinking	Semantic, thinking
Early (0–33)	3/18 (16.7%)	10/18 (55.6%)	11/18 (61.1%)
Middle (34–67)	1/20 (5.0%)	8/20 (40.0%)	8/20 (40.0%)
Late (68–100)	3/19 (15.8%)	9/19 (47.4%)	11/19 (57.9%)
Total	7/57 (12.3%)	27/57 (47.4%)	30/57 (52.6%)

outputs and deterministically select 20 from each training stage: early (steps 0–33), middle (34–67), and late (68–100). Selection is independent of previous anchor validity, first-error clarity, response length, and new diagnosis results. JSD step-max and both semantic configurations use the same 60 responses and step boundaries. This sample represents the annotated pool rather than all training outputs.

Training anchors and comparison configurations. Qwen3-8B receives the problem, full reference solution, and numbered steps of the original student response. The reference is used only for diagnosis; the student response is not rewritten. Both configurations share the prompt, step boundaries, and acceptance criteria for valid, complete step predictions. Thinking-disabled diagnosis uses temperature 0.7 and top-p 0.8; thinking-enabled diagnosis uses temperature 0.6 and top-p 0.95. Both use top-k 20 and an 8,192-token output cap. This compares two mode-specific configurations, not a compute-matched intervention changing only the thinking switch. The thinking-enabled semantic anchors are those used during SELD training; the non-thinking configuration provides a comparison on the same responses. JSD step-max selects the step containing the maximum token-level teacher–student JSD. All three methods use the same anchor-plus-predecessor transition for coverage evaluation.

Human labels and metrics. The primary analysis uses the revised human first-error annotations: 57 responses have a clear first error, two have no substantive error, and one has insufficient information for a clear judgment. Anchor validity requires passing format, range, and termination-state checks; it does not imply correct localization. First-error-step coverage requires a token intersection between the annotated first-error step and the transition formed by the predicted anchor and its predecessor. Its denominator includes all 57 clear-error responses, counting localization failures as misses. Coverage of a step is weaker than coverage of the actual erroneous expression and does not establish corrective teacher feedback. We also report the stricter criterion that the transition contain the entire annotated first-error step.

Coverage on training outputs. Table 11 shows that semantic diagnosis without and with thinking exceeds JSD step-max by 35.1 and 40.4 percentage points, respectively. Both semantic configurations have higher coverage in each stage. JSD selects step 0 on 39/60 responses (65.0%), indicating a strong preference for the start of the response. This position statistic is distinct from the rate of selecting before the annotated error. Semantic anchor validity remains much higher than coverage: without thinking, valid outputs number 20/20, 20/20, and 19/20 across stages (59/60 overall); with thinking, they number 20/20, 20/20, and 18/20 (58/60 overall).

Comparing the two semantic modes, seven cases change from uncovered to covered and four change in the opposite direction, giving a gain of 5.3 percentage points (paired bootstrap 95% interval $[-5.3, 15.8]$; exact McNemar $p = 0.55$). The interval includes zero, so a stable advantage is not established. Full-step containment is 29/57 (50.9%) with thinking and 26/57 (45.6%) without it. Using the earlier annotations, the corresponding step-intersection counts are 33/58 and 26/58 ($p = 0.092$); the direction is unchanged, but the magnitude depends on annotation version. The small stage-specific samples do not establish a trend in localization quality over training. Differences in the sampling population and coverage definition preclude interpreting the contrast with the structured 159-rollout audit as a measured decline during training. Anchor validity and first-error coverage do not by themselves demonstrate effective correction.

Table 12: Anchor selection on the same 159 rollouts. All rates use denominator 159. Invalid or abstaining outputs count as localization misses and neither causal nor harmless. On + Full is the primary semantic configuration; decoding details are in Appendix D.

Distributional signals							
Selector		**Valid anchor**	**Exact first-error**	**Within ±1**	**Causal ↑**	**Harmless ↓**	**Invalid**
JSD step-max		100.0%	32.1%	56.0%	52.8%	43.4%	0.0%
JSD-mean		100.0%	27.7%	52.8%	40.3%	57.9%	0.0%
NLL-jump		100.0%	23.3%	54.7%	64.2%	33.3%	0.0%
Semantic diagnosis (Qwen3-8B)							
Thinking	**Reference**	**Valid anchor**	**Exact first-error**	**Within ±1**	**Causal ↑**	**Harmless ↓**	**Invalid**
Off	None	96.9%	40.9%	61.6%	69.8%	22.6%	3.1%
Off	Full	96.9%	40.3%	62.9%	72.3%	20.1%	3.1%
On	None	98.1%	**51.6%**	69.8%	79.9%	12.6%	1.9%
On	Full	98.7%	**51.6%**	**72.3%**	**83.6%**	**10.7%**	1.3%

Table 13: Gold first-error location under the anchor-to-region rule. Each row partitions all 159 rollouts. Prefix indicates eligibility; invalid outputs have no localized support.

