STEPS: SELECTIVE ON-POLICY SELF-DISTILLATION FOR REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy self-distillation (OPSD) uses a model as its own teacher under privileged context, providing token-level supervision on the model's own reasoning trajectories. However, persistent all-token guidance can degrade training in our math RL setting. Such guidance may unnecessarily constrain non-critical tokens and reinforce biases induced by privileged information unavailable at inference. Motivated by these risks, we propose STEPS, a selective OPSD framework that controls the location, direction, and duration of distillation. STEPS identifies critical spans in student rollouts, applies forward KL to key spans or reverse KL to error spans, and gradually transitions to pure GRPO after a short distillation phase. Across four models from three families, STEPS improves mean accuracy over GRPO. On Qwen3-8B, it raises average accuracy across four math benchmarks and GPQA-Diamond by an absolute 2.76% with a strong annotator. Using the training policy as its own annotator yields an absolute mean gain of 1.90%, with an estimated 2.7% training-time overhead amortized over 300 steps. Controlled ablations support semantic span selection, class-specific KL directions, and the hand-off to GRPO. Together, these results support brief, targeted self-distillation as an efficient complement to reward-driven reasoning training. Our code is available at https://anonymous.4open.science/r/trace-749F/.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR), exemplified by GRPO (Shao et al., 2024), improves reasoning through outcome-level feedback (DeepSeek-AI et al., 2025; Team Olmo et al., 2025). However, a trajectory-level reward does not directly identify which intermediate decisions require correction (Lightman et al., 2024; Kazemnejad et al., 2025). On-policy self-distillation (OPSD) offers a complementary source of supervision: the model itself, or a policy snapshot, acts as a teacher conditioned on privileged context such as verified solutions or environment feedback (Hübotter et al., 2026; Zhao et al., 2026). Its predictions provide token-level guidance along the student's own rollouts, making fine-grained self-supervision available without a separate teacher model (Shenfeld et al., 2026).

However, a teacher's prediction under privileged context is useful only insofar as it improves the student's ability to solve problems without that context (Penaloza et al., 2026; Yang et al., 2026). Applying this guidance uniformly across tokens and maintaining it throughout training may impose constraints beyond those needed for useful correction (Kim et al., 2026; Li et al., 2026a). In our reproduced SDPO and SRPO runs on Qwen3-8B math RL, responses shorten, per-token actor entropy rises, and validation accuracy falls by an absolute 18–19% from early peaks, while GRPO remains stable under the same data and optimizer stack, as shown in Figure 1 and Appendix D.6. These observations raise a central question: how can OPSD retain useful corrective guidance while preserving the role of verifier-grounded learning?

The key idea is to use self-distillation as a brief correction phase focused on semantically important reasoning spans. Prior work on token importance shows that a subset of positions can retain much of the benefit of distillation (Xu et al., 2026), motivating selective guidance instead of uniform coverage. Limiting its duration gives the auxiliary privileged signal a defined role in training: support early correction before optimization returns fully to verifier rewards. This also motivates analyzing

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

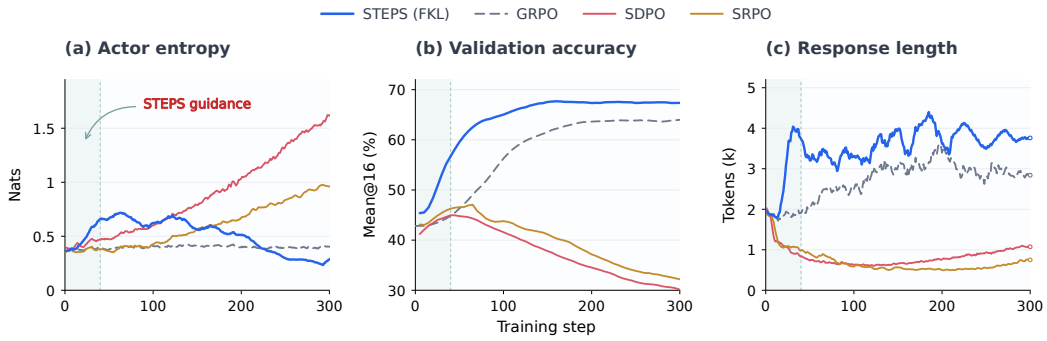

Figure 1: **Training diagnostics on Qwen3-8B.** (a) Per-token actor entropy, (b) OpenThoughts held-out validation accuracy (mean@16), and (c) response length. In these runs, SDPO and SRPO exhibit rising entropy and shorter responses alongside validation degradation, while STEPS (FKL on key spans) sustains its validation gains. Shading marks the initial STEPS guidance phase.

how much privileged-gradient variation the student encounters during the guided phase (Yang et al., 2026).

We introduce **STEPS (Span-Targeted Early-Phase Self-Distillation)**, a selective OPSD framework that controls the location, direction, and duration of guidance. An annotator identifies key spans in correct rollouts and error spans in incorrect rollouts. Its span masks determine where distillation is applied, while its coarse diagnostic labels provide privileged context for the self-teacher. We consider forward KL (FKL) on key spans and reverse KL (RKL) on error spans as alternative configurations. STEPS caps span coverage at 25% and decays the KL coefficient to zero by step 40, while restoring GRPO on selected spans. Training then proceeds with pure GRPO, without annotation or teacher forward passes. The annotator can also be the training policy itself.

Our analysis motivates these choices through local gradient behavior and conditional exposure to privileged guidance. For fixed teacher targets, FKL provides non-vanishing logit pressure on teacher-supported tokens as their student probability approaches zero, whereas RKL's pressure is scaled by student probability and depends on the centered student–teacher log-ratio, as established by Corollary 3 in Appendix B.1. These identities motivate testing KL directions by span class. For the FKL branch, Proposition 5 in Appendix B.2 characterizes how restricted coverage and finite-duration decay limit cumulative privileged-gradient variation. Proposition 7 in Appendix B.3 gives conditions under which selected key spans retain a positive learning signal.

Experiments on four models from three families support this selective use of self-distillation. On Qwen3-8B, STEPS with a strong annotator improves the average accuracy across four math benchmarks and GPQA-Diamond by an absolute 2.76% over GRPO. Online self-annotation yields an absolute mean gain of 1.90%, with an estimated 2.7% training-time overhead amortized over 300 steps. FKL on key spans performs better on Qwen3-8B, while RKL on error spans performs better on Qwen3-1.7B, demonstrating an empirical dependence on the tested model. Additional Llama and SmolLM3 results extend the evaluation across families. Controlled mask, class-by-direction, and hand-off ablations probe these design choices through matched comparisons, with the hand-off experiment evaluating the coupled decay of KL and restoration of GRPO.

Our contributions are threefold:

- **Method.** We introduce STEPS, a selective OPSD framework that combines semantic span selection, class-specific KL directions, and a brief distillation phase followed by pure GRPO.
- **Theoretical analysis.** We characterize local KL-gradient asymmetry and derive conditional bounds on privileged-gradient exposure and retained key-span signal.
- **Empirical evidence.** Across four models from three families, STEPS improves mean accuracy over GRPO. It also demonstrates higher training-step efficiency: SmolLM3 reaches 71.84 AVG at step 40, exceeding GRPO's 69.45 at step 325. Matched ablations support the design choices, while online self-annotation retains gains at low estimated amortized cost.

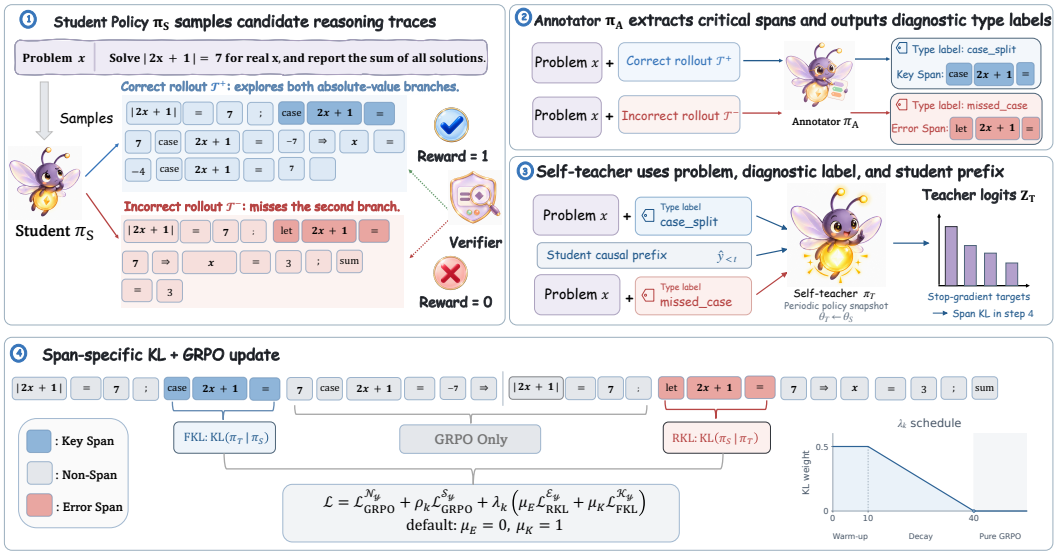

Figure 2: **STEPS pipeline.** The student samples a rollout and receives a verifier outcome. An annotator supplies span masks and a coarse diagnostic label; the self-teacher conditions on this label and the student's causal prefix. Span-specific KL is combined with GRPO, which is restored on selected spans as KL decays to zero.

2 Related Work

Knowledge distillation transfers teacher predictions to a student (Hinton et al., 2015); on-policy variants supervise student-generated sequences using teacher logits (Gu et al., 2024; Agarwal et al., 2023). OPSD methods (Hübotter et al., 2026; Zhao et al., 2026) remove the separate-teacher requirement, but long-horizon collapse has been attributed to epistemic suppression (Kim et al., 2026) and privileged-information leakage (Yang et al., 2026), with sample-level routing (Li et al., 2026a) and advantage reweighting (Yang et al., 2026) as concurrent mitigations. Analyses of token-level distillation (Li et al., 2026b; Xu et al., 2026), critical reasoning tokens (Lin et al., 2024), and step-level credit assignment (Kazemnejad et al., 2025) motivate studying where guidance is applied. STEPS shares the long-horizon-collapse diagnosis but operates at token level via explicit span localization, treats divergence direction as a per-token-class action in Section 3, and identifies a teacher-coverage-gap regime where advantage reweighting can damp correct non-canonical gradients, as formalized in Proposition 9. Standard RLVR baselines (Shao et al., 2024; Yu et al., 2025; Team Olmo et al., 2025) are the post-training framework we build on; Appendix D.1 gives the extended citation map and compares the paradigms in Table 8.

3 STEPS: Selective On-Policy Self-Distillation

STEPS uses self-distillation to guide selected reasoning decisions during an initial phase of RLVR, then continues with pure GRPO. The method separates three choices: which positions receive teacher guidance, which KL direction is applied, and how guidance is withdrawn as reward-driven learning takes over. Figure 2 summarizes the training pipeline.

3.1 Span Annotation and Self-Teacher Construction

Locating reasoning spans. The verifier indicates whether a rollout succeeds, but does not locate the reasoning steps to reinforce or correct. Given a problem x, the student policy $\pi_S = \pi_\theta$ samples a rollout $\hat{y}$, which receives a verifier outcome $R(x, \hat{y}) \in \{0, 1\}$. During the distillation phase, we split the response into numbered segments and ask an annotator π_A to identify key reasoning spans in correct rollouts or localized error spans in incorrect rollouts. Segment indices are projected to token positions, yielding key and error sets $\mathcal{K}_y$ and $\mathcal{E}_y$. Let $\mathcal{S}_y = \mathcal{K}_y \cup \mathcal{E}_y$ denote all annotated positions

and $\mathcal{N}_y$ the remaining response positions. We cap coverage at $|\mathcal{S}_y| \leq \alpha|\hat{y}|$, with $\alpha = 0.25$. The annotator may be an external model or the current student policy; its outputs are discrete annotations and receive no gradient.