Anchor	Transition	Prefix	Suffix	Invalid
Semantic	108 (67.9%)	32 (20.1%)	17 (10.7%)	2 (1.3%)
JSD step-max	71 (44.7%)	33 (20.8%)	55 (34.6%)	0 (0.0%)

Thinking and reference access. In the structured-rollout audit, the thinking-enabled configurations reach 51.6% exact first-error accuracy with or without the reference, compared with 40.3% and 40.9% when thinking is disabled (Table 12). Within the thinking-enabled mode, adding the reference raises Causal from 79.9% to 83.6% but leaves exact accuracy unchanged; the reference benefit is not statistically resolved. These are mode-specific decoding configurations, so their differences do not isolate the thinking switch at equal compute.

D.1 Selection-Pattern Diagnostics

Analysis population. The audit evaluates nine configurations on the same 159 rollouts, yielding 1,431 method–rollout pairs. All rates and paired comparisons use the full population. Each prediction is checked against the available step indices and gold first-error annotation; identical selected steps receive the same role label across methods. Figure 3 summarizes the resulting role and offset distributions.

Signed offsets and uncertainty. The offset is $\Delta_{\text{step}} = \hat{s} - s^\star$. We use 10,000 paired rollout-level bootstrap resamples (seed 20260913). Semantic has an early-selection rate 23.9 points below JSD (95% CI $[-32.1, -16.4]$); the late-selection difference is 3.1 points (CI $[-6.3, 12.6]$). Its direct-consequence rate is 11.3 points higher (CI $[3.1, 19.5]$). Of 75 valid but non-exact semantic predictions, 50 are direct consequences. Of 108 JSD misses, 69 are harmless. These exploratory intervals are unadjusted; causal-chain selection and exact recovery remain distinct outcomes.

Anchor-to-support diagnostic. For a valid anchor h, the gold first error $s^\star$ lies in the transition if $h \in \{s^\star, s^\star + 1\}$, in the eligible prefix if $h \geq s^\star + 2$, and in the excluded suffix if $h < s^\star$. Table 13 applies this rule to all 159 rollouts. The paired transition-inclusion difference is 23.3 points (95% CI $[12.6, 34.0]$). Invalid outputs are reported separately. Prefix eligibility does not guarantee retention after sparse selection, and transition membership does not establish corrective gradients. No realized training fallback weights are reconstructed.

What do incorrect JSD selections contain? Among the 108 JSD top-1 misses, 34 select valid reasoning, 21 formatting, six repaired exploration, and eight final-answer restatement (69 harmless selections). Another 32 select a direct consequence, six an independent later error, and one a causal precursor. Valid reasoning and formatting account for 55 of these misses, illustrating how large distributional scores can occur away from substantive errors.

Aggregation and diagnostic context. JSD-mean averages token scores within each step, providing a comparison with maximum aggregation, which can favor longer steps. The observed exact first-error accuracy difference is -4.4 points for mean relative to max (Appendix D.2). Reference access is varied within each semantic mode using Qwen-recommended settings. Distributional selectors use token probabilities, while semantic diagnosis additionally uses generated reasoning and, in the full-reference condition, the reference solution.

D.2 Selector Definitions and Additional Results

The full comparison includes JSD step-max, sampled-gap step-max, Adaptive-EMA, JSD-mean, and NLL-jump alongside four semantic configurations (Table 14). NLL-jump adapts the adjacent-token uncertainty-increment signal of UAG (Yin et al., 2024), not the full adaptive-guidance pipeline. JSD-mean differs from JSD step-max by -4.4 points in exact first-error accuracy (95% paired CI $[-12.6, 3.8]$; $p = 0.371$). This comparison shows no clear exact-localization advantage for mean aggregation in the present audit.

Token scores versus step aggregation. Let S_k be a nonempty step and a_t its token-level score. Step-max and step-mean aggregate that score as

$$A_k^{\max} = \max_{t \in S_k} a_t, \qquad A_k^{\text{mean}} = \frac{1}{|S_k|} \sum_{t \in S_k} a_t, \qquad \hat{k} = \arg\max_k A_k. \tag{19}$$

Thus, the token signal and its aggregation rule are separate choices. JSD-max and JSD-mean use $a_t = d_t$ from Eq. 1; Sampled-gap step-max uses $a_t = g_t$ from Eq. 20. Peak-based selectors break tied step maxima by the smallest step ID, matching the replay implementation, and do not apply smoothing or threshold detection.

Sampled-gap signal. For each generated token y_t, the shared signed log-probability gap is

$$g_t = \log p_T(y_t \mid x, y_{<t}) - \log p_S(y_t \mid x, y_{<t}), \tag{20}$$

which is large where the teacher is more confident than the student on the sampled token. The audit's **Sampled-gap step-max** uses this raw signed signal: it selects the step with the largest within-step maximum, without EMA smoothing, prefix skipping, or threshold detection. For the threshold construction used by Adaptive-EMA, a fixed-threshold precursor smooths the signal, scans after a skipped prefix, and returns the first token meeting the absolute, increment, and persistence conditions below, then maps that token to its step. This differs from maximum aggregation. The smoothing is

$$\tilde{g}_t \;=\; \lambda_{\text{EMA}}\, g_t + (1 - \lambda_{\text{EMA}})\, \tilde{g}_{t-1}, \tag{21}$$

with $\tilde{g}_1 = g_1$. The threshold-based selector scans $\tilde{g}_t$ from a skip prefix (the larger of the persistence window w and a 25%-of-length prefix, further extended past the chat-template / `<think>` prelude) and returns the first index $t^\star$ such that all three of the following hold:

$$\tilde{g}_{t^\star} \;\geq\; \tau_{\text{abs}}, \tag{22}$$

$$\tilde{g}_{t^\star} - \tilde{g}_{t^\star - 1} \;\geq\; \tau_{\text{jump}}, \tag{23}$$

$$\tfrac{1}{w} \sum_{t^\star \leq t < t^\star + w} \tilde{g}_t \;\geq\; \tau_{\text{abs}}. \tag{24}$$

We use the global defaults λ_{EMA}=0.3, w=8, τ_{abs}=0.8 nats, τ_{jump}=0.5 nats. The selected token is mapped to the enclosing step under the segmentation rules in Appendix C.2.

Table 14: All nine configurations. Every metric uses all 159 rollouts. Invalid includes abstentions and counts as no causal or harmless selection. Semantic labels specify Thinking / Reference.

Selector	Valid anchor	Exact first-error	Within ± 1	Causal	Harmless	Invalid
JSD step-max	100.0%	32.1%	56.0%	52.8%	43.4%	0.0%
Sampled-gap step-max	100.0%	22.6%	57.2%	58.5%	37.1%	0.0%
Adaptive-EMA	28.9%	4.4%	13.2%	18.2%	9.4%	71.1%
JSD-mean	100.0%	27.7%	52.8%	40.3%	57.9%	0.0%
NLL-jump	100.0%	23.3%	54.7%	64.2%	33.3%	0.0%
Semantic Off / None	96.9%	40.9%	61.6%	69.8%	22.6%	3.1%
Semantic Off / Full	96.9%	40.3%	62.9%	72.3%	20.1%	3.1%
Semantic On / None	98.1%	51.6%	69.8%	79.9%	12.6%	1.9%
Semantic On / Full	98.7%	51.6%	72.3%	83.6%	10.7%	1.3%

Adaptive-EMA signal. The Adaptive-EMA selector keeps the same per-token gap g_t, the same EMA smoothing $\tilde{g}_t$, the same persistence-window test, and the same skip-prefix as the threshold-based Sampled-gap variant, but replaces the two fixed thresholds with data-adaptive ones derived from the rollout itself. Let $\mathrm{med}(\cdot)$ and $\mathrm{MAD}(\cdot)$ denote the median and the median-absolute-deviation, write $c = 1.4826$, and let $\delta_t = \tilde{g}_t - \tilde{g}_{t-1}$. Define rollout-level estimates

$$\hat{\sigma}_g = c \cdot \mathrm{MAD}(\tilde{g}), \tag{25}$$

$$\hat{\sigma}_\delta = c \cdot \mathrm{MAD}(\delta), \tag{26}$$

and set the thresholds as

$$\tau_{\mathrm{abs}} = \max\big(\mathrm{med}(\tilde{g}) + k_{\mathrm{abs}}\, \hat{\sigma}_g,\, \tau^{\min}_{\mathrm{abs}}\big), \tag{27}$$

$$\tau_{\mathrm{jump}} = \max\big(\mathrm{med}(\delta) + k_{\mathrm{jump}}\, \hat{\sigma}_\delta,\, \tau^{\min}_{\mathrm{jump}}\big). \tag{28}$$

We use k_{abs}=2.0, k_{jump}=1.5, $\tau^{\min}_{\mathrm{abs}}$=0.3, $\tau^{\min}_{\mathrm{jump}}$=0.15. Compared with the fixed-threshold Sampled-gap variant, the only change is the threshold rule: rather than a single global $(0.8,\, 0.5)$ pair, each rollout calibrates its own thresholds from its own gap distribution, so the same selector adapts to flat versus spiky gap profiles across rollouts.

NLL-jump signal. Following the uncertainty-increment signal of Yin et al. (2024), NLL-jump uses the student's negative log probability of each generated token and the increment between adjacent tokens:

$$u_t = -\log p_S(y_t \mid x, y_{<t}), \qquad \Delta u_t = u_t - u_{t-1}, \quad t \geq 2. \tag{29}$$

We select the token with the largest increment and use the step containing that token as the anchor:

$$t^\star = \arg\max_{2 \leq t \leq T} \Delta u_t, \qquad \hat{k} \text{ such that } t^\star \in S_{\hat{k}}. \tag{30}$$

This is a student-side token-level distributional signal, not a difference between teacher-scored step means or full-vocabulary predictive entropy.

Role-assessment protocol. All methods in Table 14 use all 159 rollouts and the same annotated step-role labels. Semantic selects one precursor, 82 first errors, and 50 direct consequences, totaling 133/159 (83.6%). Its two invalid outputs remain in the denominator. Identical selected steps have identical labels across methods.

Position and role preferences. JSD-mean selects before the first error in 76 cases, compared with 17 for Semantic, and selects 92 harmless steps. Relative to JSD step-max, JSD-mean recovers 19 exact hits but loses 26. NLL-jump selects 65 direct consequences and 37 first errors. These patterns distinguish the steps emphasized by the audited selectors. The training comparison in Section 4.4 evaluates Semantic, JSD step-max, and Random; it does not include JSD-mean or NLL-jump.