Turning annotations into teacher guidance. Each annotation provides both span locations and a coarse diagnostic label, such as `case_split` or `missed_case`. These outputs have separate roles: the locations determine where distillation is applied, while the label forms a private diagnostic prefix c for the self-teacher. The teacher parameters θ_T are copied from the student and periodically synchronized during the distillation phase. At position t, the student and teacher distributions are

$$\pi_{S,t} = \pi_\theta\big(\cdot \mid x, \hat{y}_{<t}\big), \qquad \pi_{T,t} = \pi_{\theta_T}\big(\cdot \mid x, c, \hat{y}_{<t}\big).$$

Both distributions are evaluated on the student's causal prefix. Only the coarse diagnostic description is added as privileged context; span locations are used in the loss mask. Teacher logits and span masks are treated as stop-gradient targets. Appendices C.3–C.4 provide the prompts, annotation schema, and segment-to-token alignment.

3.2 Span-Specific Distillation

Once the spans and teacher distributions are available, STEPS specifies the KL objective on each span class. We consider forward KL (FKL) on key spans and reverse KL (RKL) on error spans. FKL can lift teacher-supported alternatives even when their student probabilities are small, whereas RKL weights its centered log-ratio correction by the student's current probability. These distinct local effects motivate evaluating the directions by span class; Section 4 formalizes the distinction. Binary switches $\mu_E, \mu_K \in \{0, 1\}$ enable the corresponding branches and are fixed for each training run. The default Qwen3-8B configuration is $(\mu_E, \mu_K) = (0, 1)$; $(1, 0)$ enables the RKL-on-error alternative. The empirical comparison is reported in Section 5.

For a position set $\mathcal{A}$, let $\mathcal{L}^{\mathcal{A}}_{\mathrm{GRPO}}$ denote the GRPO token loss summed over $\mathcal{A}$ and normalized by the full rollout length $|\hat{y}|$. The per-rollout objective at training step k is

$$\begin{aligned}\mathcal{L}^{(k)}(\hat{y};\theta) = \mathcal{L}^{\mathcal{N}_y}_{\mathrm{GRPO}} + \rho_k\, \mathcal{L}^{\mathcal{S}_y}_{\mathrm{GRPO}} \\ + \frac{\lambda_k}{|\hat{y}|}\left[\mu_E \sum_{t\in\mathcal{E}_y} D_{\mathrm{KL}}\big(\pi_{S,t} \,\|\, \mathrm{sg}(\pi_{T,t})\big) + \mu_K \sum_{t\in\mathcal{K}_y} D_{\mathrm{KL}}\big(\mathrm{sg}(\pi_{T,t}) \,\|\, \pi_{S,t}\big)\right].\end{aligned} \tag{1}$$

Here sg denotes stop-gradient, λ_k controls distillation strength, and ρ_k controls the restoration of GRPO on annotated positions. Non-span positions retain the GRPO objective throughout training. The GRPO weight on an annotated span is governed by ρ_k, independently of whether its KL branch is enabled. All loss components use the same sequence normalization, and empty sets contribute zero. The implementation clips per-vocabulary KL contributions at $\tau = 0.05$ before summation (Zhao et al., 2026), as detailed in Appendix C.1; Equation (1) displays the underlying divergences analyzed in Section 4.

3.3 Distillation Decay and GRPO Handoff

The span mask localizes teacher guidance, while the schedule limits how long that guidance remains active. We hold the KL coefficient at w_0 for an initial interval, then decrease it linearly:

$$\lambda_k = w_0 \begin{cases} 1, & k < t_{\mathrm{start}}, \\ 1 - \dfrac{k - t_{\mathrm{start}}}{T_{\mathrm{decay}}}, & t_{\mathrm{start}} \leq k \leq t_{\mathrm{start}} + T_{\mathrm{decay}}, \\ 0, & k > t_{\mathrm{start}} + T_{\mathrm{decay}}. \end{cases} \tag{2}$$

We couple this decay to $\rho_k = 1 - \lambda_k / w_0$, so that GRPO is gradually restored on the same annotated positions as teacher guidance is withdrawn. Our configuration uses $w_0 = 0.5$, $t_{\mathrm{start}} = 10$, and

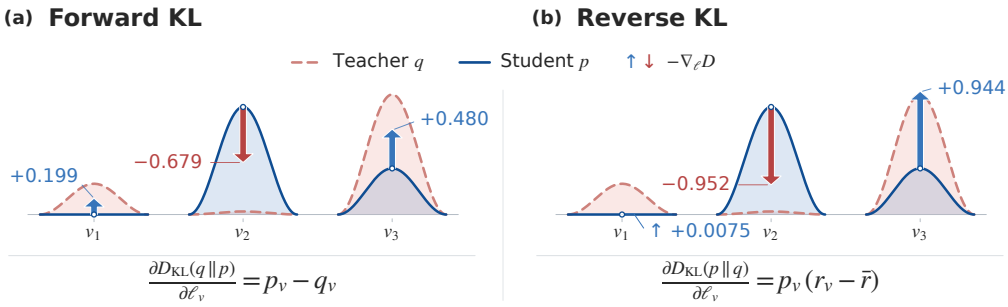

Figure 3: **KL gradient asymmetry.** Smooth peaks illustrate discrete token probabilities $p = (0.001, 0.699, 0.300)$ and $q = (0.200, 0.020, 0.780)$. The six arrows show negative logit gradients of unclipped KL with fixed teacher q, on a shared scale independent of probability heights; they do not represent probability displacements. Here $r_v = \log(p_v/q_v)$ and $\bar{r} = \sum_v p_v r_v$.

$T_{\text{decay}} = 30$. At step 40, $\lambda_k = 0$ and $\rho_k = 1$, and Equation (1) reduces to GRPO over all response positions. Annotation and teacher forward passes are then skipped for the remainder of training.

During the active phase, the teacher is synchronized from the student every $N = 10$ steps. Algorithm 1 gives the update procedure; training and evaluation protocols are detailed in Appendix C. Together, the mask and schedule confine privileged guidance to selected positions within a finite training window. Section 4 analyzes the local KL effects and the trade-off between limiting privileged guidance and retaining useful training signals.

4 Theoretical Analysis

We analyze the unclipped losses with stop-gradient teacher targets and masks. The KL identities explain the local role of divergence direction; the exposure and signal bounds characterize selective guidance for the FKL-on-key branch. Full statements and proofs are deferred to Appendix B.

4.1 KL Gradient Asymmetry

At a fixed rollout position, let $p_v = \pi_S(v) > 0$, $q_v = \pi_T(v) > 0$, $r_v = \log(p_v/q_v)$, and $\bar{r} = \sum_v p_v r_v$. For student logits ℓ_v, the two KL directions yield

$$\frac{\partial D_{\mathrm{KL}}(q\|p)}{\partial \ell_v} = p_v - q_v, \qquad \frac{\partial D_{\mathrm{KL}}(p\|q)}{\partial \ell_v} = p_v(r_v - \bar{r}). \tag{3}$$

For fixed $q_v > 0$ and bounded $|\bar{r}|$, as $p_v \to 0$, FKL retains upward logit pressure approaching q_v, while RKL pressure vanishes. RKL instead weights its centered log-ratio correction by student probability, with downward pressure exactly when $r_v > \bar{r}$. Figure 3 illustrates these local differences, motivating comparisons of KL direction by span class; the preferred configuration is determined empirically.

Consistent with the local mechanism above, STEPS-FKL yields the largest post-update key-token log-probability lift on the shared held-out token set in Table 1. Appendix D.2 provides the metric definition and detailed measurement protocol. Figure 6 complements this aggregate measure with token-level comparisons against GRPO.

Table 1: Post-update key-token log-probability lift.

Method	STEPS-FKL	STEPS-RKL	RLSD	GRPO	SRPO	SDPO
Δ_{lift} (nats)	**0.145**	0.078	0.068	0.054	0.027	0.019

4.2 Privileged-Gradient Exposure

To analyze the extent and duration of guidance, let $\mathcal{F}_k$ contain the training history, problem, rollout, and selected mask. Let $\delta_{k,t}$ be the per-token FKL gradient centered over privileged context conditional on $\mathcal{F}_k$ (Yang et al., 2026). Under the uniform score-operator bound C_s, coverage cap α, and a bound $\bar{V}$ on the expected average conditional teacher variance over selected positions, Appendix B.2 establishes

$$\mathcal{E}_K := \sum_{k=1}^{K} \lambda_k^2 \, \mathbb{E}\left[\frac{1}{|\hat{y}_k|} \sum_{t \in \mathcal{K}_{y_k}} \|\delta_{k,t}\|_2^2 \right] \le C_s^2 \alpha \bar{V} \sum_{k=1}^{K} \lambda_k^2. \tag{4}$$

The mask limits coverage through α, while decay limits cumulative weighting through $\sum_k \lambda_k^2$. The finite schedule in Section 3 makes this upper bound independent of the total training horizon. It controls centered conditional variation, not shifts in the conditional teacher mean or overall training stability.

4.3 Key-Span Signal Bound

Restricting guidance is useful only if the selected positions retain informative updates. Conditional on $\mathcal{F}_k$, suppose every nonempty selected key set has expected precision at least q_K. Relative to a common $\mathcal{F}_k$-measurable oracle loss-gradient direction $\tilde{g}_k$, assume expected per-token alignment, conditional on $\mathcal{F}_k$ and critical status, is at least $\gamma_K > 0$ for true-critical positions and $-B_K$, $B_K \ge 0$, for false positives. With $p_K = \mathbb{E}[|\mathcal{K}_y|/|\hat{y}|]$ and the sequence-normalized FKL gradient $g_k^{\mathrm{sel}} = \lambda_k |\hat{y}|^{-1} \sum_{t \in \mathcal{K}_y} g_{\mathrm{FKL}}(t)$, Appendix B.3 gives

$$\mathbb{E}\left[\langle g_k^{\mathrm{sel}}, \tilde{g}_k \rangle\right] \ge \lambda_k p_K \left[q_K \gamma_K - (1 - q_K) B_K \right]. \tag{5}$$

The bound makes selection quality explicit: selected spans retain positive expected alignment when true-critical contributions outweigh false-positive effects and $\lambda_k p_K > 0$. Together, the two bounds describe how limited guidance can retain useful signal under explicit selection and alignment assumptions.

5 Experiments

5.1 Experimental Setup

Model, data, evaluation. We evaluate Qwen3-8B and Qwen3-1.7B (Yang et al., 2025), Llama-3.2-3B-Instruct (Meta, 2024), and SmolLM3-3B (Bakouch et al., 2025). Training uses the OpenThoughts-114k math subset (Guha et al., 2025) (up to 30K problem–solution pairs), verl (Sheng et al., 2025), H100 GPUs, and DAPO clip-higher (Yu et al., 2025). Evaluation covers four math benchmarks—MATH-500 (Lightman et al., 2024; Hendrycks et al., 2021), AIME 2024/2025 (Mathematical Association of America, 2024; 2025), and AMC 2023 (Mathematical Association of America, 2023)—and GPQA-Diamond (Rein et al., 2023) for out-of-distribution generalization. Thinking-mode `avg@k` averages per-problem accuracy over k samples. Accuracy scores are expressed in %; gains and their confidence intervals are absolute differences on this scale. Cross-family runs reuse the configuration; hyperparameters, decoding settings, and model-specific details are in Apps. C and C.7.

Baselines and checkpoint selection. Baselines share the training stack: GRPO (Shao et al., 2024) with recent improvements (Team Olmo et al., 2025; Khatri et al., 2025), SDPO (Hübotter et al., 2026), SRPO (Li et al., 2026a), and RLSD (Yang et al., 2026). Qwen3 baselines use the best learning-rate/batch-size grid setting. We select the checkpoint with the highest OpenThoughts validation mean@16 among the evaluated checkpoints of each run. Held-out benchmark scores are never used for checkpoint selection. Coefficients and grids are in Apps. C.1 and D.5; degradation diagnostics appear in Fig. 1 and App. D.6.

For Qwen models, the main STEPS, SDPO, and SRPO runs enable thinking for both the student and teacher during training. All methods in Table 2 use thinking-enabled evaluation. Tables 13 and 14

Table 2: **Main results with thinking-enabled evaluation.** STEPS, SDPO, and SRPO use symmetric training. AVG is the unweighted five-benchmark mean. Pink AVG cells flag runs with later validation degradation in Fig. 4; all scores use validation-selected checkpoints. **Bold** / underline: best / second-best trained methods per model; base rows are references. Appendix D.4 reports cell-wise bootstrap 95% CIs.

Method	MATH-500	AIME 24	AIME 25	AMC 23	GPQA-D	AVG
Qwen3-8B						
Base (ref.)	96.80	76.25	67.50	95.94	58.27	78.95
+ GRPO	97.30	77.08	68.96	96.56	53.85	78.75
+ SDPO	95.13	52.50	41.25	91.41	39.90	64.04
+ SRPO	96.48	69.17	54.17	93.12	37.31	70.05
+ RLSD	96.68	75.42	68.54	97.03	57.26	78.99
+ STEPS (RKL on $\mathcal{E}_y$)	97.48	78.54	70.42	**97.81**	56.44	80.14
+ STEPS (FKL on $\mathcal{K}_y$)	**97.83**	**80.21**	**73.54**	97.66	**58.33**	**81.51**
Qwen3-1.7B						
Base (ref.)	91.35	43.75	35.00	86.56	37.18	58.77
+ GRPO	90.33	44.58	35.00	81.56	37.18	57.73
+ SDPO	81.73	16.25	17.50	51.25	27.41	38.83
+ SRPO	86.02	24.17	23.75	62.81	35.82	46.51
+ RLSD	91.12	**45.42**	32.92	84.38	**38.89**	58.55
+ STEPS (RKL on $\mathcal{E}_y$)	**92.05**	44.92	**38.33**	**86.82**	38.67	**60.16**
+ STEPS (FKL on $\mathcal{K}_y$)	91.00	44.83	36.30	83.38	36.77	58.46

examine asymmetric training and non-thinking evaluation on Qwen3-8B; these are Qwen-specific thinking-mode ablations.

5.2 Main Results

Results across model scales. Table 2 shows that on Qwen3-8B, STEPS-FKL with the strong annotator raises five-benchmark AVG from GRPO's 78.75% to 81.51%, an absolute gain of 2.76% with a paired 95% CI of $[0.41, 4.91]$%. GPQA-Diamond accuracy is 58.33, preserving the base model's 58.27 within evaluation resolution, compared with GRPO's 53.85. On Qwen3-1.7B, STEPS-RKL reaches 60.16% versus GRPO's 57.73%, an absolute gain of 2.43%, while STEPS-FKL reaches 58.46%. FKL on key spans leads at 8B, RKL on error spans at 1.7B. Cell-wise and paired intervals are in Apps. D.4 and E.2.

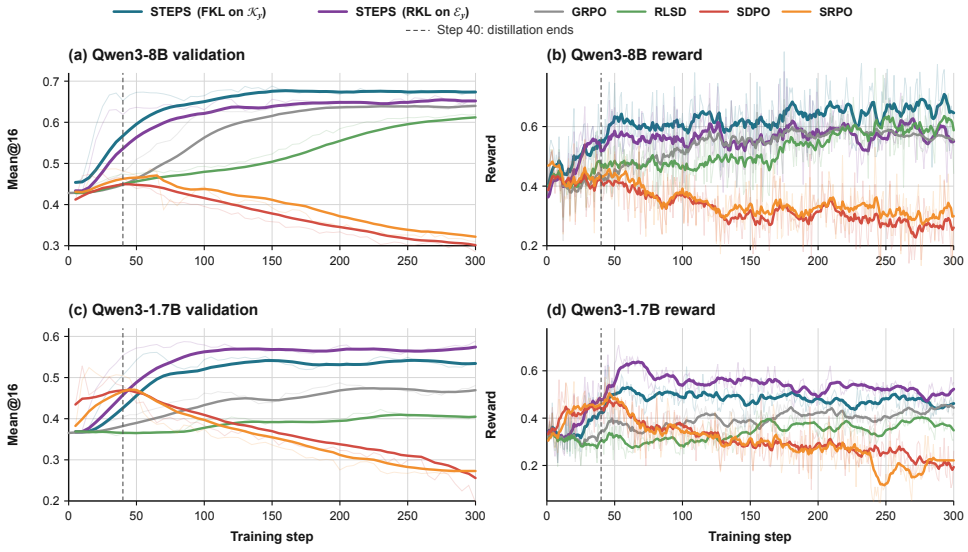

Figure 4: **Training dynamics.** OpenThoughts validation mean@16 (left) and reward (right): Qwen3-8B (top), Qwen3-1.7B (bottom). EMA α=0.85. Markers at step 40 indicate the handoff to pure GRPO.

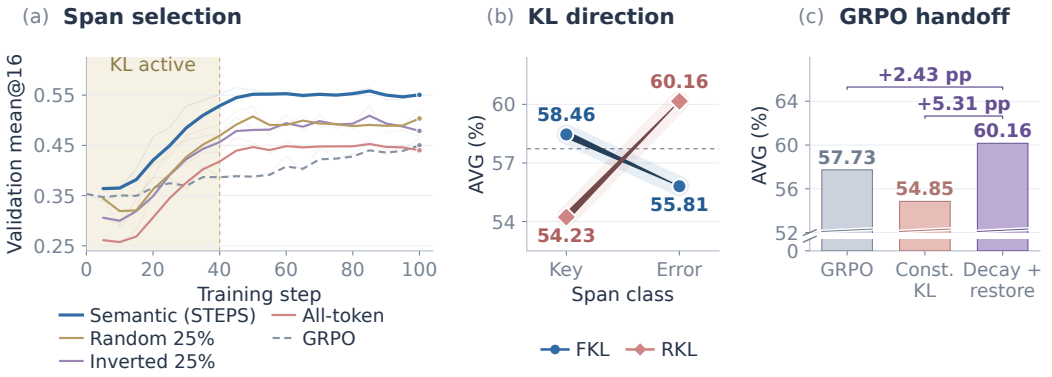

Figure 5: **Design ablations on Qwen3-1.7B.** (a) Span selection: validation mean@16 (EMA $\alpha = 0.55$), shared decay, and 25% selective coverage; shading marks the KL-active phase. (b) Span class $\times$ KL direction; the dashed line denotes GRPO. (c) GRPO, constant KL ($\lambda = 0.5$), and KL decay with GRPO restoration. Panels (b,c) report single-seed five-benchmark AVG. Appendix E.1 details matched settings.

Training dynamics. Figure 4 tracks validation accuracy and training reward. Gains persist after distillation ends at step 40 and training continues with pure GRPO; SDPO and SRPO degrade after early peaks in these reproduced runs.

Cross-family results. Table 3 reports 30.13% AVG for STEPS-RKL with online self-annotation on Llama-3.2-3B-Instruct, improving over GRPO and SDPO by 1.48 and 0.50 percentage points, respectively. On SmolLM3-3B, STEPS-RKL reaches 71.84% at step 40, exceeding GRPO's selected step-325 checkpoint (69.45%) by 2.39 percentage points. These results extend the benefits of selective self-distillation beyond Qwen and demonstrate substantially improved training-step efficiency on SmolLM3, where STEPS attains higher accuracy with far fewer training steps.

5.3 Ablations

Span Selection. Figure 5(a) shows that with RKL and decay fixed on Qwen3-1.7B, semantic error spans outperform random and inverted non-critical masks at matched 25% coverage, supporting localization beyond sparsity alone. Table 15 in Appendix E.1 also reports lower held-out AVG for the all-token control than GRPO: 56.68% versus 57.73%.

KL Direction. Unlike the two configurations in Tab. 2, Figure 5(b) separates span class and KL direction under the matched annotator, coverage, and schedule detailed in Appendix E.1. FKL performs better on key spans and RKL on error spans, supporting class-specific objectives rather than a uniform KL direction.

Table 3: **Cross-family results.** Ranking notation follows Tab. 2; SmolLM3 checkpoint steps are shown.

Model / Method	MATH-500	AIME 24	AMC 23	GPQA-D	AVG
Llama-3.2-3B-Instruct					
Base (ref.)	43.7	6.2	21.3	16.8	22.00
+ GRPO	52.1	**6.7**	29.5	26.3	28.65
+ SDPO	**54.4**	6.3	30.4	27.4	29.63
+ STEPS-RKL (online-self)	53.5	6.1	**30.8**	**30.1**	**30.13**
SmolLM3-3B					
Base (ref.)	90.16	52.50	78.69	44.85	66.55
+ GRPO (step 325)	94.35	54.31	79.42	49.73	69.45
+ SDPO	83.14	35.29	48.37	35.13	50.48
+ STEPS-RKL (step 40)	**96.23**	**56.12**	**82.31**	**52.69**	**71.84**

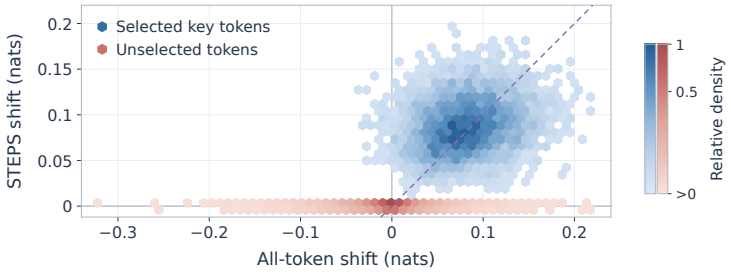

Figure 6: **Token-level update localization.** Post-update log-probability shifts relative to GRPO: all-token guidance (horizontal) versus STEPS (vertical). Blue and coral hexagons denote selected key and unselected tokens; density is normalized within each group. The dashed line denotes equal shifts. Appendix D.2 gives diagnostic details.

GRPO Handoff. Figure 5(c) shows that at fixed coverage α=0.25, holding λ=0.5 lowers Qwen3-1.7B AVG from 60.16 to 54.85, supporting the complete handoff design: λ controls both distillation and the restoration of GRPO on selected spans. The full coverage and duration sweep is in App. E.1.

Token-Level Updates. Localizing the KL loss does not automatically localize its effects, since model parameters are shared across token positions. Figure 6 therefore examines realized token-level log-probability changes relative to GRPO. On selected key tokens, STEPS retains positive shifts, with a slightly larger mean than all-token guidance (0.088 versus 0.082 nats). On unselected tokens, STEPS remains close to GRPO, whereas all-token guidance produces a wider spread of positive and negative shifts. This small-scale diagnostic complements the selection ablation: the observed updates preserve additional reinforcement at selected positions while limiting departures from GRPO elsewhere.

5.4 Annotation and Cost

Table 4 shows that the strong annotator achieves the highest AVG for Qwen3-8B FKL-on-key. Online self-annotation uses the training policy without an external annotator, reaching 80.65%, an absolute gain of 1.90% over GRPO. Annotation details and training dynamics are in App. D.3.
For this online-self configuration, annotation and teacher forward occur only in the first 40 steps. Component timings in Appendix E.3 yield an estimated training-time overhead of 2.7% amortized over 300 steps for online self-annotation.

Table 4: **Annotator comparison** (Qwen3-8B, FKL on key spans). Five-benchmark AVG and absolute gain over GRPO (%). Paired confidence intervals are reported in Appendix E.2.

Annotator	AVG	Δ (%)
None (GRPO)	78.75	—
Strong API	**81.51**	**+2.76**
Online-self	80.65	+1.90
Mid (32B)	79.84	+1.09
Static-self (8B)	79.67	+0.92

6 Discussion and Conclusion

STEPS addresses the training collapse observed in SDPO and SRPO under our mathematical reasoning setup. By concentrating span-specific KL guidance on critical reasoning spans during an initial correction phase and then transitioning to GRPO, STEPS maintains stable training while improving reasoning performance. Experiments across four models from three families show gains in mean accuracy, including with online self-annotation. On SmolLM3, STEPS achieves 71.84 AVG at step 40, exceeding GRPO's 69.45 at step 325 and demonstrating effective learning with substantially fewer training steps.