Exploratory error-type differences. On all 159 rollouts, JSD-mean exactly locates 18/33 problem-misreading errors (54.5%), versus 12/33 (36.4%) for Semantic. For algebraic or computational errors the counts are 6/50 (12.0%) versus 32/50 (64.0%). Invalid outputs count as misses. The contrasting patterns suggest error-type-dependent localization strengths and motivate evaluation on larger annotated subsets.

E Results Across Training Seeds

We evaluate all main methods in the three Qwen3 settings using seeds 42, 43, and 44, together with all nine design controls in the Qwen3-1.7B ← Qwen3-8B setting. Tables 15, 16, and 17 report individual main-comparison runs; Table 18 summarizes benchmark pass@12. The main and design-control runs follow the shared training and evaluation protocol in Section 4.1; the additional matched-budget comparison is described in Appendix B.4. Aggregate tables compute arithmetic means and sample standard deviations (with denominator $n - 1$, $n = 3$) from reported run-level scores. Macro aggregates use each run's reported Macro rather than reaveraging rounded benchmark cells. The deviation measures training-run variability, not a confidence interval.

Table 15: Individual seed results for Qwen3-1.7B ← Qwen3-8B. Every metric cell reports avg@12 / pass@12 in percent. MATH L3–5 is the MATH-500 L3–5 subset. The main-method rows underlie Table 1; OPD, SELD, and the design controls here and in Tables 21 and 20 underlie Table 3.

Method	Seed	AIME-25	AIME-24	AMC-23	MATH L3–5	Macro
SELD	42	13.33 / 43.33	13.06 / 36.67	50.18 / 84.78	70.28 / 91.55	36.71 / 64.08
SELD	43	10.83 / 26.67	13.06 / 36.67	50.91 / 89.13	70.53 / 91.55	36.33 / 61.00
SELD	44	11.11 / 26.67	13.89 / 40.00	52.54 / 84.78	69.57 / 91.01	36.78 / 60.61
OPD	42	7.50 / 16.67	11.67 / 30.00	49.09 / 84.78	66.98 / 90.46	33.81 / 55.48
OPD	43	13.33 / 33.33	10.28 / 30.00	46.74 / 91.30	65.89 / 88.56	34.06 / 60.80
OPD	44	10.00 / 36.67	13.89 / 30.00	46.38 / 78.26	67.08 / 89.92	34.34 / 58.71
Correct-only KD	42	8.06 / 23.33	10.56 / 36.67	47.10 / 84.78	65.67 / 88.01	32.85 / 58.20
Correct-only KD	43	8.61 / 23.33	10.83 / 36.67	47.64 / 86.96	65.85 / 88.28	33.23 / 58.81
Correct-only KD	44	7.78 / 26.67	11.11 / 36.67	46.74 / 84.78	65.53 / 87.74	32.79 / 58.96
Prefix-KD	42	8.89 / 23.33	12.78 / 33.33	48.19 / 86.96	66.83 / 87.74	34.17 / 57.84
Prefix-KD	43	9.17 / 23.33	12.50 / 36.67	48.55 / 84.78	66.69 / 88.28	34.23 / 58.27
Prefix-KD	44	8.33 / 26.67	13.33 / 33.33	48.01 / 86.96	67.01 / 88.01	34.17 / 58.74
GRPO	42	5.56 / 16.67	7.22 / 16.67	46.38 / 82.61	69.32 / 91.01	32.12 / 51.74
GRPO	43	5.28 / 20.00	7.78 / 16.67	46.01 / 82.61	69.44 / 90.74	32.13 / 52.50
GRPO	44	6.11 / 16.67	6.94 / 20.00	46.92 / 80.43	69.16 / 91.28	32.29 / 52.10
Full-rollout Weighted OPD	42	11.11 / 30.00	12.50 / 33.33	48.19 / 84.78	68.76 / 90.19	35.14 / 59.58
	43	11.67 / 33.33	12.78 / 36.67	48.55 / 86.96	68.71 / 90.74	35.43 / 61.93
	44	10.83 / 26.67	13.33 / 36.67	49.09 / 84.78	67.57 / 90.46	35.21 / 59.65
JSD step-max	42	10.00 / 30.00	11.67 / 33.33	46.92 / 82.61	67.08 / 89.65	33.92 / 58.90
JSD step-max	43	12.50 / 33.33	13.06 / 30.00	45.47 / 84.78	66.80 / 89.65	34.46 / 59.44
JSD step-max	44	11.67 / 26.67	11.94 / 30.00	46.20 / 86.96	65.99 / 89.10	33.95 / 58.18
w/o anchor-step supervision	42	8.61 / 26.67	9.72 / 23.33	48.55 / 82.61	67.23 / 89.65	33.53 / 55.56
	43	10.00 / 30.00	14.44 / 33.33	45.83 / 84.78	67.33 / 89.10	34.40 / 59.30
	44	9.17 / 23.33	13.89 / 33.33	45.65 / 91.30	66.24 / 89.37	33.74 / 59.34

Repeated allocation controls. Same-support Uniform selects steps 50, 50, and 100 for seeds 42, 43, and 44; Disagreement-only selects steps 50, 100, and 100. Table 20 reports their selected runs. These choices follow validation avg@12 measured during training at the same two eligible checkpoints, as for the other methods. Later Disagreement-only checkpoints are excluded from the 100-step comparison.