These results support treating privileged self-distillation as a targeted, temporary correction mechanism for reasoning RL. Where guidance is applied, which KL direction it uses, and when it is withdrawn are important design choices. STEPS makes these choices explicit, connecting early corrective supervision with subsequent reward-driven learning. This perspective motivates adaptive span selection and guidance schedules, as well as extensions to broader reasoning tasks.

AI USE STATEMENT

We used ChatGPT and Codex to assist with manuscript editing and language polishing. The authors take responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

Section 3 specifies the objective and schedule. Appendix B states the theoretical assumptions and derivations. Appendix C provides training and decoding configurations, the update algorithm, prompts, and span-to-token alignment; Appendix D reports additional results and diagnostics. Paired evaluation uncertainty and component-level timing are reported in Appendices E.2 and E.3.

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. arXiv preprint arXiv:2306.13649, 2023.

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

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

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In International Conference on Learning Representations, volume 2024, pp. 32694–32717, 2024.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. OpenThoughts: Data recipes for reasoning models. arXiv preprint arXiv:2506.04178, 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015.

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

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. VinePPO: Refining credit assignment in RL training of LLMs. In Proceedings of the 42nd International Conference on Machine Learning, 2025.

Devvrit Khatri, Lovish Madaan, Rishabh Tiwari, Rachit Bansal, Sai Surya Duvvuri, Manzil Zaheer, Inderjit S Dhillon, David Brandfonbrener, and Rishabh Agarwal. The art of scaling reinforcement learning compute for LLMs. arXiv preprint arXiv:2510.13786, 2025.

Jeonghye Kim, Xufang Luo, Minbeom Kim, Sangmook Lee, Dohyung Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. Why does self-distillation (sometimes) degrade the reasoning capability of LLMs? arXiv preprint arXiv:2603.24472, 2026.

Gengsheng Li, Tianyu Yang, Junfeng Fang, Mingyang Song, Mao Zheng, Haiyun Guo, Dan Zhang, Jinqiao Wang, and Tat-Seng Chua. Unifying group-relative and self-distillation policy optimization via sample routing. arXiv preprint arXiv:2604.02288, 2026a.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. arXiv preprint arXiv:2604.13016, 2026b.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In The Twelfth International Conference on Learning Representations, 2024.

Zicheng Lin, Tian Liang, Jiahao Xu, Qiuzhi Lin, Xing Wang, Ruilin Luo, Chufan Shi, Siheng Li, Yujiu Yang, and Zhaopeng Tu. Critical tokens matter: Token-level contrastive estimation enhances LLM's reasoning capability. arXiv preprint arXiv:2411.19943, 2024.

Changze Lv, Jie Zhou, Wentao Zhao, Jingwen Xu, Shihan Dou, Zisu Huang, Muzhao Tian, Xiaohua Wang, Zhengkang Guo, Yang Liu, et al. Learning query-specific rubrics from human preferences for DeepResearch report generation. arXiv preprint arXiv:2602.03619, 2026.

Mathematical Association of America. AMC 12 2023 problems, 2023. https://artofproblemsolving.com/wiki/index.php/AMC_12_Problems_and_Solutions.

Mathematical Association of America. AIME 2024 problems, 2024. https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions.

Mathematical Association of America. AIME 2025 problems, 2025. https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions.

Meta. Llama-3.2-3B-Instruct model card. https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct, 2024.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models. arXiv preprint arXiv:2602.04942, 2026.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. arXiv preprint arXiv:2311.12022, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300, 2024.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. arXiv preprint arXiv:2601.19897, 2026.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. In Proceedings of the Twentieth European Conference on Computer Systems, pp. 1279–1297, 2025.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. arXiv preprint arXiv:2512.13961, 2025.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin 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 (Volume 1: Long Papers), pp. 9426–9439, 2024.

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

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.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled RLVR. arXiv preprint arXiv:2604.03128, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. In Advances in Neural Information Processing Systems, volume 38, pp. 113222–113244, 2025.

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.

A Notation and Preliminaries

Notation. Let π_θ be a language model parameterized by θ. Given a problem x, $y = (y_1, \ldots, y_T)$ is generated autoregressively; $R(x, y) \in \{0, 1\}$ is a binary verifier; c denotes privileged information (a verified reasoning trace, environment feedback, or coarse type label). The student is $\pi_S(\cdot \mid x) := \pi_\theta(\cdot \mid x)$ and the privileged-context teacher is $\pi_T(\cdot \mid x, c) := \pi_{\theta_T}(\cdot \mid x, c)$ (same parametric family, with parameters periodically synchronized from the student and a different prompt). Our goal is to train π_θ to maximize verifier reward without incurring the epistemic suppression and OOD degradation observed under all-token OPSD (Kim et al., 2026; Yang et al., 2026).

GRPO baseline and all-token OPSD. Group Relative Policy Optimization (Shao et al., 2024) samples G rollouts $\{y^{(i)}\}_{i=1}^G \sim \pi_S(\cdot \mid x)$ and assigns each rollout the sequence-level advantage $A^{(i)} = (R^{(i)} - \mu_G)/\sigma_G$; all tokens in $y^{(i)}$ share that advantage. When the group has uniform reward, the zero-variance convention sets $A^{(i)} \equiv 0$ on every token; Remark 4 describes the local auxiliary signal in this dead zone. All-token OPSD (Hübotter et al., 2026; Zhao et al., 2026) densifies this signal by adding token-level KL guidance $D(\pi_T(\cdot|x, c, \hat{y}_{<t}) \,\|\, \pi_S(\cdot|x, \hat{y}_{<t}))$ whose support spans every response position regardless of semantic role, with gradients flowing only through π_S.

Forward and reverse KL. For a fixed rollout prefix, let $q_t := \pi_T(\cdot \mid x, c, \hat{y}_{<t})$ be the teacher distribution and $p_t := \pi_S(\cdot \mid x, \hat{y}_{<t})$ be the student distribution. We call $\mathrm{KL}(q_t \| p_t)$ forward KL (FKL) and $\mathrm{KL}(p_t \| q_t)$ reverse KL (RKL), always relative to teacher-to-student distillation. FKL and RKL are not interchangeable: their per-token gradients depend asymmetrically on p_t and q_t, as formalized in Lemma 2 in Appendix B.1, so the divergence direction is itself a per-token-class design decision rather than a global hyperparameter.

Critical-span structure. For each rollout y we receive a span partition $\{1, \ldots, |y|\} = \mathcal{E}_y \sqcup \mathcal{K}_y \sqcup \mathcal{N}_y$ from the privileged annotator described in Section 3.1, where $\mathcal{E}_y$ (error spans, empty on $R{=}1$ rollouts) and $\mathcal{K}_y$ (key spans, empty on $R{=}0$ rollouts) jointly form $\mathcal{S}_y = \mathcal{E}_y \cup \mathcal{K}_y$ and cover at most $\alpha = 0.25$ of response tokens; $\mathcal{N}_y$ contains all remaining (non-span) tokens.

B Theoretical Statements and Proofs

We use the notation of §4.1 throughout. All distributions are on a finite vocabulary $\mathcal{V}$ unless stated otherwise. KL derivatives refer to the idealized unclipped objectives with stop-gradient teacher targets and masks.

Remark 1 (Span-restricted decomposition)**.** With stop-gradient masks, the partition $\mathcal{S}_y \sqcup \mathcal{N}_y = \{1, \ldots, |\hat{y}|\}$ is fixed during differentiation. Splitting the sequence-normalized token sum in Eq. (1) therefore gives $\nabla_\theta \mathcal{L} = \nabla_\theta \mathcal{L}_{\text{span}} + \nabla_\theta \mathcal{L}_{\text{nonspan}}$. The supports are disjoint in token positions, not in model parameters.

B.1 KL Gradient Asymmetry

At a fixed rollout position, let $p_v = \pi_S(v) = \pi_\theta(v) > 0$ and $q_v = \pi_T(v) > 0$ denote the student and teacher probabilities on the same causal prefix. Define $D_F = D_{\mathrm{KL}}(q\|p) = \mathrm{KL}_F(\pi_T, \pi_\theta)$ and $D_R = D_{\mathrm{KL}}(p\|q) = \mathrm{KL}_R(\pi_T, \pi_\theta)$, with q fixed during the update. The notation $r(v)$ used in the derivation below is equivalent to r_v.

Lemma 2 (Softmax gradient identities)**.** Let ℓ_v be the student logit, $r_v = \log(p_v/q_v)$, and $\bar{r} = \sum_v p_v r_v$. Then

$$\frac{\partial D_F}{\partial \ell_v} = p_v - q_v, \qquad \frac{\partial D_R}{\partial \ell_v} = p_v(r_v - \bar{r}).$$

Proof of Lemma 2. For $\pi_\theta(v) \propto \exp(\ell_v)$ with $Z := \sum_u \exp(\ell_u)$,

$$\frac{\partial \pi_\theta(v')}{\partial \ell_v} = \pi_\theta(v')\big(\mathbb{1}[v = v'] - \pi_\theta(v)\big), \qquad \frac{\partial \log \pi_\theta(v')}{\partial \ell_v} = \mathbb{1}[v = v'] - \pi_\theta(v). \tag{6}$$

Reverse KL. $\mathrm{KL}_R(\pi_T, \pi_\theta) = \mathrm{KL}(\pi_\theta \| \pi_T) = \sum_{v'} \pi_\theta(v')(\log \pi_\theta(v') - \log \pi_T(v'))$. Differentiating:

$$\begin{aligned}
\frac{\partial \mathrm{KL}_R}{\partial \ell_v} &= \sum_{v'} \frac{\partial \pi_\theta(v')}{\partial \ell_v}(\log \pi_\theta(v') - \log \pi_T(v')) + \sum_{v'} \pi_\theta(v') \frac{\partial \log \pi_\theta(v')}{\partial \ell_v} \\
&\overset{(6)}{=} \sum_{v'} \pi_\theta(v')\big(\mathbb{1}[v = v'] - \pi_\theta(v)\big) r(v') + \sum_{v'} \pi_\theta(v')\big(\mathbb{1}[v = v'] - \pi_\theta(v)\big) \\
&= \pi_\theta(v) r(v) - \pi_\theta(v)\bar{r} + \pi_\theta(v) - \pi_\theta(v) \\
&= \pi_\theta(v) \cdot \big(r(v) - \bar{r}\big).
\end{aligned}$$

This proves Eq. (3).

Forward KL. $\mathrm{KL}_F(\pi_T, \pi_\theta) = \mathrm{KL}(\pi_T \| \pi_\theta) = -\sum_{v'} \pi_T(v') \log \pi_\theta(v') + \text{const}$. Differentiating:

$$\frac{\partial \mathrm{KL}_F}{\partial \ell_v} = -\sum_{v'} \pi_T(v')\big(\mathbb{1}[v = v'] - \pi_\theta(v)\big) = -\pi_T(v) + \pi_\theta(v) \cdot 1 = \pi_\theta(v) - \pi_T(v).$$

This proves Eq. (3). □

Corollary 3 (Pointwise KL asymmetry)**.** Equation (3) distinguishes two local regimes:

- **Under-allocated teacher-supported tokens.** For fixed $q_v > 0$ and bounded $|\bar{r}|$, as $p_v \to 0$, FKL produces upward logit pressure of magnitude $q_v - p_v \to q_v$, whereas the RKL magnitude $p_v|r_v - \bar{r}| = O(p_v \log(q_v/p_v))$ vanishes.
- **Student-overweighted tokens.** When $p_v > q_v$, FKL produces downward logit pressure of magnitude $p_v - q_v$. RKL produces downward pressure precisely when $r_v > \bar{r}$, with magnitude $p_v(r_v - \bar{r})$. The condition $p_v > q_v$ alone implies neither this direction nor a larger RKL magnitude.