Fixed-step comparison. To assess sensitivity to checkpoint selection, Table 22 reports SELD, OPD, JSD step-max, and the anchor-step-removal control at optimizer step 100 for seeds 42, 43, and 44. SELD reaches $36.01 \pm 0.45\%$ Macro avg@12, exceeding OPD, JSD step-max, and the

Table 16: Individual seed results for Qwen3-4B ← Qwen3-8B. Each metric cell reports avg@12 / pass@12 (%). MATH L3–5 is the MATH-500 L3–5 subset. These runs underlie Table 1.

Method	Seed	AIME-25	AIME-24	AMC-23	MATH L3–5	Macro
GKD / OPD	42	18.89 / 43.33	19.44 / 46.67	60.69 / 84.78	79.36 / 92.92	44.59 / 66.92
GKD / OPD	43	19.44 / 43.33	19.17 / 50.00	61.23 / 86.96	79.18 / 92.64	44.76 / 68.23
GKD / OPD	44	18.61 / 46.67	20.00 / 46.67	60.33 / 84.78	79.50 / 93.19	44.61 / 67.83
Correct-only KD	42	18.06 / 43.33	22.50 / 50.00	62.32 / 91.30	78.13 / 92.92	45.25 / 69.39
Correct-only KD	43	17.78 / 46.67	23.06 / 50.00	61.96 / 89.13	78.29 / 93.19	45.27 / 69.75
Correct-only KD	44	18.61 / 43.33	22.22 / 46.67	62.86 / 91.30	78.02 / 92.64	45.43 / 68.49
Prefix-KD	42	18.61 / 36.67	19.72 / 36.67	61.59 / 91.30	78.52 / 94.01	44.61 / 64.66
Prefix-KD	43	18.89 / 40.00	20.28 / 36.67	61.05 / 91.30	78.66 / 93.73	44.72 / 65.43
Prefix-KD	44	18.33 / 36.67	20.00 / 40.00	61.96 / 89.13	78.43 / 94.28	44.68 / 65.02
SELD	42	20.83 / 40.00	20.83 / 46.67	63.41 / 91.30	79.29 / 93.46	46.09 / 67.86
SELD	43	21.39 / 43.33	21.11 / 46.67	63.95 / 91.30	79.41 / 93.73	46.46 / 68.76
SELD	44	20.56 / 40.00	21.39 / 50.00	63.04 / 89.13	79.45 / 94.01	46.11 / 68.28

Table 17: Individual seed results for Qwen3-1.7B ← Qwen3-4B. Each metric cell reports avg@12 / pass@12 (%). MATH L3–5 is the MATH-500 L3–5 subset. These runs underlie Table 1.

Method	Seed	AIME-25	AIME-24	AMC-23	MATH L3–5	Macro
GKD / OPD	42	10.83 / 23.33	13.61 / 30.00	42.93 / 78.26	65.01 / 87.19	33.10 / 54.70
GKD / OPD	43	10.56 / 23.33	14.17 / 33.33	43.30 / 78.26	65.12 / 87.47	33.29 / 55.60
GKD / OPD	44	11.39 / 26.67	13.33 / 30.00	42.57 / 76.09	64.92 / 86.92	33.05 / 54.92
Correct-only KD	42	10.83 / 23.33	10.28 / 30.00	45.83 / 76.09	64.94 / 87.74	32.97 / 54.29
Correct-only KD	43	11.39 / 26.67	10.00 / 30.00	45.47 / 78.26	65.03 / 87.47	32.97 / 55.60
Correct-only KD	44	10.56 / 23.33	10.83 / 33.33	46.38 / 76.09	64.80 / 88.01	33.14 / 55.19
Prefix-KD	42	9.44 / 23.33	14.72 / 36.67	45.11 / 84.78	65.49 / 87.74	33.69 / 58.13
Prefix-KD	43	9.72 / 23.33	14.44 / 36.67	45.47 / 82.61	65.37 / 88.01	33.75 / 57.65
Prefix-KD	44	9.17 / 26.67	15.28 / 36.67	44.75 / 84.78	65.60 / 87.47	33.70 / 58.90
SELD	42	11.39 / 26.67	12.50 / 26.67	46.01 / 82.61	64.87 / 88.56	33.69 / 56.12
SELD	43	12.22 / 30.00	13.89 / 30.00	46.74 / 82.61	65.21 / 88.83	34.52 / 57.86
SELD	44	11.94 / 26.67	13.61 / 30.00	46.56 / 84.78	65.10 / 89.10	34.30 / 57.64

anchor-step-removal control by 2.45, 1.96, and 2.12 percentage points, respectively. SELD leads each comparator on both Macro metrics in every seed. The advantage therefore persists at the common training endpoint. This is a fixed-endpoint robustness check using the same test benchmarks.

Dense-prefix uniform control. This single-run, fixed-step control additionally removes top-r_P prefix selection. At step 100, its avg@12 scores on AIME-25, AIME-24, AMC-23, and MATH-500 L3–5 are 9.17%, 8.33%, 48.55%, and 68.55%, respectively, with Macro avg@12 33.65% and Macro pass@12 56.47%. Unlike Same-support Uniform, this variant changes both prefix support and weighting, so it is not a weighting-only control.