Proof of Corollary 3. For fixed $q_v > 0$ and bounded $|\bar{r}|$, $r_v = \log(p_v/q_v) \to -\infty$ and $p_v|r_v - \bar{r}| = O(p_v \log(q_v/p_v)) \to 0$, whereas $p_v - q_v \to -q_v$. The signs and magnitudes in the second case follow directly from the two derivatives in Eq. (3). □

For a direct Euclidean logit step, $\ell_v^{(+)} = \ell_v - \eta\, \partial D/\partial \ell_v$ with $\eta > 0$. Thus RKL lowers ℓ_v when $r_v > \bar{r}$, and FKL lowers every logit in a token set whose entries satisfy $p_v > q_v$. These are pointwise logit statements: softmax normalization couples probability changes, and neither monotone token-set mass nor the direction of a shared-parameter optimizer update follows.

Remark 4 (Auxiliary signal in GRPO dead zones)**.** If every rollout in a group is correct, the zero-variance convention sets all GRPO advantages to zero. During $\lambda_k > 0$, the default FKL branch still has local logit gradient $\lambda_k |\hat{y}|^{-1}(p_v - q_v)$ at each selected key position by Eq. (3), which is nonzero as a vector when $p \neq q$. This unclipped, position-wise statement does not ensure a nonzero shared-parameter gradient; the auxiliary signal vanishes when $\lambda_k = 0$.

B.2 Privileged-Gradient Exposure

Following the decomposition of Yang et al. (2026), fix a step k and condition on $\mathcal{F}_k$: the training history, problem x_k, rollout $\hat{y}_k$, and selected mask m_k. Thus the current student and teacher parameters, score vectors, response length, and mask are fixed. The context c follows its conditional distribution given $\mathcal{F}_k$; it need not be independent of the mask. Define $\bar{\pi}_{T,k,t}(v) := \mathbb{E}_c[\pi_T(v \mid x_k, c, \hat{y}_{k,<t}) \mid \mathcal{F}_k]$. For the default FKL-on-key action $(\mu_E, \mu_K) = (0, 1)$, the centered privileged-gradient deviation is

$$\delta_{k,t} := -\sum_v \big(\pi_T(v \mid x_k, c, \hat{y}_{k,<t}) - \bar{\pi}_{T,k,t}(v)\big) \nabla_\theta \log \pi_S(v \mid x_k, \hat{y}_{k,<t}), \tag{7}$$

with $\mathbb{E}_c[\delta_{k,t} \mid \mathcal{F}_k] = 0$ and matching conditional variance

$$V_{k,t} := \sum_v \mathrm{Var}_c[\pi_T(v \mid x_k, c, \hat{y}_{k,<t}) \mid \mathcal{F}_k] \geq 0. \tag{8}$$

Here $V_{k,t} = 0$ if and only if the teacher distribution is almost surely constant under this conditional context distribution; this does not imply that it ignores privileged information across masks or rollouts.

Assume that, uniformly over steps and positions, every zero-sum vector a satisfies the score-operator bound

$$\left\| \sum_v a_v \nabla_\theta \log \pi_S(v) \right\|_2^2 \le C_s^2 \sum_v a_v^2, \qquad \sum_v a_v = 0. \tag{9}$$

This is an analytical assumption, distinct from optimizer gradient clipping.

Proposition 5 (Cumulative exposure bound)**.** Let $m_{k,t} = \mathbb{1}[t \in \mathcal{K}_{y_k}]$ select the active FKL keyspan positions in rollout $\hat{y}_k$. Under Equation (9), the cumulative weighted second moment satisfies

$$\begin{aligned} \mathcal{E}_K &:= \sum_{k=1}^{K} \lambda_k^2 \, \mathbb{E}\left[\frac{1}{|\hat{y}_k|} \sum_t m_{k,t} \|\delta_{k,t}\|_2^2 \right] \\ &\le C_s^2 \sum_{k=1}^{K} \lambda_k^2 \, \mathbb{E}\left[\frac{1}{|\hat{y}_k|} \sum_t m_{k,t} V_{k,t} \right]. \end{aligned} \tag{10}$$

If coverage is at most α and the expected variance averaged over selected positions is bounded by $\bar{V}$ at every step, with empty masks contributing zero, then

$$\mathcal{E}_K \le C_s^2 \alpha \bar{V} \sum_{k=1}^{K} \lambda_k^2,$$

which is the bound reported in Eq. (4).

Proof. Let $a_v = \pi_T(v \mid x_k, c, \hat{y}_{k,<t}) - \bar{\pi}_{T,k,t}(v)$, so $\sum_v a_v = 0$ and $\delta_{k,t} = -\sum_v a_v \nabla_\theta \log \pi_S(v)$. Applying the score-operator bound and averaging over context conditional on $\mathcal{F}_k$ gives

$$\mathbb{E}_c[\|\delta_{k,t}\|_2^2 \mid \mathcal{F}_k] \le C_s^2 \sum_v \mathbb{E}_c[a_v^2 \mid \mathcal{F}_k] = C_s^2 V_{k,t}.$$

The operator bound controls the full vocabulary sum, including its cross terms. Since the mask and response length are $\mathcal{F}_k$-measurable, summing over selected positions, taking the outer expectation, and then summing with the deterministic weights λ_k^2 yields Eq. (10) by the tower property.

For the coverage bound, let $\mathcal{M}_k = \{t : m_{k,t} = 1\}$ and define $\bar{V}_{k,\mathrm{span}} = |\mathcal{M}_k|^{-1} \sum_{t \in \mathcal{M}_k} V_{k,t}$, with value zero for empty masks. Nonnegativity, $|\mathcal{M}_k|/|\hat{y}_k| \le \alpha$, and the uniform assumption $\mathbb{E}[\bar{V}_{k,\mathrm{span}}] \le \bar{V}$ imply

$$\begin{aligned} \mathbb{E}\left[\frac{1}{|\hat{y}_k|} \sum_t m_{k,t} V_{k,t} \right] &= \mathbb{E}\left[\frac{|\mathcal{M}_k|}{|\hat{y}_k|} \bar{V}_{k,\mathrm{span}} \right] \\ &\le \alpha \, \mathbb{E}[\bar{V}_{k,\mathrm{span}}] \le \alpha \bar{V}. \end{aligned}$$

Substitution into Eq. (10) gives Eq. (4). The joint expectation is not factored: no independence of mask, coverage, and context is assumed, and the variance bound concerns selected positions rather than the rollout average.

Finally, the finite schedule in Eq. (2) satisfies $0 \le \lambda_k \le w_0$ and $\lambda_k = 0$ beyond $t_{\mathrm{start}} + T_{\mathrm{decay}}$. Hence $\sum_{k=1}^{K} \lambda_k^2 \le (t_{\mathrm{start}} + T_{\mathrm{decay}}) w_0^2$ for every K, making the exposure upper bound independent of the total training horizon for a fixed guidance schedule. □

Remark 6 (Scope of the exposure bound)**.** We bound the cumulative conditional second moment. The vector mean $\mathbb{E}_c[\delta_{k,t} \mid \mathcal{F}_k]$ vanishes by centering, whereas the first absolute moment $\mathbb{E}_c[\|\delta_{k,t}\| \mid \mathcal{F}_k]$ need not vanish. This variance bound does not control mask-dependent shifts in the

conditional teacher mean, all privileged-information effects, or downstream optimization error. If the context is determined by the rollout and mask, the conditional variance may be zero even while privileged information affects the update. Finite exposure is therefore not a no-harm or generalization guarantee.

B.3 Key-Span Signal Bound and Extensions

Use the conditioning $\mathcal{F}_k$ and fixed teacher targets defined in Appendix B.2. At each selected position, $g_{\mathrm{FKL}}(t) = \nabla_\theta D_F(t)$ is the per-token gradient before sequence normalization. The following proposition states the default FKL-on-key case; the proof retains the error-span RKL and joint-branch extensions.

Proposition 7 (Key-span signal lower bound)**.** At step k, let $\tilde{g}_k$ be a common $\mathcal{F}_k$-measurable oracle loss-gradient direction, and let $g_{\mathrm{FKL}}(t)$ denote the per-token FKL gradient. Suppose every nonempty selected key set has expected true-critical fraction at least q_K, conditional on $\mathcal{F}_k$. For each selected position, assume that its expected alignment with $\tilde{g}_k$, conditional on $\mathcal{F}_k$ and its critical status, is at least $\gamma_K > 0$ for true-critical positions and at least $-B_K$, $B_K \geq 0$, for false positives. Define

$$p_K = \mathbb{E}\left[\frac{|\mathcal{K}_{y_k}|}{|\hat{y}_k|}\right], \qquad g_k^{\mathrm{sel}} = \frac{\lambda_k}{|\hat{y}_k|} \sum_{t \in \mathcal{K}_{y_k}} g_{\mathrm{FKL}}(t),$$

with empty sets contributing zero. Then

$$\mathbb{E}\left[\langle g_k^{\mathrm{sel}}, \tilde{g}_k \rangle\right] \geq \lambda_k p_K \left[q_K \gamma_K - (1 - q_K) B_K\right].$$

For $\lambda_k p_K > 0$, this lower bound is positive whenever $q_K > B_K/(\gamma_K + B_K)$.

Proof. We work under binary span weights $w_t \in \{0, 1\}$ as in §3.1. Under action profile $(\mu_E, \mu_K) \in \{0, 1\}^2$, the per-step selected gradient is

$$g_k^{\mathrm{sel}} = \frac{\lambda_k}{|\hat{y}_k|} \left(\mu_E \sum_{t \in \mathcal{E}_{y_k}} g_{\mathrm{RKL}}(t) + \mu_K \sum_{t \in \mathcal{K}_{y_k}} g_{\mathrm{FKL}}(t) \right),$$

i.e., a sequence-normalized sum matching Eq. (1). Equivalently, each per-set mean is multiplied by its fraction of the full rollout. Empty sets contribute zero.

Condition on $\mathcal{F}_k$ and use the same $\mathcal{F}_k$-measurable oracle direction $\tilde{g}_k$ at every selected position. For a class $C \in \{E, K\}$, let $Z_t^C \in \{0, 1\}$ indicate true class membership at a selected position and write $g_E = g_{\mathrm{RKL}}$, $g_K = g_{\mathrm{FKL}}$. The conditional margin assumptions are

$$\mathbb{E}[\langle g_C(t), \tilde{g}_k \rangle \mid \mathcal{F}_k, Z_t^C] \geq Z_t^C \gamma_C - (1 - Z_t^C) B_C.$$

For each nonempty selected set $\mathcal{C}_{y_k}$, require the uniform conditional precision bound

$$\frac{1}{|\mathcal{C}_{y_k}|} \sum_{t \in \mathcal{C}_{y_k}} \Pr(Z_t^C = 1 \mid \mathcal{F}_k) \geq q_C.$$

Taking expectation over Z_t^C and summing the margins against the common direction gives

$$\mathbb{E}\left[\left\langle \frac{1}{|\hat{y}_k|} \sum_{t \in \mathcal{C}_{y_k}} g_C(t), \tilde{g}_k \right\rangle \middle| \mathcal{F}_k \right] \geq \frac{|\mathcal{C}_{y_k}|}{|\hat{y}_k|} \big(q_C \gamma_C - (1 - q_C) B_C \big).$$

Taking the outer expectation and writing $p_C := \mathbb{E}[|\mathcal{C}_{y_k}|/|\hat{y}_k|]$ at step k yields

$$\mathbb{E}[\langle g_k^{\mathrm{sel}}, \tilde{g}_k \rangle] \geq \lambda_k \Big[\mu_E p_E \big(q_E \gamma_E - (1 - q_E) B_E \big) + \mu_K p_K \big(q_K \gamma_K - (1 - q_K) B_K \big) \Big].$$

The default $(\mu_E, \mu_K) = (0, 1)$ gives the proposition. The lower bound is positive if every active class with positive coverage has a positive bracket and at least one has $\lambda_k p_C > 0$. Independence between token positions, coverage, and context is unnecessary because the precision and margin assumptions hold uniformly under the conditioning. □

Remark 8 (On the oracle direction)**.** $\tilde{g}_k$ is an unobservable verifier-grounded loss-gradient direction. Prop. 7 is a conditional lower bound under the stated precision and alignment assumptions. Alignment with different token-specific directions $\tilde{g}(t)$ would not imply the sequence-level statement; the shared direction is essential. Neither the local softmax identities nor the empirical token-lift diagnostic verifies these assumptions. This is a structural complement to the conditional exposure bound, not an optimality theorem for the trained network or a guarantee for the clipped optimizer step.