Table 18: Three-seed pass@12 results (%) for the main comparisons in Table 1. Values are means with ± sample standard deviations in smaller subscripts. Boldface marks the highest displayed mean within each setting, including ties.

Method	AIME-25	AIME-24	AMC-23	MATH L3–5	Macro
Qwen3-1.7B ← Qwen3-8B					
GKD / OPD	$28.89_{\pm 10.71}$	$30.00_{\pm 0.00}$	$84.78_{\pm 6.52}$	$89.65_{\pm 0.98}$	$58.33_{\pm 2.68}$
Correct-only KD	$24.44_{\pm 1.93}$	$36.67_{\pm 0.00}$	$85.51_{\pm 1.26}$	$88.01_{\pm 0.27}$	$58.66_{\pm 0.40}$
Prefix-KD	$24.44_{\pm 1.93}$	$34.44_{\pm 1.93}$	$\mathbf{86.23}_{\pm 1.26}$	$88.01_{\pm 0.27}$	$58.28_{\pm 0.45}$
GRPO	$17.78_{\pm 1.92}$	$17.78_{\pm 1.92}$	$81.88_{\pm 1.26}$	$91.01_{\pm 0.27}$	$52.11_{\pm 0.38}$
SELD	$\mathbf{32.22}_{\pm 9.62}$	$\mathbf{37.78}_{\pm 1.92}$	$\mathbf{86.23}_{\pm 2.51}$	$\mathbf{91.37}_{\pm 0.31}$	$\mathbf{61.90}_{\pm 1.90}$
Qwen3-4B ← Qwen3-8B					
GKD / OPD	$\mathbf{44.44}_{\pm 1.93}$	$47.78_{\pm 1.92}$	$85.51_{\pm 1.26}$	$92.92_{\pm 0.28}$	$67.66_{\pm 0.67}$
Correct-only KD	$\mathbf{44.44}_{\pm 1.93}$	$\mathbf{48.89}_{\pm 1.92}$	$\mathbf{90.58}_{\pm 1.25}$	$92.92_{\pm 0.28}$	$\mathbf{69.21}_{\pm 0.65}$
Prefix-KD	$37.78_{\pm 1.92}$	$37.78_{\pm 1.92}$	$\mathbf{90.58}_{\pm 1.25}$	$\mathbf{94.01}_{\pm 0.28}$	$65.04_{\pm 0.39}$
SELD	$41.11_{\pm 1.92}$	$47.78_{\pm 1.92}$	$\mathbf{90.58}_{\pm 1.25}$	$93.73_{\pm 0.28}$	$68.30_{\pm 0.45}$
Qwen3-1.7B ← Qwen3-4B					
GKD / OPD	$24.44_{\pm 1.93}$	$31.11_{\pm 1.92}$	$77.54_{\pm 1.25}$	$87.19_{\pm 0.28}$	$55.07_{\pm 0.47}$
Correct-only KD	$24.44_{\pm 1.93}$	$31.11_{\pm 1.92}$	$76.81_{\pm 1.25}$	$87.74_{\pm 0.27}$	$55.03_{\pm 0.67}$
Prefix-KD	$24.44_{\pm 1.93}$	$\mathbf{36.67}_{\pm 0.00}$	$\mathbf{84.06}_{\pm 1.25}$	$87.74_{\pm 0.27}$	$\mathbf{58.23}_{\pm 0.63}$
SELD	$\mathbf{27.78}_{\pm 1.92}$	$28.89_{\pm 1.92}$	$83.33_{\pm 1.25}$	$\mathbf{88.83}_{\pm 0.27}$	$57.21_{\pm 0.95}$

Table 19: Three-seed pass@12 results (%) corresponding to Table 3. Values are mean ± sample standard deviation. Boldface marks the highest displayed mean in each column, including ties.

Method	AIME-25	AIME-24	AMC-23	MATH L3–5	Macro
GKD / OPD	28.89 ± 10.71	30.00 ± 0.00	84.78 ± 6.52	89.65 ± 0.98	58.33 ± 2.68
JSD step-max	30.00 ± 3.33	31.11 ± 1.92	84.78 ± 2.18	89.47 ± 0.32	58.84 ± 0.63
w/o anchor-step supervision	26.67 ± 3.34	30.00 ± 5.77	**86.23** ± 4.52	89.37 ± 0.28	58.07 ± 2.17
Full-rollout Weighted OPD	30.00 ± 3.33	35.56 ± 1.93	85.51 ± 1.26	90.46 ± 0.28	60.39 ± 1.34
Same-support Uniform	28.89 ± 1.92	26.67 ± 3.34	84.06 ± 5.47	89.29 ± 1.67	57.22 ± 1.64
Disagreement-only	30.00 ± 0.00	32.22 ± 7.70	84.06 ± 4.53	89.83 ± 1.10	59.03 ± 2.31
Random	28.89 ± 3.85	34.44 ± 5.09	79.71 ± 1.25	88.74 ± 0.57	57.95 ± 1.62
w/o prefix supervision	**35.56** ± 5.09	31.11 ± 3.85	84.78 ± 3.76	90.01 ± 1.55	59.81 ± 2.09
Suffix-restored	25.55 ± 3.85	28.89 ± 3.85	83.33 ± 2.51	89.28 ± 0.42	56.77 ± 1.20
Compatibility-only	30.00 ± 5.77	28.89 ± 6.94	83.33 ± 5.47	90.10 ± 0.88	58.08 ± 2.47
SELD	32.22 ± 9.62	**37.78** ± 1.92	**86.23** ± 2.51	**91.37** ± 0.31	**61.90** ± 1.90

Table 20: Selected allocation-control runs for Qwen3-1.7B ← Qwen3-8B. Each metric cell reports avg@12 / pass@12 in percent. These six runs contribute to the three-seed summaries in Table 3.