B.4 Per-token damping under teacher coverage gap

Proposition 9 (Per-token damping)**.** Let $w_t = \pi_T(y_t|x, c, y_{<t})/\pi_S(y_t|x, y_{<t})$ be the unclipped positive-advantage reweighting factor considered here, with $A > 0$. If $\pi_T(y_{t^\star}|c) \le \delta$ and $\pi_S(y_{t^\star}) \ge p_0 > 0$ at some position $t^\star$, then $\hat{A}_{t^\star} := A \cdot w_{t^\star} \le A\delta/p_0$.

The condition is local — when $\delta < p_0$, one token with low teacher probability suffices for damping — and can arise in math reasoning, where the privileged context is typically a single canonical solution and the student may sample valid alternatives. Independent of Yang et al. (2026)'s mutual-information ill-posedness; the two viewpoints are complementary.

Proof. By definition,

$$w_{t^\star} = \frac{\pi_T(y_{t^\star} \mid x, c, y_{<t^\star})}{\pi_S(y_{t^\star} \mid x, y_{<t^\star})} \le \frac{\delta}{p_0},$$

where the inequality applies the two probability bounds in the proposition. Multiplying both sides by $A > 0$: $\hat{A}_{t^\star} = A \cdot w_{t^\star} \le A \cdot \delta/p_0$. □

Remark 10 (On the unclipped form)**.** The proposition is stated for the unclipped reweighting factor $w_t = \pi_T/\pi_S$. If this factor is clipped to $[1 - \epsilon_w, 1 + \epsilon_w]$, with $0 < \epsilon_w < 1$, the per-token deviation is bounded. Under clipping, the weight is bounded below by $1 - \epsilon_w$, so the unclipped upper bound δ/p_0 need not hold; the qualitative phenomenon — that w_t still damps correct non-canonical tokens whenever $\pi_T(y_t) < \pi_S(y_t)$ — persists but is bounded by the clip.

Remark 11 (Why this is not a global collapse theorem)**.** Prop. 9 is a per-token statement: at any single position satisfying the hypotheses with $\delta < p_0$, the reweighting factor damps the positive-advantage term on that token. We do not prove that this damping aggregates to global collapse of the training objective. The aggregate training behavior depends on the remaining token updates, clipping, and data; it is an empirical question rather than a consequence of this bound.

C Experimental Details

C.1 Training Hyperparameters

Common settings. The Qwen3-8B experiments share the following settings: Qwen3-8B base, OpenThoughts-114k math subset (up to 30K problem–solution pairs), H100 GPUs, AdamW optimizer with weight decay 0.01 and gradient-clip norm 1.0, max prompt length 2048, vLLM rollout at $T = 0.6$, train batch size 32 with $G = 8$ rollouts per problem, mini-batch size 32, learning rate 1×10^{-5} with 10-step linear warm-up, and 300 total training steps. Max response length is 32768 for Think-student training and 8192 for NoThink-student ablations. The cross-family runs reuse the optimizer, rollout, and evaluation settings, with model-specific details in App. C.7. Method-specific deltas are listed in Tab. 5 (baselines) and Tab. 6 (STEPS); best-checkpoint indices used for Tab. 2 are STEPS step 250, GRPO step 250, SDPO step 35, SRPO step 65 (collapse onset).

Table 5: Method-specific hyperparameters for the four baselines. "—" marks parameters not applicable to the method. All baselines use AdamW with the common settings above and learning rate 1×10^{-5}.

	GRPO	SDPO (Hübotter et al., 2026)	SRPO (Li et al., 2026a)	RLSD (Yang et al., 2026)
Distillation divergence	—	JSD (α=0.5)	FKL	— (no KL grad)
Distillation top-K	—	100	100	—
Teacher / EMA	—	frozen (EMA 0)	EMA 0.05	sync N=10
λ schedule	—	constant	constant	1.0→0 over 50 steps
Other	DAPO ε=0.2/0.28, IS-clip 2	correct-sibling rollouts	entropy suppression β=1	teacher/student-ratio advantage weighting

Table 6: STEPS-specific hyperparameters (default $(\mu_E, \mu_K) = (0, 1)$, FKL on $\mathcal{K}_y$ only).

Category	**Parameter**	**Value**
Loss	Distillation divergence (default)	FKL on $\mathcal{K}_y$
	Distillation top-K	100
	Per-vocab KL clip (τ)	0.05
	Coverage cap (α)	0.25
	DAPO $\varepsilon_{\text{low}}/\varepsilon_{\text{high}}$	$0.2/0.28$
Schedule	KL decay start (t_{start})	10
	KL decay window (T_{decay})	30
	Initial KL weight (w_0)	0.5
	Teacher sync interval (N)	10
Routing	Default (μ_E, μ_K)	$(0, 1)$
	Thinking mode	symmetric (student / teacher both Think)
	Annotator	`qwen3.5-plus`, thinking disabled

C.2 Algorithm

Algorithm 1 gives the per-mini-batch update loop. When $\lambda_k = 0$, both annotation and the teacher forward are skipped before either is invoked, and all response tokens receive the GRPO objective. Otherwise, the annotator π_A runs once per rollout to produce span masks and a coarse type label, the teacher π_T then forwards in thinking mode conditioned on $(x, c^{(i)})$, and the routed mixed-KL loss is added to the GRPO loss with weight λ_k during the warm-up / decay window; after step $t_{\text{start}} + T_{\text{decay}}$ the method reduces to pure GRPO.

Algorithm 1 STEPS (one mini-batch update)

Require: policy π_θ, teacher π_T, annotator π_A, current step k

Sample G rollouts $\{\hat{y}^{(i)}\}_{i=1}^{G} \sim \pi_S(\cdot|x)$; compute $R^{(i)}$ and GRPO advantages $\hat{A}^{(i)}$
Compute λ_k from Eq. (2); set $\rho_k = 1 - \lambda_k / w_0$
if $\lambda_k = 0$ **then**
    Update θ using GRPO on all response tokens ▷ skip annotation and teacher forward
else
    for each rollout $\hat{y}^{(i)}$ **do**
        $\mathcal{E}_{\hat{y}^{(i)}}$ or $\mathcal{K}_{\hat{y}^{(i)}}$, type label $\tau^{(i)} \leftarrow \pi_A(x, \hat{y}^{(i)}, R^{(i)})$
        Project character spans to token mask $m^{(i)}$, cap at $\alpha = 0.25$
    end for
    Forward π_T in Think mode with $(x, c^{(i)})$ on each causal prefix $\hat{y}^{(i)}_{<t}$, where $c^{(i)} = \tau^{(i)}$
    Forward π_S in Think mode on the same causal prefixes (default symmetric configuration)
    Compute routed KL and GRPO losses per Eq. (1)
    Update θ with $\mathcal{L}^{(\mathcal{N})}_{\text{GRPO}} + \rho_k \mathcal{L}^{(\mathcal{S})}_{\text{GRPO}} + \lambda_k[\mu_E \mathcal{L}^{(\mathcal{E})}_{\text{RKL}} + \mu_K \mathcal{L}^{(\mathcal{K})}_{\text{FKL}}]$
end if
if $k \bmod N = 0$ **and** $k < t_{\text{start}} + T_{\text{decay}}$ **then**
    $\theta_T \leftarrow \theta$ ▷ periodic teacher sync while future KL remains
end if

C.3 Annotator and Teacher Prompts

The privileged annotator π_A (default: `qwen3.5-plus` API) receives a (problem, rollout, verifier-outcome) triple and returns a JSON span specification. Annotator thinking mode is disabled; teacher thinking mode is enabled. The four prompt templates used for these roles are reproduced below.

Annotator π_A — wrong-rollout error-span prompt

You are a mathematics error locator. The student's solution is shown as numbered segments `[0]`, `[1]`, `[2]`, ... Identify the segments that contain the root-cause reasoning errors.

Output (strict JSON):
`{ error_spans: [{ segment_ids: [...], error_type }] }`

`error_type` ∈ {`missed_case`, `illegal_step`, `wrong_constraint`, `wrong_equivalence`, `premature_conclusion`, `format_error`, `arithmetic_slip`, `sign_error`, `off_by_one`}.

Rules. Each span is 1–3 consecutive `segment_ids`. Identify 0–3 spans, focusing on root-cause errors rather than downstream propagation of an upstream mistake. For omissions, select the segment immediately before the gap. Return an empty list if no segment can be confidently identified as root-causal — do not guess. Output only the JSON; do not explain, rewrite, or solve the problem.

Annotator π_A — correct-rollout key-span prompt

You are a mathematics reasoning analyst. The student's solution is shown as numbered segments `[0]`, `[1]`, `[2]`, ... Identify the segments that carry the decisive reasoning steps — the critical decisions or insights that make this solution work.

Output (strict JSON):
`{ key_spans: [{ segment_ids: [...], step_type }] }`

`step_type` ∈ {`case_split`, `boundary_check`, `invariant`, `substitution`, `constraint_use`, `symmetry`, `construction_step`, `final_verification`, `key_formula`, `insight`}.

Rules. Each span is 1–3 consecutive `segment_ids`. Identify 0–3 spans, skipping routine algebra, restatements, and final-answer formatting. Return an empty list if no segment is confidently critical — do not pad with low-importance steps. Output only the JSON; do not explain or solve the problem.

Teacher π_T — wrong-rollout diagnostic prefix

`{ problem }`

[Private diagnostic context. Do not mention this context in the response.]

A sampled solution to this problem contains a local reasoning issue of the following type:

- `{ error_descriptions }`

When solving, be especially careful around this type of issue. Do not reveal, quote, or locate the diagnostic context. Solve the problem independently and end with a boxed answer.

Teacher π_T — correct-rollout diagnostic prefix

`{ problem }`

[Private diagnostic context. Do not mention this context in the response.]

A successful solution to this problem relies on reasoning steps of the following type:

- `{ step_descriptions }`

When solving, make the corresponding reasoning explicit enough to support the answer, but do not use checklist language or refer to this context. End with a boxed answer.

The teacher therefore receives only the original problem x plus a coarse type label; no rollout text, span locations, or span content are passed in as additional privileged context. As in standard on-policy distillation it then evaluates token logits causally on the sampled prefix $\hat{y}_{<t}$.

C.4 Span-to-Token Alignment

Given character-level spans returned by π_A over the response text, we project to a token boolean mask $m \in \{0,1\}^{|\hat{y}|}$ as follows: for each token y_t we compute its character interval $[c_t^-, c_t^+)$ by greedy alignment of single-token decodes against the original response (avoiding decode–re-encode round-trip drift); a token is marked if its interval intersects any span. We then enforce $|\{t : m_t = 1\}| \le \alpha|\hat{y}|$ with $\alpha = 0.25$ by keeping the top-weight tokens.

C.5 Decoding Configurations

We use two decoding protocols. The **main thinking-mode protocol** uses T=0.6, p=0.95, k=20, min p=0, and max_new_tokens= 38912, with `avg@8` on MATH-500 and GPQA-Diamond and `avg@16` on AIME and AMC. It applies to the Qwen3 main results, annotator-quality and asymmetric-training ablations, and OOD evaluation. The **non-thinking protocol** disables thinking, uses max_new_tokens= 8192 and `avg@8` on all benchmarks, and applies only to the NoThink-evaluation ablation in Table 14.

C.6 Training and Evaluation Modes on Qwen Models

We distinguish three choices in the Qwen experiments: the student's rollout mode during training, the self-teacher's conditioning mode, and the student's decoding mode during evaluation. Table 7 summarizes the main configuration and the Qwen3-8B thinking-mode ablations.

Table 7: Qwen3 training and evaluation modes. The NoThink-evaluation row reuses the asymmetric-training checkpoints.

Configuration	Training student	Teacher	Evaluation
Main STEPS, SDPO, and SRPO	Think	Think	Think
Asymmetric ablation in Table 13	NoThink	Think	Think
NoThink-eval ablation in Table 14	NoThink	Think	NoThink

GRPO has no teacher; its row in Table 13 reuses the reference from Table 2.

C.7 Cross-Family Configuration Details

Llama-3.2-3B-Instruct and SmolLM3-3B use RKL on error spans, with the optimizer and rollout settings of Appendix C.1 and the main decoding protocol of Appendix C.5. Llama uses online self-annotation; SmolLM3 uses the strong annotator, `qwen3.5-plus`. Checkpoint selection follows the validation mean@16 rule in Section 5.1.

For both families, AVG covers MATH-500, AIME 2024, AMC 2023, and GPQA-Diamond, omitting AIME 2025 from the Qwen3 five-benchmark mean. The reported SmolLM3 checkpoints are step 40 for STEPS and step 325 for GRPO.

D Additional Results

D.1 Extended Related Work and Paradigm Comparison

Table 8 supplements Section 2 by distinguishing teacher provenance, privileged information, and the token support of supervision. External-teacher OPD/GKD (Gu et al., 2024; Agarwal et al., 2023) provides dense guidance, whereas SDPO and SRPO use self-teachers conditioned on feedback or sibling rollouts (Hübotter et al., 2026; Li et al., 2026a). SRPO routes objectives by sample correctness; RLSD reweights verifier advantages with teacher–student differences (Yang et al., 2026). STEPS instead localizes class-specific KL with semantic masks.

Unlike process reward models (Lightman et al., 2024; Wang et al., 2024), STEPS uses the self-teacher's logits without training an auxiliary reward model. Query-specific rubric methods (Lv et al., 2026) share one rubric across responses to a query; STEPS uses a rollout-specific diagnostic prefix for teacher conditioning and separate span masks for loss localization, without per-query rubric training.

Table 8: **Comparison of post-training paradigms. Bold** marks STEPS's distinctive choices.

Method	Signal	Teacher	Extra info	KL support
SFT	labels	none	gold response	demo tokens
GRPO	verifier R	none	none	no teacher KL
OPD/GKD	logits	external / frozen	info-symmetric	all tokens
SDPO	logits	self, EMA	feedback + sibling	all tokens
SRPO	R + logits	self, EMA	feedback + sibling	branch-weighted
RLSD	$R \times \tilde{w}_t$	self, sync N=10	verified trace r	all tokens
STEPS	**R + span KL**	**self, sync N=10**	**coarse type only**	**critical spans**

D.2 Key-token Lift and Credit Assignment

Both the key-token lift analysis in Table 1 and the token-level update localization analysis in Figure 6 use Qwen3-8B and follow the experimental setup described in Section 5.1 and Appendix C.1, including the starting checkpoint, number of training updates, and training-batch configuration.

Key-token lift. To examine whether the local effect in Corollary 3 is reflected in actual updates, we measure post-update log-probability lift on a common held-out set of teacher-supported key-span tokens. Let $\mathcal{T}_{\text{lift}}$ denote this set, selected before the update by requiring $t \in \mathcal{K}_y$ and $\pi_T^{(k)}(\hat{y}_t) > \pi_{\theta_k}(\hat{y}_t)$. We define

$$\Delta_{\text{lift}} = \mathbb{E}_{t \in \mathcal{T}_{\text{lift}}}\big[\log \pi_{\theta_{k+1}}(\hat{y}_t) - \log \pi_{\theta_k}(\hat{y}_t)\big].$$

Table 9 shows the largest lift for STEPS-FKL, consistent with the local mechanism. This diagnostic measures realized token-probability changes, not the oracle alignment assumed in Proposition 7.

Table 9: Post-update key-token log-probability lift on a shared held-out token set.

Method	STEPS-FKL	STEPS-RKL	RLSD	GRPO	SRPO	SDPO
Δ_{lift} (nats)	+0.145	+0.078	+0.068	+0.054	+0.027	+0.019
vs. GRPO	+168%	+44%	+26%	—	−50%	−65%

Token-level update localization. Figure 6 compares paired post-update log-probability shifts relative to GRPO, $d_t^{(m)} = \log p_t^{(m)} - \log p_t^{(\text{GRPO})}$, where m is STEPS or all-token guidance. The diagnostic comprises 22,715 tokens from eight responses to four problems. The horizontal and vertical axes show all-token and STEPS shifts, respectively. Blue and coral hexagons represent selected key tokens and unselected tokens, with density normalized separately within each group. Points above the dashed equality line have larger shifts under STEPS.

Selected-token mean shifts are 0.088 nats for STEPS and 0.082 nats for all-token guidance; unselected-token mean absolute shifts are 0.00052 and 0.02425 nats, respectively. In this diagnostic sample, STEPS retains slightly larger mean shifts on selected tokens while limiting departures from GRPO elsewhere. Since model parameters are shared across positions, this localization is an observed property of the realized updates, not an automatic consequence of masking the KL loss.

Token-level credit assignment. Figure 7 visualizes token-level credit on a Diophantine case study ($6^a = 1 + 2^b + 3^c$) for one incorrect and one correct rollout. Credit profiles are measured per-token update magnitudes for RLSD, SDPO, and STEPS (one optimization step on the same rollout, same seed); the right panels report concentration inside annotator spans versus outside.



Figure 7: Token-level credit assignment on the Diophantine case study. Black boxes denote annotator spans overlaid for all methods; only STEPS uses them as loss masks. Credit is normalized update magnitude, and the concentration ratio is mean credit inside boxes divided by outside. Rollout A shows the optional RKL-on-E_y ablation; Rollout B shows the default FKL-on-K_y configuration.

D.3 Annotator Quality and Dynamics

Table 10 reports the benchmark-level results summarized in Table 4. Online-self reuses the actor weights through vLLM on the same H100 during the initial 40-step phase. Static-self uses the fixed Qwen3-8B base checkpoint, zero-shot with the same annotation template.

Table 10: **Full annotator comparison on Qwen3-8B (FKL on key spans).** AVG is the five-benchmark mean; Δ is the absolute gain over GRPO, expressed in %. Best per column is **bold**. Strong API: `qwen3.5-plus`; mid: Qwen3-32B; static-self: the Qwen3-8B base checkpoint. Appendix E.2 reports the online-self paired 95% CI, $[-0.25, 3.75]\%$, which includes zero.

Annotator	MATH-500	AIME 24	AIME 25	AMC 23	GPQA-D	AVG	Δ (%)
None (GRPO)	97.30	77.08	68.96	96.56	53.85	78.75	—
Strong API	**97.83**	**80.21**	**73.54**	**97.66**	**58.33**	**81.51**	**+2.76**
Online-self	97.55	78.54	72.29	97.03	57.84	80.65	+1.90
Mid (32B)	96.83	79.08	68.61	97.27	57.43	79.84	+1.09
Static-self (8B)	97.34	75.92	72.19	96.11	56.78	79.67	+0.92

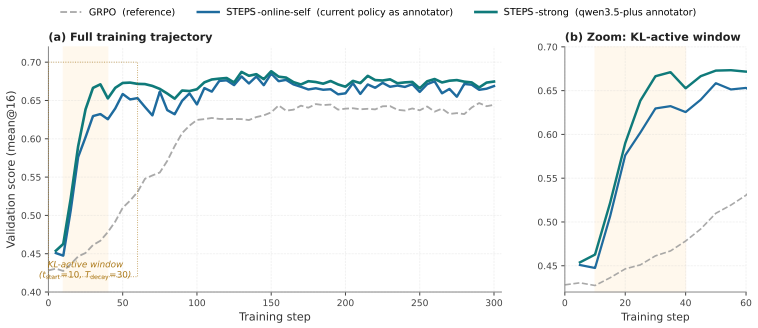

Figure 8: **Annotator ablation dynamics.** OpenThoughts held-out validation mean@16 for the annotators in Table 4: (a) full training trajectory; (b) early-phase zoom. Shading marks distillation decay at steps 10–40; auxiliary computation ends after step 40. Curves use EMA smoothing ($\alpha = 0.85$).

D.4 Evaluation Protocol Sanity Checks

The Qwen3-8B base scores reported in the "base" row of Tab. 2 are independently re-graded with an ensemble OR grader (`verl/math_dapo.py` ∨ HF `math-verify` ∨ AIME-int compare ∨ normalized-string compare); across all five eval cells the strict-vs-ensemble difference is $\leq 0.03\%$, i.e. smaller than the bootstrap CI width on a problem set of this size. The headline strict-grader numbers are therefore not driven by grader choice. Tab. 11 reports cell-wise bootstrap 95% confidence intervals over problems for the main held-out results. These intervals are intended for calibration of benchmark uncertainty; because AIME contains only 30 problems, we avoid treating individual AIME-column gaps as standalone significance claims.

Table 11: **Cell-wise bootstrap 95% confidence intervals** for Tab. 2. Each cell reports mean accuracy with a bootstrap interval over problem IDs. AVG is the unweighted mean over the five benchmark means; its shading follows Tab. 2.

Method	MATH-500	AIME 24	AIME 25	AMC 23	GPQA-D	AVG
Qwen3-8B base (ref.)	96.80 [95.48, 98.00]	76.25 [62.50, 87.92]	67.50 [54.17, 80.42]	95.94 [92.50, 98.75]	58.27 [52.21, 64.20]	78.95 [75.04, 82.86]
GRPO	97.30 [96.03, 98.45]	77.08 [62.71, 89.38]	68.96 [54.58, 82.08]	96.56 [92.19, 99.38]	53.85 [48.04, 59.66]	78.75 [74.67, 82.83]
SDPO	95.13 [93.65, 96.48]	52.50 [38.33, 66.67]	41.25 [26.04, 56.67]	91.41 [84.69, 96.25]	39.90 [34.85, 44.95]	64.04 [59.58, 68.50]
SRPO	96.48 [95.13, 97.65]	69.17 [54.58, 82.08]	54.17 [39.58, 67.50]	93.12 [87.50, 97.50]	37.31 [32.32, 42.42]	70.05 [65.87, 74.23]
RLSD	96.68 [95.35, 97.90]	75.42 [61.67, 87.08]	68.54 [55.21, 81.46]	97.03 [93.12, 99.38]	57.26 [51.58, 62.82]	78.99 [74.96, 82.98]
STEPS (RKL on $\mathcal{E}_y$)	97.48 [96.18, 98.55]	78.54 [64.17, 90.42]	70.42 [56.25, 83.33]	97.81 [94.38, 99.69]	56.44 [50.06, 62.69]	80.14 [76.08, 84.15]
STEPS (FKL on $\mathcal{K}_y$)	97.83 [96.65, 98.83]	80.21 [66.67, 92.50]	73.54 [59.58, 86.25]	97.66 [94.06, 99.69]	58.33 [52.53, 64.14]	81.51 [77.57, 85.45]

D.5 SDPO/SRPO Re-Implementation Sweep

Tab. 12 lists the configurations we tried for the SDPO and SRPO baselines. The entries are our re-implementations of the published objectives under the training setup of this study. Each row gives the values explored and which value was selected for the headline rows of Tab. 2.

Evaluation setting. We evaluate the published SDPO and SRPO objectives under a shared long-horizon math RLVR setup with Qwen3 Think rollouts. The sweep covers two model scales and more than 12 configurations; Figures 4 and 9 document the observed validation degradation, response shortening, and entropy rise.

Figure 9 shows two failure symptoms across the sweep, length collapse and entropy rise, appearing within 150–200 steps; the third (validation collapse) is the headline observation of §5.2. The checkpoints reported in Tab. 2 maximize validation mean@16 within their runs and fall in the early training window.