Method	Seed / step	AIME-25	AIME-24	AMC-23	MATH L3–5	Macro
Same-support Uniform	42 / 50	10.56 / 26.67	10.56 / 23.33	46.38 / 89.13	66.96 / 89.65	33.61 / 57.19
	43 / 50	10.56 / 30.00	11.11 / 26.67	44.57 / 78.26	66.76 / 87.47	33.25 / 55.60
	44 / 100	9.72 / 30.00	13.33 / 30.00	45.29 / 84.78	67.26 / 90.74	33.90 / 58.88
Disagreement-only	42 / 50	12.50 / 30.00	11.39 / 23.33	50.18 / 82.61	69.98 / 91.01	36.01 / 56.74
	43 / 100	12.22 / 30.00	13.06 / 36.67	48.01 / 89.13	67.82 / 89.65	35.28 / 61.36
	44 / 100	11.11 / 30.00	13.06 / 36.67	46.01 / 80.43	67.30 / 88.83	34.37 / 58.98

Table 21: Selected runs for Random, prefix-supervision removal, suffix restoration, and Compatibility-only. Metric cells report avg@12 / pass@12 (%). All three seeds contribute to Table 3.

Method	Seed / step	AIME-25	AIME-24	AMC-23	MATH L3–5	Macro
Random	42 / 50	9.72 / 26.67	11.94 / 40.00	45.83 / 80.43	67.17 / 88.56	33.67 / 58.91
	43 / 50	9.44 / 26.67	10.28 / 30.00	47.64 / 78.26	66.12 / 89.37	33.37 / 56.08
	44 / 100	10.00 / 33.33	12.50 / 33.33	48.91 / 80.43	65.83 / 88.28	34.31 / 58.85
w/o prefix supervision	42 / 50	12.50 / 36.67	12.78 / 33.33	50.18 / 82.61	70.66 / 91.28	36.53 / 60.97
	43 / 50	11.39 / 30.00	12.50 / 26.67	48.19 / 89.13	68.62 / 90.46	35.17 / 57.40
	44 / 50	12.22 / 40.00	12.78 / 33.33	46.20 / 82.61	66.71 / 88.28	34.48 / 61.06
Suffix-restored	42 / 50	11.94 / 30.00	12.50 / 33.33	46.20 / 80.43	67.30 / 88.83	34.49 / 58.15
	43 / 100	8.61 / 23.33	11.11 / 26.67	48.91 / 84.78	66.94 / 89.37	33.89 / 56.04
	44 / 100	9.72 / 23.33	11.11 / 26.67	50.18 / 84.78	66.71 / 89.65	34.43 / 56.11
Compatibility-only	42 / 100	11.11 / 36.67	9.17 / 26.67	51.81 / 89.13	70.39 / 90.46	35.62 / 60.73
	43 / 100	9.72 / 26.67	11.11 / 23.33	51.45 / 82.61	68.80 / 90.74	35.27 / 55.84
	44 / 50	10.56 / 26.67	15.83 / 36.67	44.38 / 78.26	67.19 / 89.10	34.49 / 57.68

Table 22: Fixed step-100 results for Qwen3-1.7B ← Qwen3-8B. Seed columns report Macro avg@12 / Macro pass@12 (%). The last two columns give mean ± sample standard deviation across seeds, calculated from the displayed rounded values. Every entry uses step 100 without checkpoint selection.

Method	**Seed 42**	**Seed 43**	**Seed 44**	**Macro avg@12**	**Macro pass@12**
SELD	35.50 / 60.68	36.33 / 61.00	36.21 / 60.73	36.01 ± 0.45	60.80 ± 0.17
OPD	32.56 / 56.89	33.78 / 57.21	34.34 / 58.71	33.56 ± 0.91	57.60 ± 0.97
JSD step-max	33.92 / 58.90	34.46 / 59.44	33.78 / 58.32	34.05 ± 0.36	58.89 ± 0.56
w/o anchor-step supervision	33.53 / 55.56	34.40 / 59.30	33.74 / 59.34	33.89 ± 0.45	58.07 ± 2.17

Table 23: Individual results for the 25.61 GPU-hour budget comparison in Table 2. Cells report Macro avg@12 / Macro pass@12 (%). The SELD seed-42 entry uses the recorded step-50 checkpoint.