D.6 Three-Symptom Failure Pattern of All-Token OPSD Baselines

Figure 9 shows the training diagnostics of all-token OPSD baselines and the SRPO inner mechanism.

Table 12: SDPO/SRPO re-implementation sweep on Qwen3-8B + OpenThoughts math. "Main setting" lists the configuration used in Table 2, including symmetric training. Checkpoints maximize OpenThoughts validation mean@16 within each run; held-out benchmarks are not used for selection.

Factor	Values tested	Main setting
Learning rate	$\{10^{-6},\ 5\times10^{-6},\ 10^{-5}\}$	1×10^{-5}
Batch size	$\{32,\ 256\}$	32
EMA rate	$\{0,\ 0.05\}$	SDPO 0 / SRPO 0.05
Distillation top-K	$\{20,\ 100\}$	100
Divergence	{FKL, RKL, JSD}	SDPO JSD (α=0.5) / SRPO FKL
Thinking mode	symmetric / asymmetric	symmetric (student / teacher Think)
Per-vocab KL clip	$\{0,\ 0.05\}$	0.05

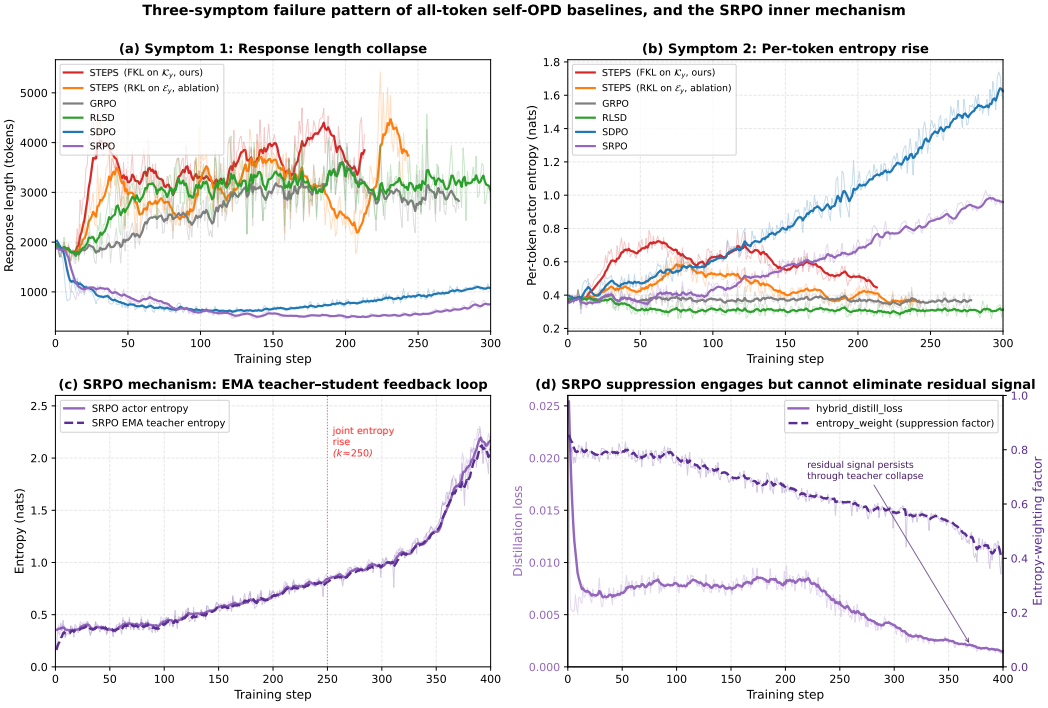

Figure 9: **All-token OPSD training diagnostics and SRPO dynamics.** (a) Response length; (b) per-token actor entropy. SDPO and SRPO produce shorter responses as entropy rises. (c) SRPO actor and EMA-teacher entropy co-vary, consistent with a possible feedback loop (Kim et al., 2026), without establishing causality. (d) Entropy-aware suppression reduces its weight but leaves nonzero distillation loss during degradation. Curves use EMA smoothing ($\alpha = 0.80$); faint traces show raw values.

D.7 Thinking-Mode Ablations on Qwen Models

Tables 13 and 14 report the Qwen3-8B ablations defined in Table 7; decoding settings are centralized in Appendix C.5.

Asymmetric training. STEPS with FKL on key spans achieves a five-benchmark AVG of 81.51 under the default symmetric configuration, compared with 75.79 under asymmetric training when both are evaluated with thinking enabled. Although the asymmetric configuration has a lower AVG than the GRPO reference (75.79 vs. 78.75), it retains a GPQA-Diamond improvement (56.69 vs. 53.85).

Table 13: **Asymmetric-training ablation on Qwen3-8B.** Student NoThink, teacher Think; evaluation Think. The GRPO row is reused from Table 2 as a reference and has no teacher.

Method	MATH-500	AIME 24	AIME 25	AMC 23	GPQA-D	AVG
Qwen3-8B base, Thinking ON (ref.)	96.80	76.25	67.50	95.94	58.27	78.95
+ GRPO	**97.30**	**77.08**	**68.96**	**96.56**	53.85	**78.75**
+ RLSD	95.85	73.75	65.42	95.31	**56.88**	77.44
+ STEPS (FKL on $\mathcal{K}_y$)	94.88	70.83	62.50	94.06	56.69	75.79
+ STEPS (RKL on $\mathcal{E}_y$, ablation)	93.95	68.33	60.42	93.12	56.44	74.45
+ SRPO	96.62	68.75	52.71	92.50	35.54	69.22
+ SDPO	94.88	50.42	38.75	90.62	38.26	62.59

Non-thinking evaluation. Reevaluating the same asymmetric-training checkpoints with the non-thinking protocol gives STEPS-FKL an AVG of 60.95%, exceeding the corresponding GRPO reference of 58.73% by an absolute 2.22%. Relative to thinking-enabled evaluation, this protocol also shortens the generation budget and uses fewer samples on AIME / AMC; the comparison therefore reflects these joint decoding changes.

Table 14: **Non-thinking evaluation of the asymmetric-training configuration.** Student NoThink, teacher Think; evaluation NoThink on Qwen3-8B. GRPO is a teacher-free reference. Best per column is **bold**.

Method (NoThink eval)	MATH-500	AIME 24	AIME 25	AMC 23	GPQA-D	AVG
Qwen3-8B base, Thinking OFF (ref.)	84.20	27.08	19.17	68.12	48.11	49.34
+ GRPO	91.47	**43.75**	29.58	80.00	**48.86**	58.73
+ RLSD	90.88	43.33	33.75	82.50	45.22	59.14
+ SDPO	78.85	15.83	11.67	58.75	28.66	38.75
+ SRPO	62.52	8.33	5.00	40.94	36.49	30.66
+ STEPS (RKL on $\mathcal{E}_y$, ablation)	86.88	26.67	25.83	80.00	48.36	53.55
+ STEPS (FKL on $\mathcal{K}_y$)	**92.25**	42.08	**36.25**	**87.81**	46.34	**60.95**

D.8 Negative Results

Generic rubrics and free-form critique (negative). A pre-STEPS exploration phase tested query-level rubric generators (Rubric-ARM-style (Lv et al., 2026)) and free-form LLM critique as alternative teacher conditioning. Both underperformed span-localized routing in our setup while incurring substantially higher annotator cost. The negative finding motivated the per-rollout type-label design adopted in STEPS.

All-token forward KL plus decay (negative). We further tested whether the same FKL signal applied to all response tokens, with the same decay schedule, would suffice. Under matched warm-up and decay, full-token FKL still degraded GPQA-Diamond OOD relative to the span-localized variant by step 100, indicating that span localization (not just decay) is necessary.

E Additional Controlled Experiments and Accounting

E.1 Coverage and Distillation Duration

Table 15 reports the coverage sweep on Qwen3-1.7B. Scores are the unweighted five-benchmark AVG used for the Qwen3 main results; GRPO is 57.73. The decayed runs use the same decay-to-GRPO schedule, and the constant-weight run uses $\lambda = 0.5$. Among the four measured coverage settings, $\alpha = 0.25$ performs best. These single-seed results characterize sensitivity in this configuration.

The constant-weight comparison evaluates the complete hand-off design. In Eq. (1), $\rho_k = 1 - \lambda_k / w_0$ restores GRPO on selected spans as KL decays. Holding $\lambda = w_0 = 0.5$ therefore also holds $\rho_k = 0$, rather than restoring the span-token GRPO term. The drop from 60.16 to 54.85 cannot be attributed

Table 15: **Coverage and duration sweep on Qwen3-1.7B.** The constant-KL comparison uses $\alpha = 0.25$. The $\alpha = 1.00$ row applies KL to all tokens.

Coverage α	Decayed KL	Constant KL ($\lambda = 0.5$)
0.10	58.20	—
0.25	**60.16**	54.85
0.50	56.90	—
1.00	56.68	—

solely to persistent KL independently of that coupling. Likewise, the all-token result 56.68 uses matched decay and differs from GRPO by -1.05%.

The class-by-direction experiment in Figure 5(b) uses the same seed, annotator, coverage, schedule, and checkpoint-selection rule across the four cells. It separates the span-class and divergence-direction changes that are coupled in the original two-corner comparison. The values are single-seed evaluation results, not averages over independent training seeds.

E.2 Paired Evaluation Uncertainty

Table 16 reports paired bootstrap comparisons for Qwen3-8B. Evaluation outcomes are paired by problem ID between methods, using 10,000 resamples. The target statistic is the difference in AVG, the unweighted mean of the five benchmark accuracies. Pairing preserves the correspondence between methods on each evaluated problem. Table 11 reports cell-wise intervals for individual benchmark accuracies, while the paired intervals quantify uncertainty in method differences.

Table 16: **Paired evaluation differences on Qwen3-8B.** Absolute accuracy differences and their confidence-interval endpoints are expressed in %. Strong uses `qwen3.5-plus`; online-self reuses the current policy as annotator.

Comparison	Δ AVG	Paired 95% CI
STEPS-strong − GRPO	+2.76	[0.41, 4.91]
STEPS-strong − RLSD	+2.52	[0.17, 3.67]
STEPS-online-self − GRPO	+1.90	[−0.25, 3.75]

The online-self interval includes zero, so we report its observed mean improvement without a significance claim. These intervals quantify uncertainty from evaluation problems conditional on the trained checkpoints. They do not measure variability across training seeds.

E.3 Online-Self Timing Breakdown

Table 17 reports the online-self timing components at three model sizes. Annotation and the teacher forward are additional operations during the early distillation phase, confined to the first 40 training steps; when $\lambda_k = 0$, Algorithm 1 skips both operations. The 4B column is a timing measurement only and is not an additional model in the effectiveness evaluation.

Table 17: **Online-self timing.** Component times are seconds per KL-active step. The last row reports overhead amortized over a 300-step run. Accelerator counts differ across columns, so absolute seconds cannot be compared directly across model sizes.

Component	1.7B	4B	8B
Shared rollout / forward / actor	124.4	117.3	131.2
Annotator generation	15.8	16.7	17.2
Teacher forward	6.7	7.4	9.4
Auxiliary total	22.5	24.1	26.6
Overhead across 300 steps	2.4%	2.7%	2.7%

Amortized overhead is computed over 300 training steps, using a 40-step allowance for auxiliary work:

$$\text{overhead}(\%) = 100\,\frac{40\,t_{\text{aux}}}{300\,t_{\text{shared}}}. \tag{11}$$

This yields 2.41%, 2.74%, and 2.70% for 1.7B, 4B, and 8B, respectively, before rounding. The denominator is shared training work over all 300 steps, not the duration of a KL-active step. The 40-step allowance covers the initial phase; the scheduled nonzero-KL window is specified separately in Eq. (2).

This timing breakdown is an overhead accounting, not a compute-matched comparison. It applies to online-self annotation and does not estimate the cost or latency of the strong API annotator.