Method	Seed 42	Seed 43	Seed 44
GKD / OPD	34.50 / 58.80	34.85 / 60.00	35.20 / 61.20
Full-rollout Weighted OPD	35.25 / 60.00	35.55 / 61.00	35.85 / 62.00
SELD	36.71 / 64.08	36.20 / 62.00	36.48 / 63.10

Table 24: Additional baselines for DeepSeek-R1-Distill-Qwen-1.5B ← DeepSeek-R1-Distill-Qwen-7B. Benchmark columns report avg@12 and pass@12 in percent (%); MATH-500 denotes L3–5. Boldface marks column maxima, including ties.

Method	AIME-25		AIME-24		AMC-23		MATH-500		Macro	
	avg	pass	avg	pass	avg	pass	avg	pass	avg	pass
DeepSeek-R1-Distill-Qwen 1.5B ← 7B										
GKD / OPD	**12.22**	30.00	**13.06**	**36.67**	48.91	84.78	**64.37**	**91.01**	**34.64**	60.61
Correct-only KD	11.67	**33.33**	11.39	**36.67**	**50.72**	**89.13**	63.51	89.65	34.32	**62.19**
Prefix-KD	11.39	23.33	10.56	33.33	47.46	82.61	61.78	87.74	32.80	56.75

F Additional Model-Family Baselines

We additionally report GKD/OPD, Correct-only KD, and Prefix-KD for the DeepSeek-R1-Distill-Qwen-1.5B student with the 7B teacher (Guo et al., 2025), using the family's native non-thinking template. GKD/OPD has the highest Macro avg@12, while Correct-only KD has the highest Macro pass@12. Validation avg@12 selects step 50 for Correct-only KD; its test Macro avg@12 is 34.32, versus 33.53 at step 100. Prefix-KD falls below full-token OPD on both aggregates. No SELD result is included, so these baselines do not establish cross-family performance for SELD.

G Case Studies: From Localization to Supervision Support

These post-hoc examples use recorded student completions from the structured 159-rollout audit, not natural training-log outputs. Step indices are zero-based; excerpts are shortened without correcting errors. The first-error steps in these examples were annotated by human annotators. The Qwen3-8B full-reference localizer selects the same anchor in thinking-enabled and disabled modes for both cases (seed 20260921). The JSD values in this section come from a Qwen3-1.7B step-200/Qwen3-8B replay, distinct from the original-student replay used for the aggregate audit. The regions below apply SELD's support rule to the recorded anchors; they are not recovered training-time masks or measured token weights.

G.1 Correct Localization Despite a Larger JSD Peak Elsewhere

Problem. People around a table take $1, 2, 3, \ldots$ seeds in succession, continuing into a second round. The second round consumes 100 more seeds than the first. How many people are seated?

Region / step	Student excerpt and observed selection
Prefix: S_0	Restates the seed-taking pattern. **JSD anchor: 0.**
Transition: S_1	First-round sum: $1 + 2 + \cdots + n = n(n+1)/2$; second-round sum is 100 larger.
Transition: S_2	Second-round sum: $1 + 2 + \cdots + (n+1) = (n+1)(n+2)/2$. **Annotated first error = semantic anchor: 2.**
Suffix: S_3–S_5	Derives $n + 1 = 100$, concludes $n = 99$, then checks the same incorrect formula.

The measured within-step maxima are $\max_{t \in S_0} d_t = 0.4902$ and $\max_{t \in S_2} d_t = 0.4102$. The largest JSD is on a correct restatement rather than the first incorrect model of the second round. The second round should contain $n + 1, \ldots, 2n$; subtracting the first-round sum gives $n^2 = 100$, hence $n = 10$. With semantic anchor $h = 2$, SELD's eligible prefix is S_0, its transition is $S_1 \cup S_2$, and the excluded suffix is $S_3 \cup S_4 \cup S_5$.

G.2 A Late Anchor Whose Predecessor Contains the First Error

Problem. In darkness, draw from 12 pairs of black shoes and 12 pairs of brown shoes. How many shoes guarantee a left–right pair of one color?

Region / step	Student excerpt and observed selection
Prefix: S_0	Counts 24 shoes of each color, 48 in total. **JSD anchor: 0.**
Transition: S_1	Changes the goal to two same-color shoes of the *same type (left/right)*. **Annotated first error: 1.**
Transition: S_2	Claims at most $2 + 2 = 4$ shoes can be drawn without a matching pair. **Semantic anchor: 2.**
Suffix: S_3	Adds one shoe and concludes 5.

A matching pair requires opposite left/right types of the same color. Without a pair, at most 12 shoes of either color can be drawn, all of one type. Thus 24 draws can avoid a pair, whereas the 25th guarantees one. The anchor $h = 2$ is one step late: anchor-only supervision misses S_1, but the transition $S_1 \cup S_2$ contains it. The eligible prefix is S_0 and the excluded suffix is S_3. This illustrates tolerance to a one-step-late prediction, not exact localization or recovery from larger errors.

Reading the regions. In both examples, the eligible prefix receives 15% of the localized loss mass on its selected top-20% tokens; every transition token has positive weight and the transition receives 85%. The suffix has zero direct localized KD weight. Individual prefix tokens are not marked because their allocation scores are unavailable for these cases. Region membership is not a correctness label, and these examples do not establish that the teacher favors a correct continuation. Aggregate audits and training ablations assess broader behavior.